Entropy, analogy and paradigm structure

Matías Guzmán Naranjo
Heinrich-Heine Universität Düsseldorf

1 Introduction

Recent work on paradigm organization has focused on the question of how speakers can deduce the complete paradigm of a lexeme given that they only encounter a limited number of inflected forms of that lexeme (Ackerman et al., 2009). This is also known as the paradigm cell filling problem (PCFP). These studies have proposed entropy (Shannon, 1948) as a measurement of paradigm complexity/predictability. The basic idea is that one can calculate the conditional entropy between two cells of a paradigm, which measures the predictability between cells (i.e. how much information does Cell 1 provide about Cell 2 of a paradigm). Bonami & Beniamine (2016) have even expanded this approach to work for multiple cells.

Entropy-based approaches have a serious of limitations, however. First, entropy is not a normalized metric, which makes it unreliable for comparing different systems/languages. Second, many studies have convincingly shown that the inflection class of a lexeme is predictable from its phonology and semantics (Bybee & Slobin, 1982; Skousen, 1992; Eddington, 2002; Matthews, 2005; Blevins et al., 2017), which is information that entropy cannot easily take into account.

Using the Russian nominal inflection system as an example, I will argue that analogical classification (i.e. class assignment on the basis of similarity) offers a convincing solution to the PCFP, and that accuracy metrics are a better measurement of predictability/paradigm complexity than entropy.

1.1 Materials and methodology

From the Grammatical Dictionary of Russian by Zaliznyak (1977), I extracted all nouns (43412) with their complete paradigm (including the prepositional case). I then converted the extracted forms to a phonological transcription using epitran (Mortensen et al., 2018). This phonological transcription is not perfect but it is a reasonable approximation of the Russian system. For the present study I did not consider stress but this feature could be easily included into the models.

Many accounts of Russian nominal inflection have been proposed in the literature, each suggesting a different analysis of the inflection classes found in the Russian system (Fraser & Corbett, 1995, for a well known example). To sidestep these discussions, I extracted the inflection class of each noun automatically with a surface-based method. The method is as follows:

1. find the non-continuous phonological sub-sequence common to all cells in the paradigm of a lexeme (from now on the *stem*),
2. remove this sub-sequence from each cell. In cases of discontinuous sub-sequences add a separation mark (-),
3. the result in each cell is the *marker* for that cell,
4. the inflection class of the lexeme is the set of markers for all cells.

Because this method makes no assumptions about underlying representations, it is very conservative and thus it produces the maximum possible number of inflection classes. As an

	sgular		plal	
Cell	form	marker	form	marker
NOM	fʲirma	-a	fʲirmɨ	-ɨ
GEN	fʲirmɨ	-ɨ	fʲirm	-ø
DAT	fʲirm’e	-’e	fʲirmam	-am
ACC	fʲirmu	-u	fʲirmɨ	-ɨ
INS	fʲirmoj; firmou	-oj; -ou	fʲirmami	-am’i
PRE	fʲirm’e	-’e	fʲirmax	-ax

Table 1: Markers for *фирма*.

example we consider the lexeme *фирма* (‘firm’). The phonological transcription of *фирма* is *fʲirma*, the longest common substring (stem) is *fʲirm*, and the resulting markers are in Table 1.

the analogical models in the next section consider both phonological and semantic information of the *stem*. To include semantic information I used the pre-trained semantic vectors provided by Kutuzov & Kuzmenko (2017) using word2vec.[1]. To match a lexeme to a semantic vector I used the NOM.SG cell. From the dataset I only kept those nouns for which there was a semantic vector.

In order balance the dataset I only considered 1000 nouns for each class and removed all nouns belonging to classes with fewer than 20 nouns. Limiting the maximum number of nouns to 1000 helps the model avoid overestimating a couple of very frequent classes. The final dataset contained 17275 nouns, with 79 different inflection classes. This step also removes certain errors in the inflection class induction, as well ass irregular/suppletive forms.

