The PCFP and the Zipfian frequency distribution: The Median Threshold Hypothesis and French conjugation

Gilles Boyé, Gauvain Schalchli

The PCFP and the Zipfian frequency distribution: The Median Threshold Hypothesis and French conjugation

Gilles Boyé & Gauvain Schalchli
Université Bordeaux-Montaigne & CLLE-ERSSàB

Background

PCFP: Predictions vs predictability

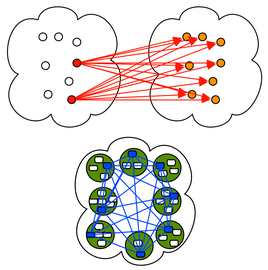
The Paradigm Cell Filling Problem (PCFP, Ackermann & al., 2009):

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

- Most answers to the PCFP revolve around the low conditional entropy conjecture but entropy is a figure that summarise the ease/difficulty of making predictions.
  - low entropy ensures high predictability but it is a prediction about predictions
- We propose to measure the reliability of a given predictive model
  - using SWIM (Boyé, 2015, 2016, 2024; Boyé & Schalchli, 2019) to predict inflected surface forms, we will evaluate its precision and recall

Paradigm Cell Filling with SWIM

- Preparing cells and forms:
  - Making optimal paradigms (Boyé & Schalchli 2016, 2019)
    - grouping cells with compatible content
  - Abstracting analogies (Bonami & Boyé 2014, Boyé 2016)
    - phonological transformations between pairs of cells
    - statistical distributions for output
  - Generating forms (Boyé 2015, 2016)
    - using all known forms
    - filling all cells
- Extracting paradigms:
  - Cliquing forms
    - inflectional forms are all analogically connected
  - Comparing candidates
    - the best clique is the largest and the most cohesive
  - Filling the zero-entropy cells (Boyé, 2019, 2024)



OUTLINE

The effects of frequency on memory in general have long been documented. Furthermore, Ambridge et al. (2015) has demonstrated that token frequency has an effect on linguistic behaviour at large. In morphology, researchers (e.g. Bybee, 1995; Corbett et al., 2001; Herce, 2016) have studied its links with inflectional irregularities.

Recently, Blevins et al. (2017) have outlined the implications of this Zipfian frequency distribution (Zipf, 1932) on the modelling of inflectional behaviour.

But, despite the recognition of this effect, most studies in this domain ignore frequencies in their analyses (e.g. Stump and Finkel, 2013; Stump, 2015; Bonami and Boyé, 2014; Bonami, 2014; Beniamine, 2018; Pellegrini, 2023).

Yet other works, (e.g. Boyé, 2016; Herce, 2022; Sims-Williams, 2022), confirmed the importance of this effect on the predictability of inflected forms (a.k.a the Paradigm Cell Filling Problem, PCFP, Ackerman et al. 2009).

In this context, we propose:

- to make predictions rather than measure predictability
- to make predictions based on corpus data
- to take frequency into account

We conducted 2 experiments:

1. predict all the forms for the lexemes present in the dataset
2. from half of the dataset predict the other half

Experiments

Methodology

- Our study is based on:
  - a corpus extracted lexicon with token frequencies, Lexique4Linguists, 100k forms, (L4L, Schalchli, 2022) based on OpenSubtitles2016 (Lison and Tiedemann, 2016).
  - a reference database of inflected forms, Flexique (Bonami et al., 2014)
- To test the Median Threshold Hypothesis (Schalchli, 2021):
  - we cut the L4L at its median frequency (head 51k forms, tail 49k forms)
  - we compare the predictions made by the generalisation extracted from the head, the tail and the whole lexicon.

Results

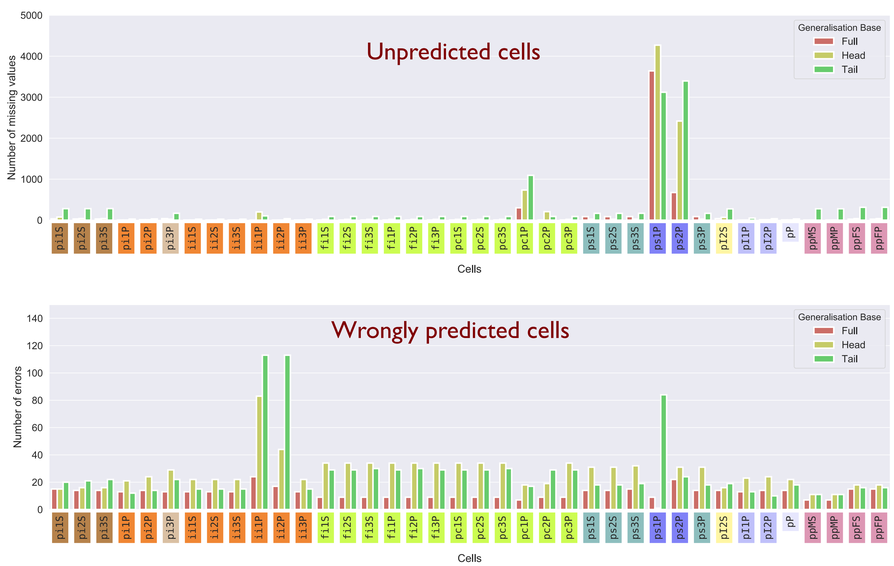
- 1st experiment: control=Flexique, without simple past & imperfective subjunctive

