An analogical approach to the typology of inflectional complexity

Ackerman and Malouf (2013) propose the distinction between Enumerative (E) complexity and Integrative (I) complexity. The first is the complexity in morphosyntactic distinctions and the way languages encode them, while the second one is the difficulty a paradigm poses to the speakers of a language in terms of implicative relations.

E-complexity has received considerable attention in the typological literature (Baerman, Brown, and Corbett, 2015; Dressler, 2011, for an overview), but so far little work has gone towards developing implementations that can automatically and systematically quantify the E-complexity of a language. The large majority of work on E-complexity relies on handcrafted linguistic analysis, which are not always commensurable across researchers and languages.

At the same time, while there are multiple computational proposals for capturing I-complexity (Ackerman and Malouf, 2013; Bonami and Beniamine, 2016; Cotterell et al., 2019; Guzmán Naranjo, 2020; Marzi et al., 2018), most studies have looked at a relatively small samples and the emphasis has not been on cross-linguistic comparison (although see Cotterell et al., 2019). This means that we still do not have a picture of how I-complexity varies across languages and systems. For example, one still open question is how verb and noun paradigms compare crosslinguistically in terms of I-complexity.

While recent studies have proposed well formalized methods for exploring I-complexity (Bonami and Beniamine, 2016; Guzmán Naranjo, 2020), these implementations are very costly in terms of dataset size and quality as well as computational time, which makes them unsuitable for large scale typological studies. In this talk we present a new hybrid method. We focus pair-wise analogies as proposed by Bonami and Beniamine (2016), but instead of calculating the conditional entropy we fit a classifier to calculate predictive accuracy as suggested by Guzmán Naranjo (2020).

Our approach consists of two steps. First, we find all optimal pairwise alternations for all pairs of cells in a given paradigm, in a manner similar to Beniamine, 2018, and as exemplified in Table 1. Second, we perform analogical classification to predict, from the form in a predictor cell, which form is expected in a predicted cell. To do this, we first compute an edit distance matrix between the forms of all lexemes in the predictor cell. For a given predictor form, we predict as an output the result of applying to it the compatible alternation that is the most prevalent among its five nearest neighbors in the distance matrix. We check how frequently that prediction is correct, giving us an accuracy score for analogical prediction. We average these accuracies over all ordered pairs of cells to quantify the I-complexity of the system: the higher the accuracy, the less complexity the system has.

For E-complexity we present a new metric we call fragmentation. The fragmentation of an alternation is the number of variables in it. This captures the idea that the alternation Xa⇋Xo is simpler than the alternation XaYi⇋XoYi.

We applied this method to a combination of the Unimorph dataset (Kirov et al., 2018) as of the 5th of January of 2021, and another 12 datasets, for a total of 110 datasets (including nouns, verbs and adjectives). To control for the fact that different datasets have different sizes, for each dataset, we repeated the process for subsets of up to 200, 500, 1000, 2000, and 5000 lexemes.

Table 2 shows a subset of mean I-complexity for some of the languages in our dataset. A consistent result is that as datasets increase in size, I-complexity rapidly decreases. For example, while the mean predictability of of the 200 lexeme Hungarian noun dataset was of 0.87, the mean predictability of the 2000 lexeme dataset was of 0.94. Including more nouns increases the number of different inflection classes but it also makes predictions easier because the classifiers have more reliable neighbors. If there are enough items in a dataset, a large number of classes does not increase the I-complexity of the system for speakers. From a practical perspective this shows that we cannot reliably estimate the complexity of an inflectional system based on a small tables of inflection classes.

Regarding E-complexity, we find that there are no clear correlations between I-complexity and E-complexity. Cells with very high E-complexity are sometimes very easy to predict, and in other cases they are very difficult to predict. Similarly, the number of cells in a paradigm did not correlate with I-complexity. Both dimensions of complexity are, as far as we can tell, independent from each other.

lexeme	cell 1	cell 2	proportion
lexeme 1	pata	pato	Xa $\leftrightharpoons$ Xo
lexeme 2	mata	mato	Xa $\leftrightharpoons$ Xo
lexeme 3	kis	ki	Xs $\leftrightharpoons$ X
...			

