The distribution of zero forms in nominal and verbal inflection: A lexeme-based approach

Anonymous

ABSTRACT

This study examines zero forms, i.e. the absence of an exponent, in nominal and verbal inflectional morphology across languages. It explores the UniMorph dataset to provide an overview of the distribution of zero forms and to contribute to the debate on the role of coding efficiency in typology. Coding efficiency refers to the phenomenon that more frequent (grammatical) elements tend to be zero-marked or shorter than comparable less frequent ones. The results of this study show that there is no strong preference for inflectional values to be encoded by zero forms. However, those morphosyntactic values that are more likely to be encoded by zero forms correspond to the values with higher frequencies in language use. The findings of this study thus provide evidence for coding efficiency to play some role in the distribution of zero forms. However, the high degree of variation found across lexemes and languages suggests that a number of other, potentially language-specific factors play an equally important role.

Keywords: lexeme-based typology, corpus typology, zero marking, UniMorph

INTRODUCTION 1

The present paper examines the distribution of zero forms in nominal and verbal inflectional morphology. In typology, "zero marking"

plays an important role for coding efficiency or form-frequency effects in morphosyntax. Form-frequency effects go back to the early findings by Zipf (1935) that more frequent lexical elements tend to be shorter than less frequent ones. There is crosslinguistic evidence that also in inflectional morphology, more frequent or predictable markers tend to be shorter or at least not longer than comparable less frequent markers (Greenberg 1966; Guzmán Naranjo and Becker 2021a; Stave *et al.* 2021; Haspelmath 2008b; Haspelmath *et al.* 2014; Haspelmath 2021; Haspelmath and Karjus 2017). Such effects can be subsumed under the term of coding efficiency. The coding of grammatical expressions is efficient, as it allows to save efforts with maximal benefits of successful transfer of information in the production and processing of speech (cf. Levshina 2022, for an overview of efficiency in language and communication).

Usually, zero forms are grouped with shorter (as opposed to longer) markers, and it is explicitly or implicitly assumed that zero forms are used to express highly frequent morphosyntactic functions similarly to shorter markers (e.g. Bybee 2011; Croft 2003, Ch. 4; Diessel 2019, Ch. 11; Greenberg 1966, 32-37; Haspelmath 2008a, 2008b, 2021; Song 2018, Ch. 7). However, a crosslinguistic quantitative overview of the distribution of zero forms is still not available. The objective of this paper is to start filling this gap.

To do so, I analyze the distribution of zero forms in the UniMorph dataset (McCarthy *et al.* 2020), which is a crosslinguistic database containing inflectional paradigms of single lexemes. I first provide some theoretical background on coding efficiency in Section 2 and introduce a working definition of zero forms in Section 3. Section 4 describes the dataset, the extraction of zero forms, and presents examples of zero forms. I then analyze the distribution of zero forms in the UniMorph dataset in Section 5, examining which cells and values of nominal and verbal inflection paradigms are crosslinguistically most likely to be expressed by zero forms. This gives us an overview of the morphosyntactic functions that are most robustly associated with zero forms across languages. As we will see, some cells are indeed more likely than other cells to be expressed by zero forms. Nevertheless, the occurrence of the latter will be shown to be language-specific and lexeme-dependent to a large extent. I then analyze whether zero forms are similarly distributed in nominal and verbal inflection paradigms,

showing that they are more likely to occur in nominal paradigms. In Section 6, I discuss the result of the present study with a special focus on the hypothesis that the distribution of zero forms can be accounted for by coding efficiency. Section 7 concludes.

ZERO FORMS AND CODING EFFICIENCY 2

The modern understanding of coding efficiency started out with Zipf (1935), who showed that more frequent words tend to be shorter than less frequent words. In typology, Greenberg (1966) was one of the first linguists to relate the frequency of certain values of grammatical categories in corpora to their formal markedness. An “unmarked” value in this sense is characterized by the absence of an exponent, which is contrasted with a “marked” value that is expressed by an overt exponent. Greenberg (1966) applied markedness to various areas of grammar, making use of a crosslinguistic sample. For instance, he showed how the markedness of singular and plural (and dual) forms of nouns, verbs, and adjectives is reflected in their distribution in corpora from various languages (Greenberg 1966, 32-37). Thus, he noted that the “unmarked” number value, singular, is substantially more frequent than the usually “marked” number values of plural and dual in corpus data from different languages.

Taking up Greenberg's findings and doing away with the concept of markedness, Haspelmath (2008a,b) showed that the length, complexity or availability of grammatical markers can be accounted for by their frequency in language use. In a more recent study, Haspelmath (2021, 2) proposed the following form-frequency correspondence hypothesis:

(1) *The grammatical form-frequency correspondence hypothesis*
When two grammatical construction types that differ minimally (i.e. that form a semantic opposition) occur with significantly different frequencies, the less frequent construction tends to be overtly coded (or coded with more segments), while the more frequent construction tends to be zero-coded (or coded with fewer segments), if the coding is asymmetric. (Haspelmath 2021, 2)

This hypothesis includes the assumption that zero forms pattern with shorter forms in that they are used for comparatively frequent expressions. Applying this to inflectional morphology, we should thus expect zero forms to express highly frequent values of morphosyntactic features. In fact, the hypothesis predicts that more frequent constructions or expressions have a preference for zero (and shorter) forms.[1] There is substantial evidence for such effects of coding efficiency between comparable grammatical expressions, they usually only concern the length of forms in terms of shorter vs. longer forms.[2] The participation of zero forms, though, has not yet been quantitatively examined in detail. There are some indications from the literature, however, suggesting that coding efficiency and frequency may not be a suitable explanation for the distribution of zero forms.

For instance, Stolz and Levkovych (2019) give a qualitative overview of the distribution of zero forms in inflectional paradigms from the perspective of canonical morphology, laying the grounds to include the "absence of material exponence (AOME)" as a non-canonical phenomenon in inflection morphology. To do so, they examine zero forms in inflectional paradigms of 11 typologically diverse languages. Stolz and Levkovych (2019, 396-397) note that "[f]rom the small number of cases discussed above it transpires that frequency might not always be the most powerful factor to make a given word-form or category a candidate for AOME."

Guzmán Naranjo and Becker (2021a) come to a similar conclusion based on a quantitative analysis of the association between the length of inflection markers and their type frequency in the UniMorph database. While they find effects of coding efficiency, their results suggest that the occurrence of zero forms does not follow that of shorter forms. They note that a simple Poisson model predicting the length

[1] Zero forms are often not mentioned explicitly regarding this assumption about efficiency the typological literature. A notable exception is Haiman (1983), who explains the distribution of zero forms in reflexive and reciprocal marking in terms of economy, i.e. efficiency.

[2] A few examples of quantitative approaches to form-frequency effects in grammar are: Stave *et al.* (2021) for the length and frequency of morphemes in general, Haspelmath *et al.* (2014) for the expression of causal and non-causal alternations, Haspelmath (2008c) for reflexive marking, Haspelmath and Karjus (2017) for number marking and Ye (2020) for (in)dependent possessor marking.

of markers on the basis of their type frequency strongly overestimates the occurrence of zero forms. In other words, based on the frequency information, many more zero forms are predicted than observed.[3]

Another example is the occurrence of zero forms for person and number indexing on verbs. Several quantitative typological studies (Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010) find that zero forms for person indexing are rather uncommon across languages; they do not find evidence for a paradigmatic preference of third person (singular) being expressed by a zero form on the verb. However, all three studies show that if a verbal index corresponds to a zero form, it is more likely to express third person (singular) than first or second person.

Arguing for efficiency pressures in diachronic processes to account for crosslinguistic patterns, Seržant and Moroz (2022) also mention zero forms in verbal indexing. Analyzing the length of the indexes on verbs in a typological sample, they argue for an attractor state in which the lengths of different indexes are associated with their frequencies in language use. Seržant and Moroz (2022, 6) also note that "[…] articulatory efficiency plays an important role here: the more expected the sign is the shorter it is. Nevertheless, zero is not preferred." Interestingly, they nevertheless motivate the crosslinguistic avoidance of zero forms through the concept of efficiency, although they refer to two other types of efficiency: processing and planning efficiency. Seržant and Moroz (2022, 7) hypothesize that an overt exponent facilitates processing on the addressee's side. They also propose that avoiding zero forms makes planning more efficient on the hearer's side, "[…] because it provides a straightforward link from meaning to coding, while zero is inherently ambiguous by being linked to various meanings and domains" (Seržant and Moroz 2022, 7). Whether or not the avoidance of zero forms can be accounted for by processing or planning efficiency requires proper psycholinguistic testing. The relevant point here is that coding efficiency does not seem to be applicable

[3] In order to deal with the overestimated number of zero forms, they fit a Hurdle Poisson model that can take into account that zero forms are distributed differently from non-zero forms, resulting in a substantially better fit. Overall, the Hurdle Poisson model predicts zero forms to have a very low probability of 0.02 (Guzmán Naranjo and Becker 2021a, 6).

to the frequency distribution of zero forms in person indexing in the same way as it is in other contexts.

3 TOWARDS A DEFINITION OF ZERO FORMS

The discussion and use of zero forms, or zero morphs, has a long tradition in morphology and in linguistics in general. It goes back to Pāṇini, who introduced the idea of zero morphs for morphemes that lack a phonetic representation as the outcome of morphological rules (Robins 1997, 181-182). The concept of zero morphs for linguistic analysis was also widely applied in later work of structuralists, e.g. (Jakobson [1939] 1983; Bloomfield 1933; Bloch 1947; Saussure 1916).[4] Starting with Haas (1957), linguists began to criticize the assumption of zero morphs in the structuralist tradition and argued for stricter criteria to define zero morphs in order to avoid the assumption of excessive linguistic structure (e.g. Sanders 1988; Mel'cuk 2002; McGregor 2003). The potential danger is that the linguist may postulate a zero morph for any single morphosyntactic function that does not correspond to an overt exponent. As Anderson (1992, 30) noted, it "leads to the formal problem of assigning a place in the structure (and linear order) to all of those zeros".[5] Others, e.g. Contini-Morava (2006) and Mithun (1986) used data from typologically diverse languages to show that the absence of phonetic material can indeed correspond to the absence of a morphosyntactic feature rather than to zero marking in some cases.

In line with those more cautious approaches to zero morphs, this study uses the notion of "zero forms" as a descriptive shorthand for the absence of material exponence of a given morphosyntactic function (cf. Stolz and Levkovych 2019) In other words, I do not assume

[4] Cf. McGregor (2003) for more details on the history of linguistic zero.

[5] For a discussion on issues related to the use of zero morphs in morpheme-based, segmental approaches to morphology, see Anderson (1992); Pullum and Zwicky (1991); Blevins (2016).

the presence of a zero morph, but understand zero forms as the absence of exponence expressing a certain morphosyntactic function in addition to the lexical content of a word form. This also means that zero forms can only occur in contrast to at least one other, overtly coded morphosyntactic function of the same inflection paradigm.

Thus, for the purposes of the present study, it is necessary to identify the invariable, lexical parts, i.e. stems, as well as the potential exponents of an morphosyntactic function in an inflected word form. This conforms with the basic intuition that we want to separate the segments that convey the word's lexical meaning from the segments that convey morphosyntactic information (cf. Matthews 1972). For the purposes of the present study, I define stems, exponents, and zero forms as shown in (2), (3) and (4), respectively.

(2) *Stem*
The stem expresses the lexical content of a word form; it corresponds to the longest common substring shared by all inflected forms of word.

(3) *Exponent*
An exponent encodes the morphosyntactic function (or combination thereof) of a word form; it corresponds to the phonetic material outside of the stem of a word form.

(4) *Zero form*
A zero form is made up of the stem only; it corresponds to a word form that does not have any additional exponent for the given morphosyntactic function (or combination thereof) that it expresses.

This operationalization of stems, exponents, and zero forms has the practical advantage that it does not require any morphological analysis particular to a single language or paradigm. It is a solution to identify the segments that contribute inflectional information that can be applied automatically and consistently to the crosslinguistic UniMorph dataset used in this study (cf. Section 4).

Besides practical considerations, this solution is also based on theoretical grounds and follows the definition of stems by Beniamine and Guzmán Naranjo (2021); Bonami and Beniamine (2021) and Guzmán Naranjo and Becker (2021a). Despite much theoretical work

on the role and identification of stems in morphology, Bonami and Beniamine (2021) note that "there is no agreed upon method for identifying which part of an inflected word is a stem, and that the heuristics used by morphologists in that area are neither systematic nor principled enough".[6] The authors compare two types of stem identification based on prioritizing two different principles, namely to avoid stem allomorphy and to avoid discontinuous stems. Since those two principles are in conflict with each other many times, every approach to stem identification needs to rank them in some way to resolve such conflicts. Bonami and Beniamine (2021) compare the two methods of either adhering to the first or the second principle, resulting in what they call "unique discontinuous stems" (no stem allomorphy allowed) and "continuous stem sets" (no discontinuous stems allowed). While the first method of unique discontinuous stems allocates all the variation of word forms to the exponents, leading to more exponent allomorphy, the second method of continuous stem sets keeps exponent allomorphy minimal, but leads to a high degree of stem allomorphy, since all variation that is enclosed by stem segments has to be included in the stems. What this shows is that neither approach creates more allomorphy; they simply allocate it differently. Obviously, which of the two approach is more useful depends on the research question at hand.

