Form predictability and cell frequency: a behavioural study

Maria Copot maria.copot@etu.u-paris.fr
Olivier Bonami olivier.bonami@linguist.univ-paris-diderot.fr

Background

Much recent work in morphology moving away from the traditional view of exceptionality as a binary property (Prasada & Pinker, 1993; O'Donnell, 2015; Yang, 2016) and takes it to be a continuum (Bybee & Slobin, 1982; Rumelhart and McClelland, 1986; Smolensky, 1995; Albright, 2002; Blevins, 2016; Herce, 2019). The renewed interest in paradigmatic structure and information theory has provided a useful framework for thinking about word form exceptionality in a quantitative fashion: a word form's exceptionality can be operationalised in terms of surprisal or entropy involved in predicting it from another member of its paradigm (the intuition behind Albright, 2002; Albright & Hayes, 2003; made more explicit in Ackerman & Malouf, 2013; Bonami & Beniamine, 2016), which in turn can be derived from the type frequency of the patterns that exist between the two cells.

It is also increasingly a matter of interest that paradigmatic form predictability interacts with various frequency measures, for reasons to do with linguistic processing and learnability (Milin et al., 2009; Divjak, 2019). The more high-frequency a word is, the more it can afford to have an unpredictable form, because its frequency ensures that its phonological form is highly active in memory and thus easily accessible. On the flip side, low frequency words are more likely to be easily predictable from other members of the paradigm: if a word is already syntagmatically uncertain (low-frequency words are tautologically an unexpected way to continue the average utterance), it's unlikely to tolerate additional uncertainty on the paradigmatic axis (Filipović Đurđević & Milin, 2018).

Building on this, Copot & Bonami (2021) show in a corpus study that the frequency in use of a word is negatively correlated to its paradigmatic predictability (at parity of lexeme frequency, the word with the target meaning that is most easily accessible will be employed by speakers), but this relationship is moderated by the frequency of all members of the lexeme's paradigm (high-frequency lexemes and word forms will have representations in memory that are more independent of the pattern they instantiate, and so can be accessed through more direct retrieval, rather than have to be produced as the result of analogy), and by the frequency of the cell (if a cell is very frequent, it will rarely need to be predicted, but rather it will form the basis of prediction).

Motivation

Following Copot & Bonami (2021)'s findings, we perform a behavioural experiment on the interaction between word frequency, cell frequency (the summed frequency of all words filling a particular paradigm cell across lexemes), and paradigm predictability (how expected the form filling one cell of a paradigm is given the rest of the makeup of that paradigm).

The corpus study employed average paradigmatic predictability (the average of a form's predictability based on each of its other paradigm members) as a predictor, so while it appears that form predictability matters on average, more work is necessary to establish the impact that paradigmatic predictability has on language processing. On this matter, we can ask 1) Are speakers sensitive to individual relationships of paradigmatic predictability between two cells/word forms in a larger paradigmatic system? Previous research on this topic has looked at small two-cell subsystems (the English past tense, the English plural) - claims concerning the paradigmatic nature of predictability would be stronger if evidence for them could be found in more complex systems. 2) Is the effect of paradigmatic predictability bidirectional? Or, as per Jun & Albright (2017), are predictability relationships only exploited when predicting from the base form?

Moreover, the corpus study used token frequency of a word as the variable to be predicted as a function of paradigmatic word form predictability, lexeme frequency and cell frequency, but token frequency interacts in complex ways with lexeme and cell frequency. To disentangle such interactions, we employ pseudoword stimuli - this enables us to isolate the effect of paradigmatic predictability and cell frequency.

Methods

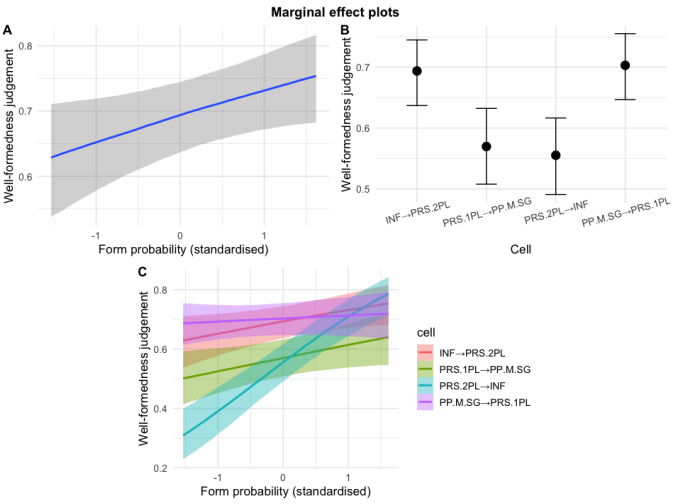
To tackle these questions, we implement a modified version of Jun and Albright (2017)'s methodology with French data. Experimental items are sentences containing the same pseudolexeme twice, in two different inflected forms. Participants are asked to use a continuous, unmarked slider to express a well-formedness judgement on the second inflected form, under the assumption that the first form belonged to the same lexeme. The hypothesis is that the more predictable the second form is from the first, the higher the score it will