On the resulting dataset I trained several analogical models using a multilayer perceptron following Guzmán Naranjo (2019).[2] For every cell in the paradigm I trained models predicting that cell from: (i) one other cell, (ii) two other cells, (iii) one cell and stem information,[3] (iv) and two cells and stem information.

For evaluation I used accuracy because this intuitively corresponds to our intuition of what predictability means Using Kappa scores or any other similar metric would also work. The important point is that accuracy metrics are normalized and therefore allow for comparisons across different models (even across different systems and languages).

[1] More precisely the `ruwikiruscorpora-func_upos_skipgram_300_5_2019` semantic vector data-set downloaded `http://rusvectores.org/en/models/`, accessed 17.06.2019.

[2] Each model had three hidden layers (with n*4, n*2 and n neurons respectively, where n = number of classes) with tanh activation. For all models, the learning rate was kept at 0.001, the momentum at 0.8 and dropout at 1. The models were trained with an Nvidia Titan Xp donated by the NVIDIA Corporation.

[3] The stem information consisted of the last four segments of the stem plus the semantic information in the semantic vectors (only looking at the nominative singular form).

	Predictor											
Predicted	NOM.SG	GEN.SG	DAT.SG	ACC.SG	INS.SG	PRE.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL	INS.PL	PRE.PL
NOM.SG	1	0.77	0.77	0.97	0.83	0.47	0.56	0.79	0.46	0.64	0.46	0.46
GEN.SG	0.7	1	0.89	0.7	0.84	0.63	0.63	0.74	0.57	0.73	0.57	0.57
DAT.SG	0.77	0.9	1	0.78	0.9	0.64	0.61	0.71	0.58	0.71	0.58	0.58
ACC.SG	0.82	0.61	0.61	1	0.7	0.41	0.48	0.64	0.4	0.66	0.4	0.4
INS.SG	0.77	0.76	0.77	0.75	1	0.51	0.52	0.74	0.44	0.61	0.44	0.44
PRE.SG	0.62	0.67	0.77	0.6	0.74	1	0.68	0.71	0.55	0.72	0.55	0.55
NOM.PL	0.52	0.67	0.58	0.52	0.57	0.6	1	0.64	0.38	0.89	0.38	0.38
GEN.PL	0.67	0.66	0.62	0.65	0.67	0.5	0.54	1	0.4	0.67	0.4	0.4
DAT.PL	0.97	0.99	1	0.97	0.99	0.99	1	0.96	1	0.97	0.99	0.99
ACC.PL	0.41	0.51	0.42	0.51	0.46	0.42	0.71	0.48	0.24	1	0.24	0.25
INS.PL	0.97	0.99	1	0.97	0.99	0.99	1	0.96	0.99	0.97	1	0.99
PRE.PL	0.97	1	1	0.97	0.99	0.99	1	0.96	0.99	0.97	0.99	1

Table 2: Cell predictability without stem information.

1.2 Results

Table 2 shows the accuracy score for the model predicting each cell using only one other cells as predictor. This result is comparable to the use of entropy to measure implicational relations between cells of a paradigm. This table shows that some cells in the paradimg can perfectly predict other cells. For example, DAT.SG completely predicts DAT.PL. Similarly, these results show that the ACC.SG cell is the best overall predictor of other cells in the paradigm.

However, it is also clear that most cells are not completely predictable from only one other cell. Table 3 shows how the results change once the analogical models also consider the information in the stem of the nouns. The effect is a very clear improvement.

