Matías Guzmán Naranjo* and Laura Becker

Coding efficiency in nominal inflection: expectedness and type frequency effects

https://doi.org/10.1515/lingvan-2019-0075
Received November 1, 2019; accepted December 21, 2020

Abstract: Since (Zipf, George Kingsley. 1935. *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press; Zipf, George Kingsley. 1949. Human behavior and the principle of least effort. *Journal of Consulting Psychology* 13(3)), it has been known that more frequent lexical items tend to be shorter than less frequent ones, and this association between the length of an expression and its frequency has been applied to various grammatical patterns (syntactic, morphological, and phonological) and related to predictability or expectedness in the typological literature. However, the exact interactions of frequency and expectedness, their effect on shortening, and the mechanisms involved, are still not well understood. This paper proposes the Form-Expectedness Correspondence Hypothesis (FECH), taking into account not only the frequency of expressions but their overall structure and distribution, and explores the FECH in the domain of nominal inflection from a quantitative perspective.

Keywords: coding efficiency; entropy; form-expectedness correspondence hypothesis; inflection morphology; quantitative typology

1 Introduction

Since Zipf's findings concerning the exponential relation between the frequency of words and their length (Zipf 1935, 1949), a number of typological studies have examined this relation and extended it to grammatical markers and syntactic constructions, using frequency as an explanation for crosslinguistic trends. In his 1966 work, Greenberg observes that there is a universal trend against marking all features of a single category. For instance, most languages either have zero-marking in the singular; or, in languages in which the singular is morphologically marked, it is almost always marked by less, or at least not more linguistic material than the plural. Greenberg (1966, pp. 37–38) notes for languages with case marking that the so-called direct cases (nominative, accusative, vocative) are generally shorter or zero-expressed, as opposed to oblique cases (possessive, genitive, dative), which are all morphologically marked (by more material).

Greenberg (1966) also showed that the unmarked features, which are expressed by less or at least not more linguistic material than the marked ones, occur more frequently than their marked counterparts. Similarly, frequency as a driving mechanism of grammaticalization and as one of the explanations for the crosslinguistically attested patterns, especially in terms of the systematic distribution of formally unmarked or more marked forms, has received a lot of attention in the typological literature (e.g. Bybee 2001, 2007; Comrie 1986; Croft 2003; Diessel 2007; Du Bois 1987; Haiman 1983, 1985; Haspelmath 2008a, 2008b; Haspelmath 2021; Haspelmath et al. 2014; Hawkins 2004, 2014).

We can distinguish three types of frequency-related accounts that all resort to efficiency as the explanatory factor. Previous studies have argued that frequent expressions are subject to shortening over time. In turn,

We thank Natalia Levshina, Steven Moran, Olivier Bonami, Anne Abeillé, and Maria Copot, and two anonymous reviewers for their useful comments and suggestions. This work was partly supported by a public grant overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (reference: ANR-10-LABX-0083).

***Corresponding author: Matías Guzmán Naranjo**, LLF, CNRS, Université de Paris, Paris, France, E-mail: mguzmann89@gmail.com
Laura Becker, University of Freiburg, Erlangen, Germany

shortening over time has been explained in two different ways. One line of argumentation is based on efficient speech production, which involves phonetic erosion or reduction of frequent phonetic material over time due to automatization in speech production (e.g. Bybee 2001, 2007). In addition, it has been argued that, if the information conveyed by a given expression is more frequent, the speaker can assume that it will be more expected by the hearer. Thus, the speaker can afford a shorter/shortened expression without loss of information (Comrie 1986; Croft 2003; Du Bois 1987; Haiman 1983, 1985; Haspelmath 2008a, 2008b; Haspelmath 2021; Haspelmath et al. 2014; Hawkins 2004, 2014). The opposite way of looking at those form-frequency effects is to argue that less frequent markers are longer than more frequent markers, because less frequent concepts are not as expected as frequent ones and thus may lead to the development of a longer marker (Croft 2003, p. 116).

This relation between frequency and length of expression has been labelled the form-frequency correspondence universal in Haspelmath (2021), stating that "[l]anguages tend to have shorter forms for more frequent meanings". Number marking provides robust evidence for such form-frequency correspondences in the nominal inflectional domain; that the frequency and the length of singular < plural < dual markers is negatively correlated to their respective frequencies (SG > PL > DU) has been demonstrated for various languages, with very few exceptions. For case marking, the picture is less clear. The relation between subject (nominative) and direct object (accusative) cases has often been discussed as one example of form-frequency correspondences (Croft 2003; Diessel 2007, 2019; Greenberg 1966; Haspelmath 2021). Even though it is not clear to what extent the frequency distribution of cases is comparable across languages (Downing and Stiebels 2012), studies like Du Bois (1987) or Primus (1999) have argued and shown that subjects (nominative/absolutive markers), are more frequent than direct objects (accusative/ergative markers). However, as far as we are aware, there are no studies exploring how form-frequency correspondence effects interact with marker variation across inflection classes, e.g. when grammatical categories or combinations thereof have more than one exponent.

In Czech, for instance, in some inflection classes, the nominative singular corresponds to the stem of the lexeme (1a), while in other inflection classes it is the genitive plural which corresponds to the stem (1b). This means that not only can one cell of the paradigm be expressed by markers of different lengths, but also the relation of shorter/longer markers between cells can be reversed.

(1) Czech
   a. *koš-ϕ* 'basket-SG.NOM' vs. *koš-**ů*** 'basket-PL.GEN'
      (Janda and Townsend 2000, pp. 17–20)
   b. *žen-**a*** 'woman-SG.NOM' vs. *žen-ϕ* 'women-PL.GEN'
      (Janda and Townsend 2000, pp. 17–20)

Therefore, in the present study, we address the nature and possible efficiency-based explanations of frequency effects in nominal inflection markers, exploring whether frequency and related effects also hold for concrete inflectional exponents rather than abstract grammatical functions. Since different exponents of the same cell arguably have the same meaning, this allows us to disentangle frequency effects from potential semantic confounds.

