Segmentation strategies for inflection class inference

Inflection classes (hereafter ICs) can be defined as groups of lexemes that inflect in similary ways. Descriptions of IC systems take many forms, from flat to hierarchical structures, and depend on theoretical assumptions and on the data at hand. Building on previous work from Brown and Evans (2012), Lee and Goldsmith (2013) and Bonami (2014), we describe an unsupervised strategy for automatically inferring IC systems from paradigmatic data. This allows us to explore the notion of an IC by providing a quantitative and reproducible basis for typological and methodological comparisons and gives us a way of measuring how linguistic theories fit large scale linguistic data. To systematise the linguist's work in inducing ICs, we face three challenges: (i) defining what form an IC system should take, (ii) deciding what kinds of abstractions should be made from the data in order to infer ICs, (iii) choosing an evaluation metric for candidate ICs and use it to cluster lexemes into an optimal structure of classes.

1 What is an inflection class system?

IC systems as flat partitions Flat descriptions of IC systems consist of a partition of lexemes into classes. To determine the partition, the first step is to determine how similar two lexemes must be to be considered members of a same class. A strict definition of similarity as identity yields a system of MICRO-CLASSES, where all lexemes in a class share the exact same generalisations and principal parts (as in Stump and Finkel's (2013) *plat*). Different micro-classes are likely to share some generalisations with each other, which deviates from the canonical description of ICs as externally heterogeneous (Corbett, 2009). At the opposite end of the spectrum, a system of MACRO-CLASSES distinguishes large groups of lexemes that share many of their paradigmatic generalisations, but not all. Those could maximise external heterogeneity, but are not internally homogeneous, deviating again from a canonical definition of IC. The description must then account for the internal heterogeneity of the MACRO-CLASSES: Traditional grammar usually lists macro-classes and considers variation inside the classes as deviations or exceptions.

IC systems as hierarchies Several authors (Corbett and Fraser, 1993; Dressler and Thornton, 1996; Kilani-Schoch and Dressler, 2005; Haspelmath and Sims, 2010; Brown and Hippisley, 2012; Lee and Goldsmith, 2013) adopt a hierarchical view of IC systems (Fig. 1), in which fine-grained ICs inherit properties from higher classes. In such hierarchies, the root represents the whole system, and the leaves are the MICRO-CLASSES. A distinguished level in between can constitute the MACRO-CLASSES.

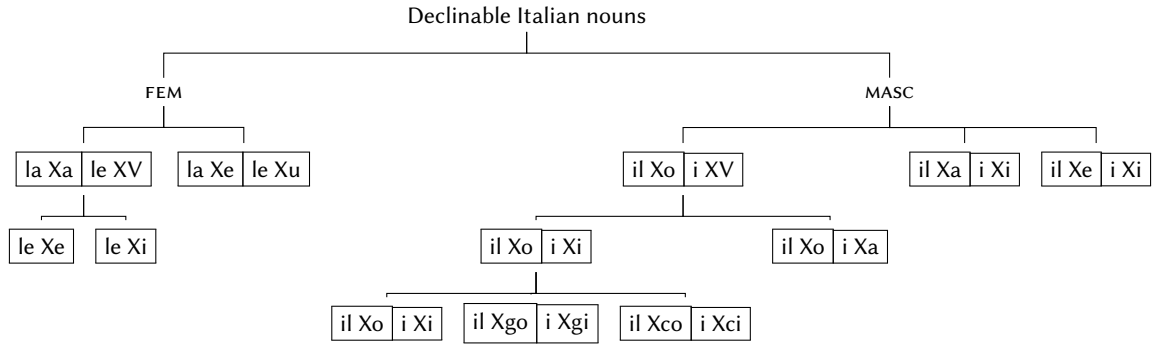


Figure 1: Hand-crafted hierarchical IC system (Dressler and Thornton, 1996)

Two problems arise when describing hierarchical IC systems. First, many competing criteria could be used to distinguish the macro-class level: semantic (animacy), phonological ("verbs ending in *-xy* after a dental"), morphosyntactic (gender) etc. In this study we focus on structural properties of the paradigm. Second, the definition of the MICRO-CLASSES strongly depends on theoretical assumptions, especially the role given to morphophonology, the basic units posited by the model and segmentation choices. We intend to investigate what morphological generalisations are most relevant to describe IC systems.

