Spare us the surprise: the interplay of paradigmatic predictability and frequency

Maria Copot *Olivier Bonami*
Université de Paris Université de Paris

1 Background

There is a generally recognised relationship between frequency and the paradigmatic predictability of a word form: word forms that are paradigmatically unpredictable (such as suppletive or otherwise highly irregular forms) tend to be frequent, while infrequent word forms tend to be highly paradigmatically predictable. In other words, it is within very frequent lexemes or very frequent paradigm cells that we tend to find unpredictable word forms. When unpredictability of form is encountered outside these contexts, there is a diachronic push to regularise it (the praeterite of English HELP used to be *holp*, now regularised to *helped*), or for the whole context to fall out of use (see the ongoing decline of the Italian *passato remoto*).

This relationship is rooted in the communicative function of language, and the way this interacts with memory: the more high-frequency a syntactic 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, 2019).

Syntagmatic predictability is well known to facilitate access during language use (Hale, 2001; Levy, 2008; Frank, 2013). Less is known about the role of word form predictability (though see Milin et al. (2009) for an overview from a paradigmatic perspective). We set out to test the hypothesis that, at parity of lexeme frequency, less paradigmatically predictable word forms will be used less frequently, as they are less easily accessible than their counterparts (a host of causal factors can be invoked here: for example, less predictable forms are a case of rare behaviour, so their neighbourhood will be less populated, making access more difficult). This general tendency is predicted to be absent for very high lexeme frequency: as very frequent lexemes are overall highly accessible, so are their individual phonological word forms.

2 Motivation

With the goal of better understanding the relationship between frequency and predictability, we perform a corpus study in which we attempt to predict the frequency of a word type based on the frequency of the lexeme it belongs to, and its paradigmatic predictability. We hope to provide an operationalisation of form predictability that is

- empirical: it is derived bottom-up from morphological data, and lines up with how predictability is characterised in other domains of language.
- paradigmatic: it makes use of the paradigmatic structure of morphological data, in a way that emulates emerging evidence about how speakers exploit said paradigmatic structure in language use.

- continuous (as a corollary property, falling out from the other two). We therefore make the falsifiable prediction that we don't expect the effect of predictability to be categorical.

The hypothesis we wish to test is that at parity of lexeme frequency, words that are less paradigmatically predictable will be used less frequently, since they are more difficult to access than their predictable counterparts, due to the low type frequency of the pattern they instantiate. Because we are investigating the effect of predictability and lexeme frequency at the level of individual words, we expect that predictability will be weighted differently at different levels of lexeme frequency: if the overall frequency of a lexeme is high, then its predictability will matter less – since frequent word forms have their own representation in the mental lexicon, the speaker does not need to predict them in order to use them, but rather they just need to retrieve them from memory. We expect all these to show up as gradient effects, partly due to the impact of all sorts of other factors on the accessibility of mental representations, but chiefly because we believe the effect to truly be gradient.

3 Operationalising Predictability

We adopt an information-theoretic view of paradigmatic predictability (Ackerman et al., 2009), whereby a word is predictable inasmuch as its shape is unsurprising given the rest of its paradigms and the distribution of inflectional patterns in the language. More precisely, we use the Qumín package (Beniamine, 2018) to identify, for all pairs of words (w_1, w_2) filling the same two paradigm cells (c_1, c_2), the alternation pattern relating these two cells. From this we can estimate the conditional probability of the word in c_2 having the shape w_2 given that the word in c_1 has the shape w_1, based on the statistical distribution of patterns relating c_1 and c_2. The PARADIGMATIC SURPRISAL of w_2 in c_2 given knowledge of c_1 is the negative logarithm of this conditional probability: the more frequent the pattern relating w_1 and w_2 is among viable alternatives, the lower the surprisal. We use the paradigmatic surprisal of a word form filling a paradigm cell given knowledge averaged over all possible predictors as our estimation of paradigmatic predictability.

In the present case study on French, all calculations rely on applying Qumín to the full paradigm of the 4951 nondefective verbs in the Flexique database (Bonami et al., 2014).

4 Surprisal and frequency

In order to investigate the relationship between form predictability (operationalised as paradigmatic surprisal) and frequency (at the level of the lexeme, the cell, and the word form), we perform a corpus study on the French verbal system. For frequency data, we extracted word form and lexeme frequency from FrCoW (Schäfer & Bildhauer, 2012). Whenever lexeme annotations were missing, we converted the token into the most appropriate lexeme given the POS tag using Levenshtein distance.