Moreover, the extension from grammatical functions to concrete exponents allows us to explore the effects of the expectedness of a single marker, which does relate to its frequency but also to the different functions it expresses. Thus, in addition to the frequency of a marker, we can examine the effect of its distribution in nominal paradigms and of its relation to other, competing markers expressing the same grammatical function on its length. This is important, as recent work (e.g. Haspelmath 2021; Hawkins 2014) has suggested that frequency may not have a direct effect on the length of the expression of a given grammatical category. Instead, high frequency leads to the expression being highly expected, which in turn leads to its efficient shortening over time. From this proposal it follows that frequency does not necessarily have a direct impact on the length of an expression if other properties of the structure of the system can cause an expression to be highly expected, potentially leading to low-frequency markers becoming shorter (cf. Diessel 2007; Givón 1983; Haiman 1983).

Morphological paradigms are an interesting testing ground for comparing frequency with other structure-related measures of expectedness in their impact on marker length and thus efficiency. Morphological and psycholinguistic work on paradigms has shown that there are complex relations between the cells of a paradigm, which aid speakers to learn and produce new forms (Ackerman et al. 2009; Ackerman and Malouf 2013, 2016; Aronoff 1994; Bonami and Beniamine 2016; Cotterell et al. 2019). Thus, the expectedness of a given marker does not only depend on its overall frequency (token frequency), but also on how many lexemes it is used with (type frequency), on how many other markers can express the same cell of the paradigm, and on its overall distribution across cells (the number of functions that this marker can express).

Although the typological literature makes use of the terms *expected* and *predictable* interchangeably, we will only use the former. By expected and expectedness we mean the following: marker *A* is highly expected in a given context *C*, if *A* has a high frequency in *C* compared to the alternatives in that context. This can be due to the overall high token frequency of A, but also due to its high type frequency or a small number of alternative, competing markers (we introduce marker/cell entropy, as well as marker/cell flexibility as additional measures of expectedness in Section 3). With this in place, we propose the following Form-Expectedness Correspondence Hypothesis (FECH):

(2) *Form-Expectedness Correspondence Hypothesis:*
Grammatical markers which are more expected tend to be expressed by less or at least not more material than grammatical markers which are less expected.

In this paper, we present a quantitative, crosslinguistic approach to examining the relevance of the FECH for inflectional morphology, comparing the effect of type frequency with the effect of other measures suggested in the morphological literature that can serve as predictors of the length of nominal inflection markers. After presenting our method and materials in Section 2 and introducing the relevant expectedness measures in Section 3 we examine the relation between the latter and the length of different nominal inflection markers across languages. Section 4 presents our models and results. As we will discuss in Section 5.1, our results support the FFEC, i.e. that expectedness is at least one of the driving forces of the effects on marker length. Finally, Section 5.2 shows that our findings concerning type frequency may extend to token frequency as well.

2 Methods and materials

2.1 Dataset

For the present study, we used the UniMorph database (Kirov et al. 2018), a large-scale morphological database with complete inflection paradigms of verbs, nouns, and adjectives for 111 languages. From these, we selected the languages with nominal inflection paradigms. Our final dataset contains 60 languages.[1] The languages of the final dataset together with the number of available lexemes, forms and mean marker lengths can be found in the supplementary materials; the final dataset contains a total of 12,365 observations.

The consequence of using a morphological database is that we can only examine the type frequency of inflection markers. There are two reasons for this decision: First, examining Zipfian token frequency effects in low frequency markers would require large, morphologically annotated corpora which do not exist for enough different languages. Second, even in a relatively large corpus low frequency lexemes do not appear in all their inflected cells (Bonami and Beniamine 2016; Janda and Tyers 2018; Kettunen and Airio 2006).

2.2 Case and number markers

The definition of an inflection marker is not uniform in the literature, since it is often difficult to define morphemes and especially morpheme boundaries (see Baerman 2015). A straightforward analysis of what

1 See Appendix 1 for a detailed description of the data selection process.

constitutes stems, affixes, and morphemes may not always be available. Motivations for morpheme boundaries are even more problematic when non-affixal markers such as stem alternations, reduplication, or tones are involved. For example, the German word *Haus* 'house' has the plural *Häuser*, in which both the stem umlaut /aʊ/ → /ɔɪ/ and the suffixal marker /ɐ/ are used to express the plural meaning. Instead of a morphemic approach, we propose a language-independent definition of an inflectional marker that can be fully operationalized. Our approach is partially based on the work by Beniamine (2018) and follows a mostly Word and Paradigm perspective of morphology (Ackerman and Malouf 2016; Blevins 2016). We define the stem of a lexeme as the phonological material common to all its inflected forms. We define a marker to be the phonological material which distinguishes a single inflected form from the stem. Since this definition is language agnostic, we extracted markers automatically for all languages in the dataset.[2]

3 Frequency, expectedness and paradigm structure

As mentioned in Section 1, frequency resulting in more efficient coding has been associated with expectedness in a way that expected forms tend to be shorter than less expected ones. The FECH, formulated in (2), captures this relation between the length of markers and their expectedness.

In the literature on morphological complexity there have been several proposals for measuring uncertainty and expectedness, mostly in terms of principal parts (Stump and Finkel 2013) and in terms of entropy and conditional entropy (Ackerman et al. 2009; Ackerman and Malouf 2013, 2016; Bonami and Beniamine 2016; Hume and Mailhot 2013). In this paper, we use entropy and type frequency to quantify the concept of expectedness. Entropy is a summary metric that captures the distribution of values in a random variable. For instance, given the random variable X with its possible value being x_1, x_2, …, or x_n, the entropy $H(X)$ captures how much uncertainty we have about the true value of X.[3]

In order to test the FECH in nominal inflection we examine the relation between the following six different predictors and the length of inflection markers: (i) relative marker frequency, (ii) relative cell frequency, (iii) marker flexibility, (iv) cell flexibility, (v) cell entropy, and (vi) marker entropy. We take a cell to be a full set of morphosyntactic features. This means that if nouns of a given language inflect for case and number, the instrumental singular counts as a single cell, while singular or instrumental do not make up cells on their own.