2 Two accounts of inflection

Exponents or implicative relations Recent research in morphology has been polarised around the distinction between exponent-based accounts (as illustrated in ex. 3) and those which focus on implicative relations (as illustrated in ex. 4) (Stump and Finkel, 2013; Ackerman and Malouf, 2013; Bonami, 2014). The former links morphosyntactic properties with their manifestation in a form and has been mainly associated with constructive approaches (in the sense of Blevins, 2006), which account for the combination of smaller units into wordforms. The latter link surface wordforms to one another and have been associated with abstractive approaches, which account for the relationships between well-formed words (Blevins, 2006).

(1) English noun: *bag* (SG), *bags* (PL)

(2) /bæg/ (SG), /bægz/ (PL)

(3) /bæg/ + $\left\{ \begin{matrix} /z/\text{(PL)} \\ \emptyset\ \text{(SG)} \end{matrix} \right\}$

(4) /bæg/ (SG) $\rightleftharpoons$ /bægz/ (PL)

The segmentation issue In exponent-based accounts, a segmentation strategy is explicitly needed to map subword units to morphosyntactic properties. When looking at implicative relations, no subword unit need be postulated, but a segmentation strategy is nonetheless implicit in order to determine what changes or not between two wordforms. While most descriptions take the segmentation for granted, we argue that it can have a significant impact on the analysis. Segmentation heuristics range from completely local to completely global strategies. Global approaches infer a segmentation from the whole paradigm of each lexeme. They have a natural link with constructive accounts, as they usually attempt to minimise the number of stems per lexeme. Local approaches infer a segmentation for each pair of forms in the paradigm, which closely corresponds to the abstractive view. However, it is entirely possible to study implicative relations after a global segmentation. Figure 2 shows the concurrent abstractions produced by a local and a global segmentation when looking at relations between pairs of forms.

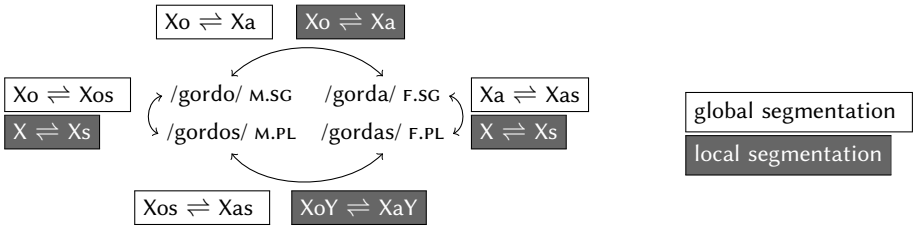


Figure 2: Four implicative rules with different segmentations for the Spanish adjective GORDO 'fat'

The two heuristics are usually combined, for example by computing a global segmentation in groups of cells considered as sharing a same stem allomorph Montermini and Bonami (2013). In our work, we evaluate the relative impact on IC system inference of choosing either an entirely global or local heuristic.

3 How to choose among candidate ICs

Given abstractions over the paradigms, we are faced with as many possible ICs as there are possible groupings of lexemes. To choose the best of these candidate ICs, we need to choose a scoring function.

Distances An obvious choice stems from our definition of ICs in terms of similarity. Similarity can be evaluated by a distance measure (Brown and Evans, 2012; Bonami, 2014). The upside of such approaches is that fast clustering algorithms based on distances are available. On the other hand, they only evaluate the quality of clusters, rather than of the system, and are thus unable to evaluate at which point merging decisions become detrimental to the system (e.g. to locate MACRO-CLASSES).

Paradigm entropy A second intuition is familiar to the Word and Paradigm approach; a system of ICs should make it easy to predict a lexeme's forms given its class, so the paradigm entropy (in the sense of Ackerman and Malouf, 2013) of a pair of forms should decrease when the IC is known. This property is

not a good criterion to build ICs because groups of very different lexemes that cannot undergo the same operations have an entropy of 0, although their members are far from similar.

Description length A third intuition is that an inventory of IC is better than another one if it can serve as the basis for a more economical description of the inflection system as a whole. The notion of DESCRIPTION LENGTH (hereafter DL; Rissanen, 1984) provides a theoretically grounded way to assess the economy of such a description. We shall adopt this approach. Its strength lies in its ability to evaluate the quality of the whole system. This will allow us to identify the macro-classes in the hierarchy as the nodes above which no increase in DL is possible. Other works in morphology have used a similar mesure: Sagot and Walther (2011) evaluate concurrent hand-made descriptions relatively to their DL, but do not produce the description themselves; Goldsmith (2001) minimises DL to optimally segment raw forms into morphemes and Lee and Goldsmith (2013) apply it to a task akin to ours, although starting from different assumptions and data, and using a more naive segmentation strategy and DL definition than ours.

