Inferring inflection classes with description length

Sacha Beniamine, Olivier Bonami, Benoît Sagot

Inferring inflection classes with description length

Sacha Beniamine[1]*, Olivier Bonami*[1]*, and Benoît Sagot*[2]
[1] Université Paris Diderot, Laboratoire de linguistique formelle
[2] Inria

ABSTRACT

We discuss the notion of an *inflection class system*, a traditional ingredient of the description of inflection systems of nontrivial complexity. We distinguish systems of *microclasses*, which partition a set of lexemes in classes with identical behavior, and systems of *macroclasses*, which group lexemes that are similar enough in a few larger classes. On the basis of the intuition that macroclasses should contribute to a concise description of the system, we propose one algorithmic method for inferring macroclasses from raw inflectional paradigms, based on minimisation of the description length of the system under a given strategy of identifying morphological alternations in paradigms. We then exhibit classifications produced by our implementation on French and European Portuguese conjugation data and argue that they constitute an appropriate systematisation of traditional classifications. To arrive at such a convincing systematisation, it was crucial for us to use a *local* approach to inflection class similarity (based on pairwise comparisons of paradigm cells) rather than a *global* approach (based on the simultaneous comparison of all cells). We conclude that it is indeed possible to infer inflectional macroclasses objectively.[1]

Keywords:
morphology,
MDL,
inflection classes

[1] Work reported here has been presented at the First Quantitative Morphology Meeting (Belgrade, June 2015), at the 9th *Décembrettes* conference (Toulouse, December 2015), and at workshops organized by Université Paris Diderot and Labex EFL. We thank the audiences at these events and three anonymous reviewers for their comments. This work was partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

INTRODUCTION

The concept of INFLECTION CLASS is central to many analyses of inflection systems, both in theoretical linguistics (see among many others Matthews 1972; Carstairs 1987; Wurzel 1989; Aronoff 1994; Dressler and Thornton 1996; Corbett 2009) and in psycholinguistic studies (see among others Milin *et al.* 2009; Veríssimo and Clahsen 2014). Inflection class systems are commonly taken to be a classification of lexemes according to their INFLECTIONAL REALISATIONS. While such a broad characterization is largely agreed upon, there are alternative ways of applying it. Some authors (e.g. Stump and Finkel 2013) insist on strict identity of inflectional realisations, leading to systems with a large number of small classes. Many others follow traditional descriptions in defining a small number of classes based on broad similarity, and allowing some amount of variability within each class. Despite this uncertainty as to the characterization of classes a partition of the lexicon into classes is often taken for granted as a starting point for analysis, rather than explicitly argued for.

In this paper, we show that inflection classes can be deduced in a systematic and motivated way from raw paradigms, without introducing any preconception about organizing principles other than similarity. Our approach is abstractive in nature (in the sense of Blevins 2006), and is intended to systematize the strategies of descriptive morphologists in finding inflection classes.

The strategy is systematic enough to allow for full computational implementation.[2] We use the minimum description length principle (Rissanen 1984) to balance similarity within classes and dissimilarity between classes. This presupposes that we have a way of assessing similarity between overall inflection patterns. In this paper, we will consider two different but closely related ways of assessing that similarity. Under a GLOBAL approach, inflection patterns are determined by comparing all of the inflected forms of a lexeme simultaneously; under a LOCAL approach, the overall characterization is deduced from pairwise comparisons of paradigm cells. We propose a simple procedure for identifying patterns that can be applied either locally or globally,

[2] The full code to replicate the classifications discussed in this papier is available at `http://drehu.linguist.univ-paris-diderot.fr/qumin/`.

and show that a local approach captures the kinds of generalizations that descriptive morphologists rely on for classification.

The paper is organized as follows. First, we explore alternative definitions of inflectional classes and inflectional realisations. Then, we present a strategy for inferring inflection classes from raw paradigms in two steps: deducing inflectional realisations from the forms, and classes from realisations. In the third section, we present the detailed algorithms devised to perform each of these two steps, and describe the description length measure we use. The final section discusses results on both French and European Portuguese.

1 WHAT ARE INFLECTION CLASSES?

In this section, we argue that the apparent consensus on inflectional classification masks important differences between accounts that often rest on unstated theoretical assumptions, especially the role given to morpho-phonology, the basic units posited by the model, segmentation strategies, and the definition of similarity. For this reason, there is no agreed upon method to rigorously infer the classes from raw paradigms.

1.1 *Two definitions of inflection classes*

Following Aronoff (1994, p. 64) and Carstairs-McCarthy (1994, p. 639), we could define an inflection class as "a set of lexemes whose members each select the same set of inflectional realisations". We illustrate the definition with the twelve classes of Latin nouns in Table 1, as presented by Stump and Finkel (2013).

According to this definition, an inflection class system is an exhaustive partition of the set of lexemes in several non-overlapping classes. All members of a class have the exact same inflectional behavior. For example, classes (2a) and (2b), although they share all their other realisations, are distinct in that (2b) shows no affixal realisation for nominative and vocative singular.

While they match both Aronoff and Carstairs-McCarthy's definition of inflection classes, the 12 inflection patterns identified as rows in Table 1 do not correspond to the traditional characterisation of the Latin system. The tradition distinguishes only five classes, which group together some of the rows. Within those classes, as Dressler *et al.*

Table 1: Latin noun endings organized by declensions

		Singular						Plural					
Declension		NOM	VOC	ACC	GEN	DAT	ABL	NOM	VOC	ACC	GEN	DAT	ABL
First	(1)	a	a	am	ae	ae	ā	ae	ae	ās	ārum	īs	īs
Second	(2a)	us	e	um	ī	ō	ō	ī	ī	ōs	ōrum	īs	īs
	(2b)	–	–	um	ī	ō	ō	ī	ī	ōs	ōrum	īs	īs
	(2c)	um	um	um	i	ō	ō	a	a	a	ōrum	īs	īs
Third	(3a)	s	s	em	is	ī	e	ēs	ēs	ēs	um	ibus	ibus
	(3b)	–	–	–	is	ī	e	a	a	a	um	ibus	ibus
(i-stems)	(3c)	is	is	em	is	ī	e	ēs	ēs	ēs	ium	ibus	ibus
	(3d)	s	s	em	is	ī	e	ēs	ēs	ēs	ium	ibus	ibus
	(3e)	e	e	e	is	ī	ī	ia	ia	ia	ium	ibus	ibus
Fourth	(4a)	us	us	um	ūs	uī	ū	ūs	ūs	ūs	uum	ibus	ibus
	(4b)	ū	ū	ū	ūs	ū	ū	ua	ua	ua	uum	ibus	ibus
Fifth	(5)	ēs	ēs	em	ēī	ēī	ē	ēs	ēs	ēs	ērum	ēbus	ēbus

This table follows Stump and Finkel (2013, p. 183, Table 7.1). We omit the locative, reorder the paradigm cells and add numbering to facilitate reference to specific declensions.

(2008, p. 52) remind us, “not all nouns of one class inflect in exactly the same way”. For example, while some lexemes of the third declension end in *-ium* in the genitive plural (3c, 3d, 3e), others end in *-um* (3a and 3b). Rather than identity, then, members of the traditional classes display a strong degree of similarity.

This example is representative of a general observation that traditional descriptions of inflection systems distinguish a small number of broad classes, comprising both common patterns seen as regular and less common patterns seen as deviating from the regular situation. This leads some authors, such as Brown and Hippisley (2012, p. 4), to adopt a less strict criterion in the definition of inflection classes, which are then seen as “classes of lexemes that share similar morphological contrasts”.

The existence of two alternative definitions of inflection classes is sometimes the source of confusion. For instance, it is notable that, after proposing a definition of classes based on identity of realisation,

Carstairs-McCarthy's (1994)'s account of Latin nouns relies on mere similarity: starting from six classes (out of 8), he proposes to merge some of them so as to have a system of three classes. Whatever one may think of the motivation for such merges, the absence of a clear distinction between the two notions of inflectional classification makes such proposals hard to evaluate.

In a series of influential publications, Dressler and colleagues (Dressler *et al.* 1987; Dressler and Thornton 1996; Kilani-Schoch and Dressler 2005) propose not to choose between the two strategies and provide separate names for the two types of classes: MICROCLASSES are small uniform classes whose members have identical realisations, while MACROCLASSES are large classes exhibiting some amount of internal variation.[3]

Dressler and Thornton (1996) define microclasses and macroclasses as the two extremes in an inflection class hierarchy which may also contain classes of intermediate grain. This is very similar in spirit, if not in the details of execution, to inflection class hierarchies customarily used in Network Morphology (Corbett and Fraser 1993; Brown and Hippisley 2012).

Under this view, microclasses correspond to Aronoff and Carstairs-McCarthy's definitions. There is little doubt that, given a set of paradigms and some way of abstracting inflectional realisations from the paradigms, one can deduce a unique system of microclasses appropriately describing the system. The situation of macroclasses is more uncertain. Given that macroclasses are defined in terms of similarity, and that similarity is a gradual and multidimensional notion, there are various ways to partition a system into macroclasses, among which it is not obvious which should be chosen, short of a quantitative evaluation of the complexity of the resulting grammar (Walther 2013). For instance, there is no obvious way of deciding whether the Latin first and second declensions should be considered to form one class

[3] Our use of the term 'microclass' differs from that of Dressler and coauthors in one minor way. For Dressler and Thornton (1996), "An isolated paradigm is a paradigm which differs morphologically or morphophonologically from all other paradigms; it does not form a microclass of its own but is considered a satellite to the most similar microclass." In our usage, isolated paradigms are just microclasses of cardinality 1. This is of little theoretical consequence, but dramatically increases the number of microclasses for some systems.

(as in Carstairs-McCarthy 1994), because they inflect similarly in the dative, ablative, and locative plural, or two, because the realisations are distinct everywhere else in the paradigm.

Since the validity of microclasses is not under question, we will focus our attention on the status of macroclasses. We will take a system of inflection classes to be a partition of lexemes into classes, and attempt to infer a system of macroclasses from observed paradigms. Notice that we focus on the inference of macroclasses rather than a full hierarchy of classes of variable granularity. While this is definitely an interesting endeavour (see for instance Brown and Evans 2012; Lee and Goldsmith 2013; Bonami 2014 for some proposals), it calls for a different methodology, and does not directly help us evaluate which partition in the hierarchy should correspond to the level of macroclasses.

Defining macroclasses requires a definition of inflectional realisations from which the similarities follow, and a criterion to decide the appropriate level of generality. In the following two sections, we first describe some possible ways of defining inflectional realisation before investigating the possible criteria with which to define macroclasses.

1.2 *Macroclasses follow from inflectional realisations*

1.2.1 Circularity of inflectional realisation definitions

Any enterprise in inflectional classification starts with the identification of inflectional realisations. The heuristics used for that purpose are seldom made explicit, although they are rarely obvious. For instance, it is customary to assume that inflectional variability combines the use of different patterns of stem allomorphy and different affixal exponents, although deciding on the exact boundary between stem and exponent is far from being a trivial matter. Carstairs-McCarthy's (1994)'s work on inflection classes is commendable for its explicitness in such matters. We thus propose to explore it in some detail.

Carstairs-McCarthy's strategy relies on two central decisions. First, inflection classes are defined purely in terms of affixal exponence, and abstract away from stem allomorphy. Thus inflectional realisations are considered affixes, and any alternation that is not affixal is ignored. Second, segmentation choices are justified by the desirability of the inflection class system they yield. This is motivated by the goal of testing whether inflection class systems satisfy the 'No Blur

Table 2: Latin masculine declensional endings without and with thematic vowels

	Singular						Plural			
Declension	NOM	VOC	ACC	GEN	DAT	ABL	N/V	ACC	GEN	D/A
First	a	a	am	ae	ae	ā	ae	ās	ārum	īs
Second	us / –	e	um	ī	ō	ō	ī	ōs	ōrum	īs
Third (cstem)	s / – / ēs	s / – / ēs	em	is	ī	e	ēs	ēs	um	ibus
(mixed)	s	s	em	is	ī	e	ēs	ēs	ium	ibus
(istems)	is	is	im > em	is	ī	ī > e	ēs	īs > ēs	ium	ibus
Fourth	us	us	um	ūs	uī	ū	ūs	ūs	uum	ibus

	Singular						Plural			
Declension	NOM	VOC	ACC	GEN	DAT	ABL	N/V	ACC	GEN	D/A
First	–	–	m	ī	ī	V:	ī	:s	:rum	īs
Second	s / –	e	m	ī	V:	V:	ī	:s	:rum	īs
Third (cstem)	s / –	s / –	m	s	ī	e	ēs	ēs	um	bus
(mixed)	s	s	m	s	ī	e	ēs	ēs	um	bus
(istems)	s	s	m	s	ī	V: > e	ēs	:s > ēs	um	bus
Fourth	s	s	m	s	ī	V:	:s	:s	um	bus

This table is adapted from Carstairs-McCarthy (1994, pp. 749-750). The symbol > means 'tends to be replaced by'. Slashes separate affixes which are distributed on a partly phonological, partly arbitrary basis. In each column, shades of gray highlight repeated affixes when they violate the No Blur Principle.

Principle', according to which any affix realising some paradigm cell must be either a class identifier (i.e. specific to that class) or a class default (i.e. common to all those classes that do not possess a class identifier). By Carstairs-McCarthy's reasoning, if a system can be described with classes that satisfy the No Blur Principle, then that classification should be used, even if there are alternative classifications that do not satisfy the principle.

Carstairs-McCarthy explores two alternative segmentations of the affixes of masculine latin nouns, reproduced in Table 2. We show blur in columns using grayed cells. The traditional analysis, presented at the top of the table, presents some blur: for instance, in the dative plural, *-īs* is neither a class identifier (it is common to two classes) nor a default (since the other possible affix, *-ibus*, is not an identifier either). In an alternative analysis, presented at the bottom of the table, theme vowels are taken to be part of stems rather than affixes. This analysis

shows twice as many blurred columns. Therefore, Carstairs-McCarthy prefers the first analysis. Whatever one may think of the relative merits of the two analyses and the relevance of the No Blur Principle, it is worth noting that Cartairs-McCarthy's heuristic for choosing a segmentation leads to circularity: inflection classes are taken to be sets of words displaying the same set of inflectional realisations, but what counts as an inflectional realisation is decided on the basis of the desirability of the resulting inflection class system. Such circularity is particularly vivid in Carstairs-McCarthy's paper, but we suspect that it is present in many descriptions that do not make their segmentation heuristics explicit. This dependency of the realisations on the classes is problematic in the context of an abstractive approach to inflectional classification, where the realisations are the starting point for the inference of classes.

More generally, despite relevant attempts (e.g. Montermini and Boyé 2012; Spencer 2012), there is no agreed upon systematic strategy to decide where to place the boundary between stem and affix; and as Blevins (2005) and Blevins (2006) argues, in some systems, there is just no coherent way of making such a decision. From this observation we conclude that a systematic method for inferring inflectional realisations should not rely on a preexisting segmentation into stems and affixes. Given this, one possible way forward is to explore different segmentation strategies and rely on Occam's razor to decide which is optimal (Sagot and Walther 2011; Walther and Sagot 2011). Another, which we pursue here, is to take whatever alternation is seen in the data at face value, irrespective of how (un)systematic it is or whether it affects peripheral rather than central segments of the alternating forms.

1.2.2 Global and local alternation patterns

To avoid making any undermotivated decision as to the boundary between affixal exponence and stem allomophy, we define inflectional realisation in terms of the alternation patterns relating the different forms in the paradigm of a lexeme to each other. Interestingly, wherever paradigms have a more than two cells, there are at least two strategies for identifying such patterns. We illustrate this with the small sample of the French adjectival lexicon in section A of Table 3.

Table 3: Alternative segmentation choices for a subset of French adjectives

	A. Paradigms			B. Stem and exponents, global		
lexemes	M.SG	F.SG/PL	M.PL	M.SG	F.SG/PL	M.PL
NORMAL	nɔʁmal	nɔʁmal	nɔʁmo	Xal	Xal	Xo
VERT	vɛʁ	vɛʁt	vɛʁ	X	Xt	X
BLEU	blø	blø	blø	X	X	X
	C. Patterns, local			D. Patterns, global		
	M.SG ∼ F.SG/PL	M.SG ∼ M.PL	F.SG/PL ∼ M.PL	M.SG ∼ F.SG/PL	M.SG ∼ M.PL	F.SG/PL ∼ M.PL
NORMAL	X ∼ X	Xal ∼ Xo	Xal ∼ Xo	Xal ∼ Xal	Xal ∼ Xo	Xal ∼ Xo
VERT	X ∼ Xt	X ∼ X	Xt ∼ X	X ∼ Xt	X ∼ X	Xt ∼ X
BLEU	X ∼ X	X ∼ X	X ∼ X	X ∼ X	X ∼ X	X ∼ X

In this table, gray cells highlight patterns that are common to two lexemes, showing that only local patterns capture the simiarity between NORMAL and BLEU.