The relative *marker frequency* and relative *cell frequency* correspond to the total number of nouns that a marker or a cell occurs within the dataset (within languages), measured in parts per million in order to make the counts comparable across languages. Including *cell frequency* is important, as not all nouns in a given language can always inflect for all the cells of the nominal paradigm. Regarding *cell frequency*, one would expect that cells which can be expressed by a larger number of nouns should have shorter markers than cells which are only available for a smaller number of nouns.[4] The *relative marker frequency* is counted by cell, i.e. if two different cells of the paradigm use the same, syncretic marker, it is treated as distinct markers according to their cell resulting in different marker frequencies. Table 1 shows this for Modern Greek, where the marker *-is* is used in different cells of the paradigm, namely as the plural marker for nominative, accusative, and vocative. In this case, the marker *-is* receives three different frequency counts for each of its functions in the paradigm. The *marker flexibility* corresponds to how many cells of the paradigm the marker expresses. In the case of Modern Greek *-is*, the flexibility count is 3. The *flexibility* of a cell refers to the number of different markers which can occur in that cell (discontinuous markers are counted in as different markers).

2 For details on the marker extraction process see Appendix 1.

3 See Shannon (1948) for the original proposal of entropy in information theory and Ackerman et al. (2009: 63) for a detailed description of entropy in the context of morphology.

4 Diverging cell frequencies within single languages mostly occur with defective lexemes.

Table 1: Distribution of the nominal inflection marker *-is* in Modern Greek (Holton et al. 2004: 34).

Cell	Form	Stem	Marker
NOM.SG	Koureas	Koure	as
NOM.PL	Koureis	Koure	**is**
ACC.SG	Kourea	Koure	a
ACC.PL	Koureis	Koure	**is**
GEN.SG	Kourea	Koure	a
GEN.PL	Koureon	Koure	on
VOC.SG	Kourea	Koure	a
VOC.PL	Koureis	Koure	**is**

For instance, the nominative singular in Icelandic can be expressed by the six markers shown in (3), so that its *cell flexibility* count is 6.

(3) Markers for the nominative singular cell in Icelandic
  a. hest-**ur** 'horse-NOM.SG'
  b. himmin-**n** 'sky-NOM.SG'
  c. snjó-**r** 'snow-NOM.SG'
  d. penn-**i** 'feather-NOM.SG'
  e. húf-**a** 'cap-NOM.SG'
  f. borð-ϕ 'table-NOM.SG'
  (Kress 1982, pp. 56–79)

The *marker entropy* captures a marker's distribution across different cells in the paradigm. Intuitively, this metric tells us how well a marker determines the cell that it expresses, i.e. how strongly it is associated with a single cell or function of the nominal inflection paradigm. If, for example, marker *-X* only appears in CELL-1, then it has an entropy of 0. If, however, it appears (in equal proportions) in CELL-1, CELL-2, CELL-3 and CELL-4 it will have an entropy of 2. A lower entropy means that a marker is strongly associated with few cells, and thus speakers can easily deduce the cell being expressed by knowing the marker. Meanwhile, higher entropy means that a marker is more ambiguous as to which cells it expresses, because it is distributed more evenly across many different cells. *Cell entropy* is the opposite of *marker entropy*: It captures how well a cell determines the marker by which it is expressed. A cell which is overwhelmingly expressed by a single marker will have a low entropy, while a cell which is often expressed by different markers or which is expressed by different markers to a similar extent will have a high entropy.

Assuming that not only frequency but also the structure-related metrics of expectedness are relevant for efficient coding in nominal inflection, we would expect to see that predictors such as *marker flexibility* and *marker entropy* also have an effect on the length of the marker.

4 Results

In order to examine the relation between marker length and the six expectedness measures described in the previous section, we used a Hamiltonian Monte Carlo process with STAN (Carpenter et al. 2017) to fit a series of models to our data. We used the BRMS interface with R (Bürkner 2017, 2018).[5] The final model was a Hurdle Poisson model fitted with the formula given in 4:

5 For details on model specification and evaluation see Appendix 2.

(4)
```
marker_length  1 + marker_frequency + cell_frequency +
marker_flexibility + cell_flexibility + marker_entropy +
cell_entropy + (1 | language) + (1 | language:cell),
hurdle  1 + (1 | language)
```

For the model fit, all predictors were standardized in order to make them comparable in terms of effect sizes. Additionally, we divided our dataset into training and test datasets with 90 and 10% of the data respectively. All model fit statistics reported were carried out on the test dataset. The population-level effects (fixed effects) are the estimates for the six predictors. The results of the model are shown in Table 2. The intercept corresponds to the expected length of a marker when all predictors are equal to 0. The intercept corresponds to 2.16 segments, since the coefficients of a Poisson model are in a logarithmic scale. This is the grand mean; mean marker lengths for individual languages are modulated by the group-level intercepts. Negative values in the estimates of the population-level effects mean that there is a negative correlation between the predictor and the outcome. The positive value for *cell entropy*, on the other hand, means that larger predictor values are correlated with longer markers.

The model summary in Table 2 shows the following main findings: As expected, a higher *marker frequency* predicts the marker to be slightly shorter. *Cell frequency* is also negatively correlated with marker length. This means that a marker which expresses a more frequent cell in a given language will be shorter on average than a marker which occurs in a less frequent cell. For *cell flexibility*, the model shows that markers which appear in cells expressed by many different markers can be shorter than markers which appear in highly specialized cells. *Cell entropy* has an effect in the opposite direction: Low-entropy cells have slightly shorter markers.

The Hurdle intercept indicates how likely zero segments are for a given marker: an estimate of −3.78 means that there is a low probability of 0.02 that the model will predict a zero marker. The model additionally allows variation by language of the Hurdle intercept (not shown here), which means that for some languages zero markers are estimated to be more or less likely.

We used the Bayesian R^2 (Gelman et al. 2019) as a measure of model fit. A Bayesian R^2 value ranges from 0 to 1 and captures the variance explained by the model. Table 2 reports the R^2 value for the performance of the model on the training and test datasets, which are relatively high in both cases. In other words, we can predict marker length from the predictors included in the model very well.

Table 2: Model coefficients.