One of the questions discussed by the authors is what types of stems are more helpful in addressing the 'Inflected Word Recognition Problem' (IWRP), i.e. understanding what allows speakers to draw inferences from a word's form to its content. This results in the task of separating the lexical and the inflectional parts of a word form, and Bonami and Beniamine (2021) note that "[i]n terms of the IWRP, the answer is quite simple. Sets of continuous stems are by definition less useful than a unique discontinuous stem: the unique discontinuous stem identifies exactly that part of the word that has no exponential value, while stem allomorphs blur the distinction between exponential and nonexponential material." As the identification of zero forms

[6] Cf. Blevins (2003); Bonami (2012); Brown (1998); Maiden (1992); Montermini and Bonami (2013); Pirrelli and Battista (2000); Spencer (2012); Stump (2001); Stump and Finkel (2013) for work on stem identification and stem allomorphy.

relies on separating lexical segments from exponents of morphosyntactic information in word forms, the IWRP is of high relevance to this study, which provides the theoretical grounds for the definition of stems given in (2).

Furthermore, this study will largely follow a word and paradigm approach to inflection (cf. Anderson 1992; Blevins 2016; Matthews 1972; Stump 2001; Zwicky 1985). This approach bases morphological analyses on the paradigmatic relation between different word forms, representing the different morphosyntactic functions a given word can have. The exponent of a cell in an inflection paradigm is determined through the relation of the word form to the forms used for the other cells of the paradigm. The word and paradigm approach has a very important practical advantage. It allows to refrain from further segmentation of exponents into morphemes which may require language-specific insights and which may not always be desirable or useful (cf. Blevins 2005, 2006). Although morphological segmentation analyses may sometimes be uncontroversial, there are many cases where a morpheme analysis is less than clear. Various examples are given in Spencer (2012), one of them being the Spanish subjunctive verb form *cantaríamos* 'we would sing'. A number of theoretical motivations exist to segment this word form into morphemes in five different ways, i.e. as (i) *cant-a-r-í-a-mos*, (ii) *canta-ríamos*, (iii) *cant-a-ría-mos*, (iv) *canta-r-í-a-mos* and (v) *cantar-í-amos* (Spencer 2012, 93). The fact that these profoundly varying morphological analyses are motivated in the literature suggests that such morpheme segmentations are always, whether explicitly or implicitly, theoretically guided. Moreover, it is likely that the segmentation into morphemes in lesser-studied languages involves even more theoretical uncertainty, given that we may know much less about the morphological structure and its diachrony than for languages like Spanish.

Let us consider a simple example of stem and exponent identification. The paradigm of English nouns consists of two cells: the singular form and the plural form. Given the paradigmatic relation between the singular form */deɪ/* (*day*.SG) and the plural form */deɪz/* (DAY.PL), we can identify the string */deɪ/* as the stem, i.e. the phonetic material that both forms of the paradigm share. Since the form filling the plural cell includes the additional material */z/*, we can establish */z/* as the exponent of the plural. In the singular cell, on the other hand, the

form does not include any material other than what was identified as the stem. We can therefore treat the form of the singular cell of *day* in English as a zero form.

As will be shown in more detail in Sections 4.3 and 4.4, cells of paradigms need not consist of a single morphosyntactic function but can combine the values of different morphosyntactic features. For instance, the inflection paradigms of German nouns combine the morphosyntactic features of case and number. While nouns are inherently specified for gender, each word form is also specified for number and case so that each cell of the paradigm corresponds to a number-case combination, e.g. dative plural.

For the purposes of this study, I will not distinguish between an exponent for plural number and one for dative case. Instead, I treat the material in addition to the stem in the dative plural cell as the exponent of the dative-plural function. In case no additional phonetic material is used, as e.g. in the nominative singular cell, this cell is then analyzed as a zero form (examples from German are discussed in more detail in Section 4.3). Put differently, I do not assign zero forms to single abstract morphosyntactic values but to the relevant value combinations of the inflection paradigms. The theoretical reason to do so lies in exponents of morphosyntactic functions being defined based on the relations between the forms of the different cells of the inflection paradigm, which combine these function. This also reflects the morphological reality of many if not most languages in that morphosyntactic functions are usually not marked in isolation but often occur in combinations, and, as mentioned above, it is not always trivial to justify a segmental analysis. The practical reason is that, as noted and illustrated for a Spanish verb form above, there is still no language-independent and theory-independent way of segmenting distinct morphosyntactic exponents, and those segmentations are not (yet) automatizable. Since automatic processing is indispensable for the purposes of the present study, no further segmentation of different morphosyntactic exponents will be carried out.

The segmentation into stems and exponents is often additionally complicated by inflection classes, which make use of different exponents for the cells of the paradigms. Sections 4.3 and 4.4 show in more detail how the present approach deals with variation in the exponents due to inflection classes, with the shared exponence of different val-

ues, with stem alternations as well as with suppletive forms.

4 DATASET AND SEGMENTATION

4.1 *The UniMorph dataset*

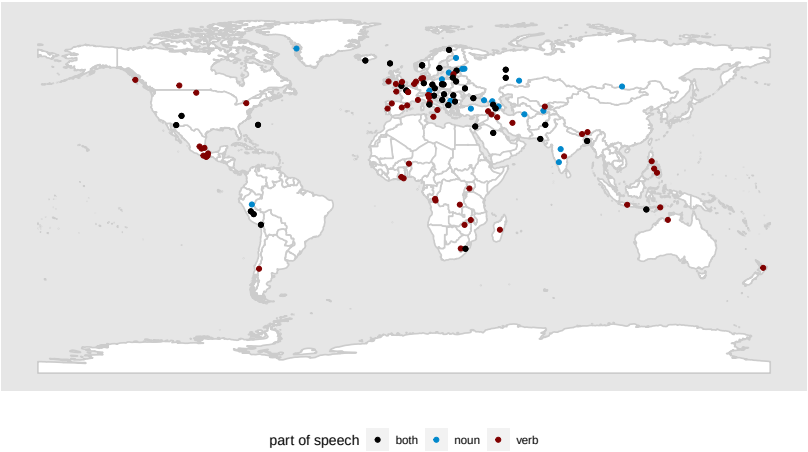
The data used in the present study comes from the UniMorph database (McCarthy *et al.* 2020), a large-scale crosslinguistic database of complete inflection paradigms of nouns, verbs, and adjectives for single lexemes from 167 languages. For this study, I used the verbal paradigms of 104 languages and the nominal paradigms of 61 languages. Since some languages are featured with both nominal and verbal paradigms, the total number of languages analyzed in this study is 141.[7] Figure 1 shows the geographical distribution of the languages in the sample; the blue dots represent languages with nominal paradigms, the red dots show languages with verbal paradigms. A language with both nominal and verbal paradigms is represented by a black dot. While the sample is clearly not a properly balanced typological sample in the strict sense, it does include languages from all six macro areas (Africa, Eurasia, Papunesia, Australia, North America and South America), which ensures that typological and areal diversity is captured at least to a certain extent.

4.2 *Preprocessing*

Since the database is somewhat biased towards languages spoken in Eurasia (mostly Indo-European languages), I only included languages with paradigms for more than 30 lexemes from this area. For languages from other macro areas, especially from Africa or the Americas, I did not apply this threshold of 30 lexemes in order to include more Non-Indo-European languages and to keep the dataset as diverse

[7] More details about the languages and the number of lexemes for which verbal and nominal paradigms were available is provided in the file haszero.csv in the supplementary materials.

Figure 1: Map of the dataset.



as possible. For nouns, the only languages of the dataset with less than 30 lexemes are Kodi (13) and Greenlandic (23). For verbs, these include Sotho (26), Mapudungun (26), and Murrinpatha (29). Besides this threshold, I excluded languages on the basis of unclear or faulty annotations in the original datasets, some of which were annotated only automatically with no manual checks. This led to the inclusion of the final 61 languages for nominal and 104 languages for verbal paradigms.

For certain languages, the UniMorph database already provided the verbal and nominal forms in a phonological transcription. For most other languages, however, forms were given in the standard orthographic representation. This can of course be problematic, especially in languages such as French, where the orthographic representation continues to make many distinctions that are no longer realized in the spoken language. For this reason, whenever possible, I added a phonological transcription using Epitran (Mortensen *et al.* 2018). Epitran currently has modules to transcribe 28 of the languages used here.[8]

While not perfect, Epitran offers a somewhat more realistic representation of the forms occupying the different cells of inflection paradigms. Table 1 illustrates this by showing the transcriptions gen-

[8] For more information, see the file preprocessing.txt in the supplementary materials.

erated with Epitran for the French verb *allumer* 'light something, turn on (light)'. The rows show seven TAM combinations; for each of these, the first row contains the form in their orthographic representation, and the second row shows the phonological transcriptions generated with Epitran.[9]

	1SG	2SG	3SG	1PL
PRS.IND	*allume*	*allumes*	*allume*	*allumons*
	alym	alym	alym	alymɔn
PST.IPF.IND	*allumais*	*allumais*	*allumait*	*allumions*
	alymɛ	alymɛ	alymɛ	alymjɔn
PST.PFV.IND	*allumai*	*allumas*	*allumat*	*allumâmes*
	alymɛ	alyma	alyma	alymɑm
FUT	*allumerai*	*allumeras*	*allumera*	*allumerons*
	alymɛrɛ	alymɛra	alymɛra	alymɛrɔn
PRS.COND	*allumerais*	*allumerais*	*allumerait*	*allumerions*
	alymɛrɛ	alymɛrɛ	alymɛrɛ	alymɛrjɔn
PRS.SUBJ	*allume*	*allumes*	*allume*	*allumions*
	alym	alym	alym	alymjɔn
PST.SUBJ	*allumasse*	*allumasses*	*allumât*	*allumassions*
	alymas	alymas	alymɑ	alymasjɔn

Table 1: Parts of the inflection paradigm of the French verb *allumer* 'turn on (light)'

The pre-processing of the data included other minor and language-specific corrections, e.g. deleting "!" occurring with imperative forms or deleting "?" following the interrogative verb forms in the Turkish data. Some datasets, e.g. Norwegian, contained alternative forms for certain lexemes; in those cases, the first form was systematically chosen. In addition, I manually adapted the cell annotations provided by UniMorph. For instance, many cells with language-specific values or value combinations were originally coded as "LGSPEC" for "language-specific". Whenever possible, I resolved such generic labels using the

[9] Other forms such as imperative and nonfinite forms are omitted in Table 1 for reasons of space.

information provided in the source or in reference grammars. Other steps of manual cleaning included resolving inconsistencies in the annotations across languages; for instance, the value “indefinite” was coded as “INDF” in some languages and as “NDEF” in others. In such cases, I changed the annotation to a single label for a given value in all languages.[10]

4.3 *Segmentation and extraction of zero forms*

In order to analyze the distribution of zero forms, I automatically segmented the forms following the method developed in Beniamine and Guzmán Naranjo (2021) and Guzmán Naranjo and Becker (2021a). As mentioned in Section 3, the segmentation or analysis follows a word and paradigm approach to morphology in that whole forms are paired with morphosyntactic functions according to their distribution across the inflectional paradigms. To give an example, Table 2 shows parts of the present tense paradigm of the French verb *allumer (alyme)* from Table 1.

Comparing the forms of the different cells of the paradigm, we can analyze the string of *alym* as the longest common substring between all forms of the paradigm, i.e. the stem. The string of *alym* also corresponds to the exponent of a number of inflected forms in the paradigm, which are shaded in gray in Table 2. For the purposes of the present paper, the forms of these cells are analyzed as zero forms.

In the case of French *allumer*, the stem corresponds to a continuous segment. This does not necessarily have to be the case. Consider the forms of the German noun *Klos (klos)* ‘dumpling’ in Table 3, shown in the phonological transcription generated with Epitran. In the case of *Klos*, Table 3 shows that the longest common substring does not have to be continuous; due to the umlaut process in the plural forms, the stem of *Klos* is analyzed as consisting of the three consonants *kls*. As can be seen in Table 3, the vowel which is traditionally included in the stem changes from /o/ in the singular to /ø/ in the plural. Because of this vowel difference across the cells of the paradigm, the vowels

[10] For more details on the language-specific pre-processing steps, see the file preprocessing.txt in the supplementary materials.

cell	form	stem	exponent
PRS.IND.1SG	alym	alym	-
PRS.IND.2SG	alym	alym	-
PRS.IND.3SG	alym	alym	-
PRS.IND.1PL	alymɔn	alym	-ɔn
PRS.COND.1SG	alymɛrɛ	alym	-ɛrɛ
PRS.COND.2SG	alymɛrɛ	alym	-ɛrɛ
PRS.COND.3SG	alymɛrɛ	alym	-ɛrɛ
PRS.COND.1PL	alymɛrjɔn	alym	-ɛrjɔn
PRS.SUBJ.1SG	alym	alym	-
PRS.SUBJ.2SG	alym	alym	-
PRS.SUBJ.3SG	alym	alym	-
PRS.SUBJ.1PL	alymjɔn	alym	-jɔn

Table 2: Segmentation of French *alyme* 'turn on (light)'

are analysed as a part of the cells' exponents, respectively. Therefore, lexemes such as *Klos* in German do not have zero forms.

cell	form	stem	exp	form	stem	exp
	Klos 'dumpling'			*Kreuz* 'cross'		
NOM.SG	klos	kls	-o-	krȯ͡yts	krȯ͡yts	-
ACC.SG	klos	kls	-o-	krȯ͡yts	krȯ͡yts	-
DAT.SG	klos	kls	-o-	krȯ͡yts	krȯ͡yts	-
GEN.SG	kloses	kls	-o-es	krȯ͡ytses	krȯ͡yts	-es
NOM.PL	kløsə	kls	-ø-ə	krȯ͡ytsə	krȯ͡yts	-ə
ACC.PL	kløsə	kls	-ø-ə	krȯ͡ytsə	krȯ͡yts	-ə
DAT.PL	kløsən	kls	-ø-ən	krȯ͡ytsən	krȯ͡yts	-ən
GEN.PL	kløsə	kls	-ø-ə	krȯ͡ytsə	krȯ͡yts	-ə

Table 3: Inflection paradigms of two German nouns

The other noun given in Table 3, *Kreuz (krȯ͡yts)* 'cross', is an example with no stem alternations. Here, we see that the forms of the nominative, accusative, and dative singular cells correspond to the longest

common substring, i.e. the stem. Thus, these cells are expressed by zero forms, as there is no additional overt material to encode their morphosyntactic functions.