The first and most familiar strategy consists of identifying the similarities and differences between forms GLOBALLY. This is indicated for our toy example in section B of Table 3: in each row, the substring common to all paradigm cells has been replaced by a variable. An alternative strategy, often invoked by proponents of implicative approaches to morphology (e.g. Blevins 2005, 2006; Ackerman *et al.* 2009; Bonami and Beniamine 2015), consists of identifying LOCAL similarities between pairs of paradigm cells. This is indicated in section C of Table 3, where each column now corresponds to a different pair of cells. Note that since we are dealing with a mostly concatenative system, both strategies can be seen as amounting to proposing a segmentation of words into constant ('stems') and variable ('affixes') subparts. However, in the global approach the constant part is common to the whole paradigm, whereas in the local approach it is particular to one pair of cells: for instance, M.SG *normal* is segmented into *norm*+*al* for purposes of comparison with the M.PL, but not for purposes of comparison with the feminine.

One way of highlighting the difference between the two strategies is to tabulate the consequences of a global strategy for the description of alternations between pairs of cells. This is done in section D of Table 3, which just sums up the information in section B of Table 3 in the forms of relations between pairs of cells. One may note that, according to the local strategy, section C of Table 3, the adjective BLEU shares inflectional characteristics with both NORMAL and VERT: like NORMAL, it does not alternate between M.SG and F; like VERT, it does not alternate between M.SG and M.PL. By contrast, according to the global strategy, VERT and BLEU share the same characteristic of not alternating between the M.SG and M.PL, but NORMAL and BLEU do not have anything in common.

If binary alternation patterns are the inflectional realisation, then microclasses are defined by vectors of patterns, where each coordinate of the vector indicates the pattern instantiated in that microclass for a different pair of paradigm cells. These vectors are represented by rows in sections C and D of Table 3. We thus conclude that in our toy example, the global and local strategies give rise to the same microclasses. However, relations of similarity among these microclasses are different. Hence the use of local or global inflection patterns to char-

acterise inflectional realisation may influence what macroclasses will be inferred.

One of the goals of this paper is to evaluate the relative perspicuity of inflectional classifications based on local and global alternation patterns. For the time being, let us comment briefly on the relationship between alternation patterns, whether global or local, and segmentation of words into stems and affixes. There is a natural relation between global patterns and stem-based segmentation. Since global patterns identify a constant subpart common to the whole paradigm, in the context of concatenative morphology, a global pattern corresponds to an analysis where each lexeme is constrained to using a single stem, and any variable element is taken to be affixal material. Interestingly, there is no such clear relation between local patterns and the classical notion of a stem. As we highlighted above, one and the same word filling one paradigm cell may be segmented differently for the purposes of comparison with two other cells. Hence, under a local pattern view, even individual paradigm cells are not associated with a unique constant substring which could be identified as a stem.

1.3 *Criteria for macroclasses*

In the preceding section we showed how different strategies for describing inflection systems, be they based on segmentation between stems and affixes or on alternation patterns, lead to different classifications. We now turn to the problem of deciding which groupings of microclasses should be considered as forming a single macroclass. We explore five strategies found in the literature: using an ad-hoc combination of criteria, the regular/irregular distinction, maximisation of inflection class heterogeneity, maximisation of internal predictability, and maximisation of descriptive economy.

1.3.1 Ad-hoc criteria

Descriptive morphologists usually motivate their classification highlighting some property or set of properties which the classes happen to differ in. For instance, Latin verb classes are characterised by the quality and length of the theme vowel in the present active infinitive: *-ā-* in the first conjugation, *-ē-* in the second, *-e-* or *-i-* in the third, and *-ī-* in the fourth. Of course, this is not the only way in which Latin conjugations contrast, and not all forms exhibit such a contrast. As any

description of Latin conjugation will note when commenting on the third conjugation, some verbs in that class have an indicative present active 1SG form similar to that of a first conjugation verb, cf. SECŌ 'cut' (INF *secāre*) vs. SERŌ 'sow' (INF *serere*); others do contrast with the first conjugation in that paradigm cell, but fail to contrast with the third conjugation, cf. CAPIŌ 'take' (INF *capere*) vs. SAEPIŌ 'surround' (INF *saepīre*). Full classification relies on an ordering of highlighted *ad-hoc* properties: in the case of Latin, tradition holds that contrasts in the infinitive are more important than contrasts in the indicative present first person singular.

There are two concerns with such a strategy for motivating a classification. First, it is unclear whether the highlighted properties are selected *post-hoc* to contrast pre-established classes, perhaps for pedagogical purposes, or whether they really play a distinguishing role. In the case at hand, it seems arbitrary that the infinitive is used to motivate the distinction between the four classes when the relevant contrast is also apparent e.g. in the present 1PL. Second, it is unclear that there is any strong motivation for the way the contrasts are prioritised.

The situation just discussed in the case of the traditional classification of Latin verbs also holds for more elaborate, thoughtful, and theoretically-informed classification attempts. We exemplify this situation by discussing , in some detail, the proposed classification of French verbs by Kilani-Schoch and Dressler (2005).

As we saw before, in Natural Morphology, macroclasses are viewed as the top-level partition in an inflection class tree (Dressler *et al.* 2008; Kilani-Schoch and Dressler 2005; Dressler and Thornton 1996). In these accounts, Macroclasses, just as classes of all granularities, are defined by implicational PARADIGM STRUCTURE CONDITION (PSCs). To study the nature of PSCs, we reproduce below those presented in Kilani-Schoch and Dressler (2005) for some classes of French verbal inflection.

(1) Macroclass I:

$$\text{Infinitive } /X+e/ \Rightarrow \begin{cases} \text{Past Participle} & = /X+e/ \\ \text{Simple Past first person} & = /X+e/ \\ \text{Singular present} & = /X/ \\ \text{Indicative present 3rd plural} & = /X/ \\ \text{Subjunctive present} & = /X/ \end{cases}$$

(2) Class I.1:

Imperfect [parl + ɛ], future [parl + ər + e].

(3) Class II.2:

$$\text{Infinitive /Xwar/} \Rightarrow \begin{cases} \text{Past Participle in /y/} \\ \text{Simple Past in /y/} \\ \text{by default, /wa/ is part of the infinitive} \end{cases}$$

We first remark that PSCs are of variable nature. They are sometimes formulated as implicative relations (Wurzel 1984; Ackerman *et al.* 2009; Stump and Finkel 2013), as is the case for macroclass I of French verbs reformulated in (1) or in class II.2 as shown in (3). These implications are sometimes relationships between two cells (if some cell is X, then some cell is Y), as in (1), and sometimes between a cell and an abstract segmented unit as in (3). Some subclasses, on the contrary, are defined by the exponence strategies they implement, as in (2) for microclass I.1.

In Kilani-Schoch and Dressler (2005)'s analysis of French verbs, the implications are frequently true for all the other classes. For example, the antecedent of the implication in (1), having an infinitive in /Xe/, is only true of the verbs in macroclass I. As a consequence, all the implications based on this premise are true of the whole system. What is implicitly defining that macroclass, then, is not the PSC but the exponent: macroclass I is the class of all verbs with an infinitive ending in /e/. The same could be said of the PSC from (3) which is true of the whole system because only verbs of the class II.2 share an infinitive ending in /-war/, revealing that it is in fact defined not by the implication but by the ending. We conclude then that, while PSCs are formulated as implications, classes are really defined by exponence strategies, mostly with a focus on the infinitive.

In light of these observations, it appears that a class is sometimes a set of lexemes having one or more common exponents (as we showed for I and II.2), sometimes a set of lexemes for which some implicative relationship between cells hold. Since macroclasses are motivated by different types of criteria, we cannot assume that they are consistently the same kind of object. If one chooses to keep both types of criteria, it is not clear how one should decide which to apply when. It seems preferable to build a class system relying only on one criterion.

1.3.2 External motivation: regularity

Another organizing principle is at work in Kilani-Schoch and Dressler's (2005)'s classification of French verbs. A core assumption of that work is a dual mechanism approach to inflection processing (see Clahsen 2006 and references therein), according to which (i) there is a categorical distinction between regular and irregular lexemes, and (ii) regular and irregular lexemes are processed differently by speakers. Whether a lexeme is regular or irregular cannot be established by examination of the synchronic inflection system, but only through assessments of productivity (only regular patterns are deemed productive) or psycholinguistic experimentation (regular and irregular lexemes should lead to measurably different learning, processing, and production). Kilani-Schoch and Dressler hold that the contrast between regulars and irregulars should be the principal criterion to distinguish macroclasses. Hence their classification makes a main distinction between two macroclasses, corresponding to the traditional first conjugation (infinitives in *-er*) vs. all other verbs.

Whatever one may think of the merits of the dual mechanism hypothesis or of the assumption that regularity in French holds only of the traditional first conjugation (see Bonami *et al.* 2008), the important point for present purposes is that Kilani-Schoch and Dressler's criterion for macroclasses is fundamentally different from the criterion used to group lexemes into microclasses. Macroclasses are No longer a generalisation over microclasses, but rather a completely different classification of lexemes, whose empirical validity cannot be established by examination of the internal structure of the synchronic system. Again, while this is a defendable position, it is unclear why one type of criterion should be privileged over another. Evidently there are multiple ways of classifying lexemes that may be relevant for different purposes, and it is not clear that there is merit in attempting to combine all such classifications in a single tree. In particular, it is an open question how exactly a broad classification based on structural similarity and contrast between inflection patterns correlates with contrasts in productivity and/or ease of processing. Presupposing a strong association between the two does not help explore the issue.

In the remainder of this paper, we will focus on approaches to inflectional classification that rely solely on examination of similarity and differences between paradigms.

1.3.3 Heterogeneity among classes

In the context of defining a canonical typology of inflection class systems, Corbett (2009, p. 4) formulates two important criteria for canonical inflection classes, respectively on distinctiveness and cohesion of classes:[4]

(4) a. "Criterion 1: In the canonical situation, forms differ as consistently as possible across inflectional classes, cell by cell."

b. "Criterion 3: Within a canonical inflectional class each member behaves identically."

According to Corbett, a canonical inflection class system is a single partition of the set of lexemes where each class is maximally cohesive internally and maximally distinct from other classes. Interestingly, Criterion 3 is reminiscent of the definition of micro-classes. It is tempting then to assume that macro-classes are defined by Criterion 1: macro-classes should be strikingly different from one another. This seems to match traditional practice, and leads to the satisfactory conclusion that a canonical system is a system where micro-classes and macro-classes coincide.

While Criterion 1 definitely captures part of the intuition behind macro-classes, we should be wary of not applying it too strictly. In any system where one paradigm cell inflects uniformly, all lexemes share at least one inflectional realisation, and this common inflectional realisation forbids perfect heterogeneity between classes. As a consequence, there is no partition that maximises distinctiveness, and hence no macroclass other than the system as a whole. Such a definition of macro-classes would then be too dependent on a rather unilluminating property of the system. Moreover, maximisation of distinctiveness does not strictly match traditional practice either. For instance, in the case of Latin nouns (1), it is not usual to suggest fusing the third and fifth declensions, despite the fact that they share the exponent *-e* in the singular ablative.

[4] Corbett's Criterion 2 refers to the shape of paradigms, and does not directly concern us here.

We thus conclude that while distinctiveness is an important property of macro-classes, it cannot be used as the sole criterion for choosing which partition should count as a partition into macro-classes.

1.3.4 Predictability within classes

Going back to Carstairs-McCarthy (1994), we find that he justifies the merging of classes into what he calls macroclasses when different affixes can be seen as suppletive allomorphs predictable from some other phonological or morphological factor (they are not competing for the speakers) (see Table 4).

This leads him to merge the first and second Latin declension (see Table 2), despite their strong dissimilarity. Indeed, the first two declensions are mostly predictable on the basis of gender. In the same way, some variations of the 3rd declension are predictable on the basis of phonological properties of the stem. These are indicated by a swung dash in Table 4. This is contrary to the intuition that macroclasses are classes of lexemes that inflect alike.

In addition, some alternations indicated by a slash in Table 4 do not correspond to systematic alternations. In this case, the classes are merged together because of the similarity of their paradigms, not because of their predictability.

Table 4: Table from Carstairs-McCarthy (1994, p. 751)

			Singular			
Declension	NOM	VOC	ACC	GEN	DAT	ABL
First/Second	a ∼ us / –	a ∼ e	am ∼ um	ae ∼ ī	ae∼ ō	ā ∼ ō
Third	s ∼ – / ēs ∼ is	s ∼ – / ēs ∼ is	em ∼ im	is	ī	e
Fourth	us	us	um	ūs	uī	ū
			Plural			
	NOM/VOC	ACC	GEN	DAT/ABL		
First/Second	ae ∼ ī	ās ∼ ōs	ārum ∼ ōrum	īs		
Third	ēs	ēs	um ∼ ium	ibus		
Fourth	ūs	ūs	uum	ibus		

Original caption: “Latin masculine nouns: third analysis, designed to remove blur. Forms separated by a swung dash are to be understood as distributed on the basis of gender (in the 1st/2nd declension) or of phonological characteristics of the stem. The distribution of forms separated by a slash is not governed in this way.”

Beyond the specific predictors used by Carstairs-McCarthy, we can see that merging paradigms according to predictability or similarity of the inflectional realisations leads to different results. Moreover, it is expected that merging together very similar paradigms is not favorable to prediction. Let us take paradigm entropy (Ackerman and Malouf 2013), the average conditional entropy of one paradigm cell given another paradigm cell, as a measure of internal predictability in a class. Unisng paradigm entropy, it becomes apparent that in fact, merging similar classes hinders predictability rather than helping it. In the case at hand, merging (2a) and (2b) in Table 1, which only differ by nominative and vocative singular, raises the difficulty of predicting these cells from any of the others, as having an accusative in *-um* and knowing that the noun is of the second declension will not guarantee that one can guess the correct nominative form. A macroclass comprising (2a) and (2b) would be justified if macroclasses are taken as similarity-based classes, but not if they are taken as classes with low paradigm entropy. On the other hand, one would not want to merge (1a) and (2a) on the basis of similarity. However, since they share few realisations, merging them would not raise the class paradigm entropy much. For example, from the accusative form, two patterns would be available to form the nominative, either *-am* → *-a* or *-um* → *-us*. This, however, does not make prediction more difficult, as only accusative forms ending in *-am* are candidates for the first pattern, and those ending in *-um* for the second one.

Devising an entire classification of macroclasses in a way that minimises the paradigm entropy in each class would lead to classifications that differ very strongly from what descriptive linguists produce. In this paper, we will rather try to find macroclasses on the basis of similarity. However, we should remember that those classes are not expected to have a lower paradigm entropy than the whole system.[5]

[5] Given several competing analyses of a system into classes on the basis of their realisations, one could prefer that which conveniently predicts other grammatical features. Corbett (1982) has argued that it is preferable to define four macroclasses of Russian nouns, rather than the three traditionally recognized, as it offers a better predictibility of gender. As a first step towards automatic inference of inflectional classification, the current study bases the inference of macroclasses strictly on wordforms. However, the model could be extended in a straightforward manner to cluster classes on the basis of other features in addi-

1.3.5 Maximisation of descriptive economy

Another approach to the problem of choosing how to define macroclasses relies on the idea that, in theory, the optimal set of macroclasses should result in the most economical description of the morphological system as a whole. This idea has been explored in particular by Sagot and Walther (Sagot and Walther 2011; Walther and Sagot 2011; Walther 2013; Sagot and Walther 2013; Walther 2016), who compare manually crafted descriptions, comprising a morphological grammar and a morphological lexicon, using a quantitative measure of their descriptive economy based on the information-theoretic notion of DESCRIPTION LENGTH (Rissanen 1978). Such an approach allowed them to compare competing accounts of a number of morphological (sub)systems in a variety of languages (French, Maltese, Khaling, and Latin), based on grammars implemented in the Alexina$_{\text{PARSLI}}$ framework, an implementation of Walther's PARSLI morphological formalism, for which see now (Walther 2016). These competing accounts can vary in different ways, one of which being the inventory of macroclasses, which roughly correspond to what they refer to as inflection patterns. For instance, Sagot and Walther (2011) compares the description lengths of four descriptions of French verbal inflection that contrast in the number of macroclasses they distinguish (from 1 to 139), in relation with different ways to dispatch morphological information between the grammar and the lexicon (e.g. lexically specified stem suppletion vs. stem alternation patterns encoded in the grammar).