Because French conjugation exhibits widespread syncretism, for many paradigm cells, it is not possible to estimate frequency reliably. We hence decided to focus on those cells where a sizeable portion of the lexicon (at least 250 lexemes) uses a form with no homograph documented in the GLÀFF (Hathout et al., 2014)). We also excluded cells out of current usage such as the past subjunctive, for which attestations might be archaic or ironic. After this filtering, 14 cells are left for modeling. Separate bayesian poisson models were fitted to each cell, each predicting token frequency based on lexeme frequency, average surprisal, and their interaction. For the reasons discussed at the start of the section, we predict

- lexeme frequency to always have a positive coefficient - tautologically the more frequent a lexeme, the more frequent the words that belong to it
- surprisal to always have a negative coefficient - once lexeme frequency is taken into account, words that are harder to predict should be used less.
- the interaction coefficient should have a positive sign - we expect that for high values of lexeme frequency, surprisal should progressively matter less, since the language user's task is to remember the form rather than predicting it.

5 Results

These predictions are largely borne out. Three cells are exceptional: the INFINITIVE, the PRESENT PARTICIPLE and the IMPERFECT 3SG. For these cells, at least one of the coefficients involving surprisal is either very small and of unexpected monotonicity, or with an effect indistinguisheable from 0. Importantly, this exceptional behaviour is attested in what are the three most frequent cells under consideration. We hypothesise that because these cells are so frequent, the decreasing importance of surprisal doesn't just hold for the most frequent lexemes but rather it applies to most items in the entire cell: because the cell (and therefore words within it) is so frequent, its word forms are easily accessible directly, which diminishes the speaker's reliance on deducing it based on other forms of its paradigm.

Cell	Lexeme freq.	Surprisal	Interaction
FUT.1SG	0.9935	–0.3783	0.0675
FUT.2SG	1.0771	–0.2306	0.0447
FUT.3SG	1.1764	–0.0261	0.0073
FUT.1PL	0.9693	–0.1932	0.0415
FUT.2PL	1.1072	–0.3368	0.0647
FUT.3PL	1.1466	–0.0040	0.0088
COND.3SG	1.2509	–1.0392	0.1835
COND.1PL	1.2544	–1.7739	0.2876
COND.2PL	1.2583	–2.7622	0.4486
COND.3PL	1.2312	–1.3889	0.2404
IPFV.3SG	1.1707	–0.0441	🟥 –0.0010
IPFV.3PL	0.9352	–0.5588	0.0959
PRS.PTCP	0.5916	🟦 0.0545	🟦 0.0053
INF	0.9438	🟥 0.0620	🟥 –0.0089

Coefficient values by cell

🟥 Unexpected coefficient sign
🟦 95% Credible interval overlaps with zero

6 Conclusion

This work proposes an operationalisation of form predictability that is empirical, gradient and inherently paradigmatic. In the corpus study described, paradigmatic surprisal appears to capture well language users' reticence to employ forms that are hard to predict at parity of lexeme frequency. The study also provides insight into the relationship between form predictability and frequency: for very frequent lexemes and paradigm cells, form predictability matters progressively less to the language user since the frequent word form, no matter how unpredictable, already has a representation in memory and does not need to rely chiefly on being deduced based on paradigmatic information.

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*, 54–82. Oxford: Oxford University Press.

Beniamine, Sacha. 2018. *Typologie quantitative des systèmes de classes flexionnelles*: Université Paris Diderot dissertation.

Bonami, Olivier, Gauthier Caron & Clément Plancq. 2014. Construction d'un lexique flexionnel phonétisé libre du français. In Franck Neveu, Peter Blumenthal, Linda Hriba, Annette Gerstenberg, Judith Meinschaefer & Sophie Prévost (eds.), *Actes du quatrième congrès mondial de linguistique française*, 2583–2596.

Filipović Đurđević, Dušica & Petar Milin. 2019. Information and learning in processing adjective inflection. *Cortex* 116. 209–227. doi:https://doi.org/10.1016/j.cortex.2018.07.020. https://www.sciencedirect.com/science/article/pii/S0010945218302375. Structure in words: the present and future of morphological processing in a multidisciplinary perspective.

Frank, Stefan L. 2013. Uncertainty reduction as a measure of cognitive load in sentence comprehension. *Topics in Cognitive Science* 5(3). 475–494. doi:https://doi.org/10.1111/tops.12025. https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12025.

Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the north American chapter of the association for computational linguistics*, https://www.aclweb.org/anthology/N01-1021.

Hathout, Nabil, Franck Sajous & Basilio Calderone. 2014. GLÀFF, a Large Versatile French Lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106(3). 1126–1177. doi:https://doi.org/10.1016/j.cognition.2007.05.006. https://www.sciencedirect.com/science/article/pii/S0010027707001436.

Milin, Petar, Dusica Filipovic Durdevic & Fermin Moscoso del Prado Martin. 2009. The simultaneous effects of inflectional paradigms and classes on lexical recognition: Evidence from serbian. *Journal of Memory and Language* 60. doi:10.1016/j.jml.2008.08.007.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In *Proceedings of the eighth international conference on language resources and evaluation*, 486–493.