Besides umlauting, Table 4 shows how the automatic segmentation into stems and exponents deals with metathesis, another process of stem alternations. The example given in Table 4 is the Hungarian noun *gyomor (ɟomor)* 'stomach', whose final segment *-or* is metathesized when certain affixes are added to the stem.[11] Again, this leads to a situation where the stem does not include the segment undergoing metathesis; only the string *ɟomo* is analyzed as the stem. This in turn leads to the nominative singular cell having the exponent *-r*. Usually, the nominative singular does not receive any morphological marking in Hungarian, as can be seen in the second example in Table 4.[12] The noun *gép (geːp)* 'machine' does not have any stem alternations across the cells of its paradigm; therefore, the nominative singular form corresponds to the longest common substring of the lexeme and is analyzed as a zero form for the purposes of the present study.

Table 4: Parts of the inflection paradigms of two Hungarian nouns

cell	**form**	**stem**	**exp**	**form**	**stem**	**exp**
	gyomor 'stomach'			*gép* 'machine'		
NOM.SG	ɟomor	ɟomo	-r	geːp	geːp	-
ACC.SG	ɟomrot	ɟomo	-r-t	geːpɛt	geːp	-ɛt
DAT.SG	ɟomornɒk	ɟomo	-rnɒk	geːpnɛk	geːp	-nɛk
INSTR.SG	ɟomorrɒl	ɟomo	-rrɒl	geːppɛl	geːp	-pɛl
TERM.SG	ɟomorig	ɟomo	-rig	geːpig	geːp	-ig
ON + ESS.SG	ɟomron	ɟomo	-r-on	geːpɛn	geːp	-ɛn
ON + ALL.SG	ɟomorrɒ	ɟomo	-rrɒ	geːprɛ	geːp	-rɛ
ON + ABL.SG	ɟomorroːl	ɟomo	-rroːl	geːprøːl	geːp	-røːl

As shown for the examples from French, German and Hungarian, I automatically segmented all forms of the dataset into stems and ex-

[11] In general, metathesis takes place when the suffix that is added to the stem has an initial vowel. However, this is not always the case; the terminative exponent *-ig* does not cause metathesis.

[12] For reasons of space, Table 4 only shows some of the singular forms.

ponents. Whenever the form of a cell of a given lexeme corresponded to the stem, I analyzed it as a zero form because of the absence of an additional overt exponent for that cell.

One more issue needs to be mentioned regarding the segmentation into stems and exponents and the extraction of zero forms. The automatic segmentation of the whole dataset resulted in 305,276 different exponents (by type of cell). Out of those, more than 50% of the exponents, namely 155,407, occurred only once in the entire dataset. I excluded them, which resulted in the total of 149,869 exponents for further analysis. Excluding exponents with single occurrences was important for two reasons. First, it dealt with suppletive forms in the data. Consider the English examples given in Table 5, were we see the two irregular verbs *know* and *think* and the regular verb *heal* for comparison. The forms of *know* and *think* only share a single consonant (*n-* and *θ-*, respectively) across all cells of their paradigms. As a consequence, the exponent ends up with all the remaining material (which would usually be analyzed as being part of an irregular stem), which means that the exponent greatly depends on the shape of the specific lexeme. Such single cases do not allow for a meaningful analysis of exponents for the purposes of the present study and have thus been removed.

cell	**form**	**stem**	**exp**	**form**	**stem**	**exp**	**form**	**stem**	**exp**
		know			*think*			*heal*	
NFIN	now	n	-ow	θɪŋk	θ	-ɪŋk	hil	hil	-
PST	nu	n	-u	θɔt	θ	-ɔt	hild	hil	-d
PTCP.PST	nown	n	-own	θɔt	θ	-ɔt	hild	hil	-d
PTCP.PRS	nowɪŋ	n	-owɪŋ	θɪŋkɪŋ	θ	-ɪŋkɪŋ	hilɪŋ	hil	-ɪŋ
PRS.3SG	nowz	n	-owz	θɪŋks	θ	-ɪŋks	hilz	hil	-z

Table 5: Inflection paradigms of two English verbs

Second, removing exponents that occurred only once in the dataset also controlled the additional marker allomorphy. As stems are not allowed to feature allophony or allomorphy for the purposes of this study, it leads to discontinuous exponents and more exponent allomorphy due to what is traditional understood as stem alternations. By removing exponents that occur only once per cell, such lexeme-specific

alternations are excluded for analysis in the present study. This is important, since one could argue that those forms may be zero forms in the traditional sense (in that they lack affixal exponents), and keeping them in the dataset would then lead to detecting fewer zero forms. By excluding them, the present study stays agnostic to such theoretically more complex cases. However, lexemes with systematic alternations such as umlauting in German (cf. Table 3) are kept in the dataset. In this case, a systematic alternation between what is traditionally analyzed as a stem vowel (e.g. /o/ vs. /ø/) does indeed contribute to the marking of inflectional information (number in this case).

In total, the final dataset contains 149,869 different exponents (by type of cell), with 513 types of cells that can be expressed by zero forms. Out of those, 293 types fall into the nominal, and 220 into the verbal domain.

4.4 *Zero forms in inflection paradigms: Examples*

This section will briefly show examples of zero forms in inflection paradigms of the nominal domain (Aymara) and of the verbal domain (Georgian).

4.4.1 Aymara

Aymara (Aymaran, Argentina, Bolivia, Chile, Peru) is a language with nominal inflection known for its subtractive morphology. The accusative singular cell is usually analyzed as being expressed by the subtraction of the final vowel of the nominative singular form (cf. Coler 2015). Table 6 illustrates this with the paradigms of three Aymara nouns; for simplicity, only the singular and non-possessive forms are shown. As can be seen in Table 6, the accusative singular form corresponds to the stem (as defined in this study), as it is the shortest common substring of all forms of the lexeme. Compared to the accusative form, the nominative form has an additional final vowel, which is also found in all other forms of the paradigm except for the inessive (INESS) and equative (EQTV) forms.

Traditionally, the nominative form with the final vowel is analyzed as the stem of the noun, while the accusative is argued to be a subtractive form, i.e. consisting of less material than the stem of

Table 6: Inflection paradigm of three Aymara nouns

cell	*anu* ‘dog’	*chaski* ‘messenger’	*luk'ana* ‘finger’
NOM.SG	**an**u	**chask**i	**luk'an**a
ACC.SG	**an**	**chask**	**luk'an**
GEN.SG	**an**una	**chask**ina	**luk'an**ana
COM.SG	**an**umpi	**chask**impi	**luk'an**ampi
BEN.SG	**an**utaki	**chask**itaki	**luk'an**ataki
PRP.SG	**an**ulayku	**chask**ilayku	**luk'an**alayku
ABL.SG	**an**uta	**chask**ita	**luk'an**ata
ALL.SG	**an**uru	**chask**iru	**luk'an**aru
INESS.SG	**an**pacha	**chask**pacha	**luk'an**pacha
EQTV.SG	**an**jama	**chask**jama	**luk'an**jama
INTER.SG	**an**upura	**chask**ipura	**luk'an**apura
PROP.SG	**an**uni	**chask**ini	**luk'an**ani
TERM.SG	**an**ukama	**chask**ikama	**luk'an**akama
VERS.SG	**an**ukata	**chask**ikata	**luk'an**akata

the lexeme (Coler 2015, 2018; Baerman *et al.* 2017). Diachronically speaking, there are valid arguments to support such an analysis. Coler (2018) provides examples of historical Aymara with accusative forms that still have the final vowel. In addition, vowel deletion is a common phonological process in Aymara. Nevertheless, aiming at a synchronic and comparable analysis across languages, I treat the accusative form as the stem of the lexeme and therefore as a zero form. In the Aymara data, the accusative corresponds to a zero form in all 1522 nouns of the dataset with no exception.

Georgian 4.4.2

Another rather unusual case of zero forms comes from verbs in Georgian (Kartvelian, Georgia). Besides a number of other theoretically interesting patterns, Georgian verbs have been cited in the typological and morphological literature for their crosslinguistically unusual 2nd person singular zero marker (e.g. Stolz and Levkovych 2019; Anderson 1992; Blevins 2016). However, not all lexemes use zero forms

in the sense of the present study to express the second person singular. Only 7 out of 48 verbal lexemes in the dataset feature a zero form in the second person singular present tense cell. Table 7 shows four examples of verb paradigms in Georgian.[13]

Table 7: Parts of the inflectional paradigm of four Georgian verbs

cell	*t'exs* 'break'	*k'vecs* 'cut off'	*gaacnobs* 'introduce'	*ak'etebs* 'make'
PRS.1SG	v**t'ex**	vk'vec	va**cn**ob	vak'**et**eb
PRS.2SG	**t'ex**	**k'vec**	**cn**ob	ak'**et**eb
PRS.1PL	v**t'ex**t	vk'**vec**t	v**cn**obt	vak'**et**ebt
PRS.2PL	**t'ex**t	**k'vec**t	**cn**obt	ak'**et**ebt
IMPERF.1SG	v**t'ex**di	vk'**vec**di	v**cn**obdi	vak'**et**ebdi
IMPERF.2SG	**t'ex**di	**k'vec**di	**cn**obdi	ak'**et**ebdi
IMPERF.1PL	v**t'ex**dit	vk'**vec**dit	v**cn**obdit	vak'**et**ebdit
IMPERF.2PL	**t'ex**dit	**k'vec**dit	**cn**obdit	ak'**et**ebdit
FUT.1SG	gav**t'ex**	šev**k'vec**	gav**cn**ob	gavak'**et**eb
FUT.2SG	ga**t'ex**	še**k'vec**	ga**cn**ob	gaak'**et**eb
FUT.1PL	gav**t'ex**t	šev**k'vec**t	gava**cn**obt	gavak'**et**ebt
FUT.2PL	ga**t'ex**t	še**k'vec**t	ga**cn**obt	gaak'**et**ebt
AOR.1SG	gav**t'ex**e	šev**k'vec**e	gav**cn**e	gavak'**et**e
AOR.2SG	ga**t'ex**e	še**k'vec**e	ga**cn**e	gaak'**et**e
AOR.1PL	gav**t'ex**et	šev**k'vec**et	gav**cn**et	gavak'**et**et
AOR.2PL	ga**t'ex**et	še**k'vec**et	ga**cn**et	gaak'**et**et

In general, Georgian verbs take a so-called preverb in some but not all of the tenses (Hewitt 1995, 148-169). When it occurs, it precedes the prefixal part of agreement marking on the verb. As we can see in Table 7, present and imperfect forms occur without the verbal prefix, while the future, aorist and perfect forms all make use of the prefix (*ga-* and *še-* in the examples in Table 7). In most TAM series, many Georgian verbs also have so-called thematic suffixes (Hewitt 1995, 143-147), as

[13] To keep it simple, I do not provide an exhaustive list of all TAM combinations but focus on those that show the relevant exponent alternations.

e.g. *-ob* in *gaacnobs* ‘introduce’ or *-eb* in *ak'etebs* ‘make’. The presence of those thematic suffixes in the aorist forms results in the absence of zero forms in most of the verbs. The thematic suffix *-eb/-ob* is part of the second person singular present form, but as it is not used in the aorist forms, the former does not correspond to the longest common substring of the verb forms. The second person singular present tense cell can thus only be expressed by a zero form with verbs that generally do not use any of the thematic suffixes. This is shown with the first two verbs in Table 7, *t'exs* ‘break’ and *kv'ecs* ‘cut off’.

5 THE DISTRIBUTION OF ZERO FORMS IN THE UNIMORPH DATA

In this section, I examine which types of cells and values from nominal (Section 5.1) and verbal (Section 5.2) inflectional paradigms are crosslinguistically most likely expressed by zero forms. To do so, I focus on the cells and values with the strongest association with zero forms. Some of the results presented in this section will be taken up in the discussion in Section 6.

5.1 *Zero forms in nominal paradigms*

5.1.1 Cells associated with zero forms

To explore which nominal cells are most likely to be expressed by a zero form, I included only those cells from the dataset with a proportion of zero forms ≥ 0.01 in at least two languages. This threshold was chosen to restrict the following analysis to the cells with a reasonable crosslinguistic probability of being expressed by zero forms. Out of 883 different nominal cells in total, the dataset contains 119 different cells that can be expressed by a zero form in at least one lexeme in the dataset. Including all 119 cells in the analysis would not be very insightful given that most of those cells are expressed by a zero form only with a handful of lexemes in the dataset.

With the threshold in place, we can focus on the relevant subset consisting of the 21 cells shown in Figure 2 that are most likely to

be expressed by a zero form. The observed proportions of zero forms still differ to a great extent across cells, though, ranging from 0.76 (accusative inanimate singular) to 0.02 (plural). The numbers above the bars in Figure 2 indicate the number of languages which allow for zero forms in a given cell; the number in brackets stands for the number of languages that have a given cell.

