Comparing Complexity Measures

Benoît Sagot

Comparing Complexity Measures

Benoît Sagot
Alpage, Inria & Université Paris–Diderot, Paris, France
benoit.sagot@inria.fr

Given the range of ways morphological complexity is quantitatively defined and used in the recent literature, understanding precisely what the term *complexity* could and should denote in this context remains an open question. In this work, we aim at providing a global picture of the major approaches to morphological complexity, discuss some of their limitations — some to be found in the literature, some being unveiled here — and suggest one path towards a more satisfying way to quantify one particular point of view on morphological complexity.

We shall leave aside basic approximations such as counting category instances (McWorther 2001; Bickel & Nichols 2005; Shosted 2006). Indeed, the very possibility to count, for example, genders and cases within morphological systems in a clear-cut and language-independent way is arguably dubious. More interestingly, complexity has been introduced within the paradigm of Information Theory, and used in a morphological context, in at least two ways, on which our work focuses: **data description length** and **predictability of a new data instance**. The intuition behind description length has been formalized by Kolmogorov (1965), and is therefore known as **Kolmogorov Complexity** (KC). The idea is to measure how random is the data at hand, e.g., a large-scale morphological lexicon. This randomness is assessed by defining complexity as the smallest quantity of information required for describing the data: KC captures any structure underlying the data, be it intuitive and meaningful or not. However, KC cannot be computed exactly. One usual way to approximate it (but not the only one, see Juola (1998)) is to use a formal description language that suits the data, and find the description of the data within that language that contains as little information as possible (information being considered here as entropy times length): one compares the **compactness of descriptions**. If the formal (morphological) description language that is used captures "meaningful" structures only, compactness results will better match human intuition. This leads to a **formalism-dependent** definition of morphological complexity (Bane 2008; Sagot & Walther 2011).

KC-based approaches are sometimes criticized because of their being formalism-dependent. Several authors have therefore used another definition of the notion of complexity, **Shannon's Entropy** (SE). One can consider and model the data (a corpus, a lexicon…) as a set of data instances (sentences/words, lexical units…) emitted by a system. Shannon's (1948) entropy is a way to assess the amount of uncertainty expected in a new data instance given a model of the system built based on previously seen data instances. This requires encoding the data as a sequence of independent and identically distributed random variables according to a probabilistic model, which is difficult in practice. Still, it has been used directly on corpora (Moscoso del Prado Martín *et al.*, 2004; Moscoso del Prado Martín, 2010; Pellegrino *et al.*, 2007, 2010), and also to model interpredictability between cells within a paradigm: this is Ackerman and colleagues' **Paradigm Cell Filling Problem** (PCFP). This formulation leads its authors to defining morphological complexity, in their case **paradigm complexity**, as the average of all conditional entropies of each cell given each other cell. This measures computes how reliable are implicative patterns for guessing one cell from another (Ackerman *et al*. 2009; Malouf & Ackerman 2010). However, Bonami *et al*. (2011) have identified four issues concerning Ackerman *et al.*'s approach, which can be rephrased as follows: type frequency is ignored (each inflection class counts just as much as another), the way the inventory of inflection classes is built alters the results, the use of manually segmented data biases the results (the boundary between stem and suffix is given), and (mor)phonology is embedded in the data, so one does not measure morphological complexity only. We have performed a set of quantitative experiments, using the same French verbal data as above, which corroborate these remarks. Using type frequency from the Le*fff* lexicon or from a corpus or ignoring type frequency, ignoring rare inflection classes or not, using manually or automatically segmented data, ignoring morphonology or not, all these parameters lead to paradigm entropies that range from 0.08 to 0.72 for the same data! But there are two deeper issues. First, this entropy-based measure is **no less formalism-dependent** than the description-length-based one mentioned above. For example, if a cell is always filled by the reduplication of the content of another cell given, the

conditional entropy of the first cell given the second is zero if reduplication is available when when guessing a cell from another one, and very high otherwise. Second issue, **paradigm size is ignored**. Let us consider French first-group (regular) verbs, as well as a simplification thereof in which only two cells, infinitive and indicative present 1pl, are preserved. It turns out that surprisingly, Ackerman *et al*.'s paradigm entropy is higher for the simplified paradigms than for the full ones, be type frequency taken into account or not. This is because many cells with local interpredictability leads to a lower average conditional entropy than only two cells with low interpredictability.