	Predictor											
Predicted	NOM.SG	GEN.SG	DAT.SG	ACC.SG	INS.SG	PRE.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL	INS.PL	PRE.PL
NOM.SG	1	0.99	0.99	1	0.99	0.98	0.98	0.98	0.98	0.98	0.98	0.98
GEN.SG	0.98	1	0.99	0.98	0.99	0.98	0.98	0.98	0.97	0.98	0.97	0.97
DAT.SG	0.99	1	1	0.99	1	0.99	0.99	0.99	0.98	0.99	0.98	0.98
ACC.SG	0.96	0.95	0.95	1	0.94	0.94	0.93	0.94	0.93	0.98	0.93	0.93
INS.SG	0.98	0.98	0.98	0.98	1	0.97	0.97	0.97	0.97	0.97	0.97	0.96
PRE.SG	0.99	1	1	0.99	1	1	0.99	0.99	0.99	0.99	0.99	0.99
NOM.PL	0.99	0.99	0.99	0.99	0.99	0.99	1	0.99	0.99	1	0.99	0.99
GEN.PL	0.99	0.98	0.98	0.99	0.99	0.98	0.98	1	0.97	0.99	0.97	0.97
DAT.PL	1	1	1	1	1	1	1	0.99	1	1	1	1
ACC.PL	0.93	0.92	0.92	0.96	0.93	0.9	0.92	0.92	0.9	1	0.9	0.9
INS.PL	1	1	1	1	1	1	1	1	1	1	1	1
PRE.PL	1	1	1	1	1	1	1	0.99	1	1	1	1

Table 3: Cell predictability including stem information.

The first thing to notice is that all cells (except those which were already at 1) in Table 3 have accuracy scores higher than the corresponding cells in Table 2. Seen in absolute terms, we can say that just knowing one form of a Russian noun (including the stem) is enough to give almost perfect predictive accuracy for 7 cells (INS.PL, INS.SG, NOM.PL, NOM.SG, PRE.PL, PRE.SG and DAT.PL), it gives a reasonable certainty for three cells (DAT.SG, GEN.PL and GEN.SG) and it gives a some certainty for the remaining two (ACC.PL and ACC.SG). It is an interesting result that these two final cells, ACC.SG and ACC.PL, are the hardest to predict from the other cells and at the same time ACC.SG is the best predictor of other cells in average.

An important point is that not all cells increased in their predictability by the same amount.

While the predictability of ACC.PL from GEN.PL increased from 0.64 to 0.79, the predictability of ACC.PL from PRE.PL increased from 0.56 to 0.78. Since in both cases we are predicting the same cell (ACC.PL), it is not the case that the stem in one model had more information than in the other model. What this shows is that the interaction between the predictor GEN.PL and the stem carries less information about ACC.PL than the interaction between the stem and PRE.PL.

It is possible that the accuracy metrics are simply restating (with a normalized metric) the same information that the information theoretic approach can already capture. The check this we can explore the correlation between the analogical models and the conditional entropy estimates as shown in Table 5.[4]. The overall correlation values for the three models and the entropy model are shown in Table 4. The entropy model and the analogical model using markers are very close to each other, while the analogical models with stem information are less so. This result is important for two reasons. First, the fact that the entropy model and the marker model capture very similar information means that, if we accept that entropy is a valid metric, accuracy is in fact a valid alternative to quantify paradigm complexity. At the same time, it is clear that adding stem information to the model does greatly change class predictability.

model	correlation
marker model	-0.95
marker + phonology + semantics model	-0.81

Table 4: Correlation with entropy values

	Predictor											
Predicted	NOM.SG	GEN.SG	DAT.SG	ACC.SG	INS.SG	PRE.SG	NOM.PL	GEN.PL	DAT.PL	ACC.PL	INS.PL	PRE.PL
NOM.SG	0.00	1.24	1.21	0.26	1.00	1.87	1.61	1.14	2.15	1.18	2.15	2.15
GEN.SG	0.88	0.00	0.43	0.77	0.68	1.30	0.95	1.02	1.74	0.78	1.74	1.74
DAT.SG	0.65	0.23	0.00	0.58	0.41	0.87	1.03	0.87	1.55	0.76	1.55	1.55
ACC.SG	0.65	1.52	1.53	0.00	1.34	2.19	1.99	1.55	2.50	1.04	2.50	2.50
INS.SG	0.74	0.78	0.71	0.69	0.00	1.40	1.48	0.95	1.93	1.03	1.93	1.93
PRE.SG	1.00	0.79	0.56	0.92	0.78	0.00	0.61	0.93	1.30	0.68	1.30	1.30
NOM.PL	1.24	0.93	1.22	1.22	1.36	1.11	0.00	1.18	1.70	0.36	1.70	1.70
GEN.PL	1.25	1.49	1.54	1.27	1.31	1.91	1.66	0.00	2.29	0.85	2.29	2.29
DAT.PL	0.14	0.09	0.09	0.09	0.18	0.16	0.07	0.17	0.00	0.07	0.00	0.00
ACC.PL	2.37	2.33	2.52	1.84	2.48	2.75	1.93	1.93	3.28	0.00	3.28	3.28
INS.PL	0.14	0.09	0.09	0.09	0.18	0.16	0.07	0.17	0.00	0.07	0.00	0.00
PRE.PL	0.14	0.09	0.09	0.09	0.18	0.16	0.07	0.17	0.00	0.07	0.00	0.00