While Sagot and Walther's work is an important inspiration for the strategy we will develop later in this paper, there are two fundamental limitations of their work. First, the fact that they rely on a specific description formalism to encode all competing accounts inevitably biases and reduces the set of possible accounts that can be compared. Second, and more importantly, they only compare a handful of manually crafted grammars. Without a way to systematically explore the space of possible descriptions, they can only draw conclusions from the relative compactness of the competing descriptions they compare.

---

tion to alternation patterns. We leave the exploration of such a possibility to a future study.

To conclude this section, we have argued that a coherent definition of macroclasses should rely on a single, well-conceived criterion to assess the level of accepted similarity. Several competing criteria are sometimes used to define macroclass membership, and most criteria used in the literature rely on more than the forms and inflectional realisations themselves. In this work, we ask whether macroclasses can be inferred from the sole examination of paradigms. This has the advantage that any preconceived idea about other properties that macroclasses have can be tested empirically. For example, we will be able to observe if we find only two macroclasses that conform to the categorical regular/irregular contrast presupposed by a dual mechanism approach to morphological processing.

2 INFERRING INFLECTION CLASSES

To automatically infer macroclasses from paradigms of raw forms, we take on two tasks, treated sequentially. First, given paradigms of forms, we want to infer all relevant alternation patterns following either a local or a global segmentation. The two segmentation strategies need to be strictly comparable. Second, given a table of alternation patterns, we attempt to infer micro- and macroclasses in a principled way.

2.1 *From forms to patterns*

The first task at hand is to infer alternation patterns from surface forms. We first describe previous work on the subject, then describe our algorithm.

2.1.1 Previous work on inflectional rule inference

A substantial amount of work has already been done on automatic inference of inflection rules from inflected forms, either in the context of modeling a speaker's knowledge of inflection (Albright and Hayes 2003, 2006) or in a Natural Language Processing context, with the goal of expanding sparse lexica (Durrett and DeNero 2013; Ahlberg *et al.* 2014; Nicolai *et al.* 2015). In this section, we review relevant aspects of these attempts.

Given a set of forms, one can formulate a large number of alternation patterns relating them. Choosing an appropriate function is an optimisation problem, seeking to minimise both the total number of

Table 5: Illustration of the alignment problem for two imaginary languages

(a) Infix language

SG	PL
to	bato
ri	bari
su	basu
ne	bane
ba	baba

(b) Prefix language

SG	PL
to	tabo
ri	rabi
su	sabu
ne	nabe
ba	baba

(c) Alignments of *baba* ∼ *ba*

Alignment						Pattern SG ∼ PL
PL	b	a	b	a		
(i) SG	_	_	b	a		_ ∼ ba_
(ii) SG	b	a	_	_		_ ∼ _ba
(iii) SG	b	_	_	a		_∅_ ∼ _ab_

patterns postulated to describe a system and to maximise the morphophonological naturality of the function. To explore the problem, let us consider two imaginary languages marking the opposition between singular and plural nouns as indicated in Table 5.

The two languages share exactly one lexeme, whose singular form is *ba* and whose plural is *baba*. There are a number of alternative ways of conceiving of the exponent of plural for that morpheme. Three prominent possibilities are (i) a *ba-* prefix, (ii) a *-ba* suffix, or (iii) an *-ab-* infix.[6] To these three possibilities correspond the three patterns listed in section (c) of Table 5, which in turn correspond to three ways of aligning the two forms. These toy languages are designed to highlight the fact that the choice of a pattern for a given lexeme is dependent on what happens in the rest of the language. In the context of language (a), where all other nouns mark the plural by prefixing *ba-*, it is clearly preferable to adopt a prefixation analysis (i); on the other hand, in the context of language (b), where all other nouns mark the plural by infixation, no descriptive linguist would doubt that the appropriate analysis for *ba*∼*baba* is an infixation analysis.

The task of deciding which alternation pattern is most relevant to relate two forms usually requires at least two steps: choosing an alignment, and abstracting a pattern from that alignment. The ambiguity can be resolved at the alignment stage by finding only one alignment or once all possible patterns are known. Note that the local and global strategies described above differ in how they perform the alignment step.

Extant approaches contrast in the way they deal with these issues. First, Durrett and DeNero (2013) infer global segmentations via

[6] Further possibilities include reduplication of the initial or final syllable.

the alignment of all forms to a base form. Ahlberg *et al.* (2014) directly align all forms of a paradigm together, also performing a global segmentation. On the other hand, Albright and Hayes (2003) and Albright and Hayes (2006) explicitly model local alternation patterns. Nicolai *et al.* (2015) compare forms locally, but only include pairs containing a designated base form, and thus do not take into account the whole array of possible alternations. Second, Durrett and DeNero (2013) and Albright and Hayes (2006) both use string alignment algorithms based on edit distance. The former perform iterated alignments to make their algorithm paradigm aware (which is why their alignment is global) while the latter optimise the similarity of aligned segments in terms of phonological features. Ahlberg *et al.* (2014) rely on transducer intersection to find the optimal alignment, and Nicolai *et al.* (2015) use the Expectation-Maximisation algorithm to learn atomic operations rather than entire alignments.

Although these studies are important sources of inspiration for the algorithm presented below, the strategies they implement are not quite appropriate for our current goals. The use of a privileged base form makes sense when trying to fill sparse paradigms as did both Durrett and DeNero (2013) and Nicolai *et al.* (2015): picking a frequent base form then allows one to reliably make inferences even for infrequent lexemes. However, while some forms can be prominent on the basis of informativeness, markedness, or other factors,here is no *a priori*otivation for favouring a base form in the identification of inflection classes. Speakers may be initially exposed to any form of a lexeme, and are able to draw inferences about the rest of that lexeme's paradigm on that basis, exhibiting no dependency on a designated base (Ackerman *et al.* 2009; Bonami and Beniamine 2016).

Likewise, Albright and Hayes's Minimum Generalisation Learner has a crucial property: the patterns it finds are gradually generalised, and generalisations at all levels are remembered. This is crucial to modeling the phenomenon of *Islands of reliability*, whereas lexemes that are phonotactically more typical of an inflection pattern are more strongly associated by speakers with that pattern. For our purposes though, it is crucial that each pair of form be associated with a single pattern, so that the lexicon is partitioned according to which pattern each lexeme instantiates. In addition, not having to keep track of all intermediate generalisations considerably reduces the algorithmic

complexity of the task, an important practical consideration when our experiments will rely on comparisons of thousands of pairs of cells for thousands of lexemes.

Finally, none of the studies we review here provide an algorithm allowing for the comparison between global and local strategies. We thus devise one that allows for strict comparison of both strategies.

2.1.2 Our pattern algorithm

To compare local and global segmentation strategies, we devise a segmentation process with two minimally different variants, which both output exactly one pattern per pair of cells. We use the same algorithm in both cases, changing only the number of forms we input.

We exemplify the algorithm on a sub-paradigm of the French verb AMENER 'bring', consisting of the three indicative present plural forms, and start with the global strategy. In that context, all forms of a paradigm are input at once, as indicated in column 1 of Table 6).

Our pattern extraction algorithm has two distinct parts. First, the input forms are left-aligned, as indicated in column 2 of Table 6. Second, all vertically identical characters are replaced by a placeholder, merging contiguous placeholders, as indicated in column 3 of Table 6. This allows us to discard constant information, and keep only the information that varies and their position in the form. We then group the resulting strings two by two to form the patterns, as indicated in column 4.

To model the local strategy, we proceed in exactly the same fashion, except for the fact that the algorithm is applied separately to each

Table 6: Plural present forms for the verb AMENER 'bring': Global pattern extraction

	1. Input	2. Left aligned forms				3. Variables
PRS.1PL	amønõ	am	ø	n	õ	…ø…õ
PRS.2PL	amøne	am	ø	n	e	…ɛ…
PRS.3PL	amɛn	am	ɛ	n		…ø…e

4. Output: patterns	
PRS.1PL ~ PRS.2PL	…ø…õ ~ …ø…e
PRS.2PL ~ PRS.3PL	…ø…e ~ …ɛ…
PRS.3PL ~ PRS.1PL	…ɛ… ~ …ø…õ

	1. Input	2. Left aligned forms					3. Variables
PRS.1PL	amønõ	a	m	ø	n	õ	…õ
PRS.2PL	amøne	a	m	ø	n	e	…e
PRS.2PL	amøne	a	m	ø	n	e	…ø…e
PRS.3PL	amɛn	a	m	ɛ	n		…ɛ…
PRS.3PL	amɛn	a	m	ɛ	n		…ɛ…
PRS.1PL	amønõ	a	m	ø	n	õ	…ø…õ

Table 7: Plural present forms for the verb AMENER 'bring': Local pattern extraction

4. Output: patterns	
PRS.1PL ~ PRS.2PL	…õ ~ …e
PRS.2PL ~ PRS.3PL	…ø…e ~ …ɛ…
PRS.3PL ~ PRS.1PL	…ɛ… ~ …ø…õ

pair of paradigm cells, rather than just once to the whole set of pairs. In the case at hand, as indicated in Table 7, this leads to three separate runs of the algorithm, leading in each case to the production of one pattern.

As we see from the tables, the local strategy produces binary alternation patterns which encode strictly local knowledge about the pair, while global alternation patterns encode knowledge about the rest of the paradigm. On this small paradigm, the choice of strategy only makes a difference for the alternation between the first and second person. The global strategy yields a pattern specific to verbs with an /ə/ in the penultimate syllable. The local strategy, on the other hand, yields a more general pattern, that also characterises verbs with no /ə/ in the penultimate syllable. This is relevant to clustering, as the global strategy, but not the local strategy, will take AMENER to exhibit a rather unusual behavior.[7]

Both strategies take the surface forms at face value and do not attempt to derive any underlying representations. Alternations are thus morpho-phonological rather than strictly morphological. There are two main reasons for this choice: First, it is not clear how to automatically abstract all regular phonology from a set of wordforms (our

[7] In fact, all French verbs except ÊTRE 'be', FAIRE 'do', DIRE 'say' and their derivative use the same pattern as AMENER.

input). Second, some regular phonological alternations do contribute to opacities in alternations, and are predictible only in one direction. Abstracting them out would be to underestimate the task speakers face when they inflect forms.

As this example illustrates, our current algorithm is able to capture stem-internal alternations that are rampant in familiar inflection systems. Actually, it is general enough to allow for multiple points of variation within the string, and hence is in principle capable of dealing with root-and-pattern morphology. On the other hand, the use of left-alignment is a clear limitation of the algorithm, making it impossible to capture systems making any use of prefixation.[8] While this is a clear limitation, it has no influence on performance on non-prefixing systems such as the ones we will explore in Section 4.

2.2 *From patterns to classes*

2.2.1 Previous work on inflection classes inference

The task of automatically inferring inflection classes has recently seen growing interest.

An early attempt at that task by Goldsmith and O'Brien (2006) used a neural network to relate features to exponents. The hope was that the hidden layer of the network would reflect inflectional classification. However, experiments on both Spanish and German failed to produce such a result. Very recently, Malouf (2017) has developed more promising uses of neural networks to model inflectional behavior, but the results cannot be interpreted straightforwardly as a partition of inflectional macroclasses.

There have also been efforts in NLP to infer microclasses from incomplete paradigms (Eskander *et al.* 2013; Monson *et al.* 2004), building on the same kinds of methods used by Dreyer and Eisner (2011) and Durrett and DeNero (2013); Nicolai *et al.* (2015) for inflectional realisation in sparse lexica.

More directly related to the present work is Brown and Evans (2012), who present an attempt at infering inflection classes for the system of Russian nouns. They evaluate redundancy between

[8] See Beniamine (2017) for a pattern inference algorithm capturing prefixation, suffixation, infixation, root-and-pattern morphology, and suprasegmental exponence, that could readily be used as a substitute for the simple algorithm used in this paper.

paradigms through a compression distance. They perform clustering on this basis using CompLearn (Cilibrasi and Vitanyi 2005). The output of CompLearn is an unrooted binary tree. Since this tree is hardly interpretable, Brown and Evans use a series of heuristics to select preferred nodes in the tree. Their approach does not rely on the abstraction of inflectional realisations. Since the compression distance is computed on forms, it captures as much, if not more, of the similarity between stems than the similarity of the inflectional material.It is then unclear whether the resulting tree encodes strictly inflectional structure. Since Brown and Evans (2012)'s goal is to validate an account of Russian noun inflection (Brown 1998), they are attempting to decide which heuristic yields an inflectional classification that is presupposed to be correct. If we do not rely on a pre-existing theory, we also lose the way to choose among such heuristics. In this paper, we thus wish to infer a partition of classes directly.

Bonami (2014) attempts to improve on Brown and Evans's (2012) strategy by inferring inflectional realisations as a separate step. He produces inflectional classification trees based on both affixes and alternation patterns, which corresponded broadly to our local and global segmentation strategies. The trees are built using distance-based agglomerative hierarchical clustering with average linkage (Sokal and Michener 1958). Unfortunately, the distances used for the alternation patterns and for the exponents are not commensurable. Moreover, the final shape of the inflectional system is a tree with no distinguished macroclass level. Indeed, since distances evaluate the fitness of one class, not the fitness of a partition, they are not an appropriate tool with which to choose a preferred partition of classes in the tree.

Lee and Goldsmith's (2013) approach is closest to ours. Starting from a representation of paradigms, they define a greedy clustering algorithm that uses the Minimal Description Length principle (Rissanen 1978) to decide which paradigms it is optimal to group together in a cell of the partition. Note that this is closely related to the use of MDL to compare inflection class systems (Sagot and Walther 2011; Walther and Sagot 2011; Walther 2013), for which see Section 1.3.5, but improves on it by using MDL as a criterion for clustering rather than using it to compare manually crafted classifications. However, Lee and Goldsmith's approach is marred by what we take as a poor choice of representation for paradigms. In their approach, paradigms

are collections of words, and words are represented by the set of characters in their orthographic forms. For instance, *delay* and *delayed* are represented by the same set {a,d,e,l,y}. This is unsatisfactory in many respects: such representations lack any plausibility as representations of the knowledge of speakers, and make it impossible to take into account important aspects of morphological structure. For instance, the character sets of *daring* ({a,d,g,i,n,r}) is closer to that of *denigrate* ({a,d,e,g,i,n,r,t}) than to that of *dare* ({a,d,e,r}).

The approach presented below can be seen as an attempt to combine ideas from Bonami on the use of alternation patterns to assess similarity between lexemes, and from Sagot and Walther and from Lee and Goldsmith on the use of the Minimal Description Principle as a criterion.

2.2.2 Our approach to inflection classes inference

Our goal is to infer a partition of macroclasses on top of microclasses directly. Doing so requires formal definitions of both of these constructs. We take microclasses to follow the strictest definition of inflection classes:

(5) A system of *microclasses* is a partition of the set of lexemes into classes which share the exact same list of inflectional realisations.

It follows that the microclasses can be transparently deduced from the inflectional realisation. We propose to define macroclasses as follows:

(6) A system of *macroclasses* is an optimal system of non-overlapping sets of microclasses.

To decide which partition is optimal, we now need a criterion to compare different partitions of a set of microclasses.

The leading idea is to look for the system of macroclasses that optimally captures the regularities in the data. Let's say we begin with a system of microclasses and wish to merge some of them into broader macroclasses. In the initial system, each microclass is described separately as having a list of patterns indexed by pairs of cells. Wherever merged microclasses have a common pattern, an optimal description will be able to mention that pattern only once by associating it with the

merged class. On the other hand, if merged classes use distinct patterns for the same cell, any description will need to disambiguate which microclass uses which pattern. Following Occam's razor, merging microclasses into a macroclass can then be seen as beneficial to concision as long as we gain more due to common patterns than we lose because of disambiguation. This follows the overall intuition of the Minimal Description Length Principle, according to which the structure best fitting a dataset is the structure allowing for the shortest description of the data. However, the reason we choose that structure is not that concision is a quality *per se*, but rather that it reflects the ability of the structure to account for regularities in the data. Thus, we decide that a partition of the set of lexemes in macroclasses is better than another one if it leads to a more concise description of the inflection class system.

In the next section, we present the probabilistic model that allows us to assess the length of a description, and the algorithm that makes use of this criterion to find the best macroclasses for a given set of microclasses.