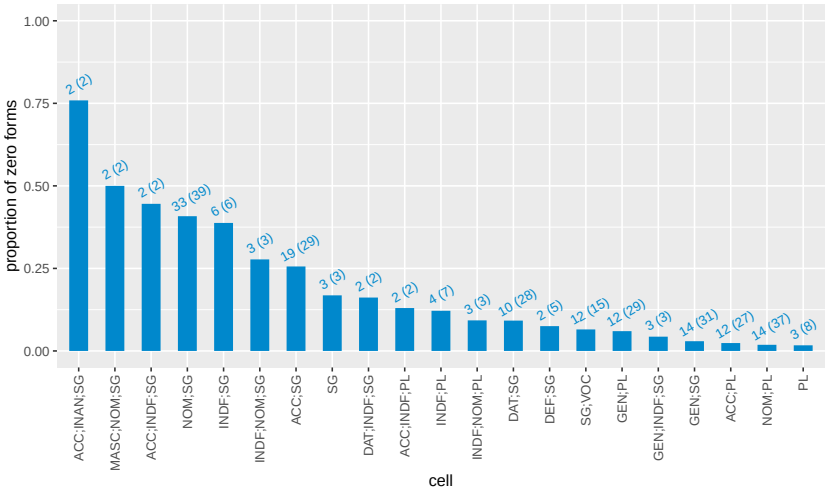


Figure 2: Nominal cells with highest proportions of zero forms

We see some value combinations that only occur in a small number of languages but that have high proportions of zero forms across lexemes. For instance, the feature combination that has the highest overall zero proportion of 0.76 is the accusative inanimate singular cell, which only occurs in two languages of the dataset (closely related Russian and Czech). The other cells with zero proportions above 0.25 are the masculine nominative singular (0.50), the accusative indefinite singular (0.45), the nominative singular (0.41), the indefinite singular (0.39), the indefinite nominative singular (0.28) and the accusative singular (0.26) cells. Except for the nominative and accusative singular cells, all other cells with comparatively high proportions of zero forms only occur in a few languages of the dataset.

Besides case and number values, we find cells including the values of inanimate, indefinite and, interestingly, definite. This means that both indefinite and definite are values that occur in cells which tend to have comparatively high proportions of zero forms. Another important

insight from Figure 2 is that only very few cells have high proportions of zero forms, once a lexeme-based approach is applied. However, the proportions in Figure 2 are likely biased by the phylogenetic relations between the languages of the dataset.

In order to account for that, we need to model the distribution of zero forms across cells. Using the noun subset of the 21 nominal cells that are most likely to be expressed by zero forms, I fitted a binomial regression model to predict the probability of zero forms based on the cell of the inflectional paradigm. I fitted the model using Stan (Carpenter *et al.* 2017) with the *brms* package (Bürkner 2017) in R (R Core Team 2021). I additionally controlled for the phylogenetic relations between the languages of the dataset using a phylogenetic regression term following the method described in Guzmán Naranjo and Becker (2021b). This term does not model the relations between languages in a categorical way but includes the information of the entire phylogenetic tree and forces the estimates of the single languages to co-vary according to the tree.[14] In other words, if two languages share many nodes of the tree, the model forces their coefficients to be very similar. If, on the other hand, two languages are not related at all, the model allows their estimates to vary freely.

Figure 3 shows the estimated probabilities of zero forms for each of the 21 cells of the noun subset.[15] The dots represent the mean values of the posterior distribution of the zero probabilities, and the error bars show the 95% uncertainty intervals. The uncertainty intervals are those intervals that 95% of the posterior distribution falls into and allow for a straightforward interpretation. This means that, given the data and the model, we can be 95% certain that the probability of zero forms will fall in that interval.

Comparing the results of the model shown in Figure 3 with the distributions given in Figure 2 reveals a few important differences. The highest probability of zero forms is predicted for indefinite singular cells at 0.62, although it had an observed proportion of zero forms of 0.39. However, the six languages with this cell belong to different language families: the Slavic branch of Indo-European (Macedonian

[14] The phyologenetic tree is taken from Glottolog (Hammarström *et al.* 2021).

[15] For more details on the model, see file code.R in the supplementary material.

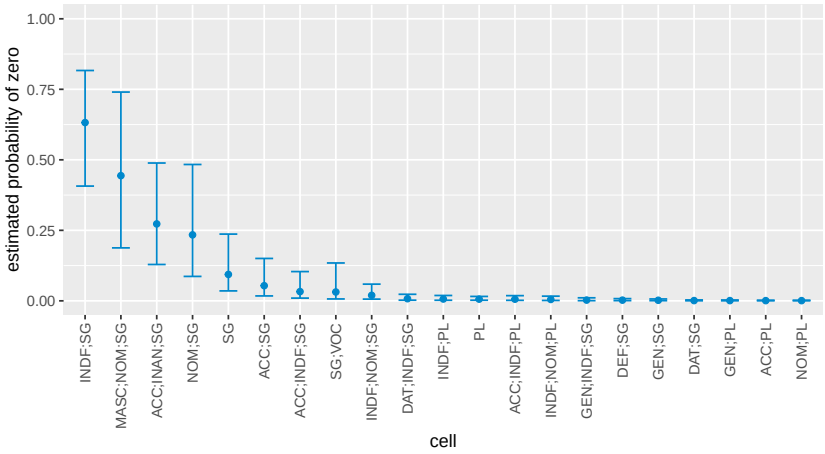


Figure 3: Conditional effects for the nominal cells with the highest proportions of zero forms

and Bulgarian), Abkhaz-Adyghe (Adyghe and Kabardian), the Semitic branch of Afro-Asiatic (Modern Hebrew) and Turkic (Tajik). While all of the six languages feature zero forms in the indefinite singular cells, in Tajik, this cell is exclusively expressed by zero forms. This results in the comparatively high predicted probability of zero forms for indefinite singular cells.

The second highest probability of zero forms is predicted for masculine nominative singular cells at 0.44. The two languages that have this cell are Yiddish and Old French, and the predicted probability closely corresponds to the observed proportions shown in Figure 2.

The cell with the third highest predicted probability (0.27) is the accusative inanimate singular cell, found in the two Slavic languages Russian and Czech. Their close phylogenetic relation also explains why the predicted probability of zero forms is very low compared to the observed proportion of zero forms of 0.76. In such cases, the fact that zero forms often occur in this cell in the two languages is accounted for by their close phylognetic relation by the model rather than the cell itself. More data from other languages is needed for those three cells with the highest probability of zero forms in order to consolidate the findings of this study, given that the number of languages with those cells is very low.

The next cell in Figure 3 is the nominative singular cell; zero forms have a predicted probability of 0.23 to occur in this cell. This is also somewhat lower than the observed proportion of 0.41. The nomina-

tive singular cell is one of the few cells that occurs in a large number of languages in the dataset and allows for zero forms in most of them. Out of 39 languages with that cell, 33 languages feature zero forms to encode the nominative singular cell. Out of the 33 languages allowing for zero forms, the two Turkic languages Tatar and Bashkir as well as Quechua (Quechuan) have exclusively zero forms in the nominative singular. Zero forms in this cell occur in languages from five different families in the dataset.[16] Even though this is probably the most crosslinguistically robust case of zero forms in nominal paradigms, it is telling that almost all of the languages allowing for zero forms in nominative singular cells are spoken in Eurasia. Moreover, this cell is not found in many languages outside of this macro area in the dataset. Therefore, it this remains to be seen in future research whether the association of the nominative singular cell with zero forms is the result of a macro areal bias, and whether the bias operates on the level of zero forms or on the level of the availability of this cell in the first place.

The cell with the next highest estimate is the singular cell, for which the model predicts zero forms to occur with a probability of 0.09. This is similar to the observed proportion of zero forms at 0.17 in the three languages O'odham (Uto-Aztecan), Zulu (Atlantic-Congo) and Tajik (Turkic). Even though the estimate is not very high, this cell is also a crosslinguistically robust candidate for zero forms.

The next cell with a comparatively high predicted probability of zero forms (0.05) is the accusative singular cell. Here, the model predictions differ to a greater extent from the observed proportion of 0.26 from 29 languages (26 of which allow for zero forms in this cell). With the exception of Aymara (Aymaran) and San Pedro Amuzgos Amuzgo (Otomanguean), all of the languages from the dataset with zero forms

[16] The families are: Quechuan (Quechua), Uralic (Finnish, Hungarian, Estonian, Northern Saami, Livonian, Votic, Ingrian), Indo-European (German, Old English, Russian, Belarusian, Polish, Slovenian, Ukranian, Serbo-Croatian, Czech, Lower Sorbian, Old Church Slavonic, Kashubian, Armenian, Latin, Pashto, Old Saxon, Urdu, Sanskrit), Turkic (Turkish, Tatar, Bashkir, Azerbaijani, Khakas, Turkmen) and Kartvelian (Georgian). The six languages in the dataset that do not show any zero forms for the nominative singular cell are: Lithuanian (Baltic), Adyghe and Kabardian (Abkhaz-Adyghe), Kannada (Dravidian), Crimean Tatar (Turkic), Aymara (Aymaran).

in the accusative singular cell are found in Eurasia.[17] Again, the model takes into account the close phylogenetic relation of most of the languages with zero forms in this cell and thus estimates the overall probability of zero forms to be much lower than observed. Hence, also for this cell, we have to assume that the observed pattern is the result of a bias from Indo-European or Eurasian languages in general.

The last cell that will be mentioned here as a potential candidate for zero forms is the vocative singular cell. It is predicted to have a very low probability of zero forms (0.03), but the upper limit of the uncertainty interval lies at 0.13, which reflects the observed proportion of zero forms. Again, even though the dataset contains 12 out of 15 languages with zero forms in the vocative singular, the distribution is not crosslinguistically robust, since all languages belong to the Indo-European family, most of which are from the Slavic branch.[18]

All other of the 21 cells tested here have estimated probabilities of zero forms of below 0.03. Based on the UniMorph data, it is therefore difficult to view them as being prone to zero marking, even though zero forms are certainly used to express those cells with some lexemes across languages.

5.1.2 Values associated with zero forms

The fact that the languages in the dataset differ to a great extent with respect to the combinations of values in single cells makes it somewhat difficult to assess the association between zero forms and crosslinguistically less common cells. While it is important to take into account cells, i.e. how the values of different grammatical features are combined, it can nevertheless be insightful to look at the association of single values and zero forms in a second step. Note that due to the definition and identification of zero forms used in this study, pulling apart

[17] The remaining languages with zero forms in the accusative singular cell are Finnish, Estonian and Northern Saami (Uralic), as well as German, Old English, Russian, Serbo-Croatian, Polish, Czech, Slovenian, Ukrainian, Lower Sorbian, Belarusian, Old Church Slavonic, Latin, Old Saxon, Urdu (Indo-European).

[18] The languages with zero forms in the vocative singular cell are: Czech, Polish, Ukranian, Macedonian, Bulgarian, Bosnian-Croatian-Serbian, Old Church Slavonic, Latvian, Romanian, Sanskrit, Pashto and Urdu. The three languages in which the vocative singular cannot be expressed by a zero form are Kashubian, Georgian and Lithuanian.

the values of cells and analysing their association with zero forms does not translate directly into the traditional analysis of an abstract feature value, e.g. singular, as being zero-coded. Given how zero forms have been extracted in this study, the singular value being expressed by a zero form refers to all cells in the dataset that encode singular (potentially besides other feature values) and that are expressed by a zero form. This method is fully faithful to surface structures and it is not designed to differentiate between a singular value that is zero-coded with the dative case value of the same cell being overtly marked.

In order to examine the association of single values with zero forms, I extracted all feature values of the nominal inflection paradigms and added up their occurrences in lexemes overall and in zero forms. The proportions of zero forms that are used for a given value are shown in Figure 4. To focus on the most likely values that occur in zero forms, Figure 4 only shows those values with proportions above 0.01, occurring in at least two languages.[19] The bars represent the proportions of zero forms of a given value; the numbers above the bars show in how many languages the value occurs in zero forms, and the numbers in brackets show the number of languages with that value in the dataset.

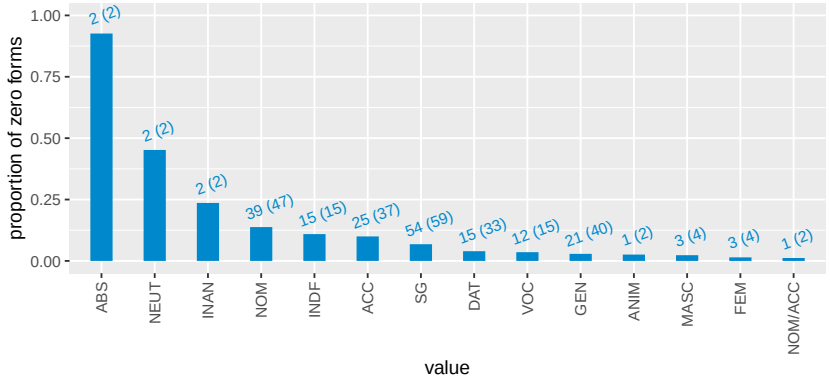


Figure 4: Nominal values with highest proportions of zero forms

The distributions in Figure 4 confirm some of the tendencies seen in the previous section but also provide new insights. Again, we find the

[19] The threshold of 0.01 is a heuristic. A higher threshold would have left us with a very low number of values. A lower threshold would have resulted in too many values, making the interpretation very complex.

values of inanimate, indefinite, nominative, accusative, singular, dative and vocative among the values that occur in cells with the highest proportion of zero forms in at least 2 languages of the dataset. Additional feature values that were not detected when considering entire cells are absolutive and neuter. The absolutive value has by far the highest proportion of zero forms with 0.93, occuring in three different cells in Shipibo-Konibo (Pano-Tacanan) and Kalaalisut (Eskimo-Aleut). It is followed by neuter, which occurs in 18 different cells in Yiddish and Old French. When considered on its own, however, we see that cells containing the neuter value are expressed by a zero form at a proportion of 0.48. While most of the values in Figure 5 are case values, we also find the gender values (neuter, feminine, masculine), inanimate and animate, indefinite, and singular.