Base	Lexemes	Input cells	Output cells	Precision	Recall
L4L Full	4946	100K	87K	99,43 %	93,4 %
L4L Head	4315	51K	113K	99,03 %	91,5 %
L4L Tail	4921	49K	138K	99,18 %	90,4 %

- 2nd experiment: control=L4L

Base	Lexemes	Input cells	Output cells	Precision	Recall
L4L Head	4746	51K	41K	99,51 %	98,4 %
L4L Tail	4746	49K	49K	99,68 %	92,3 %

The cell distribution of problems



References

Ackerman, Blevins & Malouf (2009). Parts and wholes: Implicative patterns in inflectional paradigms. In Blevins and Blevins (eds), *Analogy in grammar: Form and acquisition*, pp54-82. Oxford Scholarship Online
Ambridge, Kidd, Rowland & Theakston (2015). The ubiquity of frequency effect in first language acquisition. *Journal of child language*, **42**, 239-273.
Beniamine (2018). *Classifications flexionnelles. Étude quantitative des structures de paradigmes*. Phd thesis. Université Paris Cité.
Blevins, Milin & Ramscar (2017). The Zipfian paradigm cell filling problem. *In Perspectives on morphological organization*. 139-158.
Bonami (2014). La structure fine des paradigmes de flexion : Études de morphologie descriptive, théorique et formelle. Habilitation, Université Paris-Diderot.
Bonami, Caron & Plancq (2014). Construction d'un lexique flexionnel phonétisé libre du français. *In SHS Web of Conferences*, **8**:2583-2596. EDP Sciences.
Bonami & Boyé (2014). De formes en thèmes. In Villoing, David, Leroy (eds), *Foisonnements morphologiques. Études en hommage à Françoise Kerleroux*, pp17-45
Boyé (2015). Small World Inflectional Morphology: A fragment for French conjugation. Paper presented at Computational methods for descriptive and theoretical morphology at IMM in Vienna.
Boyé (2016). Pour une modélisation surfaciste de la flexion. Le cas de la conjugaison du français. *In SHS Web of Conferences*, **27**.
Boyé (2019). Stem spaces in abstractive morphology: A look at defectiveness in French conjugation. paper presented at ISMo19.
Boyé (2024). 20 ans après : une transition écologique pour les espaces thématiques. *In SHS Web of Conferences* **191**.
Boyé & Schalchli (2016). The Status of Paradigms. In The Cambridge Handbook of Morphology (eds. Hippisley & Stump). 206-234. Cambridge University Press.
Boyé & Schalchli (2019). Realistic data and paradigms: the paradigm cell finding problem. *Morphology*. **29:2**, 199-248. Springer Verlag.
Bybee (1995). Regular morphology and the lexicon. *Language and cognitive processes*. **10:5,** 425-255. Taylor & Francis.
Corbett, Hippisley, Brown & Marriot (2001). Frequency, regularity and the paradigm. In *Frequency and the Emergence of Linguistic Structure* (eds Bybee & Hopper).
Herce (2016). Why frequency and morphological irregularity are not independent variables in Spanish: A response to Fratini et al.(2014). Corpus Linguistics and Linguistic Theory. 12:2, 389-406. de Gruyter.
Herce (2022). Quantifying the importance of morphomic structure, semantic values, and frequency of use in Romance stem alternations. *Linguistics Vanguard*. **8:1**, 53-68. de Gruyter.
Lison & Tiedemann (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.
Pellegrini (2023). *Paradigm Structure and Predictability in Latin Inflection: An Entropy-based Approach*. Springer Nature.
Schalchli (2021). The Median Threshold Hypothesis: Measuring morphological productivity from frequency lists. paper presented at 3rd International Symposium of Morphology.
Schalchli (2022). *Lexique4Linguists*. Ortolang.
Sims-Williams (2022). Token frequency as a determinant of morphological change. Journal of Linguistics. **58:3**, 571-607. Cambridge University Press.
Stump (2015). *Inflectional paradigms: Content and form at the syntax-morphology interface*. Cambridge University Press.
Stump & Finkel (2013). *Morphological typology: From word to paradigm*. Cambridge University Press.
Zipf (1932). *Selected studies of the principle of relative frequency in language*. Harvard University Press.

CONCLUSION

- Key findings:
  - an efficient model of prediction is possible at a large scale with methods similar to standard predictability studies
  - the frequency effect allows for smaller inputs to have almost the same predictive quality as the larger ones
  - unpredicted forms are related to lack of information in the input
  - wrong predictions can be paradigmatic
    - predicting a good paradigm
    - but for the wrong lexeme
- Conclusion:
  - modelling and evaluating predictions should **both** be based on corpus data
  - models should aim at capturing speaker's inferences rather than linguists inferences
- Caveats:
  - L4L lacks reliable data for subjunctive present both for the forms and their frequency
  - the study excluded simple past and subjunctive imperfective as they are known to be difficult to use and/or interpret for naive speaker
- Further research:
  - correlating the errors with the identity and number of input cells
  - psycholinguistic/sociolinguistic experiments with speakers to establish a baseline for a linguistic kind of evaluation rather than an NLP one