3 FINDING AN OPTIMAL PARTITION

3.1 *The minimum description length principle*

Minimum Description Length (MDL) is a general framework for selecting an appropriate model of a dataset within a space of possible models (Rissanen 1984; Grünwald 2007). The underlying idea is that wherever there is structure in a dataset, that structure can be used to provide a shorter description of the dataset. Different models will capture the structure in the data to different extents. The quality of a model can thus be assessed by looking at the length of an optimal description of the data relying on the model. This will comprise both the description of the model itself, and a description of whatever aspects of the data the model was not able to describe. Optimality of the description is ensured in information-theoretic terms. The Minimal Description Length Principle then states that the best model is the model leading to the shortest description. This is supposed to embody Occam's razor: the best model is the most frugal model. For the MDL principle to make sense, it is essential that the models under consideration be strictly commensurable. MDL allows one to compare

different models written in the same formal framework, not all conceivable models, an endeavour that has been proved mathematically to be impossible.

The MDL is a general method for inductive inference, used mostly in the field of machine learning as a sound way of avoiding overfitting. In recent years, it has been used to address problems of linguistic modeling in morphology in two very different ways. As mentioned above, Sagot and Walther (2011, 2013) and Walther (2013) compare hand-designed descriptions of the same inflection system couched in the same rich formalism and use description length to decide which of these is preferable. Goldsmith (2001) then again explores automatically all possible morphological segmentations of a text (hence using a coarse-grained formalism for morphological description) and uses description length of the whole text to decide which segmentation is more likely to be correct.

In this paper, we adopt from Sagot and Walther the idea of using a description-length-based information-theoretic criterion for comparing competing accounts of a morphological system. However, we make use of this idea in a different setting; their approach, as Goldmsith's approach, is constructive in the sense of Blevins (2006); They are looking for the shortest possible grammar that generates the data within a predefined framework. This contrasts with the work reported in this paper, where we compare descriptions that are highly redundant. We make no claim that these descriptions are reasonable. We only claim that comparing them is useful to assess which set of macroclasses best represents regularities and irregularities in the data. Although this may be less familiar to linguists, this is actually the standard use of MDL in statistical inference, where descriptions are constructed for the purposes of comparing models, and do not necessarily have an inherent value.

3.2 *Modeling macroclass systems*

For the purposes of comparing inflection class systems, we thus need to define formally a family of models of inflection systems that differ in the way they group lexemes in classes, and then to assess their description length. The shape of the models we will use follows from the view of the inflectional macroclasses we argued for above. Lexemes are grouped in microclasses according to which patterns they instan-

tiate, a microclass being a class of lexemes that instantiate the exact same vector of patterns; macroclasses form a partition of the set of microclasses. A model of an inflection class system will contain the following four components:

(7) a. A specification M of which lexemes belong to which microclasses.

b. A specification C of which microclass belongs to which macroclass or CLUSTER of microclasses.

c. A specification $\mathscr{P}$ of which patterns (for each pair of paradigm cells) are instantiated in each cluster. Note that for any cluster containing more than one microclass, there will be at least one pair of cells for which two or more patterns are instantiated; otherwise there would only be one inflectional behavior and hence only one microclass in the cluster.

d. The residual information R that cannot be deduced from the assignment of a microclass to a cluster. This amounts to specifying, wherever a cluster instantiates more than one pattern for a pair of cells, which microclass in the cluster uses which pattern.

To better understand how such models can be used to compare candidate systems of macroclasses, let us consider a toy system consisting of the three French verbs AMENER 'bring', BOIRE 'drink' and DIRE 'say' in the indicative present plural. Table 8 indicates both the raw (sub)paradigms of the three verbs and the patterns abstracted from these paradigms under a local pattern inference strategy. The three verbs clearly belong to three different microclasses. Let us consider then in turn the three possible ways of grouping them into macroclasses. Table 9 provides an informal but rather detailed specification of the four components of a description of three possible classifications of this dataset. In each case, two of the three verbs are grouped together in a cluster, and the remaining third verb forms a cluster of its own.

As should be apparent from the table, the three candidate classifications do not differ in the length of a description of the assignments of lexemes to microclasses or microclasses to clusters. However they

Table 8: Subparadigms and local patterns for three French verbs in the Indicative Present Plural

	Raw data			Patterns (local strategy)		
	1PL	2PL	3PL	1PL∼2PL	1PL∼3PL	2PL∼3PL
AMENER	amənɔ̃	aməne	amɛn	…ɔ̃∼e (p_1)	…ə…ɔ̃∼…ɛ… (p_3)	…ə…e∼…ɛ… (p_6)
BOIRE	byvɔ̃	byve	bwav	…ɔ̃∼e (p_1)	…y…ɔ̃∼…wa… (p_4)	…y…e∼…wa… (p_7)
DIRE	dizɔ̃	dit	diz	…zɔ̃∼…t (p_2)	…ɔ̃∼… (p_5)	…t∼…z (p_8)

Table 9: Detailed description of three classifications of the paradigms from Table 8 in microclasses and macroclasses

Partition	{{AMENER},{BOIRE,DIRE}}	{{AMENER, BOIRE},{DIRE}}	{{AMENER,DIRE},{BOIRE}}
M	AMENER $\mapsto m_1$ BOIRE $\mapsto m_2$ DIRE $\mapsto m_3$	AMENER $\mapsto m_1$ BOIRE $\mapsto m_2$ DIRE $\mapsto m_3$	AMENER $\mapsto m_1$ BOIRE $\mapsto m_2$ DIRE $\mapsto m_3$
C	$m_1 \mapsto c_1$ $m_2 \mapsto c_2$ $m_3 \mapsto c_2$	$m_1 \mapsto c_1$ $m_2 \mapsto c_1$ $m_3 \mapsto c_2$	$m_1 \mapsto c_1$ $m_2 \mapsto c_2$ $m_2 \mapsto c_1$
$\mathscr{P}$	c_1 : 1PL ∼ 2PL : $\{p_1\}$ 1PL ∼ 3PL : $\{p_3\}$ 2PL ∼ 3PL : $\{p_6\}$ c_2 : 1PL ∼ 2PL : $\{p_1, p_2\}$ 1PL ∼ 3PL : $\{p_4, p_5\}$ 2PL ∼ 3PL : $\{p_7, p_8\}$	c_1 : 1PL ∼ 2PL : $\{p_1\}$ 1PL ∼ 3PL : $\{p_3, p_4\}$ 2PL ∼ 3PL : $\{p_6, p_7\}$ c_2 : 1PL ∼ 2PL : $\{p_2\}$ 1PL ∼ 3PL : $\{p_5\}$ 2PL ∼ 3PL : $\{p_8\}$	c_1 : 1PL ∼ 2PL : $\{p_1, p_2\}$ 1PL ∼ 3PL : $\{p_3, p_5\}$ 2PL ∼ 3PL : $\{p_6, p_8\}$ c_2 : 1PL ∼ 2PL : $\{p_1\}$ 1PL ∼ 3PL : $\{p_4\}$ 2PL ∼ 3PL : $\{p_7\}$
R	$m_2 : p_1$ $m_3 : p_2$ $m_2 : p_4$ $m_3 : p_5$ $m_2 : p_7$ $m_3 : p_8$	$m_1 : p_3$ $m_2 : p_4$ $m_1 : p_6$ $m_2 : p_7$	$m_1 : p_1$ $m_3 : p_2$ $m_1 : p_3$ $m_3 : p_5$ $m_1 : p_6$ $m_3 : p_8$

differ both in terms of assignment of patterns to clusters and in terms of residual information: because the second classification groups together two microclasses that share a pattern, the assignment of patterns to clusters is briefer (pattern p_1 is only mentioned once rather than twice), as is the residue (the clusters provide perfectly accurate information on 1PL ∼ 2PL, and hence the residue makes no mention of patterns p_1 and p_2). Hence the second classification, grouping together AMENER and BOIRE, leads to a shorter description and should be preferred over the other two.

Two more classifications have to be considered: a classification with only one macroclass, and one with one macroclass per microclass. These are illustrated in Table 10. In the first case, all of the

Table 10: Detailed description of two extreme classifications of the paradigms from Table 8 in microclasses and macroclasses

Partition	{{AMENER, BOIRE,DIRE}}	{{AMENER},{DIRE},{BOIRE}}
M (microclasses)	AMENER ↦ m_1 BOIRE ↦ m_2 DIRE ↦ m_3	AMENER ↦ m_1 BOIRE ↦ m_2 DIRE ↦ m_3
C (macroclasses)	$m_1 \mapsto c_1$ $m_2 \mapsto c_1$ $m_3 \mapsto c_1$	$m_1 \mapsto c_1$ $m_2 \mapsto c_2$ $m_3 \mapsto c_3$
$\mathscr{P}$ (patterns)	c_1 : 1PL ∼ 2PL : $\{p_1, p_2\}$ 1PL ∼ 3PL : $\{p_3, p_4, p_5\}$ 2PL ∼ 3PL : $\{p_6, p_7, p_8\}$	c_1 : 1PL ∼ 2PL : $\{p_1\}$ 1PL ∼ 3PL : $\{p_3\}$ 2PL ∼ 3PL : $\{p_6\}$ c_2 : 1PL ∼ 2PL : $\{p_1\}$ 1PL ∼ 3PL : $\{p_4\}$ 2PL ∼ 3PL : $\{p_7\}$ c_2 : 1PL ∼ 2PL : $\{p_2\}$ 1PL ∼ 3PL : $\{p_5\}$ 2PL ∼ 3PL : $\{p_8\}$
R (residue)	$m_1 : p_1$ $m_2 : p_1$ $m_3 : p_2$ $m_1 : p_3$ $m_2 : p_4$ $m_3 : p_5$ $m_1 : p_6$ $m_2 : p_7$ $m_3 : p_8$	

disambiguation is done in the residue, while in the second the same thing is done in the pattern assignment. The table gives the impression that the first description is longer, as it has both something in $\mathscr{P}$ and in R. However, it actually captures a generalisation that the other does not. In information-theoretic terms, it is a shorter description.

Going from this informal presentation to a precise measure of description length requires one to provide an explicit scheme for describing each of M, C, $\mathscr{P}$ and R as sequences of symbols. Any such sequence displays a probability distribution of the symbols via their relative frequency in the message. Information Theory (Shannon 1948) provides a way of determining the size in bits of the shortest possible encoding of that message.

Intuitively, this depends on the length of the message (all other things being equal, longer messages are longer to encode), and the frequency of the symbols within the message (symbols that occur multiple times in the message are less surprising and hence less costly). More precisely, the length of the shortest possible description of a message m is the length of message times the entropy of the distribution of the list S of symbols in the message.

$$\begin{aligned}(8)\qquad \mathrm{DL}(m) &= |m| \cdot H(m) \\ &= -|m| \cdot \textstyle\sum_{x \in S} P(x) \cdot \log_2 P(x) \\ &= -\textstyle\sum_{x \in S} \mathrm{count}(x) \cdot \log_2 \frac{\mathrm{count}(x)}{|m|}\end{aligned}$$

The appendix presents in detail the scheme we used in this paper. For present purposes, it is sufficient to note that we define the description length of an inflection system to be the sum of the description lengths of its four components.

$$(9)\qquad \mathrm{DL}(I) = \mathrm{DL}(M) + \mathrm{DL}(C) + \mathrm{DL}(\mathscr{P}) + \mathrm{DL}(R)$$

3.3 *Searching for possible partitions*

We can now define our criterion for deciding which of a set of partitions is optimal as minimisation of $\mathrm{DL}(I)$. Therefore, searching for the macroclasses could theoretically be a matter of evaluating all the possible partitions over the microclasses. This is not a realistic strategy in practice. For a system with 15 microclasses, there are more than a billion different partitions to consider. For a system such as French conjugation, with 74 microclasses, the number of partitions to consider

approaches the number of atoms in the universe (10^{80}).[9] The size of the search space entails that a full exploration of all possibilities is out of the picture. Here we use a greedy bottom-up search, which finds macroclasses from microclasses by merging repeatedly two clusters.

The algorithm can be described as follows:

(10) a. Start with a partition where each microclass is a cluster.
b. For each pair of clusters, evaluate what the DL of the system would be if the pair were to be merged.
c. Merge one of the pairs of clusters which results in a minimal DL.
d. Repeat steps (b-c) until the DL stops decreasing.

We exemplify the search with an imaginary system of five microclasses, named from A to E. Figure 1 illustrates how the algorithm proceeds. The numbers used here as description lengths are arbitrary and serve only the purpose of illustrating the algorithm.

Step (1) corresponds to the initial state, where each microclass forms its own cluster. Let us assume arbitrarily that the description length of the corresponding model is of 6 bits. In step (2), we select the pair of microclasses leading to the lowest DL. That is, we examine the 10 models obtained by putting any two microclasses in the same clusters, and pick the one whose description length is the smallest. In this instance it happens to be D and E, with a DL of 4.

We then proceed to determine again the optimal merges for the system constituting the output of step (2). In this instance, it happens that there are two optimal solutions: merging A and B or A and C both leads to models with a description length of 3.5. In such a situation, we choose one of the optimal solutions at random. Here the choice happens to be merging A and C.

[9] The number of possible partitions for a set of cardinality n is the n^{th} Bell number B_n, where $B_0 = 1$ and:

$$B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_n.$$

The Bell numbers grow very quickly—much more quickly than an exponential function, for instance.

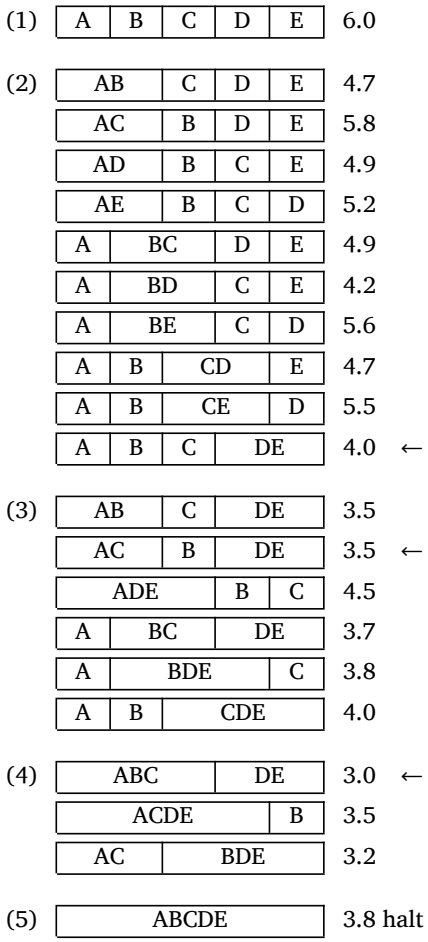


Figure 1:
Example of a run of the search algorithm

In step (4), we examine all possible merges and find that only one merge, ACB, leads to an optimal model. Finally, in step (5), we examine the result of merging the two only remaining macroclasses in a single cluster. This however leads to a description length that is longer than that of the optimal description found at step (4). This shows that merging clusters has stopped being beneficial for description length, and we conclude that the partition found at the end of step (4) is optimal.

Three important remarks about the algorithm are in order. First, there is no *a priori* guarantee that there will be several macroclasses. It is possible, if the DL continues to lower, to end up with only one

cluster. Thus this algorithm is suited to decide on an empirical basis if a system displays non-trivial macroclasses. Second, our algorithm is nondeterministic: at step (3) in the example above, we had to choose at random which classes to merge, which entails that a different choice might have taken us to a different final partition in two macroclasses. To address this issue, in the empirical studies below, we will perform multiple runs of the algorithm and check that the results are stable. Third, as with most greedy algorithms, we can only hope that the local optimum found by the algorithm indeed corresponds to the global optimum—and hence that the macroclasses we find are indeed the true macroclasses. While this is not fully satisfactory, we know of no search algorithm able to find the global optimum in a reasonable amount of time.

4 CLASSIFICATION AND RESULTS

In this section we discuss the results of applications of our algorithms to the conjugation of French and European Portuguese, and address three research questions: first, as we saw in the last section, not all datasets will lead to the emergence of a partition into macroclasses; the algorithm may terminate with the conclusion that introducing macroclasses does not lead to a more economical description. Given this, do macroclasses emerge in the systems at hand? Second, we introduced in section 2 two ways of describing inflectional realisations, relying on either a local or a global strategy. Where they emerge, how different are the macroclass systems found with both strategies? Third, how do the macroclass systems inferred by our algorithm compare to the systems posited by descriptive morphologists?

4.1 *Datasets*

Our datasets take the form of large inflectional lexica with phonemically transcribed forms.