To assess how robust those distributions are across languages, I fitted a logistic Bayesian regression model, taking into account the phylogenetic relations between the 43 languages in the subsample following the same method described in Section 5.1.1. The model predictions for the probability of zero forms being used in combination with different values are shown in Figure 5.

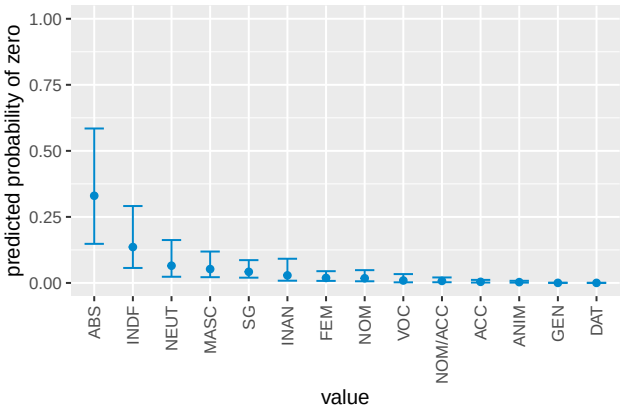


Figure 5: Conditional effects for the nominal values with the highest proportions of zero forms

The model predictions in Figure 5 confirm some of the observed proportions. The absolutive value is estimated to be the value with the highest probability of 0.33 of occurring in zero forms. Although the two languages with the absolutive value are entirely unrelated, the level of uncertainty in the estimate is fairly high, because the number

of languages with an absolutive value is so low in the dataset. Still, we can be confident that this value has a probability above 0.20 of occurring in zero forms.

The value with the second highest predicted probability (0.14) of occurring in zero forms is the indefinite value. It can occur with zero forms in all 15 languages that have this value. The languages with the indefinite value in their nominal paradigms belong to four different language families: Afro-Asiatic (Arabic, Modern Hebrew), Indo-European (Macedonian, Bulgarian, Icelandic, Swedish, Norwegian Bokmål, Norwegian Nynorsk, Faroese, Romanian, Yiddish, Bengali), Abkhaz-Adyghe (Adyghe, Kabardian) and Turkic (Tajik).

The value with the next highest estimated probability of zero forms is neuter with 0.06. The great difference to the observed proportion of 0.48 can be explained by the fact that it only occurs in two Indo-European languages (Yiddish and Old French). The neuter value is followed by the masculine and singular values, with a probability of zero forms of 0.05 and 0.04, respectively. Interestingly, the singular value, present in 59 languages, does not occur in cells expressed as zero forms in all languages; in five languages, cells including the singular value do not allow for zero forms.[20] This does not necessarily mean that these languages always use an affixal singular marker in the traditional sense. They rather lack a single longest common substring, i.e. stem, across singular and plural forms in their nominal paradigms.

The inanimate value is estimated to have a probability of 0.03 to occur with zero forms. This is again due to the fact that it occurs in zero forms in the two closely related Slavic languages Russian and Czech. The nominative value, on the other hand, occurs with zero forms in 39 out of 47 languages that have this value in their nominal paradigms.[21] Its low estimated probability to occur with zero forms of 0.02 is also due to the fact that most of the languages in which the nominative value occurs in zero forms are Indo-European. All other

[20] Those languages are: Kodi-Gaura (Austronesian), Crimean Tatar (Turkic), Bengali and Lithuanian (Indo-European) as well as Kannada (Dravidian).

[21] The 8 languages in which the nominative value does not occur in zero forms are: Arabic (Semitic), Bulgarian and Lithuanian (Indo-European), Adyghe and Kabardian (Abkhaz-Adyghe), Aymara (Aymaran), Kannada (Dravidian) and Crimean Tatar (Turkic).

values shown in Figure 5 have estimated probabilities of zero forms of 0.01 and below, which means that they may occur with zero forms occasionally in different languages but they clearly have no strong association with zero forms in the UniMorph data.

5.2 *Zero forms in verbal paradigms*

5.2.1 Cells associated with zero forms

Also for the verbal paradigms, it was necessary to subset the dataset in order to reduce the high number of different cells (3013) to the cells that allow zero forms at least to a certain extent in some languages. Therefore, the verb subset contains only those cells that have a proportion of zero forms ≥ 0.01. This leaves us with the 23 cells of verbal paradigms shown in Figure 6. The bars show the total proportions of zero forms across languages for a given cell; the numbers above the bars indicate how many languages allow for zero forms in that cell, and the numbers in brackets show how many languages in the dataset have that cell.

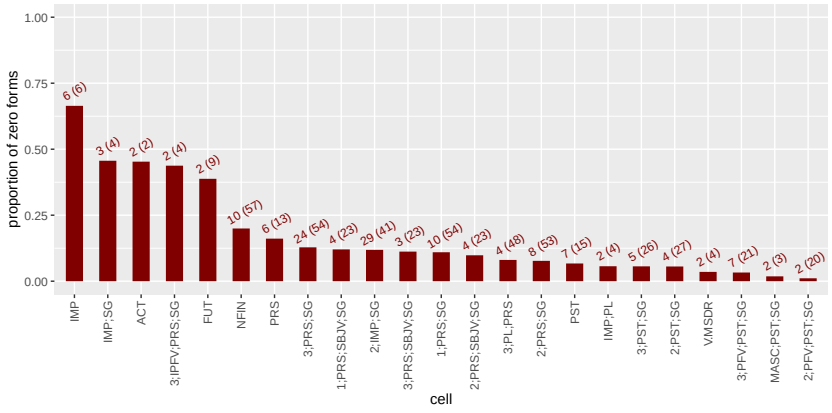


Figure 6: Verbal cells with the highest proportions of zero forms

Figure 6 shows that different imperative forms are amongst the cells with the highest proportion of zero forms, namely the imperative (IMP), imperative singular (IMP.SG), imperative second person singular (2.IMP.SG) and the imperative plural (IMP.PL) cells. Except for Tibetan, the imperative cell is found exclusively in Germanic languages: Swedish, Norwegian (Bokmål and Nynorsk), Danish and West

Frisian. In all six languages we also find zero forms, and they make up the overall highest proportion of zero forms at 0.66. The next cell is the imperative singular cell with a proportion of zero forms of 0.46 in Dutch, Old English, and Haida (Haida). The imperative second person singular cell is much more common and found in 41 languages in the dataset, 29 out of which allow for zero forms. However, the overall proportion of zero forms is much lower at 0.12. The last cell in Figure 6 with the imperative value is the imperative plural cell, which is found in four languages. It only allows for zero forms in the two Germanic languages Dutch and North Frisian at an overall low proportion of 0.06. Taken together, these four cells suggest that imperative forms are generally likely to be expressed by zero forms.

Another cell with very high proportions of zero forms at 0.45 is the the active cell. The two languages with that cell in their paradigms, Indonesian and Maori, are both Austronesian languages, so that it is very difficult to generalize from this result. Interestingly, we also find future cells with a high proportion of zero forms in 2 (out of 9) languages. The two languages with zero forms in future cells are Tibetan (Sino-Tibetan) and Cebuano (Austronesian). The use of zero forms in this cell is somewhat unexpected, given that future grams have a strong crosslinguistic tendency to be overtly expressed (Bybee *et al.* 1994, 243).

Another cell with a comparatively high proportion of zero forms at 0.20 is the nonfinite cell (NFIN), with zero forms allowed in 10 out of 57 languages that have this cell in the dataset. The nonfinite cell is a form of the verb used in combination with other finite verbs such as auxiliaries in complex verbal expressions in most of the 57 languages. The languages with zero forms occurring in this cell are mostly (except for French) languages with rather small inflectional paradigms: English, Swedish and French (Indo-European), Tagalog, Malagasy, Hiligaynon and Cebuano (Austronesian), Akan and Gã, (Kwa-Volta-Kongo), and Ganda (Bantu). In these languages, the nonfinite cell is indeed a principle part in that it serves as the base for all other cells in the paradigm. This is certainly not surprising for nonfinite cells; it is rather noteworthy that 47 out of the 57 languages with a nonfinite cell in their verbal paradigms do not allow for zero forms in this cell. In other words, in most of the languages of the dataset, nonfinite forms do actually not correspond to zero forms but have an overt exponent

coding the nonfinite value.

Another value that appears in a number of cells in Figure 6 is the present tense (PRS), as the third person singular present (3.PRS.SG), the third person singular imperfective present (3.IPFV.PRS.SG), the third person plural present (3.PL.PRS), the first person singular present (1.PRS.SG), the second person singular present (2.PRS.SG), the first person singular subjunctive present (1.PRS.SBJV.SG), the second person singular subjunctive present (2.PRS.SBJV.SG) and the third person singular subjunctive present (3.PRS.SBJV.SG). Out of those cells, only the 3.IPFV.PRS.SG cell and the PRS cells have comparatively high proportions of zero forms (0.44 and 0.16, respectively). Out of four languages with the 3.IPFV.PRS.SG cell, the two unrelated languages Macedonian (Indo-European) and Mezquital Otomi (Otomanguean) allow for zero forms. The PRS cell is found in 13 typologically diverse languages in the dataset. It occurs with zero forms in the following six languages: Nynorsk and Swedish (Germanic), Tibetan (Sino-Tibetan), Akan (Kwa-Volta-Kongo), Zarma (Songhay) and Cebuano (Austronesian). This variety of language families suggests that even though it does not appear to be very strong, the association of the present tense value with zero forms is typologically robust.

If person is specified, we mostly find cells with third persons. Out of 13 cells with a person specification in Figure 6, six cells are specified for third person, five cells for second person, and two cells for first person. Second person cells expressed by zero forms are shown to be mostly imperative forms, including subjunctive forms which can also be used to express imperatives and desired actions. The only two exceptions are the second person singular present (2.PRS.SG) and the second person singular past (2.PST.SG) cells, which have very low proportions of zero forms (0.08 and 0.06, respectively). For the second person singular present cell, except for Georgian (Kartwelian), all languages that allow for a zero form are Indo-European languages: French, Old French, Italian, Romanian, Dutch, Welsh, and Latvian. We see a similar situation for the second person singular past cell. The four languages that show zero forms are all Slavic languages (Bosnian-Croation-Serbian, Bulgarian, Macedonian and Lower Sorbian), strongly suggesting that those zero forms are a family-specific phenomenon.

Similarly to the nominal paradigms discussed in the previous sec-

tion, I fitted a Bayesian logistic regression model to predict the probability of a zero form from the type of cell, controlling for the phylogenetic relations between languages in the dataset.[22] The conditional effects for all 23 cells are shown in Figure 7.

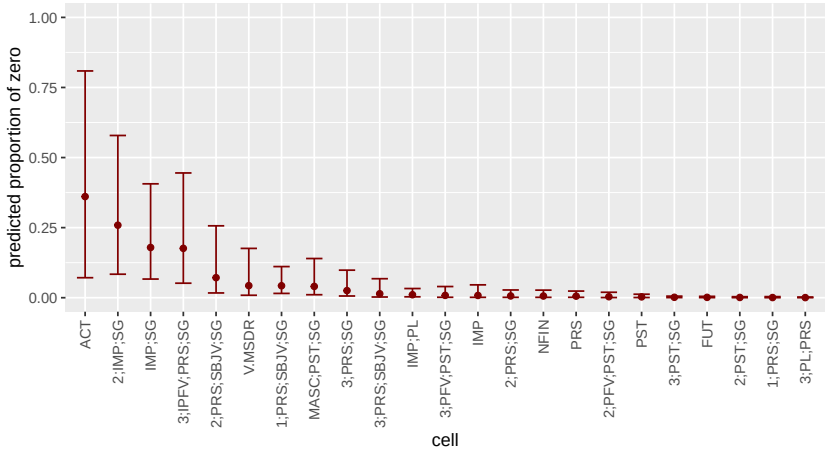


Figure 7: Conditional effects for the verbal cells with the highest proportions of zero forms

We see that the model, now taking into account the relations between languages, predicts a high probability of zero forms of 0.36 only for the active (ACT) cell. At the same time, the level of uncertainty is extremely high, which is simply due to the fact that only two languages have that cell, and that both are Austronesian languages.

The cells with the two next highest estimates are the second person singular imperative (2.IMP.SG, 0.26) and the imperative singular (IMP.SG, 0.18) cells. Again, the uncertainty intervals are very large, which makes it difficult to interpret the values as such. This nevertheless suggests that imperative forms are more likely to be expressed by zero forms compared to other cells of verbal paradigms (cf. Section 6.3 for a discussion of the association between imperatives and zero forms).