One way to abstract from paradigm size is be to build a complete directed graph that relates each cell to all other cells, each edge being weighted by the conditional probability of the target cell given the source cell. Using the Chu-Liu-Edmonds algorithm, we can extract the best spanning tree, i.e., the globally optimal paradigm structure that relates cells within a tree structure. The *sum* (and not the average) of all conditional entropies of edges retained in the best spanning define a complexity measure, which we call **Minimum Overall Paradigm Complexity** (MOPC), that can be assessed on non-segmented data or on segmented data before applying morphographemic rules. Preliminary results show MOPC's relevance: our simplified version of French has much lower a complexity than the full French first-group verbs paradigms. However, further work is required for fully validating this new measure, and especially for relating it to the compactness-based measure described above and for understanding the optimal way to deal with Bonami *et al*.'s remarks on paradigm entropy, which still hold for MOPC.

References

Ackerman, Farrell, Blevins, James P., and Malouf, Robert (2009). Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. In James P. Blevins and Juliette Blevins (ed.), *Analogy in Grammar: Form and Acquisition.* Oxford University Press, Oxford, pp. 54–82.

Bane, Max (2008). Quantifying and measuring morphological complexity. In *Proc. of the 26th West Coast Conference on Formal Linguistics*, 69–76.

Bickel, Balthasar and Nichols, Johanna (2005). Inflectional synthesis of the verb. In Martin Haspelmath, Matthew S. Dryer, David Gil and Bernard Comrie (eds.), *The World Atlas of Language Structures*, Oxford University Press, Oxford, pp. 94–97.

Bonami, Olivier, Boyé, Gilles and Henri, Fabiola (2011). Measuring inflectional complexity: French and Mauritian. Paper presented at the Workshop on Quantitative Measures in Morphology and Morphological Development San Diego, United States.

Juola, Patrick (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5:3, pp. 206–13.

Kolmogorov, A. N. (1965). Three approaches to the quantitative defintion of information. *Problems in Information Transmission* 1, pp. 1–7.

McWhorter, John (2001). The world's simplest grammars are creole grammars. *Linguistic Typology* 5, pp. 125–66.

Malouf, Robert and Ackerman, Farrell (2010). Paradigms: The low entropy conjecture. Paper presented at the Workshop on Morphology and Formal Grammar, Paris, France.

Moscoso del Prado Martín, Fermín, Kostić, Aleksandar and Baayen, R. Harald (2004). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94, pp. 1–18.

Moscoso del Prado Martín, Fermín (2011). The Mirage of morphological complexity, In *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, pp. 3524–3529.

Pellegrino, François, Coupé, C. and Marsico, E. (2007). An information theory-based approach to the balance of complexity between phonetics, phonology and morphosyntax. In *Proc. of the Annual Meeting of the Linguistic Society of America*, Anaheim, CA, USA.

Pellegrino, F., Coupé, C., and Marsico, E. 2011. A cross-language perspective on speech information rate. *Language* 87:3, pp. 539–558.

Shannon, Claude E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, p. 379–423, 623–656.

Shosted, Ryan (2006). Correlating complexity: A typological approach. *Linguistic Typology* 10, pp. 1–40.

Walther, Géraldine and Sagot, Benoît (2011). Modélisation et implémentation de phénomènes flexionnels non-canoniques. In *Traitement Automatique des Langues* 52:2, pp. 91–122.