Evaluating evaluations for the Paradigm Cell Filling Problem

Gilles Boyé
Université Bordeaux Montaigne & UMR5263

1 Introduction

Since the Paradigm Cell Filling Problem (henceforth PCFP) was formalised by Ackerman et al. (2009) in (1), there have been many studies about the complexity of answering the PCFP (e.g. Malouf & Ackerman, 2010; Bonami & Beniamine, 2016; Blevins et al., 2017).

(1) Given prior exposure to at most a subset of forms, how does a speaker produce or interpret a novel form of an item?

Most of the work has been dedicated to evaluate the entropy of inflectional systems using implicative patterns giving rise to the Low Entropy hypothesis of Malouf & Ackerman (2013). As entropy in itself is only a measure of the difficulty to make predictions, in this paper, we use a predictive model inspired from SWIM2 (Boyé, 2015) to fill out the paradigm cells of French conjugation based on vlexique2 (Beniamine et al., 2024). The aim of the study is to compare different methods of evaluation using the same predictive model with the same dataset.

2 Setting the stage

The model used in the study takes a sample of phonological forms of lexemes tagged with cell and lexeme information (the input). It uses all information available for every pair of cells to abstract possible transformations from cell_{in} to cell_{out} in a manner similar to PredSPE (Bonami & Boyé, 2014). The transformations are then used to predict the missing cells (the output).

We compare three methods to evaluate the results:

- The k-fold cross-validation method where the dataset is divided in k parts of the same size. The model is run k times using each time k-1 parts as the input and the excluded part as the gold standard for the output.
- What we call the k-fold inverse cross-validation method where the dataset is also divided in k parts of the same size and the model is run k times, but with 1 part as the input and k-1 parts as the gold standard for the output.
- What we call the ecological validation where the dataset is divided by frequency[1]. The model is run 1 time with the higher frequency forms serving as the input and the lower frequency forms serving as the gold standard for the output.

3 Comparing the methods

In all our evaluations, a form is considered correct if the predicted form is the same as the gold form with a neutralisation of the differences between ɛ/e and ɔ/o which are unreliably distinguished by French speakers in the Southern varieties (e.g. /fɔtəre/ vs /fotəre/). The evaluation

[1] We use the frequency given by vlexique2 for each inflectional form in the OpenSubtitles (Lison & Tiedemann, 2016)

of the results is limited to the forms present in the gold standard. Precision is calculated by dividing the number of correct predictions by the number of predictions (correct + incorrect). To get Recall, we compute the number of missing predictions[2], Recall is the number of correct predictions divided by the sum of this number and missing predictions.

$$\text{Precision} = \frac{\text{correct}}{\text{correct} + \text{incorrect}} \qquad \text{Recall} = \frac{\text{correct}}{\text{correct} + \text{missing}}$$

Table 1 gives the quantitative results for the 5-fold cross-validation and 5-fold inverse cross-validation.

		Average	P1	P2	P3	P4	P5
5-fold cross-validation	Precision	100.0%	100.0%	99.9%	100.0%	100.0%	100.0%
	Recall	96.0%	96.0%	96.0%	96.1 %	95.9 %	96.0%
5-fold inverse cross-validation	Precision	99.7%	99.7%	99.7%	99.7%	99.6%	99.6%
	Recall	84.5%	85.8%	86.4%	80.6%	85.3%	84.3%

Table 1: Precision and Recall for the 5-fold cross-validation and inversed cross-validation

The quantitative differences between the classic 5-fold cross-validation and the inverted one are mainly concerned with Recall (96.0% vs 84.5%) while they both share a high Precision (100.0% vs 99.7%). These relatively good evaluations are somewhat misleading as the qualitative analysis of the predictions reveals that in both cases Recall is very bad for high frequency lexemes such as ÊTRE, AVOIR, ALLER, FAIRE, DIRE, POUVOIR, VOULOIR, SAVOIR, VOIR, DEVOIR, VENIR even for very common form such as the infinitive or the past participle.

To compare with these results, we ran an ecological validation with an input containing the 20% most frequent forms and a gold standard containing the 80% least frequent ones and to present an intermediate view of the data, we also ran another ecological validation with an input containing the 50% most frequent forms and a gold standard containing the 50% least frequent ones. The result are presented in Table 2.

20% ecological validation	Precision	99.0%
	Recall	90.6%
50% ecological validation	Precision	99.7%
	Recall	97.3%

Table 2: Precision and Recall for the 20% and 50% ecological validation

For the inputs containing 20% of the dataset, on the quantitative front, the 20% ecological validation is not as good as the 5-fold inverse cross-validation on Precision (99.0% vs 99.7%) but it is better on Recall (90.6% vs 84.5%). On the qualitative front however, the ecological validation obviously avoids missing high frequency forms.

The global results (F1 Score[3]) of the 5-fold cross-validation containing 80% of the dataset are as good not 50% ecological validation (F1 97.9% vs 98.5%).

4 Discussion

Our study shows that the choice of input has a large influence on the evaluation of the same model in answering the PCFP. We argue that random k-fold cross-validation such as the one

[2] In this case, missing predictions are the number of paradigm cells for which the model did not provide content.

[3] $\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

commonly used in NLP is not a good way to evaluate PCFP-oriented models of inflection: despite its overall excellent precision and very good recall, it leads to gaps in very unnatural places in the paradigm of high frequency lexemes. The traditional k-fold cross-validation also uses a large quantity of the dataset for its training which seems to be unnecessary in our experiment as shown by the 5-fold inverse results with their precision almost as good as the normal one (with a lower recall, however). Using a frequency driven sample as the input seems better fitted quantitatively and qualitatively, in line with the Median Threshold Hypothesis of Schalchli (2021) proposing that the lexicon is split in a higher frequency part (the head) and a lower frequency part (the tail) and that the head is used to abstract generalisations. These generalisations can be tested on the tail and be used as productive to generate new data.
We think that ecological validation should be adopted as a standard method of evaluation in place of the usual cross-validation in quantitative studies in morphology in general.

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 Scholarship Online.

Beniamine, Sacha, Maximin Coavoux & Olivier Bonami. 2024. Vlexique2.0: A rich lexicon of french verbal inflection with form-level frequencies. In *21st international morphology meeting*, Vienna.

Blevins, James P, Petar Milin & Michael Ramscar. 2017. The zipfian paradigm cell filling problem. In *Perspectives on morphological organization*, 139–158. Brill.

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

Bonami, Olivier & Gilles Boyé. 2014. De formes en thèmes. *Foisonnements morphologiques. Etudes en hommage à Françoise Kerleroux* 17–45.

Boyé, Gilles. 2015. Small worlds inflectional morphology: A fragment for french conjugation. Paper presented at CMDTM 2015 (Vienna).

Lison, Pierre & Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *Proceedings of the 10th international conference on language resources and evaluation (lrec 2016)*, European Language Resources Association.

Malouf, Rob & Farrell Ackerman. 2010. Paradigm entropy as a measure of morphological simplicity. In *Workshop on morphological complexity: Implications for the theory of language*, Harvard.

Malouf, Robert & Farrell Ackerman. 2013. The low entropy conjecture: The challenges of modern irish nominal declension. *Language* 89(3).

Schalchli, Gauvain. 2021. The median threshold hypothesis: Measuring morphological productivity from frequency lists. In *Third international symposium of morphology (ismo 2021)*, 114.