For French, we rely on the verbal subset of the Flexique dataset (Bonami *et al.* 2014). It is based on the Lexique dataset (New *et al.* 2001), but the transcriptions have been corrected by hand, and the incomplete paradigms provided by Lexique have been filled semi-automatically. We ignore any defective or overabundant entries. The resulting dataset contains 5259 lexemes each containing forms for each of 51 morphosyntactic cells.

For European Portuguese, we rely on the European Portuguese pronunciation dictionary elaborated by Veiga *et al.* (2013), kindly provided and adapted by hand by Fernando Perdigão. The dataset contains 1995 lexemes, with forms for each of the 69 morphosyntactic cells. As was done for French, overabundance and defectivity are ignored.

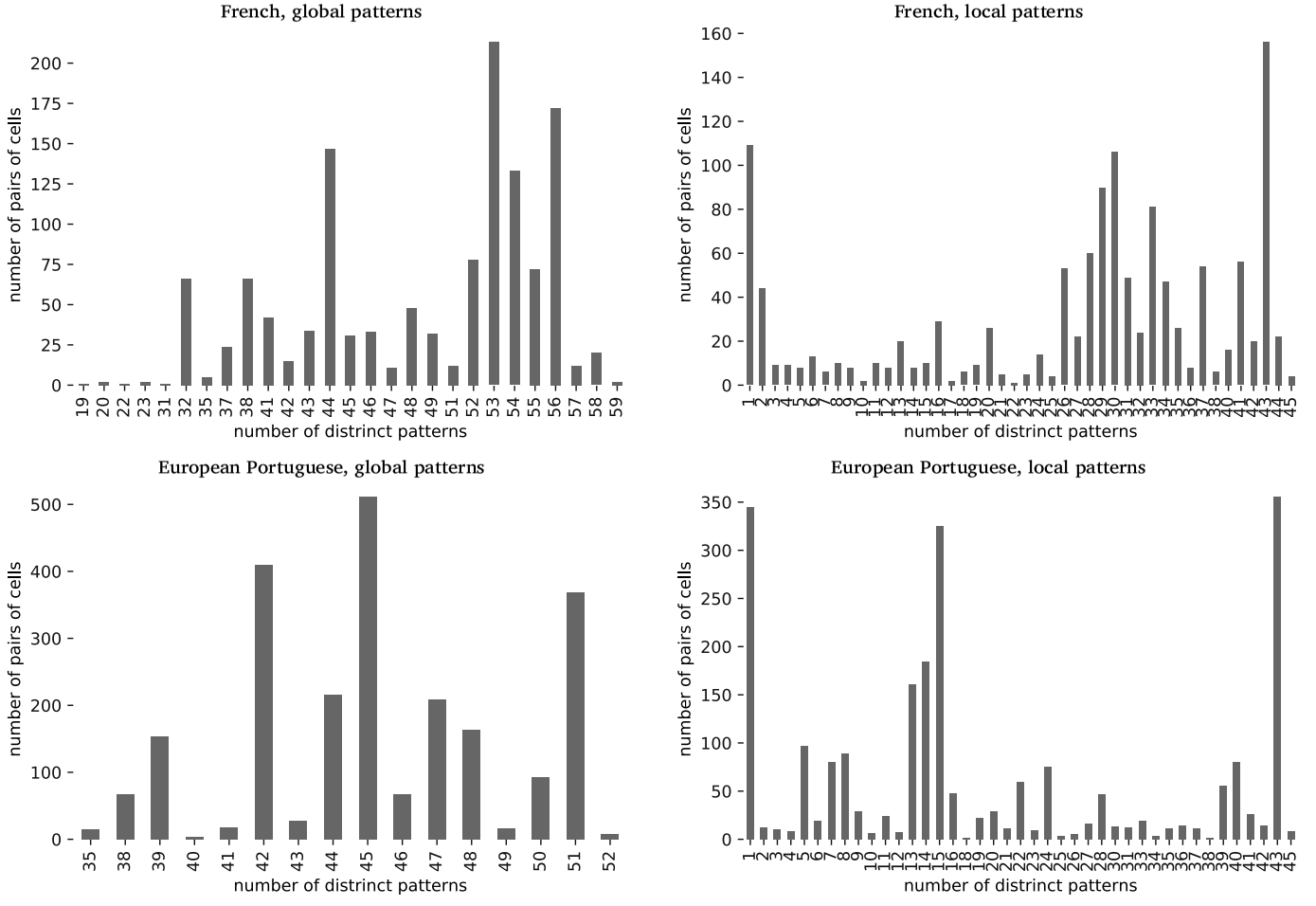
To compute phonological generalisations for on the context in which patterns are satisfied, our program also requires as input a specification of the value of each character as a vector of phonological features. We used the feature descriptions designed for the purposes of Bonami and Boyé (2014) and Bonami and Luís (2014).[10]

For both datasets, we ran the algorithm twice: once with alternation patterns found using the local strategy, and once with those found using the global strategy. The result consists of two classifications: first, that of lexemes into microclasses, then the classification of microclasses into macroclasses. The program also logs the history of the classification process as a tree of successive merges.

4.2 *Patterns*

The local and global alternation patterns differ substantially. As could be expected, the global approach results in a larger number of patterns per pair of cells, as is shown in Figure 2. This is due to the fact that any irregularity in the relation between two cells will have an impact on which patterns relate all other pairs of cells of that lexeme. For example, in a situation where two paradigm cells are identical for all lexemes, the local approach finds that generalisation, while the global approach may find more than one pattern depending on what happens elsewhere.

[10] One notable choice for the French dataset is that height distinctions between mid-vowels were neutralised by using the same feature matrices for the pairs of vowels ([e],[ɛ]), ([ø],[œ]), and ([o],[ɔ]). This is motivated by the fact that mid-vowel pronunciations are in a state of fluctuation in standard French in some positions, so that in some cases no single narrow transcription is appropriate for a given word. In examples below the neutralised vowels are noted E, Ø and O respectively.



These figures read as follows: In the French global setting, there are 213 pairs of cells which contain 53 distinct patterns, 172 pairs of cells which contain 56 distinct patterns, etc.

Figure 2: Distribution of the number of distinct patterns per pair of cells in each setting

4.3 *Microclasses*

Remember that microclasses are sets of lexemes exhibiting identical patterns for all pairs of cells. Even though the two strategies find very different patterns, in both languages, they lead to the same inventory of microclasses. This is a general property of the algorithm that is best explained by observing that two lexemes show an identical global inflectional behavior if and only if they show an identical behavior in each local context.

For French, we find 73 microclasses. The largest class contains verbs with the same inflectional behavior as AXER (3671), followed by the class of verbs behaving like AGIR (353). 60 of the classes have less than 20 members, with 15 having just one member. For European Portuguese, we find 55 microclasses, the largest of which contains verbs behaving like USAR (911), followed by that of verbs such as JOGAR (177). 43 microclasses present less than 20 members, with 15 having just one member.

Microclasses have little value as generalisations over inflectional behavior, because any small deviation between the behavior of two lexemes results in separate classes.

4.4 *Macroclasses*

Since microclasses with local and global patterns display different similarity structures, they also produce different macroclass systems. We ran the macroclass algorithm over the four different microclass systems (French and European Portuguese, local and global). The history of the algorithm can be depicted as a tree of recursive merges. Figures 3, 4, 5 and 6 show the history for both the local and global patterns. Black arcs represent merges where the description length decreased, gray arcs merges where the description length did not decrease; hence macroclasses are those clusters dominating black arcs and dominated by gray arcs. Nodes corresponding to a macroclass are labelled with the number of lexemes in the class.

We observe that with the global strategy, most microclasses do not cluster much together, while the local strategy leads to fewer macroclasses that seem more balanced. It is important to note here that the intermediate merges cannot be given a straightforward interpretation: their order does not necessarily reflect anything relevant,

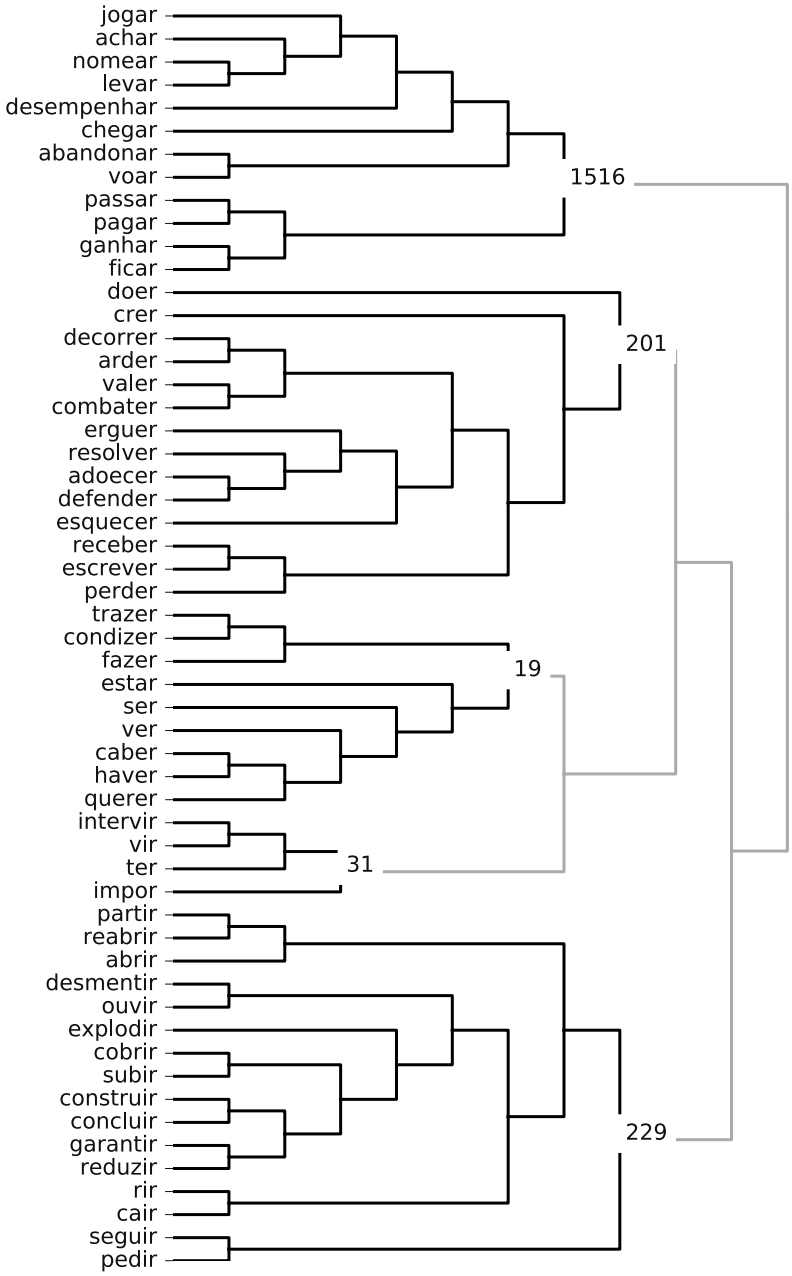


Figure 3: History of merges for European Portuguese macroclasses, on local patterns

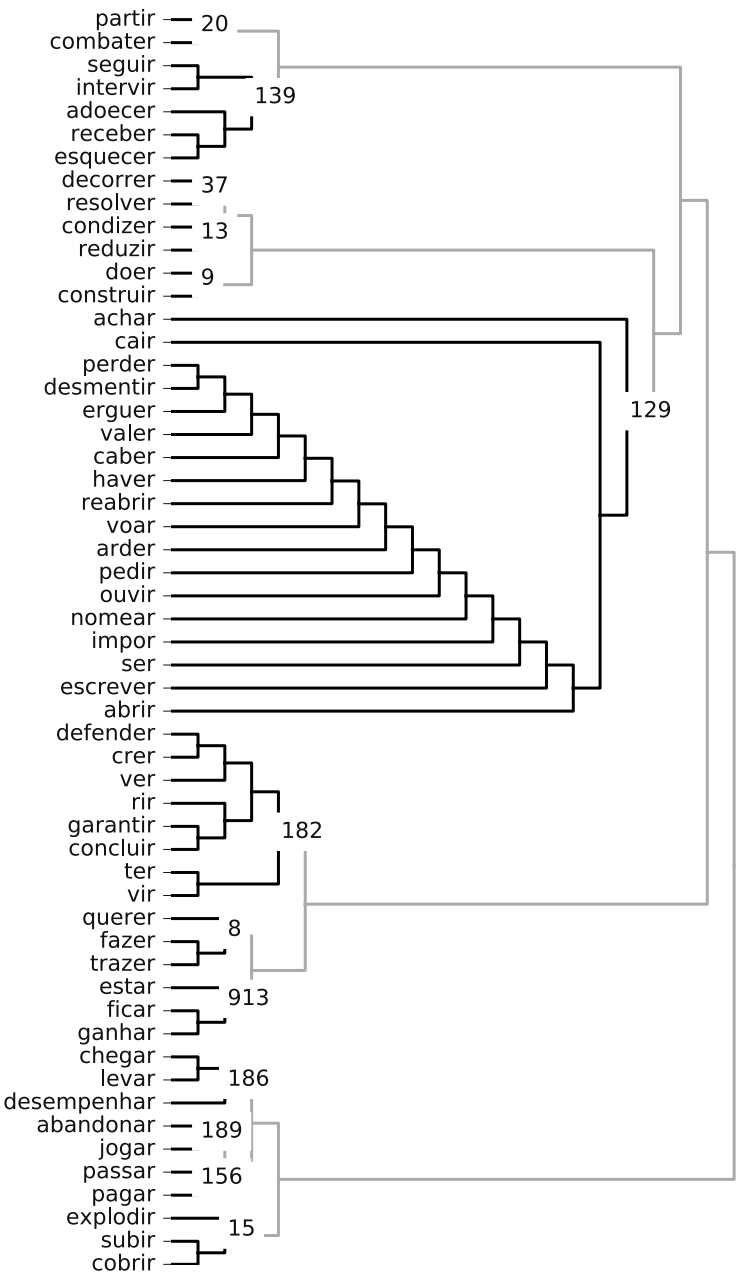


Figure 4: History of merges for European Portuguese macroclasses, on and global pattern.