The only other cell with a relatively high estimated probability of zero forms (0.18) is the third person singular imperfective present (3.IPFV.PRS.SG) cell, also with a very large uncertainty interval. This

[22] For more details on the model, see the file code.R in the supplementary materials.

indicates that third person present tense forms are comparatively more likely to be zero marked than most other cells of verbal inflection paradigms. In a way, this complements the observations from the typological literature, noting an association with zero forms for present tense (Bybee and Dahl 1989, 55; Bybee 1994, 248) and third person markers (Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010) independently. The UniMorph data suggests that it is the combination of third person and present tense / imperfective aspect that is associated with zero forms across languages.

Two other values that showed a high proportion of zero forms in the raw distributions shown in Figure 6 are future and nonfinite cells. Controlling for the phylogenetic relations between the languages in the dataset, however, shows that those two cells are not generally associated with a high probability of zero forms in the languages of the dataset.

5.2.2 Values associated with zero forms

Similarly to what was shown for nouns in Section 5.1.2, Figure 8 shows the proportion of zero forms for single values of verbal paradigms. Again, I selected only those values that occur in at least two languages and have a zero proportion of at least 0.01.

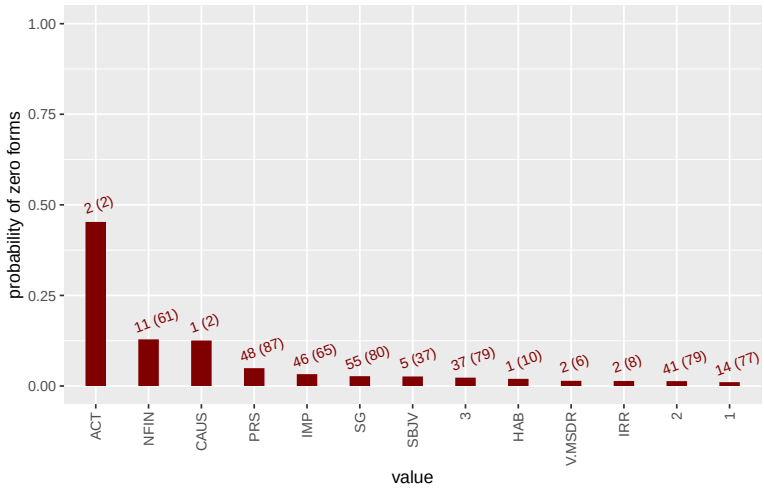


Figure 8: Verbal cells with the highest proportions of zero forms

The proportions of zero forms are very low for all but the active, nonfinite, causative, and present values. Except for the causative, those

values also figured in the cells with the highest proportions of zero forms seen in the previous section. Somewhat surprisingly, imperative, singular and third person have very low overall proportions of zero forms, although the number of languages that allow for zero forms is comparatively high.

To assess to what extent these results hold once the relations between languages are taken into account, I fitted a logistic Bayesian regression model with a phylogenetic control. The conditional effects of that model are shown in Figure 9.

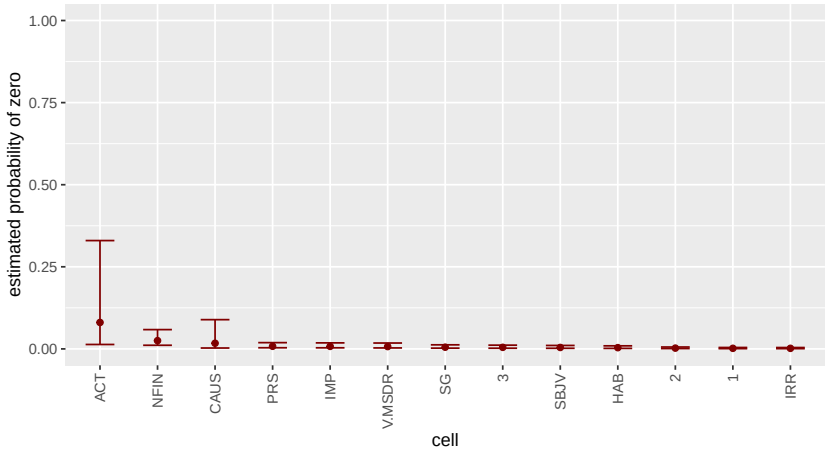


Figure 9: Conditional effects for the verbal values with the highest proportions of zero forms

It confirms the trends seen above; generally, the probability of zero forms is very low for all values. The only value that has a somewhat higher estimate is the active value. However, the estimate is again limited by the small number of languages with that feature, leading to a very high level of uncertainty. Thus, we can conclude that based on the UniMorph data, no value of verbal paradigms shows a very strong association with zero forms.

5.3 *Probability of zero forms in nominal and verbal paradigms*

A priori, we do not have any reason to expect a difference between the probabilities of zero forms in nominal and verbal paradigms. Especially under the assumption that zero forms developed for reasons

of coding efficiency, zero forms should generally be available in both domains as one of various tools to make linguistic structures and thus communication most efficient.

However, for nouns 57 out of 61 (93%) of the languages in the dataset have a zero form, while for verbs, only 76 out of 104 languages (73%) use zero forms. Already the raw proportions suggest that zero forms are generally more likely to be used in nominal than in verbal inflectional paradigms. Note that this measure does not specify how pervasive zero forms are but only registers whether or not zero forms occur in any cell of at least two lexemes in nominal or verbal paradigms within a given language in the dataset.[23]

This difference was tested using a Bayesian logistic regression model, predicting the probability of the presence of zero forms in paradigms depending on the part of speech, i.e. nouns and verbs. Again, I also added a phylogenetic regression term as a group-level effect to control for the relation between the languages. Before turning to the model results, there are three more potentially confounding factors that need to be addressed, namely the size of paradigms, the number of values expressed per cell and the number of lexemes for which inflection paradigms are available.

It could be the case that a difference between nominal and verbal paradigms stems from the fact that the verbal paradigms tend to have more cells than the nominal paradigms. The median number of cells for verbal paradigms is 36 (mean = 49), while the median size of nominal paradigms is 14 (mean = 28). Therefore, it is important to test whether a difference in the probability of zero forms is a result from the difference in paradigm size.

In a similar way, the number of values expressed per cell could be another confounding factor. One could imagine that cells with fewer or single values are more likely to be expressed by a zero form than cells that express a higher number of values. In addition, this may interact with the two domains, as the median number of values for nouns is 2 (mean = 3.1) and 4 for verbs (mean = 4.0). The other potentially confounding factor is the number of lexemes for which

[23] As was mentioned in Section 4.3, I excluded all exponents (zero and non-zero) that occurred only once in order to avoid exponents that arise from annotation errors in single lexemes.

inflection paradigms are available in the dataset. It is plausible to assume that the probability of seeing a zero form increases with more lexemes being available.

I therefore fitted 12 models that included different combinations of part of speech, the paradigm size, the number of values per cell and the number of lexemes as population-level effects (i.e. fixed effects). The performance of the 12 models was then compared to select the final model. I used approximated leave-one-out cross-validation for the comparison following the method described by Vehtari *et al.* (2017).[24] The results of the model comparisons suggest that paradigm size and the number of values per cell do not provide useful information for predicting the probability of zero forms. Thus, the best-performing model only includes part of speech and the number of lexemes as population-level effects.

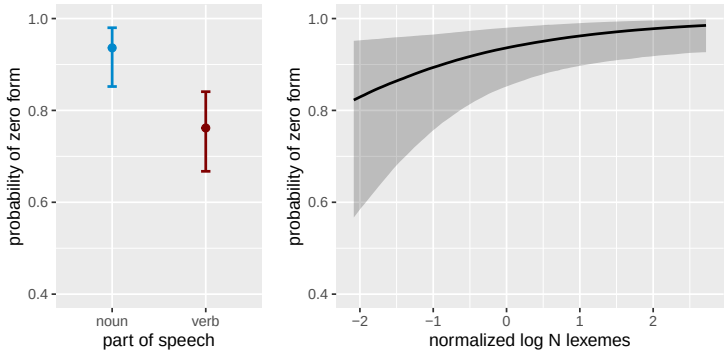


Figure 10: Conditional effects for parts of speech and number of lexemes

Figure 10 shows the conditional effects of this final model for part of speech and number of lexemes. Note that the number of lexemes is log transformed and normalized, so that it has a mean of zero and a standard deviation of 1.[25] The left plot in Figure 10 shows the effects of the part of speech. The points represent the means of the posterior

[24] The supplementary file code.R includes information on the different model specifications and the process of model comparison and selection.

[25] For both variables of paradigm size and number of lexemes, using log transformed numbers turned out to be more useful than the raw numbers, because both variables showed great variation in magnitude. Paradigm sizes range from small paradigms of only 2 cells to large paradigms with 432 cells in the dataset (overall median = 19, overall mean = 41). For number of lexemes, the small-

distribution, and the error bars delimit 95% of the posterior distribution, i.e. they show the 95% uncertainty interval. The model thus confirms that nominal paradigms are extremely likely to show zero forms, with an estimated probability of 0.93 (the uncertainty interval ranging from 0.85 to 0.97). Verbal paradigms, on the other hand, are predicted to have a slightly lower probability of zero forms at 0.70 (the uncertainty interval ranging from 0.60 to 0.80). As the intervals of nouns and verbs do not overlap, we can be certain that zero forms are more likely to occur in nominal than in verbal paradigms, given the data and the model. This difference will be taken up in the discussion in Section 6.1. At the same time, we also see a very weak effect of the number of lexemes on the probability of zero forms as well. Although the uncertainty bands are wide, we can expect to find slightly more zero forms if more data is available.

6 DISCUSSION

6.1 *Zero forms in nominal vs. verbal paradigms: explaining the difference*

The results of Section 5.3 showed that zero forms are more likely to occur in nominal than in verbal paradigms even if the paradigm size, the number of values per cell and the number of lexemes are controlled for. This calls for an explanation.

One potentially important factor is the place of inflection marking. Already Bybee *et al.* (1990) showed that suffixation is crosslinguistically more common than prefixation (or infixation, for that matter). However, comparing different domains of morphological marking, a more complex pattern emerges. Cysouw (2009) noted that the suffixing preference is much stronger for case and plural marking in the nominal domain and for TAM marking in the verbal domain. In those

---

est dataset has only data of 13 lexemes, while the largest one consists of 43258 (overall median = 391, overall mean = 1815). All three numeric variables were normalized in order to include them as predictors in the same model.

three domains, at least 80% of all exponents in his dataset were suffixes (Cysouw 2009, 2). He further showed the suffixation preference to be rather weak for person marking on the verb. Moreover, distinguishing between person marking paradigms based on the number of exponents, Cysouw (2009, 3) presented evidence for the suffixation preference to be restricted to systems with a larger number of markers (six or more). For systems with fewer exponents, his data showed a preference for prefixation.

These findings suggest that, overall, we can expect a stronger suffixation preference for inflectional marking in the nominal domain. This is relevant for the potential development of zero forms because phonetic material at the end of words is reduced at higher rates than material at the beginning of words (Bybee *et al.* 1990, 19, Hall 1988). This also relates to the insight that word-initial (or domain-initial) syllables tend to be more prominent than other syllables (e.g. Beckman 1998; Smith 2005; Cho *et al.* 2007; Kim 2004; Keating *et al.* 2003). Especially word-initial consonants tend to be strengthened and lengthened (e.g. White *et al.* 2020; Cho and Keating 2009; Fougeron 2001; Cho and Keating 2001), which is relevant here, since Bybee *et al.* (1990, 26) find that inflectional prefixes are crosslinguistically significantly more likely to have initial consonants than inflectional suffixes. Those properties may contribute to suffixes being more susceptible to phonetic reduction (to zero) over time than prefixes.

Therefore, the stronger suffixing preference in nominal inflection may facilitate the development of zero forms compared to the verbal domain, where more prefixation is used. Bybee *et al.* (1990, 22) also find that suffixes tend to have more allomorphs than prefixes, and they conclude that "prefixes appear to exhibit equal or greater stability than suffixes". This would of course require further empirical testing, but it may well be that the phonetic shape of prefixes is somewhat more stable across time than the one of suffixes, which could additionally contribute to the lower proportion of zero forms in the verbal domain.

6.2 *Coding efficiency and the distribution of zero forms*

As introduced in Section 2, the occurrence and distribution of zero forms was related to coding efficiency and form-frequency effects

in previous typological work. To be precise, the grammatical form-frequency correspondence hypothesis (Haspelmath 2021, 2) predicts more frequent morphosyntactic functions to prefer zero forms if the coding of more and less frequent and comparable functions is asymmetric. However, Section 2 also mentioned that the results from other previous studies (Bickel *et al.* 2015; Cysouw 2003; Guzmán Naranjo and Becker 2021a; Seržant and Moroz 2022; Siewierska 2010; Stolz and Levkovych 2019) point against coding efficiency as the main factor shaping the distribution of zero forms.

6.2.1 Evidence for coding efficiency to play a role

Considering the nominal and verbal morphosyntactic values that have a higher probability of being expressed by a zero form than the other morphosyntactic values (cf. Sections 5), at least some of them correspond to the values that already Greenberg (1966) identified as the more frequent values of their respective feature. In the nominal domain, we saw that the cells with the values of nominative, singular and accusative are comparatively likely to be expressed by a zero form. A new insight is that the indefinite value is fairly likely to be expressed by a zero form as well. In the verbal domain, cells including the values of active, imperative as well as the combination of third person and present tense appeared to be somewhat more likely to be encoded by a zero form than other cells.

In order to relate the findings of this study with the frequency distribution of the relevant morphosyntactic values in language use, Table 8 shows their token frequencies in the Universal Dependencies treebanks (Zeman *et al.* 2021). Table 9 shows the same for verbs.