Table 1: Analogies

lang	-n-_200	-n-_500	-n-_1000	-n-_2000	-v-_200	-v-_500	-v-_1000	-v-_2000	-adj-_200	-adj-_500	-adj-_1000	-adj-_2000
ady	0.97	0.97	0.97	0.98					0.97			
ang	0.64	0.68	0.69		0.76	0.77	0.78					
bak	0.96	0.98	0.98	0.98								
bel	0.61	0.66	0.66		0.73	0.73			0.99			
cat					0.91	0.93	0.94	0.95				
ces	0.76	0.78	0.8	0.82	0.98	0.98						
crh	0.96	0.97	0.97	0.98								
dan	0.75	0.77	0.8	0.82	1							
deu	0.6	0.68	0.7	0.73	0.77	0.81	0.81	0.85				
ell	0.69	0.74	0.75	0.78	0.71	0.77	0.8	0.8	0.98	0.98	0.98	0.98
fao	0.7	0.72	0.75	0.79	0.71	0.76	0.77		0.89	0.91		
fin	0.75	0.79	0.84	0.87	0.85	0.91	0.93	0.94	0.9	0.92	0.93	0.95
fre					0.9	0.91	0.93	0.94				
gle	0.54	0.6	0.63	0.69	0.81	0.85			0.76			
grc	0.61	0.61	0.67	0.69					0.99			
hbs	0.75	0.77	0.78	0.8	0.89	0.89	0.9	0.93	0.98	0.99	0.99	1
hun	0.87	0.92	0.93	0.94								
hye	0.92	0.94	0.94	0.95	0.95	0.95	0.95		0.95	0.97	0.96	0.97
isl	0.69	0.69	0.74	0.76	0.78	0.81	0.82					
ita					0.83	0.87	0.87	0.89				
lat	0.8	0.85	0.87	0.87	0.77	0.83	0.86	0.91				
lit	0.73	0.76	0.79	0.79	0.84				0.94			
mkd	0.84	0.9	0.92	0.92	0.87	0.91	0.92	0.94	1	1	1	1
nld					0.81	0.83	0.87	0.91	0.86	0.89	0.9	0.92
nob	0.75	0.75	0.77	0.8	0.77	0.81	0.8		0.83	0.87	0.9	
osx	0.67				0.75	0.82	0.83		0.98			
pol	0.81	0.83	0.86	0.87					0.99	0.99	0.99	1
por					0.89	0.92	0.94	0.96				
ron					0.81	0.85	0.86	0.87	0.89	0.89	0.91	
rus	0.79	0.82	0.86	0.87								
slv	0.88	0.87	0.89	0.91	0.73							
swe	0.75	0.77	0.82	0.84	0.81	0.87	0.88	0.91	0.87	0.91	0.92	0.93
tur	0.84	0.88	0.91	0.92	0.89	0.93	0.93					
xcl				0.82	0.87	0.91	0.91		0.83	0.88	0.88	

Table 2: Mean system predictability

References

Ackerman, Farrell and Robert Malouf (2013). "Morphological Organization: The Low Conditional Entropy Conjecture". In: *Language* 89.3, pp. 429–464.

Baerman, Matthew, Dunstan Brown, and Greville G. Corbett, eds. (2015). *Understanding and Measuring Morphological Complexity*. 1st ed. Oxford University Press.

Beniamine, Sacha (2018). "Classifications Flexionnelles: Étude Quantitative Des Structures de Paradigmes". Paris: Université Sorbonne Paris Cité - Paris Diderot.

Bonami, Olivier and Sacha Beniamine (2016). "Joint Predictiveness in Inflectional Paradigms". In: *Word Structure* 9.2, pp. 156–182.

Cotterell, Ryan et al. (2019). "On the Complexity and Typology of Inflectional Morphological Systems". In: *Transactions of the Association for Computational Linguistics* 7, pp. 327–342.

Dressler, Wolfgang U (2011). "The Rise of Complexity in Inflectional Morphology". In: *Poznań Studies in Contemporary Linguistics* 47.2, p. 159.

Guzmán Naranjo, Matías (2020). "Analogy, Complexity and Predictability in the Russian Nominal Inflection System". In: *Morphology* 30, pp. 219–262.

Kirov, Christo et al. (2018). "UniMorph 2.0: Universal Morphology". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. Paris, France: European Language Resources Association (ELRA).

Marzi, Claudia et al. (May 2018). "Evaluating Inflectional Complexity Crosslinguistically: A Processing Perspective". In: *Proceedings of the 11th Language Resources and Evaluation Conference* (Miyazaki, Japan). European Language Resource Association.