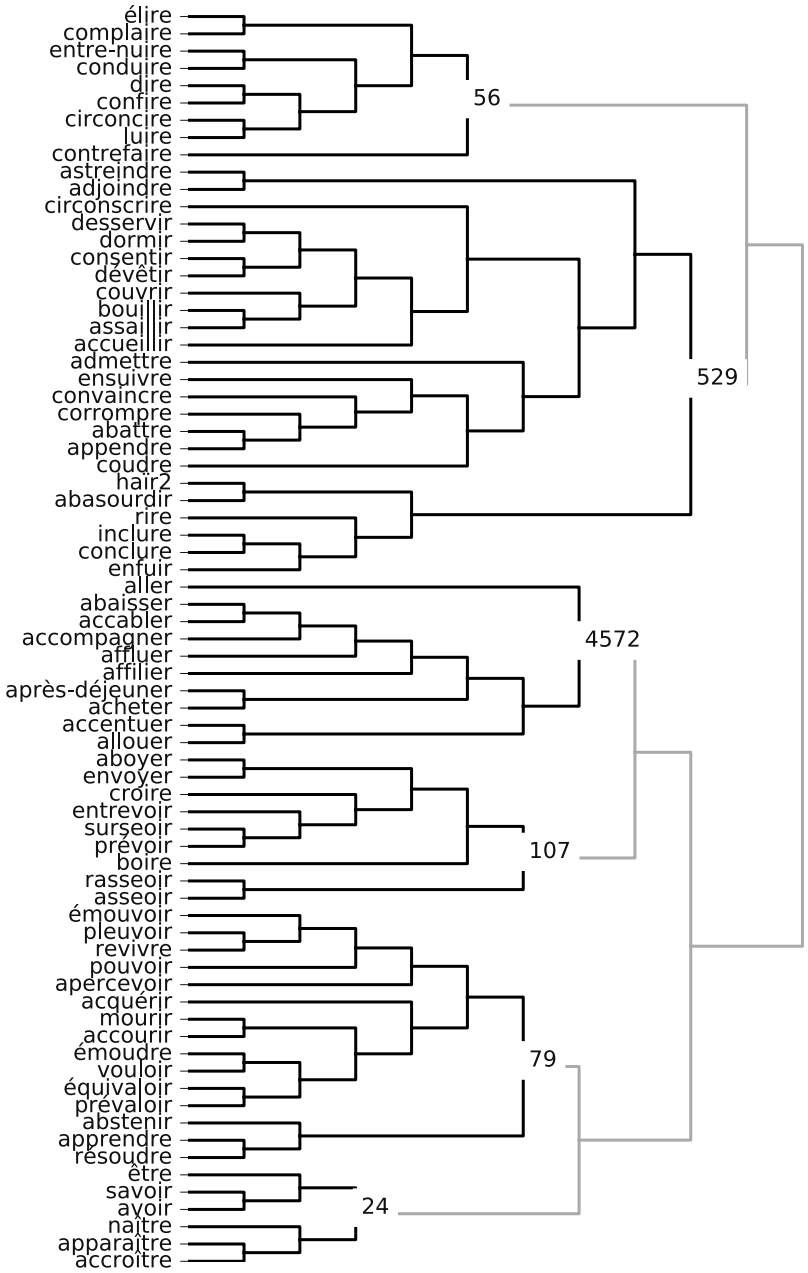


Figure 5: History of merges for European Portuguese macroclasses, on local patterns

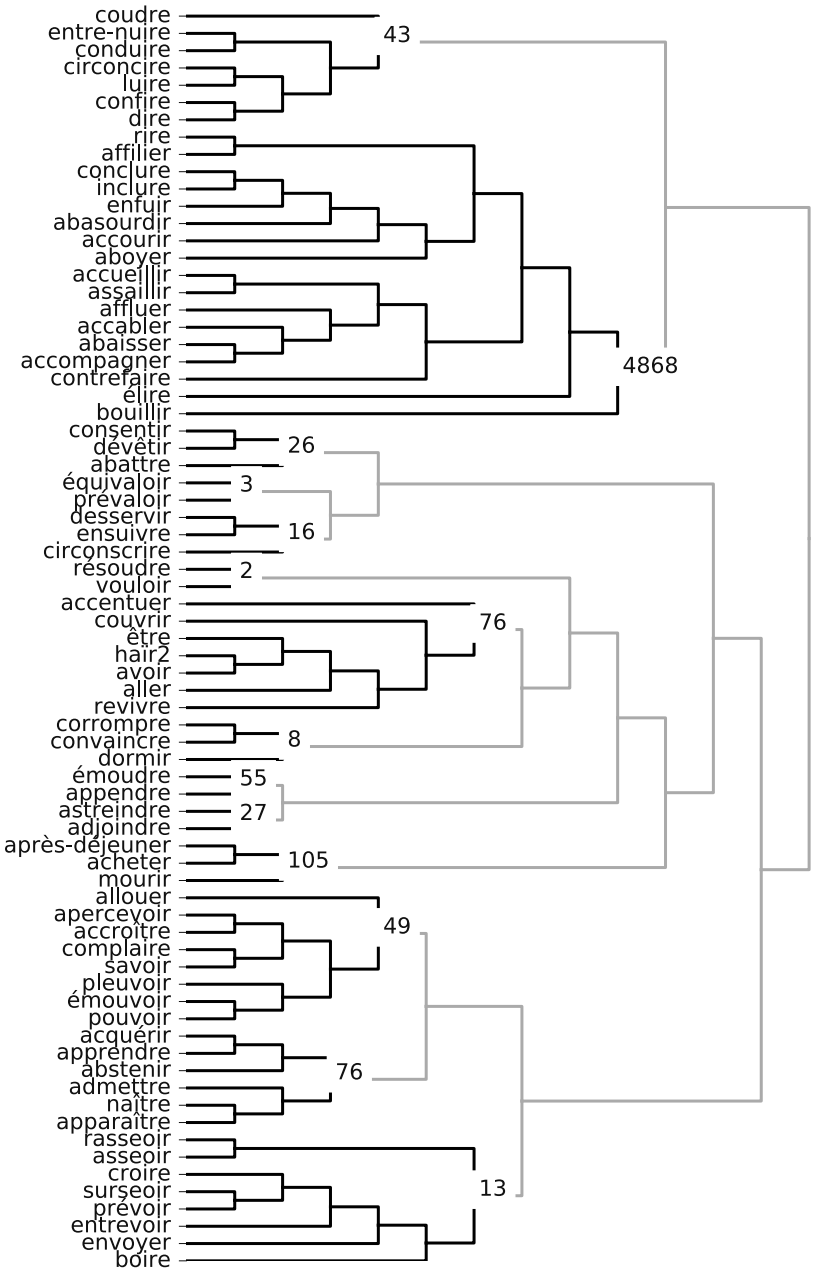


Figure 6: History of merges for European Portuguese macroclasses, on global patterns

and there is little reason to believe that they represent classes of intermediate granularity.

Remember that the greedy algorithm which we used to merge classes is nondeterministic: if we happen to encounter two competing best merges leading to the same decrease in DL, the algorithm chooses at random which to perform. To ensure the stability of our results despite this non-determinism, for each condition, we ran the classification procedure 100 times. The order of merges varied, especially at the beginning of runs, but the macroclass partition was constant over iterations. Figure 7 represents the intersection of 100 history trees for the French local patterns condition: if we consider each node as represented by the set of leaves it spans, and each edge as a pair of nodes, this history tree keeps only nodes and edges common to all 100 iterations, then adds edges (dashed in the figure) according to node spans to keep a tree structure. As can be seen on the picture, the areas of variation are small and localised at the bottom of the tree (the start of the algorithm). Results in the three other settings are similar.

In all settings, we do find non-trivial macroclasses: the clustering process stops before having merged all microclasses together. For European Portuguese verbs, we find 13 macroclasses with the global patterns and 5 with the local patterns. For French verbs, we find 14 macroclasses with the global patterns and 6 with the local patterns.

In neither condition did we find a bipartition between microclasses usually deemed regular and those that are usually deemed irregular. This suggests that a classification based on regularity and a classification based on similarity will be orthogonal to one another.

In both languages, the global strategy leads to classifications that contain numerous small macroclasses and bear no resemblance to extant classifications for these languages. Local patterns lead to fewer macroclasses, and generalisations are highly similar to traditional wisdom. This is clearly due to local patterns capturing more fine-grained similarity. We take this to suggest that our algorithm, applied under a local strategy to pattern inference, is close to operationalizing the heuristics used by descriptive linguists when designing a hand-made classification.

In French, the grammatical tradition distinguishes three conjugations. The first conjugation consists of verbs with infinitives in *-er*. The second conjugation consists of verbs with infinitives in *-ir* and exhibit-

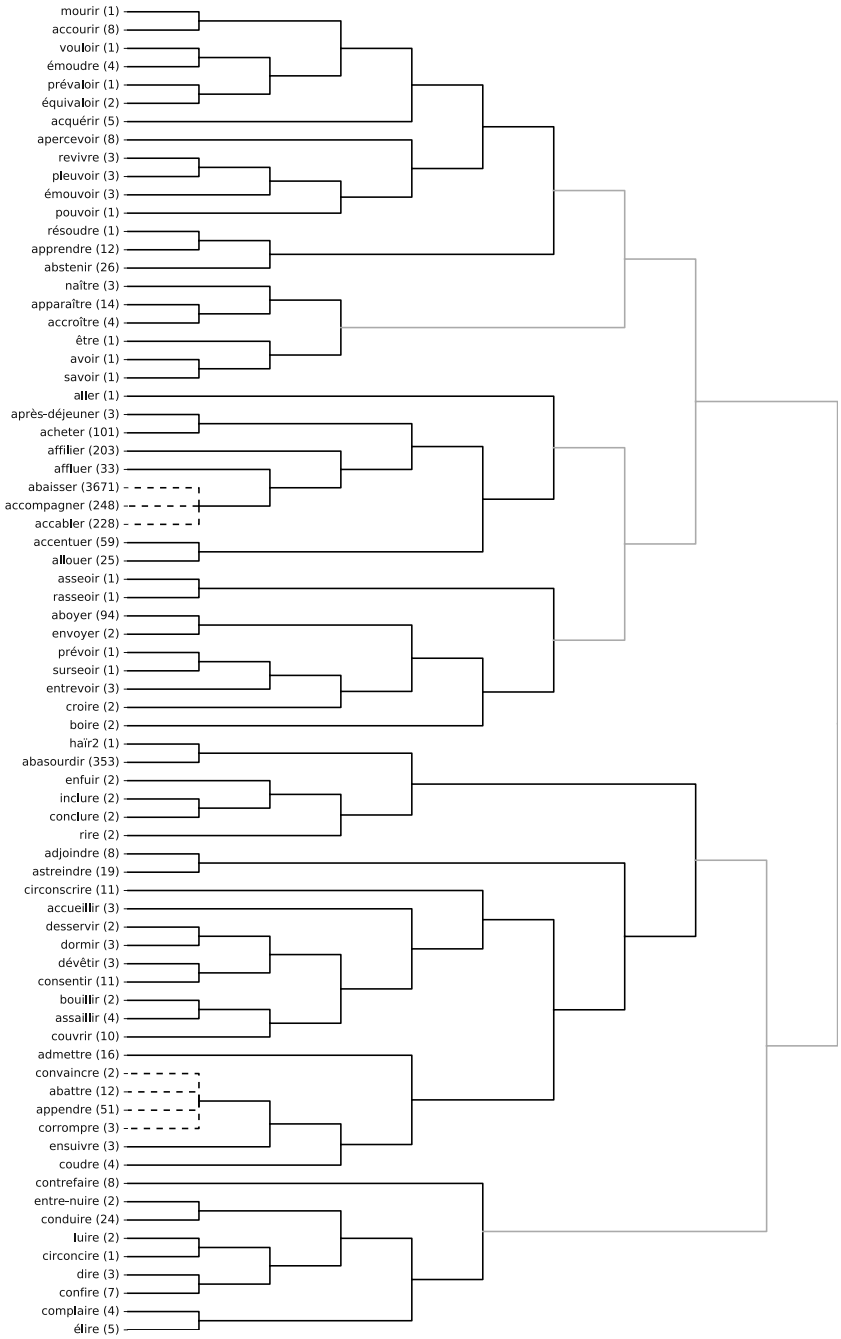


Figure 7: Intersection of the history of 100 runs for the French local patterns condition

Macroclass 1	résoudre (1), vouloir (1)
Macroclass 2	adjoindre (8), astreindre (19)
Macroclass 3	circoncire (1), conduire (24), confire (7), coudre (4), dire (3), entre-nuire (2), luire (2)
Macroclass 4	appendre (51), émoudre (4)
Macroclass 5	asseoir (1), boire (2), croire (2), entrevoir (3), envoyer (2), prévoir (1), rasseoir (1), surseoir (1)
Macroclass 6	convaincre (2), corrompre (3), dormir (3)
Macroclass 7	abstenir (26), acquérir (5), admettre (16), apparaître (14), apprendre (12), naître (3)
Macroclass 8	abaisser (3671), abasourdir (353), aboyer (94), accabler (228), accompagner (248), accourir (8), accueillir (3), affilier (203), affluer (33), assaillir (4), bouillir (2), conclure (2), contrefaire (8), enfuir (2), inclure (2), rire (2), élire (5)
Macroclass 9	accentuer (59), aller (1), avoir (1), couvrir (10), haïr2 (1), revivre (3), être (1)
Macroclass 10	accroître (4), allouer (25), apercevoir (8), complaire (4), pleuvoir (3), pouvoir (1), savoir (1), émouvoir (3)
Macroclass 11	abattre (12), consentir (11), dévêtir (3)
Macroclass 12	acheter (101), après-déjeuner (3), mourir (1)
Macroclass 13	prévaloir (1), équivaloir (2)
Macroclass 14	circonscrire (11), desservir (2), ensuivre (3)

Table 11: French macroclasses to microclass mapping (global strategy)

ing an *-i-/-iss-* stem alternation, while the third conjugation consists of all remaining verbs. Remember that Kilani-Schoch and Dressler (2005) take irregularity as a criterion in grouping the traditional second and third conjugations. See also Plénat (1987) for arguments to the effect that the second and third conjugation pattern together, at least in the formation of the simple past and past participle.

The simulations we ran both show that the traditional third conjugation is very heterogeneous, as its members always end up in different macroclasses. The global approach does not seem to capture the intuition of macroclass that has been described by linguists, showing 14 macroclasses, some of which contain a very small number of lexemes (Table 11), and none of which resembles by any stretch a traditional conjugation.

In contrast, the local strategy leads to a classification that is mostly congruent with the traditional approach (Table 12).

All verbs from the traditional first conjugation are clustered together, except ABOYER and ENVOYER, which indeed exhibit alterna-

Table 12: French: Comparison of inferred macroclasses (on local patterns) vs traditional conjugations

Macroclasses	Traditional	Lexemes
Macroclass 1	3rd conj.	circoncire (1), complaire (4), conduire (24), confire (7), contrefaire (8), dire (3), entrenuire (2), luire (2), élire (5)
Macroclass 2	3rd conj.	abstenir (26), accourir (8), acquérir (5), apercevoir (8), apprendre (12), mourir (1), pleuvoir (3), pouvoir (1), prévaloir (1), revivre (3), résoudre (1), vouloir (1), émoudre (4), émouvoir (3), équivaloir (2)
Macroclass 3	first conj.	abaisser (3671), accabler (228), accentuer (59), accompagner (248), acheter (101), affilier (203), affluer (33), allouer (25), après-déjeuner (3)
	3rd conj.	aller (1)
Macroclass 4	second conj.	abasourdir (353), haïr2 (1)
	3rd conj.	abattre (12), accueillir (3), adjoindre (8), admettre (16), appendre (51), assaillir (4), astreindre (19), bouillir (2), circonscrire (11), conclure (2), consentir (11), convaincre (2), corrompre (3), coudre (4), couvrir (10), desservir (2), dormir (3), dévêtir (3), enfuir (2), ensuivre (3), inclure (2), rire (2)
Macroclass 5	3rd conj.	accroître (4), apparaître (14), avoir (1), naître (3), savoir (1), être (1)
Macroclass 6	first conj.	aboyer (94), envoyer (2)
	3rd conj.	asseoir (1), boire (2), croire (2), entrevoir (3), prévoir (1), rasseoir (1), surseoir (1)

tions also found with some third conjugation verbs – but not in the infinitive.[11] The traditional second conjugation is so homogeneous that it is represented by only two microclasses, and their similarity with some verbs of the traditional third conjugation is large enough for them to cluster together. The verbs of the traditional third conjugation are split into different macroclasses, confirming that it has little internal homogeneity. Looking at the table, the clustering seems to be done on the basis of the infinitive ending. However, there was actually no primacy given to infinitive forms over any other in the evaluation

[11] The preference of our algorithm for this grouping is obviously due to the fact that there are many pairs of cells exhibiting a Xwa∼Xwaj alternation, while fewer pairs of cells exhibit alternations typical of the first conjugation.

Macroclasses	Traditionnal	Lexemes
Macroclass 1	first conj.	abandonar (12), achar (3), chegar (20), desempenhar (4), ficar (911), ganhar (1), jogar (177), levar (162), nomear (53), pagar (1), passar (155), voar (17)
Macroclass 2	second conj.	adoecer (1), arder (1), combater (11), crer (2), decorrer (30), defender (42), doer (3), erguer (1), escrever (8), esquecer (2), perder (1), receber (90), resolver (7), valer (2)
Macroclass 4	3rd conj.	abrir (1), cair (11), cobrir (3), concluir (28), construir (6), desmentir (5), explodir (3), garantir (90), ouvir (1), partir (9), pedir (4), reabrir (1), reduzir (11), rir (2), seguir (45), subir (9)
Macroclass 4	second conj.	impor (17), ter (9)
	3rd conj.	intervir (1), vir (4)
Macroclass 5	first conj.	estar (1)
	second conj.	caber (1), condizer (2), fazer (5), haver (1), querer (2), ser (1), trazer (1), ver (5)

Table 13: Portuguese: Comparison of inferred macroclasses (local strategy) vs traditional conjugations

of inflectional behavior. In light of this classification, it seems that the local strategy does lead to a kind of inflectional classification close to that produced by descriptive morphologists, while diverging in terms of details from the extant standard classification by highlighting previously overlooked similarities between microclasses that are prevalent enough in paradigms to emerge as classificatory criteria.