Note that the numbers and proportions given in both tables should only be taken as an approximation of the usage frequencies, since not all tokens are annotated for all features, and since the text types differ greatly across languages. The column called “N value” shows the raw number of occurrences of a given value. For instance, we find 3789088 nouns identified as having singular number across 80 languages in the Universal dependencies. Note that not all values are attested or annotated in all languages, which makes direct comparisons of numbers of occurrences difficult. Therefore, I also extracted the total number of lexemes in all languages with a given feature value, e.g. singular

Table 8: Distribution of nominal features in the Universal Dependencies

value	N value	N total	N langs	prop value
number				
singular	3789088	5144632	80	0.74
plural	1347473	5152931	88	0.26
case				
nominative	887053	3166314	57	0.28
accusative	708587	2868514	46	0.25
absolutive	9801	32220	5	0.30
ergative	1984	22465	3	0.09
dative	234759	3182335	44	0.09
genitive	862330	3070696	56	0.28
vocative	6205	1849520	24	0.003
definiteness				
definite	384650	971411	17	0.40
indefinite	458271	957243	16	0.48

(on nouns), for which the relevant morphosyntactic feature, i.e. number, is annotated. The "N total" column shows that there are 5144632 instances of nouns with a number annotation in the Universal Dependencies. The last column of Tables 8 and 9 shows the proportion of the feature value, e.g. we see that for nouns, out of all 5144632 nouns with a number annotation, the 3789088 nouns marked as singular make up a proportion of 0.74. This last "prop value" column offers the relevant proportions that can serve as a comparison between values of the same morphosyntactic feature.

Returning to the values with the highest probabilities of being encoded by a zero form, Tables 8 and 9 confirm that those values are generally more frequent in language use than other values of their morphosyntactic feature. This is what we can see for the nominal values of singular, nominative, absolutive and accusative.[26] As for the indefinite value, which also showed a comparatively high probability

[26] Table 8 also includes numbers for the ergative, dative, genitive and vocative case values for comparison.

Table 9: Distribution of verbal features in the Universal Dependencies

value	N value	N total	N langs	prop value
tense				
past	739008	1666938	80	0.44
present	857398	1667680	77	0.51
future	47114	1049995	44	0.04
mood				
Indicative	1216937	1329281	75	0.92
Imperative	37822	1331878	77	0.03
person				
1	132245	1232404	78	0.11
2	68328	1232234	73	0.06
3	1021191	1231610	76	0.83
voice				
active	1032873	1227679	40	0.84
passive	144708	1011323	52	0.14

of zero forms, the token frequencies of definite and indefinite forms in the Universal Dependencies suggest that there is no strong difference between the two values, and a more detailed analysis would be necessary to draw any conclusions.

In the verbal domain, the values of active, imperative, third person and present were shown to be part of cells with the highest probabilities of zero forms. Table 9 confirms for some of those values that they also correspond to the most frequent value of their feature. This is clearly the case for third person and active forms and to a lesser extent for present tense forms (at least in comparison to future tense).

A closer examination of the usage frequencies is necessary to paint a more detailed picture. Still, it is evident that zero forms, even though they are generally not the preferred expression for any cell in inflection paradigms, tend to occur with morphosyntactic values that are more frequent in usage than other values of the same feature.

Evidence for other factors to play a role 6.2.2

Although the values with a higher probability of zero forms largely correspond to the more frequent values in language use (cf. Section 6.2.1), there is no morphosyntactic function for which zero forms would be a preferred or most likely coding strategy. Especially for the cells and values for which the models estimated a higher probability of zero forms, we also saw very large uncertainty intervals. This means that the best estimate of the probability of zero forms for those cells contained a high level of uncertainty. Put differently, there is a large degree of variation across languages and lexemes in the data that the model cannot account for based on the information it has. This does not necessarily mean that we need to add more predictors to the models to reduce the uncertainty. It may simply point to the fact that there is a high degree of idiosyncratic variation, i.e. that the occurrence of zero forms across different cells is simply not very homogeneous across languages or lexemes and may rather be accounted for by language-specific factors. This is supported by the results of Section 5.3, showing that nominal paradigms are generally very likely (0.93) to have zero forms at least in one cell in a few lexemes. Although being lower, the probability of 0.70 of zero forms occurring with some lexemes in verbal paradigms is still relatively high. Thus, zero forms as such are common in inflectional morphology, but their distribution is not very consistent across lexemes of different languages. In addition to language-specific factors, it may even be the case that properties specific to single inflection classes or lexemes account for the occurrence of zero forms in many cases.

The development of zero forms 6.3

In order to account for the distribution of zero forms in inflectional morphology across languages, it is important to consider how zero forms develop diachronically. This section will provide a brief overview of the three different processes that can lead to the development of zero marking: differential phonetic reduction, differential non-development and reanalysis (Bybee 1994, 1985; Koch 1995; Haspelmath 2008a; Cristofaro 2019, 2021). The fact that zero forms can develop in different ways suggests that we do not, at least generally,

deal with a single process motivated by an efficient end-state that languages adapt to (cf. Cristofaro 2021).

The probably most often-cited process for the shortening of forms (and the development of zero forms) is phonetic reduction (Bybee 2003, 2007, 2015; Givón 2018; Haspelmath 2008a; Lehmann 2015). Especially Bybee (2003, 2015) has argued for phonetic reduction being a consequence of the repetition and automatization in the production that occurs in grammaticalization processes. However, Haspelmath (2008a, 207) notes that "[t]here may also be cases of differential phonological reduction of nominatives […], but it is probably very difficult to find examples of phonological reduction leading to most of the other asymmetries. Zipf's diachronic mechanism of phonological reduction is thus less important in explaining grammatical asymmetries than one might have thought". Given that we do not find many clear examples of efficiency-related phonetic reduction to zero in the literature, it is very plausible that this process does actually not account for the development of most zero forms in morphology.

Probably the main process that leads to the development of zero forms is the differential non-development of a marker (cf. Bybee 1994; Cristofaro 2019, 2021). For instance, we can imagine a scenario in which number is not marked initially on nouns. For some independent reason, plural marking could be developed and expanded from being a lexical marker that is used occasionally to then be used more and more systematically until it becomes more abstract and grammaticalized. At the same time that the plural marker develops into an inflectional exponent, the absence of it becomes more systematically associated with the singular so that at some point, the singular is expressed by a zero form. In such a scenario, the zero form develops in opposition to another exponent developing for another cell in the paradigm. Although convincing examples with diachronic data tracing the details of e.g. a developing plural exponent (and a developing zero singular form) are hard to come by, we find a number of cases in the literature for which this development has been proposed (cf. Cristofaro 2019). Importantly, in such cases it is not the final state of having a zero-expressed value that drives the diachronic process. Rather, the developing zero form is a consequence of various, potentially language-specific factors, which have led to the development of a new grammatical category with an overt exponent for one or more

values of that category, and with no overt exponent for the value in question. The zero form can simply result from the lack of a suitable source for a new grammatical marker of, e.g., the plural value in a given language. Frequency may play a role in such processes, but they certainly also involve many independent language-specific factors.

A slightly different type of differential non-development may apply to processes leading to imperatives expressed by zero forms. As was noted in Section 5.2, imperative forms are among the cells of verbal paradigms that are most likely to be expressed by zero forms. A possible explanation points to differential non-development, since the second person is highly recoverable in contexts of imperatives (as opposed to contexts of e.g. indicative forms). Thus, many languages already allow or require the use of imperatives without any second person pronoun. This in turn means that the source construction of a verbal person marker is often not available for imperative forms (Aikhenvald 2010, 147; Nikolaeva 2007, 163; Sadock and Zwicky 1985, 173).

However, the use of bare verb forms for imperatives has also been motivated by iconicity (Aikhenvald 2010, 46). Using the shortest verb form makes imperatives very direct and abrupt. This can convey urgency and reflect that imperatives usually call for an immediate reaction. If iconicity is indeed involved (at least in some contexts of imperative forms), this would be a potential example of differential non-development motivated by an efficient outcome state, where the shortest possible form can be used in a context in which successful communication is nevertheless guaranteed.

The last process that can lead to the development of zero forms in morphology is the reanalysis of an exponent as part of the stem (i.e. as belonging to the lexical material of a word form), resulting in the absence of exponence of a morphosyntactic feature or combination thereof, i.e. a zero form. This phenomenon is well known from historical linguistics as “Watkin's law” (cf. Watkins 1962; Koch 1995; Bybee 1985). The exact circumstances of this type of reanalysis are not very clear from the data or the literature either, but it is assumed that a given cell of the paradigm is used so frequently that its exponent is reanalyzed as part of the stem. At the same time, the former exponent is added to the other forms of the paradigm as well, restructuring the entire paradigm. What is clear at this point is that the frequency and predictability of the involved forms play an important role. This

makes such processes of reanalysis promising candidates for further investigations of efficiency-driven developments of zero forms.

7 CONCLUSION

This study offered a first quantitative crosslinguistic study of the occurrence and distribution of zero forms in nominal and verbal inflectional morphology. In doing so, it laid the grounds for further discussions in an ongoing debate in typology concerning the role of coding efficiency in the development and distribution of zero forms. The present study took into account the behavior of single lexemes and captured the variation across inflection classes and irregular forms. In addition, a more realistic picture of the distribution of zero forms emerged through the use of surface forms without the additional morphological segmentation of single values of morphosyntactic features.

The results of this study based on the UniMorph dataset showed that no cells, neither in nominal nor in verbal paradigms, have a strong association with zero forms. The findings further evidenced a high degree of variation across languages and lexemes in the distribution of zero forms. In general, we saw a strong crosslinguistic preference for overt exponents.

However, it could also be confirmed that, if zero forms occur, they are more likely to occur in certain cells over other cells of of inflectional paradigms. In the nominal domain, cells including the nominative, singular and indefinite values were somewhat more likely than other cells to be expressed by zero forms. In the verbal domain, cells including the imperative, the active, and the third person singular present values had a slightly higher probability of zero forms than other cells across languages. It could be shown that these values also correspond to the comparatively frequent values in the Universal Dependency treebanks, which suggests that coding efficiency does play a role in the distribution of zero forms to a certain extent. Still, the high degree of variation in the distribution of zero forms across lexemes and languages has to be taken as evidence for other, potentially language-specific factors to play an equally important role in the development and distribution of zero forms.

REFERENCES

Alexandra Aikhenvald (2010), *Imperatives and Commands*, Oxford University Press, Oxford.

Stephen R. Anderson (1992), *A-Morphous Morphology*, Cambridge University Press, Cambridge.

Matthew Baerman, Dunstan Brown, and Greville G. Corbett (2017), *Morphological Complexity*, Cambridge University Press, Cambridge.

Jill Beckman (1998), *Positional Faithfulness*, Ph.D. thesis, University of Massachusetts, Amherst.

Sacha Beniamine and Matías Guzmán Naranjo (2021), Multiple Alignments of Inflectional Paradigms, *Proceedings of the Society for Computation in Linguistics*, 4:216–227.

Balthasar Bickel, Alena Witzlack-Makarevich, Taras Zakharko, and Giorgio Iemmolo (2015), Exploring Diachronic Universals of Agreement: Alignment Patterns and Zero Marking across Person Categories, in Jürg Fleischer, Elisabeth Rieken, and Paul Widmer, editors, *Agreement from a Diachronic Perspective*, pp. 29–52, De Gruyter, Berlin.

James Blevins (2003), Stems and Paradigms, *Language*, 79(4):737–767.

James Blevins (2005), Word-Based Declensions in Estonian, in Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 2005*, pp. 1–25, Springer, Dordrecht, doi:10.1007/1-4020-4066-0_1.

James Blevins (2006), Word-Based Morphology, *Journal of Linguistics*, 42(3):531–573.

James Blevins (2016), *Word and Paradigm Morphology*, Oxford University Press, Oxford.

Bernard Bloch (1947), English Verb Inflection, *Language*, 23(4):399–418, doi:10.2307/410300.

Leonard Bloomfield (1933), *Language*, Holt, New York.

Olivier Bonami (2012), Stems in Inflection and Lexeme Formation, *Word Structure*, 5(1).

Olivier Bonami and Sacha Beniamine (2021), Leaving the Stem by Itself, in Sedigheh Moradi, Marcia Haag, Janie Rees-Miller, and Andrija Petrovic, editors, *All Things Morphology: Its Independence and Its Interfaces*, pp. 81–98, Benjamins, Amsterdam, doi:10.1075/cilt.353.05bon.

Dunstan Brown (1998), Stem Indexing and Morphonological Selection in the Russian Verb: A Network Morphology Account, in Ray Fabri, Albert Ortmann, and Teresa Parodi, editors, *Models of Inflection*, pp. 196–224, Niemeyer.

Paul-Christian Bürkner (2017), Brms: An R Package for Bayesian Multilevel Models Using Stan, *Journal of Statistical Software*, 80(1):1–28, doi:10.18637/jss.v080.i01.

Joan Bybee (1985), *Morphology. A Study of the Relation of Meaning and Form*, John Benjamins, Amsterdam.

Joan Bybee (1994), The Grammaticization of Zero: Asymmetries in Tense and Aspect Systems, in William Pagliuca, editor, *Perspectives on Grammaticalization*, pp. 235–254, Benjamins, Amsterdam.

Joan Bybee (2003), Mechanisms of Change in Grammaticization: The Role of Frequency, in Brian Joseph and Richard Janda, editors, *Handbook of Historical Linguistics*, pp. 602–623, Blackwell, Oxford.

Joan Bybee (2007), *Frequency of Use and the Organization of Language*, Oxford University Press, Oxford.