Model coefficients				
~Language (number of levels: 60)				
	Estimate	Est. Error	l-95% CI	u-95% CI
sd (Intercept)	0.26	0.04	0.20	0.34
~Language: cell (number of levels: 1,343)				
	Estimate	Est. Error	l-95% CI	u-95% CI
sd (Intercept)	0.12	0.01	0.11	0.14
Population-level effects:				
	Estimate	Est. Error	l-95% CI	u-95% CI
Intercept	0.63	0.07	0.49	0.75
Marker frequency	−0.08	0.01	−0.09	−0.06
Cell frequency	−0.06	0.01	−0.09	−0.03
Marker flexibility	−0.22	0.02	−0.26	−0.18
Cell flexibility	−0.06	0.02	−0.09	−0.02
Marker entropy	−0.29	0.02	−0.33	−0.25
Cell entropy	0.04	0.02	0.01	0.07
Hurdle intercept	−3.78	0.21	−4.20	−3.39
(Mean) training dataset Bayesian R^2	0.76	0.004	0.75	0.77
(Mean) test dataset Bayesian R^2	0.72	0.006	0.71	0.73

Figure 1 shows the group-level effects of language, revealing clear language-specific differences.[6] Languages like Azerbaijani, Belarusian, and Ancient Greek have considerably shorter markers overall; however, for Azerbaijani, as the most extreme language on the short end in our dataset, Figure 1 shows very wide credible intervals. In this case, the estimation is less precise because the dataset contains fewer datapoints (this also holds for e.g. Kashubian, Old French, or Tajik). On the other hand, languages like Quechua, Turkish, and Bengali are predicted to have comparatively longer markers. Very long markers, especially in agglutinative languages, may be accounted for by the method used in this study to extract the markers. Since the extraction was based on the segments that differ across cells, strings containing e.g. number and case markers were not further subdivided into smaller strings or markers. Quechua, for instance, is predicted to have considerably longer markers than all other languages in the dataset. Nominal inflection in Quechua marks possession, number, and case. The plural suffix *-kuna*, for instance, consists of four segments and can occur together with possessive markers and/or case markers, as in *churi-n-kuna-ta* 'child-3-PL-ACC (their children)' (Shimelman 2016: 151). As we do not further segment markers, the third person possessive marker *-n*, the plural marker *-kuna*, and the accusative marker *-ta* are treated as a single marker of 1 + 4 + 2 = 7 segments, filling the ACC.PL.POSS:3 cell.

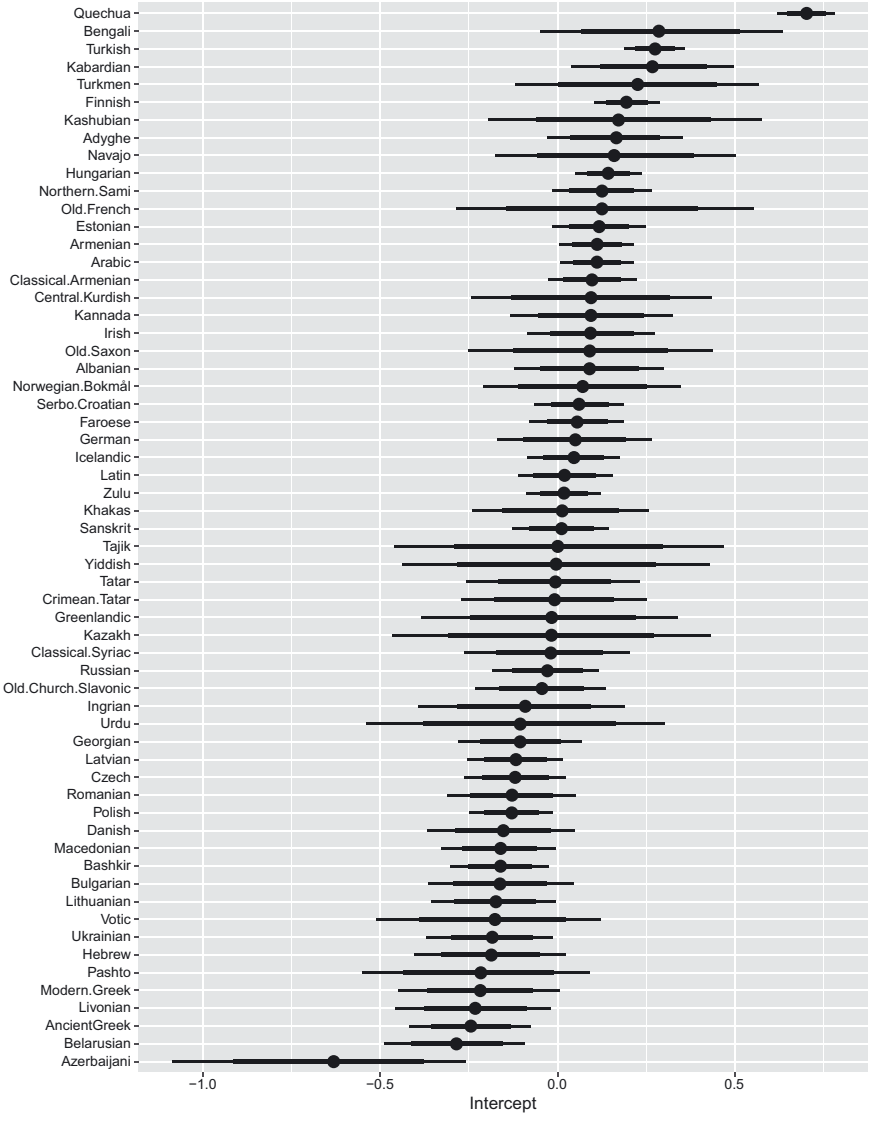

Figure 1: Language group-level effects.

6 In Figure 1, dots mark the estimate, while the 95 and 50% credible intervals are marked by thin and thick bars, respectively.

Finally, we want to explore the relative proportion of the variance explained by the individual predictors. To do this, we fitted six additional models for each of the six predictors separately, a model with all predictors except marker frequency, and a final model without population-level effects and only group-level effects. Table 3 shows the R^2 values for each of these models. Indeed, *marker frequency* can explain a portion of the variance on its own (R^2 = 0.64). However, *marker frequency* alone does not necessarily account for a much larger portion of the variance in marker length than the group-level effects alone (R^2 = 0.59). Also, the remaining predictors contribute to the overall variance explained by the model in Table 2 as well and seem to be at least comparably informative (R^2 = 0.7 for *marker flexibility* and *marker entropy*, and R^2 = 0.71 for all predictors except *marker frequency*).