4 Experiments and results

Segmentation algorithm We use a unique segmentation algorithm, feeding it either pairs of cells or the entire paradigm of a lexeme to compare the local and global approaches (see Table 1). Given a series of forms composed of phonological segments, (i) left-align the forms, matching the i-th segments of each form, (ii) build a template by keeping all i-th segments that are not identical across all aligned forms, (iii) mark identical segments as gaps in the template. For now, although it handles non-concatenativity, our algorithm enforces a left alignment, assuming that changes happen at the right margin.[1]

	PER.INF.1PL	GER	IND.PRS.3PL
ABANDONAR 'abandon'	ɐbɐ̃dunaɾmuʃ	ɐbɐ̃dunɐ̃du	ɐbɐ̃donɐ̃ũ
REABRIR 'reopen'	riɐbɾiɾmuʃ	riɐbɾĩdu	riabɾɐ̃ĩ
VOAR 'fly'	vuaɾmuʃ	vuɐ̃du	voɐ̃ũ

	Local segmentation			Global segmentation		
	PER.INF.1PL ⇌ IND.PRS.3PL	GER ⇌ IND.PRS.3PL	PER.INF.1PL⇌GER	PER.INF.1PL	GER	IND.PRS.3PL
ABANDONAR	_u_aɾmuʃ ⇌ _o_ɐ̃ũ	_u_du ⇌ _o_ũ	_aɾmuʃ ⇌ _ɐ̃du	_u_aɾmuʃ	_u_ɐ̃du	_o_ɐ̃ũ
REABRIR	_ɐ_iɾmuʃ ⇌ _a_ɐ̃ĩ	_ɐ_ĩdu ⇌ a_ɐ̃ĩ	_iɾmuʃ ⇌ _ĩdu	_ɐ_iɾmuʃ	_ɐ_ĩdu	_a_ɐ̃ĩ
VOAR	_uaɾmuʃ ⇌ _oɐ̃ũ	_u_du ⇌ _o_ũ	_aɾmuʃ ⇌ _ɐ̃du	_uaɾmuʃ	_uɐ̃du	_oɐ̃ũ

Table 1: Sample of Portugese verbal paradigms and templates with local and global segmentations.

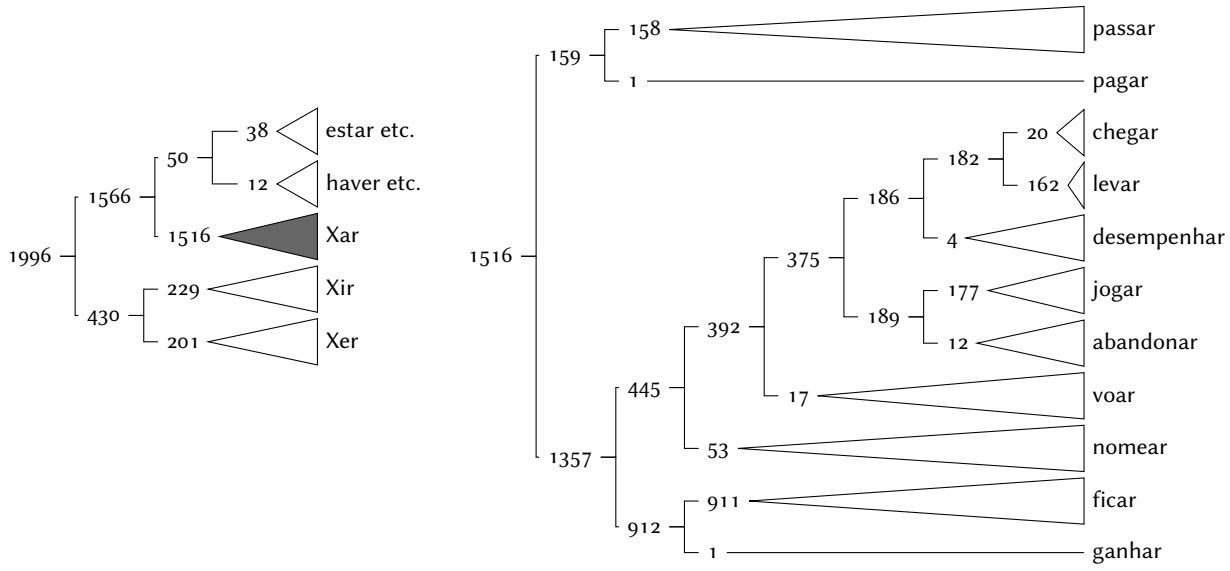
Clustering algorithm We then perform bottom-up hierarchical clustering. The algorithm starts with as many clusters as there are micro-classes.[2] At each step, it evaluates the change in DL for each possible binary merge of clusters and performs the best one. Therefore there is one less cluster at each step, until there is only one. As long as the best merge produces a decrease in DL, the new cluster's level is below or at the level of macro-classes. The macro-class level is identified as soon as the DL stops decreasing.