The picture is similar for European Portuguese. The traditional account distinguishes between three conjugations based on the infinitive. The global strategy finds 13 macroclasses with little relation to the traditional classification. The local strategy finds five macroclasses, whose content is detailed in Table 13. The first three classes clearly match the three traditional conjugations, with characteristic theme vowels in *-a*, *-e*, and *-i*. The two remaining classes are not coherent in terms of theme vowels but have other notable properties. Macroclass 4 groups verbs with a stem alternant in *-ɲ* in the indicative past imperfective, in the subjunctive, and in the present indicative 1SG. This leads to a distinctive set of alternations that sets them apart from all other macroclasses, and has a stronger effect on classification than the theme vowel, which may be *-o*, *-e* or *-i*. Macroclass 5 groups

together a set of highly irregular verbs, and exhibits maximal dissimilarity for a cluster of such a small size (19 microclasses). There is no single reason for these microclasses to be grouped together, but there is definitely no strong reason as to why they should be placed somewhere else: all of them strongly depart from regular conjugations in one way or another.

All in all, then, we observe that, under the local strategy, our algorithm produces a classification that is strongly congruent with conventional practice, and highly defensible from a linguist's perspective, while being immune to some biases of grammatical tradition, such as that of giving stronger weight to citation forms than to other paradigm cells in deciding what should be grouped together.

5 CONCLUSION

This paper has presented a method for inferring inflection classes that captures crucial intuitions and heuristics used by descriptive linguists while being entirely systematic and unambiguously applicable to any system. Our modelling strategy is computational: we start from a few leading ideas on inflectional classification and propose a computational implementation of these ideas.

We started from a distinction between inflectional microclasses and macroclasses. A system of microclasses is based on identity of inflectional behavior across lexemes: two lexemes belong to the same microclass if and only if they exhibit exactly the same alternations. A system of macroclasses groups together microclasses exhibiting *similar* rather than *identical* behavior. Since similarity is gradual and multidimensional, there is no single agreed upon strategy to choose an appropriate system of macroclasses. Many authors rely on criteria such as productivity or regularity to that effect. We proposed to ground the choice of macroclasses solely in the direct examination of paradigms of surface forms. How such a form-based classification correlates with other forms of classification is an empirical question that is best addressed once the form-based generalizations are known.

With this goal in mind, we presented an algorithm that builds on the Minimum Description Length principle to explore partitions of the set of lexemes into classes. The underlying idea is that the optimal set of macroclasses for a system is the set that leads to the most compact

description of the system; this captures the intuition that macroclasses should help the linguist or language learner by minimizing the quantity of rote learning necessary to make sense of the system.

The algorithm was applied to two datasets of French and European Portuguese conjugation, under two different strategies for representing inflectional behavior: under the local strategy, inflectional behaviour is modelled by examining pairwise similarities and differences between paradigm cells of a lexeme, while under the global strategy, it is modelled by examining the similarities and differences that hold for the whole paradigm at once.

We find that the local segmentation better captures paradigmatic structure, and produces macroclass systems that resemble those elaborated by grammatical traditions. However, we also identify some previously unidentified macroclasses. We consider the differences between our classifications and those found in the literature to be attributable to a more principled view of classification. First, we confirm that unproductive and/or irregular microclasses do not cluster together in terms of formal similarity, and hence that grouping them together, as is usual in the French tradition, is unwarranted. Second, our model does not give any privileged status to the citation form, unlike what is usually done: hence the infinitive plays no privileged role in classification. Hence inflectional characteristics that are transparent from the infinitive form, such as theme vowels, play a role in the classification only inasmuch as they result in distinct alternation patterns. Third and finally, the implemented model is able to take into account all similarities and differences between all paradigm cells among dozens of macroclasses, a task whose manual execution is not feasable. This allows previously unobserved patterns of similarity to emerge.

We make no claim as to the importance of inflectional macroclasses as an analitycal tool. Our goal was rather to establish that it is possible to devise a systematic method of inference of macroclasses from raw paradigms. Of course, a partition of the lexicon into a small set of clusters of lexemes with similar behavior is one among a variety of ways one may approach the structure of an inflectional system; the fact that is has a longstanding tradition as a pedagogical tool is not reason enough not to explore alternative forms of classification. Beniamine and Bonami (2016) is an initial attempt at inferring from surface patterns lattice-shaped classifications such as those familiar

from Network Morphology (Brown and Hippisley 2012) and HPSG approaches to morphology (Bonami and Crysmann 2016).

APPENDIX — DESCRIPTIONS OF INFLECTION SYSTEMS

This appendix presents in some detail the class of inflection system models on which we rely for macroclass inference and description length assessment.

As mentioned in the Section 2.2.2, we are not interested in finding the shortest possible description, but rather in finding the way of clustering microclasses into macroclasses that produces the largest decrease in description length. Therefore, we only need to compute the contribution to the overall description length of those parts of the description that vary when the set of macroclasses varies. The description of the set of microclasses will be constant over all possible clustering of microclasses for a given system. We include it nevertheless in the description of the inflection class system so as to be able to compare different descriptions of the same system that use different strategies for alternation pattern inference, e.g. a global or a local strategy.

The description length we define below does not take into account the number of bits needed to declare each patterns and lexemes, the name of the cells and their pairing, the contexts in which patterns apply,[12] and the description of the procedure to decode the data. None of this will vary across competing partitions, so none of it is useful to us in selecting a partition.

Following Sagot and Walther (2011), we decompose the overall description length into a number of terms, each of which encoding a distinct part of the description. We define the description length of a given description D of an inflectional system as the sum of the description lengths of the four following components, which we briefly define below: microclasses, clusters, patterns and residue:

$$\mathrm{DL}(D) = \mathrm{DL}_M(D) + \mathrm{DL}_C(D) + \mathrm{DL}_P(D) + \mathrm{DL}_R(D).$$

[12] These contexts have been replaced by placeholders when abstracting patterns, but they could be stored and generalised as in (Bonami and Beniamine 2015), and the classes of applicability could be taken into account in the residual information (for which see below).

In the remainder of this appendix, we shall use the system presented in Section 3.1, Tables 8, 9 and 10 as a running example. Diagrams and explicit descriptions correspond to the description $D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}$, which relies on the partition {{AMENER, BOIRE},{DIRE}} of the set of microclasses.

A.1 *Mapping microclasses to lexemes*

We define $\mathrm{DL}_M(D)$ as the minimum number of bits needed to describe the mapping between lexemes and microclasses in description D. If we suppose that the set of lexemes $\mathscr{L}$ is ordered in a predefined way, such a mapping can be simply expressed as a list of $|\mathscr{L}|$ microclass identifiers that is parallel to the list of $|\mathscr{L}|$ lexemes.

Let us call $\mathscr{M}$ the set of microclass identifiers. If we define $\mathrm{occ}(m)$ as the number occurrences of a given microclass identifier $m \in \mathscr{M}$, the description length $\mathrm{DL}_M(D)$ of the "microclasses" section of the description D can be defined as follows:

$$\begin{aligned}\mathrm{DL}_M(D) &= -|\mathscr{L}| \cdot \sum_{m \in \mathscr{M}} \frac{\mathrm{occ}(m)}{|\mathscr{L}|} \cdot \log_2 \frac{\mathrm{occ}(m)}{|\mathscr{L}|} \\ &= -\sum_{m \in \mathscr{M}} \mathrm{occ}(m) \cdot \log_2 \frac{\mathrm{occ}(m)}{|\mathscr{L}|}.\end{aligned}$$

Applying this definition to our running example, which contains three microclasses occurring once each, we obtain:

$$\mathrm{DL}_M\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right) - 3 \log_2 \frac{1}{3} \approx 4.75$$

A.2 *Mapping microclasses to microclass clusters*

We can also assume that the set $\mathscr{M}$ of microclasses is associated with a predefined order. We can then express the mapping from microclasses to microclass clusters by simply listing microclass cluster identifiers following the same order (the i-th cluster identifier will indicate the cluster to which the i-th microclass belongs).

In a parallel way to the above, and defining the set of microclass clusters as $\mathscr{C}$, we can then write:

$$\mathrm{DL}_C(D) = -\sum_{c \in \mathscr{C}} \mathrm{occ}(c) \cdot \log_2 \frac{\mathrm{occ}(c)}{|\mathscr{M}|}.$$

Note that the number of occurrences $\mathrm{occ}(c)$ of a cluster $c \in \mathscr{C}$ in the "clusters" part of the description corresponds to its size, i.e. the number of microclasses it contains.

Applying this definition to our running example, in which one cluster appears twice and the other appears only one time, we obtain:

$$\begin{aligned}\mathrm{DL}_C\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right) &= -2\log_2\frac{2}{3} - \log_2\frac{1}{3}\\ &\approx 2.75\end{aligned}$$

Note that this result also holds for the other two partitions, the distribution of clusters is the same:

$$\begin{aligned}\mathrm{DL}_C\left(D_{\{\{\text{AMENER}\},\{\text{BOIRE,DIRE}\}\}}\right) &= \mathrm{DL}_C\left(D_{\{\{\text{AMENER,DIRE}\},\{\text{BOIRE}\}\}}\right)\\ &= \mathrm{DL}_C\left(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}\right)\end{aligned}$$

DL_C is lower in descriptions with fewer, larger clusters, as less information is required for selecting the right cluster for each microclass. The extreme case is when there is only one cluster. In this case, the probability of this cluster is 1 and the corresponding value for DL_C is 0. Conversely, DL_C is higher when there are many smaller clusters:

$$\mathrm{DL}_C\left(D_{\{\{\text{AMENER, BOIRE,DIRE}\}\}}\right) = -3\log_2\frac{3}{3} = 0$$

$$\mathrm{DL}_C\left(D_{\{\{\text{AMENER}\},\{\text{DIRE}\},\{\text{BOIRE}\}\}}\right) = -3\log_2\frac{1}{3} \approx 4.75$$

A.3 *Relation between patterns and clusters*

For each pair of cells in the paradigm, the description associates clusters with alternation patterns used by lexemes in this cluster. This relation is not a function: several patterns can appear in a cluster, and several clusters can make use of a same pattern.

Let us call $\mathscr{K}$ the set of paradigm cells. $\mathscr{K}^2$ is then set of all n cell pairs, which we can assume is associated with a predefined order $\mathbf{k_1} \prec \mathbf{k_2} \prec \ldots \prec \mathbf{k_n}$. Let us refer to the set of alternation patterns identifiers as $\mathscr{P}$. The relation between patterns and clusters can then be encoded in the form of a sequence of pairs of the form (c, p), where $c \in \mathscr{C}$ is a cluster identifier and $p \in \mathscr{P}$ is an alternation pattern identifier. More precisely, since $\mathscr{C}$ is also supposed to be associated with a total

order, the relation between patterns and clusters can be provided as follows: first, all pairs (c, p) for the first cell pair $\mathbf{k_1}$ can be provided, ordered according to the cluster it includes; next, all pairs for $\mathbf{k_2}$ can be provided; the shift from $\mathbf{k_1}$ pairs to $\mathbf{k_2}$ pairs is visible because the cluster in the last $\mathbf{k_1}$ pair is the last cluster in (ordered) $\mathscr{C}$, whereas the cluster in the first $\mathbf{k_2}$ pair is the first cluster in $\mathscr{C}$; we then resume with $\mathbf{k_3}$ pairs, and so on.

Let us decompose $\mathrm{DL}_p(D)$ into the contribution $\mathrm{DL}_{p_c}(D)$ of cluster identifiers and the contribution $\mathrm{DL}_{p_p}(D)$ of pattern identifiers. Let us call $\mathrm{occ}_{\mathbf{k}}(c)$ (resp. $\mathrm{occ}_{\mathbf{k}}(p)$) the number of occurrences of a given cluster c (resp. of a given pattern p) in pairs of the form (c, p) associated with a given cell pair $\mathbf{k} \in \mathscr{K}$. Let us note N the total number of pairs of the form (c, p), i.e. $N = \sum_{c' \in \mathscr{C}} \mathrm{occ}_{\mathbf{k}}(c') = \sum_{p' \in \mathscr{P}} \mathrm{occ}_{\mathbf{k}}(p')$. The probability of occurrence of a given cluster identifier $c \in \mathscr{C}$ is then:

$$\begin{aligned} P(c) &= \sum_{\mathbf{k} \in \mathscr{K}^2} \frac{\mathrm{occ}_{\mathbf{k}}(c)}{\sum_{c' \in \mathscr{C}} \mathrm{occ}_{\mathbf{k}}(c')} \\ &= \frac{1}{N} \sum_{\mathbf{k} \in \mathscr{K}^2} \mathrm{occ}_{\mathbf{k}}(c). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{DL}_{p_c}(D) &= -N \sum_{c \in \mathscr{C}} P(c) \cdot \log_2 P(c) \\ &= -\sum_{c \in \mathscr{C}} \sum_{\mathbf{k} \in \mathscr{K}^2} \mathrm{occ}_{\mathbf{k}}(c) \cdot \log_2 \frac{\mathrm{occ}_{\mathbf{k}}(c)}{N} \end{aligned}$$

Similarly, he probability of occurrence of a given pattern identifier $p \in \mathscr{P}$ is:

$$\begin{aligned} P(p) &= \sum_{\mathbf{k} \in \mathscr{K}^2} \frac{\mathrm{occ}_{\mathbf{k}}(p)}{\sum_{p' \in \mathscr{P}} \mathrm{occ}_{\mathbf{k}}(p')} \\ &= \frac{1}{N} \sum_{\mathbf{k} \in \mathscr{K}^2} \mathrm{occ}_{\mathbf{k}}(p). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{DL}_{p_p}(D) &= -N \sum_{p \in \mathscr{P}} P(p) \cdot \log_2 P(p) \\ &= -\sum_{p \in \mathscr{P}} \sum_{\mathbf{k} \in \mathscr{K}^2} \mathrm{occ}_{\mathbf{k}}(p) \cdot \log_2 \frac{\mathrm{occ}_{\mathbf{k}}(p)}{N} \end{aligned}$$

The description length $DL_P(D) = DL_{P_c}(D) + DL_{P_p}(D)$ of the "patterns" section of the description can then be computed as:

$$DL_P(D) = \\ -\sum_{\mathbf{k}\in\mathscr{K}^2}\left(\sum_{c\in\mathscr{C}}\mathrm{occ}_{\mathbf{k}}(c)\cdot\log_2\frac{\mathrm{occ}_{\mathbf{k}}(c)}{N} + \sum_{p\in\mathscr{P}}\mathrm{occ}_{\mathbf{k}}(p)\cdot\log_2\frac{\mathrm{occ}_{\mathbf{k}}(p)}{N}\right)$$

Applying this definition to our running example, we obtain:

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\textsc{amener, boire}\},\{\textsc{dire}\}\}}\right) = &-2\log_2\frac{1}{2}\\ &-2\log_2\frac{1}{2}\\ &-3\log_2\frac{1}{3}\\ &-\log_2\frac{1}{3} - 2\log_2\frac{2}{3}\\ &-3\log_2\frac{1}{3}\\ &-\log_2\frac{1}{3} - 2\log_2\frac{2}{3}\\ \approx\ &14.26\end{aligned}$$

In the same fashion, we have:

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\textsc{amener}\},\ \{\textsc{boire,dire}\}\}}\right) &= \mathrm{DL}_P\left(D_{\{\{\textsc{amener,dire}\},\{\textsc{boire}\}\}}\right)\\ &= -10\log_2\frac{1}{3} - 8\log_2\frac{2}{3}\\ &\approx 20.52\end{aligned}$$

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\textsc{amener,boire,dire}\}\}}\right) &= -2\log_2\frac{1}{2} - 6\log_2\frac{1}{3}\\ &\approx 11.5\end{aligned}$$

$$\begin{aligned}\mathrm{DL}_P\left(D_{\{\{\textsc{amener}\},\{\textsc{boire}\},\{\textsc{dire}\}\}}\right) &= -16\log_2\frac{1}{3} - 2\log_2\frac{2}{3}\\ &\approx 26.52\end{aligned}$$

Unsurprisingly, the most efficient way to assign patterns to clusters is to have only one cluster, and the worst is to have as many clusters as microclasses.

A.4 *Residual ambiguity*

Since a cluster can be associated with several patterns for a same pair of cells, clustering can produce ambiguity. A complete description has to account for the information needed to disambiguate such ambiguities. As for the patterns, the necessary residual information is dispatched over each pair of cells. As it is internal to each cluster, it also has to be repeated for each of them.