5 Discussion

5.1 Type frequency, expectedness and coding efficiency

Our model shows that form-expectedness and structure related measures effects are not simply based on the frequency of abstract grammatical functions such as singular or plural, but on the frequency of form-cell pairs. We can interpret the effects of *marker flexibility* and *marker entropy* as follows: A more flexible marker appears in more cells of a paradigm, and a higher-entropy marker is distributed more evenly across the paradigms(s) of a language, which means that the speaker can expect such markers to be present in more contexts and probably more often than markers that are less flexible and/or have a lower entropy, i.e. are more skewed towards a single cell. *Cell flexibility* and *cell entropy* in our study indicated how strong its association is with single markers. In the case of *cell flexibility*, markers of cells with more markers are predicted to be shorter than markers of cells with fewer markers. This finding may relate to what has been discussed as inflectional potential of cells (e.g. Croft 2003: 97): cells with a higher token frequency allow for more inflectional distinctions than cells with a low token frequency. If this also holds for effects of inflection classes, e.g. that we find a higher number of inflection class distinctions in the singular than in the plural, it might be the case that the effect of *cell flexibility* is related to the cell token frequency. A higher degree of *cell entropy*, on the other hand, is associated with longer markers. This means markers in cells with fewer or more evenly distributed markers are predicted to be longer than markers in cells with more and less evenly distributed markers. Similarly to *cell flexibility*, this finding could also relate to the token frequency of cells: cells with a lower token frequency have a lower inflectional potential and fewer markers which are more evenly distributed across lexemes than markers of cells with a high token frequency. This in turn could lead to cells with lower token frequencies having a higher *cell entropy*, which could account for the prediction that those cells will have longer markers.

The fact that all our predictors play a role in the model is challenging for the automatization and phonetic reduction as a single explanation of coding efficiency in nominal inflection. If more frequent strategies are shorter because of efficiency in repetition, we would not expect measures of paradigm structure to play a role.

Table 3: R^2 values for additional models.

Predictor	R^2	L-95% CI	U-95% CI
Only marker flexibility	0.7	0.67	0.72
Only marker entropy	0.7	0.67	0.72
Only marker frequency	0.64	0.6	0.67
Only cell frequency	0.6	0.57	0.63
Only cell flexibility	0.6	0.57	0.63
Only cell entropy	0.6	0.57	0.63
All but frequency	0.71	0.68	0.73
No population-level effects	0.59	0.57	0.63

Instead, our results support the FECH, namely that the length of a marker is associated with its expectedness, which is only partially accounted for by its frequency.

Another important finding from Section 4 concerns the role of zero markers. The regular Poisson model largely overestimated the number of zero markers. This may suggests that the coding efficiency we find on the basis of expectedness measures accounts well for the distribution of long versus short markers. Zero markers, however, seem to underlie additional restrictions, as they are less common than predicted by a regular Poisson model fitted with our predictors.

5.2 Type versus token frequency

A methodological complication of this study was that we only considered the type frequency of markers and not their token frequency, which is traditionally argued to relate to the length of a linguistic expression/construction, as in the form-frequency correspondence universal proposed in Haspelmath (forth.).

This section will briefly show for three languages that type and token frequency of nominal inflection markers are strongly correlated. Figure 2 shows the relation between the type and token frequency of single nominal inflection markers in Czech, Finnish, and Russian. We chose these languages because the Universal Dependencies project (Nivre et al. 2019) contains comparatively large morphologically annotated treebanks for them, which are often used in typological research Berdicevskis et al. (2018), Bouma et al. (2018), Naranjo and Becker (2018), and Levshina (2019).

The log token and type frequencies shown in Figure 2 are based on the token frequencies of our markers in the UD treebanks and on their type frequencies in the UniMorph data base, respectively.

Visual inspection of Figure 2 confirms that there is a strong correlation between log type and log token frequency. Table 4 shows the posterior estimates and 95% credible intervals of the Pearson correlation coefficients for each language. The main point of divergence between type and token frequency is at the lower end of type frequency: Those are markers which only occur with a small number of nouns but which can nevertheless have a high token frequency because the nouns they occur with have a high token frequency as well. This suggests that our findings using type frequency might also hold for a model using token frequency.

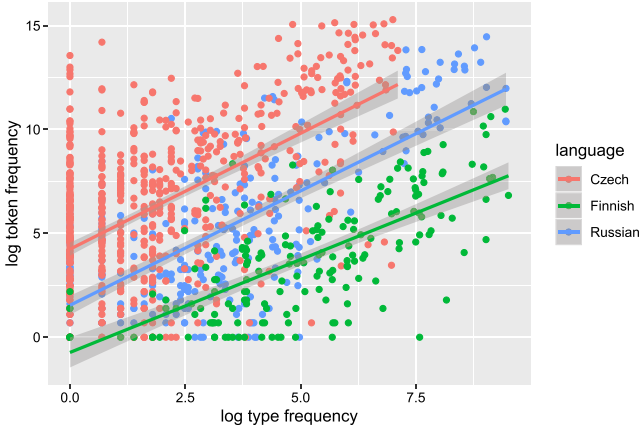

Figure 2: Token versus type frequency for Russian, Czech and Finnish inflection markers.

Table 4: Posterior estimation of correlation coefficients.

	Estimate	L-95% CI	U-95% CI
Russian	0.72	0.67	0.77
Finnish	0.71	0.63	0.78
Czech	0.57	0.52	0.62

6 Concluding remarks

In this paper, we proposed the Form-Expectedness Correspondence Hypothesis (FECH), relating the expectedness of grammatical markers to their length. The FECH is an extension of the Form-Frequency Corresponding Hypothesis. We presented a crosslinguistic quantitative study operationalizing and testing the FECH in the domain of nominal inflection, using type frequency and two measures of the distribution of markers across inflectional paradigms as predictors of the markers' length.

Our results confirm the FECH: more frequent inflection markers tend to be shorter than less frequent ones. Additionally, we showed that type frequency can only account for a portion of the variance in marker length, and that other measures of expectedness based on the structure of the paradigm and the distribution of single markers across different cells of the paradigm are also strong predictors of marker length. Thus, paradigm organization and marker distribution also have a strong impact on the length of nominal inflection markers. These results are consistent with an explanation of coding efficiency (partially) based on expectedness.

Appendix Coding efficiency in nominal inflection

A.1 Data extraction

This appendix discusses the method employed for marker extraction. We provide all data and code necessary to reproduce the results of the paper.

A.1.0.1 Data selection

From the dataset containing nominal inflection tables, we removed languages with fewer than 20 lexemes or those languages for which the extraction process only found markers that appeared with less than 10 lexemes. We also removed languages with too few lexemes, since they do not allow for reliable estimates. We removed markers with fewer than 10 attestations because the extraction process was not without errors, especially with nouns with suppletive forms in one cell of their paradigm. Removing low frequency markers mitigates the impact that such errors could have on our analysis. We additionally removed all markers that occurred with fewer than 5 lexemes for languages with 500 lexemes or fewer, and markers with fewer than 20 lexemes for languages with more than 500 lexemes.