Joan Bybee (2011), Markedness, in Jae Jung Song, editor, *The Oxford Handbook of Typology*, pp. 1–11, Oxford University Press, Oxford.

Joan Bybee (2015), *Language Change*, Cambridge University Press, Cambridge.

Joan Bybee and Östen Dahl (1989), The Creation of Tense and Aspect Systems in the Languages of the World, *Studies in Language*, 13(1):51–103.

Joan Bybee, William Pagliuca, and Revere Perkins (1990), On the Asymmetries in the Affixation of Grammatical Material, in William Croft, Suzanne Kemmer, and Keith Denning, editors, *Studies in Typology and Diachrony. Papers Presented to Joseph H. Greenberg on His 75th Birthday*, pp. 1–42, Benjamins, Amsterdam.

Joan Bybee, Revere Perkins, and William Pagliuca (1994), *The Evolution of Grammar. Tense, Aspect, and Modality in the Languages of the World*, The University of Chicago Press, Chicago.

Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017), Stan: A Probabilistic Programming Language, *Journal of Statistical Software*, 76(1):1–32, doi:10.18637/jss.v076.i01.

Taehong Cho and Patricia Keating (2001), Articulatory and Acoustic Studies on Domain-Initial Strengthening in Korean, *Journal of Phonetics*, 29(2):155–190, doi:10.1006/jpho.2001.0131.

Taehong Cho and Patricia Keating (2009), Effects of Initial Position versus Prominence in English, *Journal of Phonetics*, 37(4):466–485, doi:10.1016/j.wocn.2009.08.001.

Taehong Cho, James McQueen, and Ethan Cox (2007), Prosodically Driven Phonetic Detail in Speech Processing: The Case of Domain-Initial Strengthening in English, *Journal of Phonetics*, 35(2):210–243, doi:10.1016/j.wocn.2006.03.003.

Matt Coler (2015), Aymara Inflection, in Matthew Baerman, editor, *The Oxford Handbook of Inflection*, pp. 1–30, Oxford University Press, Oxford.

Matt Coler (2018), Subtractive Morphology & Disfixation in Aymara Case, in *Case and Agreement in Panará (... and Beyond)*, pp. 1–9, Groningen, doi:10.13140/RG.2.2.26153.03682.

Ellen Contini-Morava (2006), The Difference between Zero and Nothing: Swahili Noun Class Prefixes 5 and 9/10, in Joseph Davis, Radmila Gorup, and Nancy Stern, editors, *Advances in Functional Linguistics*, pp. 211–222, Benjamins, Amsterdam.

Sonia Cristofaro (2019), Taking Diachronic Evidence Seriously: Result-oriented vs. Source-Oriented Explanations of Typological Universals, in Karsten Schmidtke-Bode, Natalia Levshina, Susanne Maria Michaelis, and Ilya Seržant, editors, *Explanation in Typology: Diachronic Sources, Functional Motivations and the Nature of the Evidence*, pp. 25–46, Language Science Press, Berlin.

Sonia Cristofaro (2021), Typological Explanations in Synchrony and Diachrony: On the Origins of Third Person Zeroes in Bound Person Paradigms, *Folia Linguistica*, 55(s42-s1):25–48, doi:10.1515/flin-2021-2013.

William Croft (2003), *Typology and Universals*, Cambridge University Press, Cambridge, second edition.

Michael Cysouw (2003), *The Paradigmatic Structure of Person Marking*, Oxford University Press, Oxford.

Michael Cysouw (2009), The Asymmetry of Affixation, *Snippets*, 20:10–144.

Holger Diessel (2019), *The Grammar Network: How Linguistic Structure Is Shaped by Language Use*, Cambridge University Press, Cambridge.

Cécile Fougeron (2001), Articulatory Properties of Initial Segments in Several Prosodic Constituents in French, *Journal of Phonetics*, 29(2):109–135, doi:10.1006/jpho.2000.0114.

Talmy Givón (2018), *On Understanding Grammar*, Benjamins, Amsterdam.

Joseph Harold Greenberg (1966), *Language Universals: With Special Reference to Feature Hierarchies*, Mouton, The Hague.

Matías Guzmán Naranjo and Laura Becker (2021a), Coding Efficiency in Nominal Inflection: Expectedness and Type Frequency Effects, *Linguistics Vanguard*, 7(s3):20190075, doi:10.1515/lingvan-2019-0075.

Matías Guzmán Naranjo and Laura Becker (2021b), Statistical Bias Control in Typology, *Linguistic Typology*, 26(3):605–670, doi:10.1515/lingty-2021-0002.

William Haas (1957), Zero in Linguistics Description, in John Rupert Firth, editor, *Studies in Linguistic Analysis*, pp. 33–53, Blackwell, Oxford.

John HAIMAN (1983), Iconic and Economic Motivation, *Language*, 59(4):781–819.

Christopher HALL (1988), Integrating Diachronic and Processing Principles in Explaining the Suffixing Preference, in John HAWKINS, editor, *Explaining Language Universals*, pp. 321–349, Basil Blackwell, London.

Harald HAMMARSTRÖM, Robert FORKEL, Martin HASPELMATH, and Sebastian BANK (2021), *Glottolog 4.4*, Max Planck Institute for the Science of Human History, Leipzig.

Martin HASPELMATH (2008a), Creating Economical Morphosyntactic Patterns in Language Change, in Jeff GOOD, editor, *Linguistic Universals and Language Change*, pp. 185–214, Oxford University Press, Oxford.

Martin HASPELMATH (2008b), Frequency vs. Iconicity in Explaining Grammatical Asymmetries, *Cognitive Linguistics*, 19(1):1–33, doi:10.1515/COG.2008.001.

Martin HASPELMATH (2008c), A Frequentist Explanation of Some Universals of Reflexive Marking, *Linguistic Discovery*, 6(1):40–63, doi:10.1349/PS1.1537-0852.A.331.

Martin HASPELMATH (2021), Explaining Grammatical Coding Asymmetries: Form–Frequency Correspondences and Predictability, *Journal of Linguistics*, pp. 1–29, doi:10.1017/S0022226720000535.

Martin HASPELMATH, Andreea CALUDE, Michael SPAGNOL, Heiko NARROG, and Elif BAMYACI (2014), Coding Causal–Noncausal Verb Alternations: A Form–Frequency Correspondence Explanation, *Journal of Linguistics*, 50(3):587–625.

Martin HASPELMATH and Andres KARJUS (2017), Explaining Asymmetries in Number Marking: Singulatives, Pluratives, and Usage Frequency, *Linguistics*, 55(6):1213–1235, doi:10.1515/ling-2017-0026.

George HEWITT (1995), *Georgian: A Structural Reference Grammar*, Benjamins, Amsterdam.

Roman JAKOBSON ([1939] 1983), Zero Sign, in Linda WAUGH and Morris HALLE, editors, *Russian and Slavic Grammar: Studies 1931-1981*, pp. 1–14, De Gruyter, New York.

Patricia KEATING, Taehong CHO, Fougeron CECILE, and C.S. HSU (2003), Domain-Initial Strengthening in Four Languages, *Phonetic Interpretation: Papers in Laboratory Phonology*, 6.

Sahyang KIM (2004), *The Role of Prosodic Phrasing in Korean Word Segmentation*, Ph.D. thesis, University of California, Los Angeles.

Harold KOCH (1995), The Creation of Morphological Zeros, in Geert BOOIJ and Jaap VAN MARLE, editors, *Yearbook of Morphology 1994*, pp. 31–731, Springer, Dordrecht.

Christian Lehmann (2015), *Thoughts on Grammaticalization*, Language Science Press, Berlin.

Natalia Levshina (2022), *Communicative Efficiency: Language Structure and Use*, Cambridge University Press, Cambridge.

Martin Maiden (1992), Irregularity as a Determinant of Morphological Change, *Journal of Linguistics*, 28(2):285–312.

Peter Hugoe Matthews (1972), *Inflectional Morphology: A Theoretical Study Based on Aspects of Latin Verb Conjugation*, Cambridge University Press.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky (2020), UniMorph 3.0: Universal Morphology, in *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3922–3931, European Language Resources Association, Marseille, France.

William McGregor (2003), The Nothing That Is, the Zero That Isn't, *Studia Linguistica*, 57(2):75–119, doi:10.1111/1467-9582.00100.

Igor Mel'cuk (2002), Towards a Formal Concept Zero Linguistic Sign: Applications in Typology, in Sabrina Bendjaballah, Wolfgang Dressler, Oskar Pfeiffer, and Maria Voeikova, editors, *Morphology 2000: Selected Papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000*, pp. 241–258, Benjamins, Amsterdam.

Marianne Mithun (1986), When Zero Isn't There, *Annual Meeting of the Berkeley Linguistics Society*, 12(0):195–211, doi:10.3765/bls.v12i0.1882.

Fabio Montermini and Olivier Bonami (2013), Stem Spaces and Predictability in Verbal Inflection, *Lingue e linguaggio*, 2:171–190, doi:10.1418/75040.

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

Irina Nikolaeva (2007), *Finiteness: Theoretical and Empirical Foundations*, Oxford University Press, Oxford.

Vito Pirrelli and Marco Battista (2000), The Paradigmatic Dimension of Stem Allomorphy in Italian Verb Inflection, *Rivista di Linguistica*, 12(2):307–380.

Geoffrey Pullum and Arnold Zwicky (1991), A Misconceived Approach to Morphology, *Proceedings of the West Coast Conference on Formal Linguistics*, 10.

R Core Team (2021), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

R. H. Robins (1997), *A Short History of Linguistics*, Routledge, New York, fourth edition.

Jerrold Sadock and Arnold Zwicky (1985), Speech Act Distinctions in Grammar, in Timothy Shopen, editor, *Language Typology and Syntactic Description. Volume 1*, pp. 155–196, Cambridge University Press, Cambridge.

Gerald Sanders (1988), Zero Derivation and the Overt Analogue Criterion, in Michael Hammond and Michael Noonan, editors, *Theoretical Morphology*, pp. 155–175, Academic Press, San Diego, CA.

Ferdinand Saussure (1916), *Cours de Linguistique Générale*, Payot, Lausanne.

Ilja Seržant and George Moroz (2022), Universal Attractors in Language Evolution Provide Evidence for the Kinds of Efficiency Pressures Involved, *Humanities and Social Sciences Communications*, 9(1):1–9, doi:10.1057/s41599-022-01072-0.

Anna Siewierska (2010), Person Asymmetries in Zero Expression and Grammatical Functions, in Franck Floricic, editor, *Essais de Linguistique Generale et de Typologie Linguistique Offerts Au Professeur Denis Creissels à l'occasion de Ses 65 Ans*, pp. 425–438, Presses de l'École Normale Supérieure, Paris.

Jennifer Smith (2005), *Phonological Augmentation in Prominent Positions*, Taylor & Francis, Oxfordshire, doi:10.4324/9780203506394.

Jae Jung Song (2018), *Linguistic Typology*, Oxford University Press, Oxford.

Andrew Spencer (2012), Identifying Stems, *Word Structure*, 5(1):88–108, doi:10.3366/word.2012.0021.

Matthew Stave, Ludger Paschen, François Pellegrino, and Frank Seifart (2021), Optimization of Morpheme Length: A Cross-Linguistic Assessment of Zipf's and Menzerath's Laws, *Linguistics Vanguard*, 7(s3), doi:10.1515/lingvan-2019-0076.

Thomas Stolz and Nataliya Levkovych (2019), Absence of Material Exponence, *Language Typology and Universals*, 72(3):373–400, doi:10.1515/stuf-2019-0015.

Gregory Stump (2001), *Inflectional Morphology: A Theory of Paradigm Structure*, Cambridge University Press, Cambridge.

Gregory T. Stump and Rafael Finkel (2013), *Morphological Typology: From Word to Paradigm*, volume 138 of *Cambridge Studies in Linguistics*, Cambridge University Press, Cambridge.

Aki Vehtari, Andrew Gelman, and Jonah Gabry (2017), Practical Bayesian Model Evaluation Using Leave-One-out Cross-Validation and WAIC, *Statistics and Computing*, 27(5):1413–1432, doi:10.1007/s11222-016-9696-4.

Calvert Watkins (1962), *Indo-European Origins of the Celtic Verb*, Dublin Institute for Advanced Studies, Dublin.

Laurence White, Silvia Benavides-Varela, and Katalin Mády (2020), Are Initial-Consonant Lengthening and Final-Vowel Lengthening Both Universal Word Segmentation Cues?, *Journal of Phonetics*, 81:100982, doi:10.1016/J.WOCN.2020.100982.

Jingting Ye (2020), Independent and Dependent Possessive Person Forms: Three Universals, *Studies in Language*, 44(2):363–406, doi:10.1075/sl.19020.ye.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arican, órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökirmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzitis, Bruno Guillaume, Céline Guillot-Barbance, Tunga

Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşikara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, Lorena Martín-Rodríguez, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos,

Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Ivan Riabov, Michael Riessler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Saniyar, Dage Särg, Baiba Saulte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Shafi Sourov, Carolyn Spadine, Rachele Sprugnoli, Steinór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Sri Hartati Wijono, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yildiz, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane (2021), Universal Dependencies 2.9.

George Kingsley Zipf (1935), *The Psychobiology of Language: An Introduction to Dynamic Philology*, MIT Press, Cambridge, MA.

Arnold Zwicky (1985), How to Describe Inflection, in Mary Niepokuj, Mary Van Clay, Vassiliki Nikiforidou, and Deborah Feder, editors, *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, pp. 372–386, Berkeley Linguistics Society, Berkeley, CA.