Given a microclass cluster identifier $c \in \mathscr{C}$ and a pair of cells $\mathbf{k} \in \mathscr{K}^2$, the corresponding residual information is provided in the form of a set of pairs of the form (m, p), where $m \in \mathscr{M}$: such a pair means that the microclass m follows pattern p on cell pair $\mathbf{k}$. Of course, only those microclasses that belong to the cluster (identified by) c can and should be included. Since the list of microclasses included in c is a piece of information that has been already taken into account, and since microclasses have been ordered, the residual information of a given cluster c and a given cell pair $\mathbf{k} \in \mathscr{K}^2$ can be given in the form of a simple list of patterns, one for each microclass included in c, in the correct order. In such a list, each pattern p will occur with a probability $\mathrm{occ}^c_{\mathbf{k}}(p)/\mathrm{occ}(c)$, where $\mathrm{occ}^c_{\mathbf{k}}(p)$ is the number of microclasses in c that use pattern p for cell pair $\mathbf{k}$. We call $\mathscr{P}_{\mathbf{k}}(c)$ the set of patterns that are used my at least one microclass in cluster c for cell pair $\mathbf{k}$.

As a result:

$$DL(R) = \sum_{c \in \mathscr{C}} \sum_{\mathbf{k} \in \mathscr{K}^2} \sum_{p \in \mathscr{P}_{\mathbf{k}}(c)} \mathrm{occ}^c_{\mathbf{k}}(p) \cdot \log \frac{\mathrm{occ}^c_{\mathbf{k}}(p)}{\mathrm{occ}(c)}.$$

In the example above, in the first cluster, for each of the two ambiguous cells, each of the two patterns happens for only one microclass.

$$\mathrm{DL}_R(D_{\{\{\text{AMENER, BOIRE}\},\{\text{DIRE}\}\}}) = -4 \log_2 \frac{1}{2} = 4.$$

We also have:

$$\begin{aligned}\mathrm{DL}_R(D_{\{\{\text{AMENER}\},\ \{\text{BOIRE,DIRE}\}\}}) &= \mathrm{DL}_R(D_{\{\{\text{AMENER,DIRE}\},\{\text{BOIRE}\}\}}) \\ &= -6 \log_2 \frac{1}{2} = 6\end{aligned}$$

$$\mathrm{DL}_R(D_{\{\{\text{AMENER,BOIRE,DIRE}\}\}}) = -2 \log_2 \frac{2}{3} - 7 \log_2 \frac{1}{3} \approx 12.26$$

$$\mathrm{DL}_R(D_{\{\{\text{AMENER}\},\{\text{BOIRE}\},\{\text{DIRE}\}\}}) = 0$$

Unsurprisingly, while clustering maximally tends to decrease DL_P, it tends to increase ambiguity and thus DL_R, while having smaller clusters leads to less ambiguity, thus a smaller $DL(R)$. In minimizing the total description length, we seek an balance between these measures.

We can now gather all the partial DLs in Table 14 to compare each classification and recognise {{AMENER, BOIRE},{DIRE}} as the best partition according to description length.

Table 14: Description lengths for all the possible classifications of Table 8 in microclasses and macroclasses

Partition	$DL(M)$	$DL(C)$	$DL(\mathscr{P})$	$DL(R)$	total DL
{{AMENER}, {BOIRE,DIRE}}	4.75	2.75	20.52	6	34.01
{{AMENER, BOIRE},{DIRE}}	4.75	2.75	14.26	4	25.75
{{AMENER, DIRE},{BOIRE}}	4.75	2.75	20.52	6	34.01
{{AMENER, BOIRE,DIRE}}	4.75	0	11.50	12.26	28.5
{{AMENER}, {DIRE},{BOIRE}}	4.75	4.75	26.52	0	36.01

REFERENCES

Farrell ACKERMAN, James P. BLEVINS, and Robert MALOUF (2009), Parts and wholes: implicative patterns in inflectional paradigms, in James P. BLEVINS and Juliette BLEVINS, editors, *Analogy in Grammar*, pp. 54–82, Oxford University Press, Oxford.

Farrell ACKERMAN and Robert MALOUF (2013), Morphological organization: The low conditional entropy conjecture., *Language*, 89(3):429–464, doi:10.1353/lan.2013.0054.

Malin AHLBERG, Markus FORSBERG, and Manstio HULDEN (2014), Semi-supervised learning of morphological paradigms and lexicons, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden 26–30 April 2014*, pp. 569–578, ISBN 978-1-937284-78-7, doi:10.3115/v1/E14-1060.

Adam ALBRIGHT and Bruce HAYES (2003), Rules vs. analogy in English past tenses: A computational/experimental study, *Cognition*, 90:119–161, doi:10.1016/S0010-0277(03)00146-X.

Adam ALBRIGHT and Bruce HAYES (2006), Modeling productivity with the gradual learning algorithm: The problem of accidentally exceptionless generalizations, *Gradience in grammar: Generative perspectives*, pp. 185–204.

Mark ARONOFF (1994), *Morphology by Itself: Stems and Inflectional Classes*, Linguistic inquiry monographs, MIT Press, ISBN 9780262510721.

Sacha BENIAMINE (2017), Une approche universelle pour l'abstraction automatique d'alternances morphophonologiques, in *Traitement Automatique des Langues Naturelles (TALN)*, Association pour le Traitement Automatique des Langues (ATALA), pp. 77–85.

Sacha BENIAMINE and Olivier BONAMI (2016), A comprehensive view on inflectional classification, paper presented at the *Annual Meeting of the Linguistic Association of Great Britain*, Paris.

James P. BLEVINS (2005), Word-based declensions in Estonian, in Geert E. BOOIJ and Jaap van MARLE, editors, *Yearbook of Morphology 2005*, pp. 1–25, Springer.

James P. BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42:531–573, ISSN 1469-7742, doi:10.1017/S0022226706004191.

Olivier BONAMI (2014), La structure fine des paradigmes de flexion, Habilitation à diriger des recherches, Université Paris Diderot.

Olivier BONAMI and Sacha BENIAMINE (2015), Implicative structure and joint predictiveness, in Vito PIRELLI, Claudia MARZI, and Marcello FERRO, editors, *Word Structure and Word Usage. Proceedings of the NetWordS Final Conference*.

Olivier BONAMI and Sacha BENIAMINE (2016), Joint predictiveness in inflectional paradigms, *Word Structure*, 9(2):156–182.

Olivier BONAMI and Gilles BOYÉ (2014), De formes en thèmes, in Florence VILLOING, Sarah LEROY, and Sophie DAVID, editors, *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux*, pp. 17–45, Presses Universitaires de Paris Ouest.

Olivier BONAMI, Gilles BOYÉ, Hélène GIRAUDO, and Madeleine VOGA (2008), Quels verbes sont réguliers en français?, in *Actes du premier Congrès Mondial de Linguistique Française*, pp. 1511–1523, doi:10.1051/cmlf08186.

Olivier BONAMI, Gauthier CARON, and Clément PLANCQ (2014), Construction d'un lexique flexionnel phonétisé libre du français, in Franck NEVEU, Peter BLUMENTHAL, Linda HRIBA, Annette GERSTENBERG, Judith MEINSCHAEFER, and Sophie PRÉVOST, editors, *Actes du quatrième Congrès Mondial de Linguistique Française*, pp. 2583–2596, doi:10.1051/shsconf/20140801223.

Olivier BONAMI and Berthold CRYSMANN (2016), The role of morphology in constraint-based lexicalist grammars, in Andrew HIPPISLEY and Gregory T. STUMP, editors, *Cambridge Handbook of Morphology*, pp. 609–656, Cambridge University Press, Cambridge.

Olivier BONAMI and Ana R. LUÍS (2014), Sur la morphologie implicative dans la conjugaison du portugais : une étude quantitative, in Jean-Léonard LÉONARD, editor, *Morphologie flexionnelle et dialectologie romane. Typologie(s) et modélisation(s)*, number 22 in Mémoires de la Société de Linguistique de Paris, pp. 111–151, Peeters, Leuven.

Dunstan BROWN (1998), *From the general to the exceptional*, Ph.D. thesis, University of Surrey.

Dunstan BROWN and Roger EVANS (2012), Morphological complexity and unsupervised learning: validating Russian inflectional classes using high frequency data, in Kiefer FERENCE, Mária LADÁNYI, and Péter SIPTÁR, editors, *(Ir)regularity, analogy and frequency, selected papers from the 14th International morphology meeting, Budapest, 13–16 May 2010*, Current Issues in Morphological Theory, pp. 135–162, John Benjamins Publishing Co., Amsterdam, doi:10.1075/cilt.322.07bro.

Dunstan BROWN and Andrew HIPPISLEY (2012), *Network Morphology: A Defaults-based Theory of Word Structure*, Cambridge Studies in Linguistics, Cambridge University Press, ISBN 9781107005747, doi:10.1017/CBO9780511794346.

Andrew D. CARSTAIRS (1987), *Allomorphy in Inflexion*, Croom Helm linguistics series, Croom Helm, ISBN 9780709934837.

Andrew CARSTAIRS-MCCARTHY (1994), Inflection Classes, Gender, and the Principle of Contrast, *Language*, 70(4):737–788, ISSN 00978507, doi:10.2307/416326.

Rudi L. CILIBRASI and Paul M. B. VITANYI (2005), Clustering by Compression, *IEEE Transactions on Information Theory*, 51(4):1523–1545, doi:10.1109/tit.2005.844059, http://dx.doi.org/10.1109/TIT.2005.844059.

Harald CLAHSEN (2006), Dual-mechanism morphology, in Keith BROWN, editor, *Encyclopedia of Language and Linguistics*, volume 4, pp. 1–5, Elsevier.

Greville G. CORBETT (1982), Gender in Russian: an account of gender specification and its relationship to declension, *Russian Linguistics*, 2:197–232.

Greville G. CORBETT (2009), Canonical Inflectional Classes, in Fabio MONTERMINI, Gilles BOYÉ, and Jesse TSENG, editors, *Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux*, volume 1-11, Cascadilla Proceedings Project, Somerville, MA, USA.

Greville G. CORBETT and Norman M. FRASER (1993), Network Morphology: a DATR account of Russian nominal inflection, *Journal of Linguistics*, 29:113–142, ISSN 1469-7742, doi:10.1017/S0022226700000074.

Wolfgang U DRESSLER, Marianne KILANI-SCHOCH, Natalia GAGARINA, Lina PESTAL, and Markus PÖCHTRAGER (2008), On the Typology of Inflection Class Systems, *Folia Linguistica*, 40(1-2):51–74, doi:10.1515/flin.40.1-2.51.

Wolfgang U. DRESSLER, Willi MAYERTHALER, Oswald PANAGL, and Wolfgang Ullrich WURZEL (1987), *Leitmotifs in natural morphology*, volume 10, John Benjamins Publishing, doi:10.1075/slcs.10.

Wolfgang U. DRESSLER and Anna M. THORNTON (1996), Italian Nominal Inflection, *Wiener Linguistische Gazette*, 55-57:1–26.

Markus DREYER and Jason EISNER (2011), Discovering Morphological Paradigms from Plain Text Using a Dirichlet Process Mixture Model, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pp. 616–627, Association for Computational Linguistics, Stroudsburg, PA, USA, ISBN 978-1-937284-11-4.

Greg DURRETT and John DENERO (2013), Supervised Learning of Complete Morphological Paradigms, in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1185–1195, Association for Computational Linguistics, Atlanta, Georgia.

Ramy ESKANDER, Nizar HABASH, and Owen RAMBOW (2013), *Automatic Extraction of Morphological Lexicons from Morphologically Annotated Corpora*, Association for Computational Linguistics, Seattle, Washington, USA.

John GOLDSMITH (2001), Unsupervised Learning of the Morphology of a Natural Language, *Computational Linguistics*, 27(2):153–198, ISSN 0891-2017, doi:10.1162/089120101750300490.

John GOLDSMITH and Jeremy O'BRIEN (2006), Learning inflectional classes, *Language Learning and Development*, 24(4):219–250, doi:10.1207/s15473341lld0204_1.

Peter D. GRÜNWALD (2007), *Minimum Description Length Principle*, MIT press, Cambridge, MA, ISBN 978-0-262-07281-6.

Marianne KILANI-SCHOCH and Wolfgang U. DRESSLER (2005), *Morphologie naturelle et flexion du verbe français*, Tübinger Beiträge zur Linguistik, G. Narr, ISBN 9783823361619.

Jackson LEE and John A. GOLDSMITH (2013), Automatic morphological alignment and clustering, presented at the 2nd American International Morphology Meeting.

Robert MALOUF (in press), Abstractive morphological learning with a recurrent neural network, *Morphology*, 27(4):431–458.

Peter H. MATTHEWS (1972), *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, Cambridge University Press.

Petar MILIN, Dušica FILIPOVIĆ ĐURĐEVIĆ, and Fermin MOSCOSO DEL PRADO MARTÍN (2009), The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from Serbian, *Journal of Memory and Language*, 60:50–64.

Christian MONSON, Alon LAVIE, Jaime CARBONELL, and Lori LEVIN (2004), Unsupervised Induction of Natural Language Morphology Inflection Classes, in

Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON '04), pp. 52–61, doi:10.3115/1622153.1622160.

Fabio MONTERMINI and Gilles BOYÉ (2012), Stem relations and inflection class assignment in Italian, *Word Structure*, 5:69–87.

Boris NEW, Christophe PALLIER, Ludovic FERRAND, and Rafael MATOS (2001), Une base de données lexicales du français contemporain sur internet: LEXIQUE., *L'année psychologique*, 101(3):447–462.

Garrett NICOLAI, Colin CHERRY, and Grzegorz KONDRAK (2015), Inflection Generation as Discriminative String Transduction, in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 922–931, Association for Computational Linguistics, Denver, Colorado, doi:10.3115/v1/N15-1093.

Marc PLÉNAT (1987), Morphologie du passé simple et du passé composé des verbes de l' "autre" conjugaison, *ITL Review of Applied Linguistics*.

Jorma RISSANEN (1978), Modeling by shortest data description, *Automatica*, 14:465–658.

Jorma RISSANEN (1984), Universal coding, information, prediction, and estimation, *IEEE Tr. on Info. Th.*, 30(4):629–636, doi:10.1109/TIT.1984.1056936.

Benoît SAGOT and Géraldine WALTHER (2011), Non-canonical inflection : data, formalisation and complexity measures., in Cerstin MAHLOW and Michael PIOTROWSKI, editors, *Systems and Frameworks in Computational Morphology*, volume 100, pp. 23–45, Springer-Verlag, Zurich, Switzerland.

Benoît SAGOT and Géraldine WALTHER (2013), Implementing a formal model of inflectional morphology, in Cerstin MAHLOW and Michael PIOTROWSKI, editors, *Actes du Third International Workshop on Systems and Frameworks for Computational Morphology (SFCM 2013)*, volume 380 of *Communications in Computer and Information Science (CCIS)*, pp. 115–134, Humboldt-Universität, Springer-Verlag, Berlin, Germany.

Claude E. SHANNON (1948), A Mathematical Theory of Communication, *Bell System Technical Journal*, 27(3):379–423, doi:10.1002/j.1538-7305.1948.tb01338.x, `http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x`.

Robert R. SOKAL and Charles D. MICHENER (1958), A statistical method for evaluating systematic relationships, *University of Kansas Scientific Bulletin*, 28:1409–1438.

Andrew SPENCER (2012), Identifying stems, *Word Structure*, 5:88–108.

Gregory STUMP and Raphael A. FINKEL (2013), *Morphological Typology: From Word to Paradigm*, Cambridge Studies in Linguistics, Cambridge University Press, ISBN 9781107029248, doi:10.1017/CBO9781139248860.

Arlindo VEIGA, Sara CANDEIAS, and Fernando PERDIGÃO (2013), Generating a pronunciation dictionary for European Portuguese using a joint-sequence model with embedded stress assignment, *Journal of the Brazilian Computer Society*, 19(2):127–134, ISSN 0104-6500, doi:10.1007/s13173-012-0088-0.

João VERÍSSIMO and Harald CLAHSEN (2014), Variables and similarity in linguistic generalization: Evidence from inflectional classes in Portuguese, *Journal of Memory and Language*, 76:61–79.

Géraldine WALTHER (2013), *On canonicity in morphology: an empirical, formal and computational approach*, Ph.D. thesis, Université Paris Diderot.

Géraldine WALTHER (2016), Paradigm Realisation and the Lexicon, in Ferenc KIEFER, James P. BLEVINS, and Huba BARTOS, editors, *Morphological paradigms and functions*, Brill, Leiden, Pays-Bas.

Géraldine WALTHER and Benoît SAGOT (2011), Modeling and implementing non canonical morphological phenomena, *TAL*, 52(2):91–122.

Wolfgang Ulrich WURZEL (1984), *Flexionsmorphologie und Natürlichkeit. Ein Beitrag zur morphologischen Theoriebildung*, Akademie-Verlag, Berlin, translated as Wurzel (1989).

Wolfgang Ulrich WURZEL (1989), *Inflectional Morphology and Naturalness*, Kluwer, Dordrecht.