A.1.1 Preprocessing

We used Epitran (Mortensen, Dalmia, and Littell 2018) for all supported languages in our dataset (14 out of 60) to generate a phonological transcription in order to work with phonological segments rather than with the orthography.[7] We worked with the orthography for the languages in the dataset which are not supported. This could slightly distort the length of the marker in the sense that digraphs or trigraphs representing single phonemes (e.g. *sh* for /ʃ/ in English) would overestimate the number of segments a marker has. In a similar way, if a digraph such as *sh* is systematically transcribed as /ʃ/, this may cause the number of phonological segments at morpheme boundaries as in *mis.hap* to be underestimated.[8] Both issues are close to impossible to

7 For the languages in our dataset supported by Epitran, see the file "table-paper.csv" in the supplementary materials. For most supported languages (the list of supported languages can be found at https://github.com/dmort27/epitran), Epitran does a very good job of producing a phonological transcription (see Mortensen, Dalmia, and Littell 2018). For Arabic, in particular, the transliteration provided by Epitran is not very reliable, because Arabic texts usually do not include vowels. However, UniMorph includes additional diacritics for Arabic, which indicate the correct vowels.

8 We thank Steven Moran for this example.

control for without careful manual cleaning of the data, and they may lead to some noise in our data. However, both scenarios should only concern a small number of markers in the overall dataset and thus not substantially influence the findings of this paper. A number of languages used diacritics to mark suprasegmental information. In languages such as Modern Greek and BCS (Bosnian-Croatian-Serbian), this information in the orthography is lexical and independent of the morphological alternations. We manually removed the diacritics in such cases in order to avoid that diacritics would cause the detection of an artificially higher number of inflection markers.[9]

A.1.2 Marker extraction

Under the definition of stems and inflection markers given above, extracting the stem of a lexeme consists of solving the Longest Common Substring problem (Arnold and Ohlebusch 2011) for all the inflected forms of that lexeme. Once the stem is determined for each lexeme, the inflection marker of each form of that lexeme equals the additional phonological material not present in the stem of the lexeme. Levenshtein's Distance (Levenshtein 1966) offers an effective and simple way of detecting such strings. This method finds an optimal alignment between strings S1 (the stem) and S2 (an inflected form) which minimizes the number of operations (insertion, substitution, and deletion) required to transform string S1 into S2. After aligning both strings, we can define the marker for S2 as the phonological material used in the operations to transform S1 into S2 (ignoring deletion). To give an example, Table A1 shows the paradigms of German *Vorwurf* 'reproach' and *Haus* 'house' (as transcribed by Epitran) with their stems and the extracted markers. Since those two nouns have both stem alternations and affixes for number and case marking, they show how this method deals with stem alternations that occur together with affixal markers.[10]

Another example of the approach is shown for the Russian word *batɕok* 'small tank' in Table A2. This example shows that the vowel *-o-* in the nominative and accusative singular is treated as a marker for those cells instead of deletion in the other cells. The transcription in Table A2 also shows that palatalization is treated as a marker when it is contrastive, as is the case in the prepositional singular, accusative plural, and nominative plural.

Defining markers and stems in this way has two important advantages for cross-linguistic comparison. First, the implementation is relatively simple, and it can handle the most common inflection strategies (affixation and stem mutation) found in our dataset. Second, it is not necessary to detect inflection classes for single languages, since markers are defined on a lexeme-by-lexeme basis. There are two potential

Table A1: Inflectional paradigms of *forvurf* 'reproach' and *haws* 'house'.

cell	form	stem	marker	form	stem	marker
NOM.SG	forvurf	forvrf	u	haws	hs	aw
NOM.PL	forvyrfə	forvrf	y-ə	hoysər	hs	oy-ər
ACC.SG	forvurf	forvrf	u	haws	hs	aw
ACC.PL	forvyrfə	forvrf	y-ə	hoysər	hs	oy-ər
DAT.SG	forvurf	forvrf	u	haws	hs	aw
DAT.PL	forvyrfən	forvrf	y-ən	hoysərn	hs	oy-ər
GEN.SG	forvurfəs	forvrf	u-əs	hawsəs	hs	oy-ərn
GEN.PL	forvyrfə	forvrf	y-ə	hoysər	hs	oy-ər

9 Note that we did not remove diacritics which distinguish phonemes. The German umlaut, for example, is not a suprasegmental mark which is why we kept it in the dataset.

10 The method used for marker extraction is part of the R package `Paradigma` (version 0.0.1.0), which can be found at https://gitlab.com/mguzmann89/paradigma.

Table A2: Inflectional paradigm of *бачок* 'small tank'.

cell	form	stem	marker
NOM.SG	batɕok	batɕk	o
ACC.SG	batɕok	batɕk	o
GEN.SG	batɕka	batɕk	a
DAT.SG	batɕku	batɕk	u
INS.SG	batɕkom	batɕk	om
PRE.SG	batɕkʲe	batɕk	ʲe
NOM.PL	batɕkʲi	batɕk	ʲi
ACC.PL	batɕkʲi	batɕk	ʲi
GEN.PL	batɕkov	batɕk	ov
DAT.PL	batɕkam	batɕk	am
INS.PL	batɕkamʲi	batɕk	amʲi
PRE.PL	batɕkax	batɕk	ax

disadvantages, however. First, this method cannot deal with replicative processes like lengthening or reduplication. This issue becomes apparent in, e.g. Finnish or Hungarian, where this approach fails to identify consonant lengthening as a more general phonological process of a single case marker, which results in too many markers that could be collapsed into fewer. Table A3 shows this for the instrumental marker in Hungarian. The suffix-initial consonant /v/ only surfaces with vowel-final stems; a stem that ends in a consonant marks the instrumental by lengthening that consonant together with the additional segment /ɒl/.

For the purposes of this paper, this issue is relatively minor; it only applies to a small number of markers, and treating the forms that are phonologically different as different markers of the same case is arguably a representation faithful to the surface structure of the case markers.