receive. Furthermore, form predictability is expected to matter more when predicting towards more frequent cells (for which speakers have a better grip on pattern distribution), and that judgements towards less frequent cells will on average be higher (speakers are more willing to be accepting of forms in cells which they've been exposed to fewer examples of).

Experimental items varied in two dimensions: the identity of the paradigmatic cells involved, and the degree of predictability of the second form from the first. Two cell pairs were chosen based on the range of predictability values of the possible patterns of alternation to be found between them: INF⇆IND.PRS.2PL, IND.PRS.1PL ⇆ PP.M.SG. To test if the effect of predictability is bidirectional, items for each cell pair varied which cell was first vs second in the sentence (2 cell pairs * 2 directions of prediction = 4 cell conditions). To test the effect of form predictability on judgements, each experimental item had three possible versions of the second word form, differing in the degree to which the second inflected form was predictable based on the first inflected form.

A maximal bayesian zero- and one-inflated beta regression with by-participant and by-item random effects was fitted to the experimental data. Form probability was obtained using Calderone, Hathout & Bonami (2021)'s methodology.

Results & Discussion



Form predictability has a positive effect on well-formedness judgements (fig. A): on the margin, an increase of one standard deviation in form probability will lead to a 6.45% increase in well-formedness score. The result corroborates previous empirical findings on the cognitive relevance of paradigmatic predictability, and confirms the importance of treating anomaly as a matter of degree.

Participants are more generous when scoring forms in less frequent cells (fig. B). The mean scores for infinitive and past participle forms (the two most frequent verbal cells in French) are 15% lower than those for the two present indicative cells (of middling frequency for French verbs). Speakers are willing to accept rarer patterns more readily when they are less familiar with the distribution of possible patterns in a cell. Cell frequency also dictates how confident speakers are about the distribution of patterns: the importance of form predictability for well-formedness judgements is proportional to the frequency of the cell (fig. C). Contrary to Jun & Albright (2016), speakers exploit paradigmatic predictability relationships between two cells bidirectionally. In fact, once other factors have been controlled for, form predictability matters most when predicting the INF, which is both the citation form and the best overall predictor of the rest of the paradigm. Despite their opposite conclusion, Jun and Albright's results are compatible with ours: we predict that once cell frequency is taken into account, their findings can be given the same interpretation as ours.

References: Ackerman, F. & Malouf, R. 2013. 'Morphological organization: The low conditional entropy conjecture.' Language 89 – **Albright, A**. 2002. 'Islands of Reliability for Regular Morphology: Evidence from Italian'. Language 78 – **Albright, A., & Hayes, B.** 2003. 'Rules vs. analogy in English past tenses: a computational/experimental study'. Cognition 90 – **Bonami, O & Beniamine, S.** 2016. 'Joint predictiveness in inflectional paradigms'. Word Structure 9 – **Bybee, J. & Slobin, D.** 1982. 'Rules and Schemas in the Development and Use of the English past Tense'. Language. 58 – **Calderone B. & Hathout. N & Bonami, O.** 2021 'Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection' 18th SIGMORPHON Workshop **– Copot, M. & Bonami, O.** 2021. 'Spare us the surprise: the interplay of paradigmatic predictability and frequency'. Talk given at ISMo 2021 – **Divjak, D.** 2019. 'Frequency in language: Memory, attention and learning'. Cambridge UP – **Filipović Đurđević, D. & Milin, P.** 2019. Information and learning in processing adjective inflection. Cortex 116 – **Herce, B. 2019.** Deconstructing (ir)regularity. Studies in Language. – **Jun, J., & Albright, A.** 2017. 'Speakers' knowledge of alternations is asymmetrical: Evidence from Seoul Korean verb paradigms'. Journal of Linguistics 53 – **O'Donnell, T.J.** 2015. Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage**.** – **R- Rumelhart, D. E. & J. L. McClelland** 1986. `On learning past tenses of English verbs´. In Parallel Distributed Processing: Vol 2, MIT Press – **Smolensky, P.** 1995. `Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture´. In: Macdonald, C., Macdonald, G. (Eds.), Connectionism: Debates on Psychological Explanation. Blackwell.