Table 5: Conditional entropy on the Russian data-set

2 Concluding remarks

These results show that to solve the PCFP it is not enough to look at the information and predictability between markers, nor is it enough to consider the class information hidden in the stems. Both are necessary. I have shown that an analogical classifier based on a perceptron can make use of stem and marker information. With this method we can measure predictability between any number of cells, as well as making use predictors like semantic vectors, which are

[4] I calculated these following Ackerman & Malouf (2013)

hard to take into account with entropy-based approaches. Finally, this method allows us to calculate accuracy metrics, which are normalized and allow for easy model comparison.

References

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 54–82. Oxford, New York: Oxford University Press.

Ackerman, Farrell & Robert Malouf. 2013. Morphological Organization: The Low Conditional Entropy Conjecture. *Language* 89(3). 429–464. doi:10.1353/lan.2013.0054.

Blevins, James P., Petar Milin & Michael Ramscar. 2017. The Zipfian paradigm cell filling problem. In Ferenc Kiefer, James P. Blevins & Huba Bartos (eds.), *Perspectives on Morphological Organization*, 139–158. Leiden, The Netherlands: Brill.

Bonami, Olivier & Sacha Beniamine. 2016. Joint predictiveness in inflectional paradigms. *Word Structure* 9(2). 156–182.

Bybee, Joan L. & Dan I. Slobin. 1982. Rules and Schemas in the Development and Use of the English past Tense. *Language* 58(2). 265–289.

Eddington, David. 2002. Spanish gender assignment in an analogical framework. *Journal of Quantitative Linguistics* 9(1). 49–75.

Fraser, Norman M & Greville G Corbett. 1995. Gender, animacy, and declensional class assignment: A unified account for Russian. In *Yearbook of Morphology 1994*, 123–150. Springer.

Guzmán Naranjo, Matías. 2019. *Analogical classification in formal grammar* Empirically Oriented Theoretical Morphology and Syntax. Language Science Press.

Kutuzov, Andrey & Elizaveta Kuzmenko. 2017. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In Dmitry I. Ignatov, Mikhail Yu. Khachay, Valeri G. Labunets, Natalia Loukachevitch, Sergey I. Nikolenko, Alexander Panchenko, Andrey V. Savchenko & Konstantin Vorontsov (eds.), *Analysis of Images, Social Networks and Texts: 5th International Conference, AIST 2016, Yekaterinburg, Russia, April 7-9, 2016, Revised Selected Papers*, 155–161. Cham: Springer International Publishing. doi:10.1007/978-3-319-52920-2_15.

Matthews, Clive A. 2005. French Gender Attribution on the Basis of Similarity: A Comparison Between AM and Connectionist Models. *Journal of Quantitative Linguistics* 12. 262–296.

Mortensen, David R, Siddharth Dalmia & Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, .

Shannon, Claude. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27. 379–423, 623–56.

Skousen, Royal. 1992. *Analogy and structure*. Dordrecht: Springer.

Zaliznyak, Andrey Anatolyevich. 1977. *Grammatical dictionary of the Russian language*. Moscow: Russkij Jazyk.