The second drawback of this method is that, as already mentioned, it cannot directly handle suprasegmental processes. This is not a major issue for the present paper either. Even if suprasegmental patterns in nominal inflection could be identified manually or automatically, it is not clear what the predictions of the FECH are regarding this type of markers, or how one should measure their length.

Finally, because there is some degree of error associated with the extraction process, we also removed low frequency markers. Removing markers below a certain frequency threshold ensures that (i) all the markers examined are present in at least a certain number of nouns, which makes it more likely that they are not found due to faulty extraction, and that (ii) lexemes with suppletive forms are excluded. In our implementation, lexemes with suppletive forms either have no stem or a stem which will produce markers unique to that lexeme. For example, if we assume *people* is the plural form of *person*, the stem would consists of *p*, as it is the longest common substring of both forms. However, the segmentation based on this stem produces the singular and plural markers *-erson* and *-eople*, respectively, which are unique to that lexeme and which lead to the exclusion of the lexeme.

Table A3: Consonant alternations in instrumental forms in Hungarian (own knowledge).

meaning	instrumental form
ship	hɒjoː**vɒl**
flower	viraː**ggɒl**
house	haːz**zɒl**

A.2 Model specification and evaluation

This appendix discusses some issues pertaining to model specification and model evaluation. As mentioned in the paper, we used a Hamiltonian Monte Carlo process with STAN (Carpenter et al. 2017) to fit a series of models to our data. We used the BRMS interface with R (Bürkner 2017, 2018). We made sure that for all models all chains were well mixed and that there were no divergent transitions after warm up. We did not observe any autocorrelation effects, and all models converged. The final model was a Hurdle Poisson model fitted with the formula given in (1).

(1)
```
marker_length ~1 + marker_frequency + cell_frequency +
marker_flexibility + cell_flexibility + marker_entropy +
cell_entropy + (1 | language) + (1 | language:cell),
hurdle ~1 + (1 | language)
```

A Hurdle Poisson model consists of two components: a regular Poisson model, and an initial hurdle which the model must overcome. The purpose of the hurdle is to handle an either very large or very small number of zeros. In our case, we have a lower than expected number of zeros from the perspective of a Poisson model. The factor `language` was added as a group-level effect (random effect) to allow for each language to have markers of different lengths. varying slopes. We also added `cell` by `language` (`(1 language:cell)`) to the group-level effects to account for the fact that different cells (within a language) may have longer or shorter markers on average. The latter controls for potential semantic effects, i.e. that a semantic case is longer than the nominative or that a cell combining more grammatical functions is longer than a cell combining less functions.

We also explored models using a truncated Gaussian distribution, a negative binomial distribution and a geometric distribution. Additionally, we explored Poisson models with factor interactions. We performed model selection using leave-one-out cross validation (Vehtari, Gelman, and Gabry 2017). Term interactions did not improve the overall model fit and had estimates of 0, which is why we do not report on interactions. Similarly, using families other than Poisson degraded the model fit. Finally, adding non-linear terms (splines or Gaussian processes) heavily deteriorated convergence, chain mixing, and overall model fit. Adding varying slopes to the model deteriorated model fit and caused divergence in the chains. Therefore, the paper only reports on the model without varying slopes.

Because all of our predictors are drawn from related facts about our dataset, multicollinearity is a potential issue. Multicollinearity happens when two predictors are highly correlated. This issue can lead to poor estimates of the coefficients in the model. We use the Variance Inflation Factor (VIF) to assess whether collinearity is a problem for our predictors (Dormann et al. 2013; Kock 2015). A VIF below 10 indicates an acceptable level of collinearity. In our case, all the predictors had a VIF between 1 and 5, which is why we conclude that collinearity is not an issue for our model and the interpretation of the coefficient estimates.

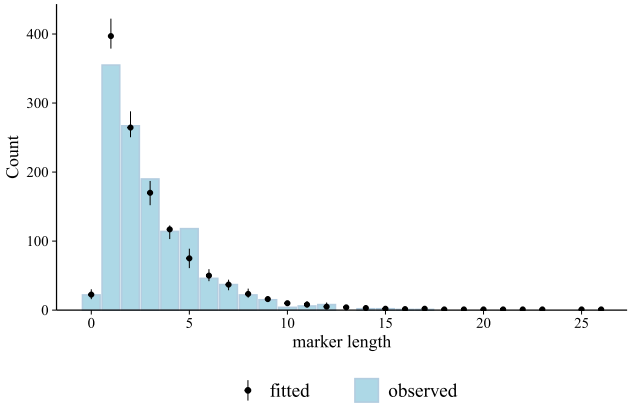

Figure A1: Posterior predictive check of the main model using ten draws.

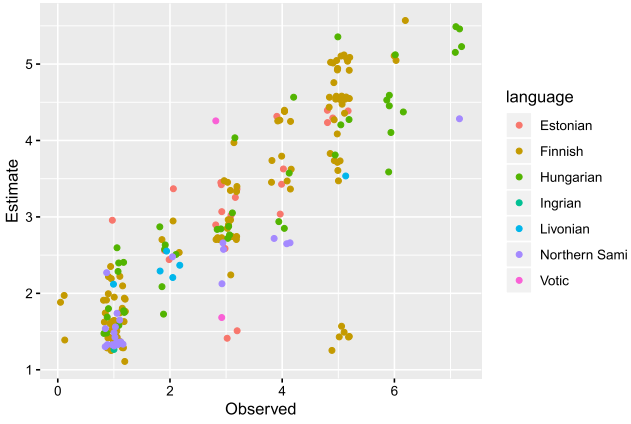

Figure A2: Predicted vs. observed marker lengths for Uralic languages.

Figures A1 and A2 serve as visualizations of the model fit. Figure A1 plots the distribution of the observed values vs. the distribution of the fitted values of the model. Overall, the distributions are very similar which means that the model has a good overall fit for the data, slightly overestimating the number of short markers and underestimating the number of very long markers. To illustrate the model performance for single markers in selected languages, Figure A2 shows the predicted vs. observed marker lengths for the Uralic languages in the test datasets. Again, we see that the model's estimation of marker length is generally very close to the observed lengths.

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: Oxford University Press.

Ackerman, Farrell & Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language* 89(3). 429–464.

Ackerman, Farrell & Robert Malouf. 2016. Word and pattern morphology: An information theoretic approach. *Word Structure* 9(2). 125–131.