Results We ran the same algorithm with both a local and a global segmentation on both European Portuguese (Veiga et al., 2013) and French (Bonami et al., 2014) verbal data. In both cases, the local algorithm produces a classification highly similar to the traditional one (Fig. 3), while the global algorithm finds numerous small and scattered macro-classes. Table 1 illustrates this by showing that the global segmentation fails to capture any similarity between the Portuguese verbal paradigms of ABANDONAR and VOAR, whereas two out of the three templates obtained using local segmentation are identical.

In our talk, we will compare our results for several languages, including French and European Portuguese. We will describe our algorithm and the implicit choices made. We will argue that information-theoretic and automatic large-scale approaches should take into account the bias introduced by hand-made data segmentations and discuss segmentation heuristics currently used in the literature. We will describe our results in more detail and elaborate on the differences produced by the two segmentation heuristics.

[1] We are currently working towards a more generic segmentation algorithm that does not rely on this assumption.

[2] Finding the micro-classes is trivial; they are groups of lexemes sharing identical templates for all cells.



(a) Overview, with macro-classes as leaves

(b) Closeup of the *Xar* class, with micro-classes as leaves

Figure 3: Portugese verbal ICs obtained with the local segmentation strategy. Nodes are annotated with the number of lexemes they contain.

References

Ackerman, F. and Malouf, R. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89:429–464.

Blevins, J. P. 2006. Word-based morphology. *Journal of Linguistics* 42:531–573.

Bonami, O. 2014. La structure fine des paradigmes de flexion. Habilitation à diriger des recherches, U. Paris Diderot.

Bonami, O., Caron, G., and Plancq, C. 2014. Construction d'un lexique flexionnel phonétisé libre du français. *In* Actes du quatrième Congrès Mondial de Linguistique Française, pp. 2583–2596.

Brown, D. and Evans, R. 2012. Morphological complexity and unsupervised learning: validating Russian inflectional classes using high frequency data, pp. 135–162. *In* F. Kiefer, M. Ladányi, and P. Siptár (eds.), Current Issues in Morphological Theory: (Ir)regularity, analogy and frequency. John Benjamins, Amsterdam.

Brown, D. and Hippisley, A. 2012. Network Morphology: a defaults based theory of word structure. Cambridge University Press, Cambridge.

Corbett, G. G. 2009. Canonical inflectional classes. *In* Selected Proceedings of the 6th Décembrettes: Morphology in Bordeaux.

Corbett, G. G. and Fraser, N. M. 1993. Network morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics* 29:113–142.

Dressler, W. U. and Thornton, A. M. 1996. Italian nominal inflection. *Wiener Linguistische Gazette* 55-57:1–26.

Goldsmith, J. 2001. Unsupervised learning of the morphology of a natural language. *Comp. Ling.* 27:153–198.

Haspelmath, M. and Sims, A. 2010. Understanding Morphology, 2nd ed. Understanding Language. Taylor & Francis.

Kilani-Schoch, M. and Dressler, W. 2005. Morphologie naturelle et flexion du verbe français. Gunter Narr Verlag, Tübingen.

Lee, J. and Goldsmith, J. A. 2013. Automatic morphological alignment and clustering. Presented at the 2nd American International Morphology Meeting.

Montermini, F. and Bonami, O. 2013. Stem spaces and predictability in verbal inflection. *Lingue e linguaggio* 2:171–190.

Rissanen, J. 1984. Universal coding, information, prediction, and estimation. *IEEE Tr. on Info. Th.* 30:629–636.

Sagot, B. and Walther, G. 2011. Non-canonical inflection: data, formalisation and complexity measures. *In* C. Mahlow and M. Piotrowski (eds.), Systems and Frameworks in Computational Morphology, volume 100 of *Communications in Computer and Information Science*, pp. 23–45, Zurich, Suisse. Springer.

Stump, G. and Finkel, R. 2013. Morphological Typology: From Word to Paradigm. Cambridge Studies in Linguistics. Cambridge University Press.

Veiga, A., Candeias, S., and Perdigão, F. 2013. Generating a pronunciation dictionary for european portuguese using a joint-sequence model with embedded stress assignment. *Journal of the Brazilian Computer Society* 19:127–134.