Information-theoretic inflectional classification

Descriptions of inflection class systems take many forms, depending on the level of documentation of the language, the theoretical preferences of the author and the goals of the classification. Traditional descriptions usually distinguish a small number of broad classes, with less common patterns seen as deviating from these classes. At the other end of the spectrum, various attempts at making sense of the structure of inflection systems presuppose a classification into fine-grained micro-classes that exhaustively partition the set of lexemes (see Stump and Finkel's (2013) *plat*).

In other words, a system of macro-classes can be conceived as exhibiting maximal heterogeneity between classes, and a system of micro-classes as displaying maximal internal homogeneity. The two types of classifications can be reconciled by assuming a hierarchically organised system of classes, where macro-classes are seen as groupings of micro-classes (Dressler and Thornton, 1996). Although there are various ways of designing such hierarchies, we limit ourselves here to tree-shaped hierarchies with monotonous inheritance of inflectional properties.

Inflectional classifications are often used to reason about the typology of inflection systems. For such reasoning to be meaningful, it is crucial that the classifications defined for different languages be commensurable. Unfortunately, the way linguists define inflection classes (be they broad classes or micro-classes) is the result of many abitrary choices often driven by traditional views about the language at hand — sometimes even views about another better studied language. Existing classifications for different languages may therefore not be commensurable. This arbitrariness is often partly controlled by relying on combinations of heuristics, notably those guiding segmentation choices. Indeed, whether two lexemes should be taken as belonging to the same class is deduced from the fact that they use the same affixal exponents in the same paradigm cells. However there is no consensus as to the exact boundary between stem allomorphy and affixal exponence. As a result, many nontrivially different classifications are equally possible.

One way of addressing this issue is by defining quantitative measures to compare competing descriptions of the same morphological system. Sagot and Walther (2011) have designed and implemented such a measure, based on the information-theoretic notion of Description Length (hereafter DL; Rissanen, 1984). Despite its usefulness, this approach suffers from relying on the manual development of the descriptions to be compared, and not being able to guarantee that the optimal description with respect to the measure has been found. Moreover, what 'optimal' means heavily depends on the measure.

Finding ways of automatically inferring a system of inflection classes from the set of inflected forms found in a large lexicon is therefore an appealing alternative. This requires a measure for assessing the inflectional similarity between lexemes. At least three strategies may be envisioned: compute a global similarity measure between paradigms, using, for instance, compression distance (Brown and Evans, 2012); for each lexeme, segment all its forms into stems and exponents independently of other lexemes, and base the similarity measure on the resulting segmented forms (Lee and Goldsmith, 2013); or compare, across lexemes, patterns of alternations between pairs of forms in the paradigm (Bonami, 2014).

In this presentation we will compare different strategies implementing two such systematised heuristics. These strategies have several features in common. First, they both take as input unsegmented surface forms, without relying on an *a priori* segmentation or any kind of morphological analysis. Second, following Bonami (2014), they both rely on alternation patterns between paradigm cells. This is because such an approach strikes a balance between fully unsupervised learning and the practice of descriptive linguists.

The first strategy applies agglomerative average linkage clustering[1] (Sokal and Michener, 1958), using the Hamming distance between vectors of patterns of alternation as the dissimilarity metric. This strategy has at least two advantages: (i) It clearly implements the notion of an inflectional micro-class as a class of lexemes with a distance of 0 (compare Brown and Evans, 2012); (ii) on French data, it

[1] Also known as Unweighted Pair Group Method with Arithmetic Mean or UPGMA.

produces a classification that is remarkably close to the traditional one, with 1st and 2nd conjugation verbs forming clear clusters, and 3rd conjugation verbs scattering around these two main classes.

At each step, such an algorithm merges the two most similar clusters, until all lexemes belong to a unique cluster. Its drawback is that it performs successive local optimisations and relies on a measure which is only relevant for identifying similar behaviours. As a result, the first merge operations, at the bottom of the resulting tree, are more reliable than further ones. Moreover, there is no obvious way to evaluate at which point merging decisions become detrimental to the system (e.g., to locate macro-classes).

This was the motivation for designing and implementing a second strategy. Reminiscent of Sagot and Walther's (2011) work, we rely on DL to define this second measure. The intuition is that an inventory of inflection classes is better than another one if it can serve as the basis for a more economic description of the inflectional system at hand. The notion of DL provides a theoretically grounded way to assess the economy of such a description. The DL of a description relying on a one-class-for-all inventory corresponds to an extensive description of the data and is not very economical. Splitting this unique class into two usually yields a decrease in DL of the corresponding description. This decrease is maximal if the way lexemes are distributed into the two classes optimally captures relevant generalisations. Further splits can then be performed until the DL can not be decreased any more.

In our case, we define the DL of a morphological description as a sum of three terms: (i) the lexicon's description length, which is the number of bits necessary to encode the mapping between each lexeme and its class; (ii) the grammar's DL, which is the number of bits necessary to define the list of patterns available for each group and each pair of cells. It constitutes a pressure towards heterogeneity between clusters, as each pattern found in two distinct clusters has to be repeated, thus increasing the DL; (iii) the residual uncertainty, which is is the number of bits necessary to encode the mapping of each lexeme to its pattern given the lexicon and the grammar. This constitutes a pressure towards homogeneity within clusters, as each differing pattern for a same cell in the cluster will add uncertainty. The number of bits is 0 when a cluster has exactly one pattern for each pair of cells.

This measure has the particular advantage of being able to express the quality of a description, and therefore that of an inventory of inflection classes. In a bottom-up algorithm, it provides a way of seeing at which points merging classes does not yield any further improvement. This could allow us to identify macro-classes. But this measure is more computationnally expensive than pairwise distances, and the development of an efficient implementation is still ongoing.

We implemented a less complex greedy top-down approach based on an appromixation of the DL as defined above. This algorithm begins with only one cluster. At each step, it successively attempts to split each cluster in two, as follows: first, it randomly splits the cluster in two equally sized clusters; second, it moves lexemes from one of these two clusters to the other, performing the optimal move as far as DL is concerned, until no DL decrease is possible.

In our talk, we will compare the results of our efforts on datasets from various languages, including English, French, and European Portuguese. We will describe our ongoing work on the implementation of an efficient bottom-up clustering algorithm with DL, and discuss our results using this variety of approach for inferring hierarchies of inflection classes, and more generally for contributing to a better understanding of the underpinnings of the notion of inflection class.

References

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

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.

DRESSLER, W. U. AND THORNTON, A. M. 1996. Italian nominal inflection. *Wiener Linguistische Gazette* 55-57:1–26.

LEE, J. AND GOLDSMITH, J. A. 2013. Automatic morphological alignment and clustering. Presented at the 2nd American International Morphology Meeting.

RISSANEN, J. 1984. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information Theory* 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.

SOKAL, R. R. AND MICHENER, C. D. 1958. A statistical method for evaluating systematic relationships. *The University of Kansas Scientific Bulletin* 38.

STUMP, G. T. AND FINKEL, R. 2013. Morphological Typology: From Word to Paradigm. Cambridge University Press, Cambridge.