Aronoff, Mark. 1994. *Morphology by itself: Stems and inflectional classes*. Cambridge and London: MIT Press.

Arnold, Michael & Enno Ohlebusch. 2011. Linear time algorithms for generalizations of the longest common substring problem. *Algorithmica* 60(4). 806–818.

Baerman, Matthew. 2015. *The morpheme: Its nature and use (Oxford Handbooks in Linguistics)*. Oxford: Oxford University Press.

Beniamine, Sacha. 2018. *Classifications flexionnelles: Étude quantitative des structures de paradigmes*. Paris: Université Sorbonne Paris Cité - Paris Diderot dissertation.

Berdicevskis, Aleksandrs, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Demberg Vera, Gary Lupyan, Taraka Rama & Christian Bentz. 2018. Using Universal Dependencies in cross-linguistic complexity research. In *Second workshop on Universal Dependencies (UDW 2018)*, 8–17. Stroudsburg, PA: The Association for Computational Linguistics.

Blevins, James P. 2016. *Word and paradigm morphology*. Oxford: Oxford University Press.

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

Bouma, Gosse, Jan Hajic, Dag Haug, Joakim Nivre, Per Erik Solberg & Lilja Øvrelid. 2018. Expletives in Universal Dependency Treebanks. In *Second workshop on Universal Dependencies (UDW 2018)*, 18–26. Stroudsburg, PA: The Association for Computational Linguistics.

Bürkner, Paul-Christian. 2017. Brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software* 80(1). 1–28.

Bürkner, Paul-Christian. 2018. Advanced bayesian multilevel modeling with the R package brms. *The R Journal* 10(1). 395–411.

Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1). 395–411.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Carpenter, Bob et al 2017. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles* 76(1). 1–32.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles* 76(1). 1–32.

Comrie, Bernard. 1986. Markedness, grammar, people, and the world. In Fred R. Eckman, Edith A. Moravcsik & Jessica R. Wirth (eds.), *Markedness*, 85–106. New York: Plenum Press.

Cotterell, Ryan, Christo Kirov, Mans Hulden & Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics* 7. 327–342.

Croft, William. 2003. *Typology and universals*, 2nd edn. Cambridge: Cambridge University Press.

Diessel, Holger. 2007. Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25(2). 108.

Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press.

Dormann, Carsten F et al 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36(1). 27–46.

Downing, Laura J. & Barbara Stiebels. 2012. Iconicity. In Jochen Trommer (ed.), *The morphology and phonology of exponence (Oxford Studies in Theoretical Linguistics)*, 379–426. Oxford: Oxford University Press.

Du Bois, John W. 1987. The discourse basis of ergativity. *Language* 63(4). 805–855.

Gelman, Andrew, Ben Goodrich, Jonah Gabry & Aki Vehtari. 2019. R-squared for Bayesian regression models. *The American Statistician* 73(3). 307–309.

Givón, Talmy. 1983. Topic continuity: The functional domain of switch-reference. In Pamela Munro & John Haiman (eds.), *Switch reference and Universal Grammar: Proceedings of a symposium on switch reference and Universal Grammar, Winnipeg, May 1981 (Typological Studies in Language 2)*, 51–82. Amsterdam: Benjamins.

Greenberg, Joseph Harold. 1966. *Language universals: with special reference to feature hierarchies*. The Hague: Mouton.

Haiman, John. 1983. Iconic and economic motivation. *Language* 59(4). 781–819.

John Haiman (ed.). 1985. *Iconicity in syntax (Typological Studies in Language 6)*. Amsterdam: Benjamins.

Haspelmath, Martin. 2008a. A frequentist explanation of some universals of reflexive marking. *Linguistic Discovery* 6(1). 40–63.

Haspelmath, Martin. 2008b. Frequency vs. iconicity in explaining grammatical asymmetries. *Cognitive Linguistics* 19(1). 1–33.

Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form-frequency correspondencies and predictability. *Journal of Linguistics* 1–29. https://doi.org/10.1017/S0022226720000535.

Haspelmath, Martin, Andreea Calude, Michael Spagnol, Heiko Narrog & Elif Bamyaci. 2014. Coding causal–noncausal verb alternations: A form-frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Holton, David, Peter Mackridge & Irene Philippaki-Warburton. 2004. *Greek: An essential grammar*. London: Routledge.

Hume, Elizabeth & Frédéric Mailhot. 2013. The role of entropy and surprisal in phonologization and language change. In Yu, Alan C. L. (ed.), *Origins of sound change: approaches to phonologization (Oxford linguistics)*, 29–47. Oxford: Oxford University Press.

Janda, Laura A. & Charles E. Townsend. 2000. *Czech*. München: Lincom Europa.

Janda, Laura A. & M. Francis Tyers. 2018. Less is more: Why all paradigms are defective, and why that is a good thing. *Corpus Linguistics and Linguistic Theory* Online Preview. 1–30. https://doi.org/10.1515/ling-2020-0252.

Kettunen, Kimmo & Eija Airio. 2006. Is a morphologically complex language really that complex in full-text retrieval? In *International Conference on Natural Language Processing (in Finland)*, 411–422.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*. European Language Resources Association (ELRA).

Kock, Ned. 2015. Common method bias in PLS-SEM: A full collinearity assessment approach. *International Journal of e-Collaboration (IJeC)* 11(4). 1–10.

Kress, Bruno. 1982. *Isländische Grammatik*. Leipzig: VEB Verlag Enzyklopädie Leipzig.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8). 707–710.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.

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

Naranjo, Matías Guzmán & Laura Becker. 2018. Quantitative word order typology with UD. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, vol. 155, 91–104.

Nivre, Joakim, Mitchell Abrams, Željko Agić, et al. 2019. Universal Dependencies 2.4. In *LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL)*. Charles University: Faculty of Mathematics and Physics.

Primus, Beatrice. 1999. *Cases and thematic roles: Ergative, accusative and active*. Tübingen: Max Niemeyer.

Shannon, Claude. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27. 379–423.
Shimelman, Aviva. 2016. *A grammar of Yauyos Quechua (Studies in Diversity Linguistics 9)*. Berlin: Language Science Press.
Stump, Gregory T. & Rafael Finkel. 2013. *Morphological typology: From word to paradigm (Cambridge studies in linguistics)*, vol. 138. Cambridge: Cambridge University Press.
Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.
Zipf, George Kingsley. 1935. *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.
Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.