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

Abstract The present paper examines the distribution of zero forms in nominal and verbal inflectional morphology. In typology, zero forms play an important role for coding efficiency and form-frequency effects in morphosyntax. Form-frequency effects reflect that more frequent (grammatical) elements tend to be shorter than comparable less frequent ones. Zero forms, i.e. the absence of an exponent to encode a morphosyntactic function, is usually assumed to pattern with shorter markers, although a crosslinguistic overview of the distribution of zero forms in inflectional morphology is still not available. This is the objective of the present study, analysing the distribution of zero forms in the UniMorph dataset, which allows for a token-based typological approach. The results show that no cells, neither in nominal nor in verbal paradigms, have a strong association with zero forms; in general, there is a strong crosslinguistic preference for overt exponents. We also find a very high degree of variation across languages and lexemes in the distribution of zero forms. Therefore, the findings of this study do not support the hypothesis of coding efficiency driving the development of zero forms. Rather, they support the hypothesis that zero forms develop as a by-product through a number of different, unrelated diachronic processes.

Keywords token-based typology, corpus typology, zero marking, quantitative morphology, UniMorph

1 Introduction

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 and 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 markers tend to be shorter or at least not longer than comparable less frequent markers (Greenberg 1966; Guzmán Naranjo & Becker 2021a; Haspelmath 2008c, 2021; Haspelmath & Karjus 2017; Haspelmath et al. 2014; Stave et al. 2021).

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 2008b, 2008c, 2021; Song 2018: Ch. 7). However, a crosslinguistic quantitative overview of the distribution of zero forms is still not available. Moreover, results from previous studies suggest that frequency and coding efficiency may not be well suited to account for the distribution of zero forms in inflectional morphology (Guzmán Naranjo & Becker 2021a; Seržant & Moroz 2022; Stolz & Levkovych 2019). The objective of this paper is to start filling this gap. To do so, I analyse 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 frequency effects and coding efficiency in Section 2 and introduce a working definition of zero forms in Section 3. Section 4 describes the dataset, the preprocessing steps and the extraction of zero forms. Finally, a few examples of zero forms will be presented. I then analyse 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, no cell of either nominal or verbal paradigms is strongly associated with zero forms, and the occurrence of zero forms appears to be very language-specific and lexeme-dependent. I then analyse whether zero forms are similarly distributed in nominal and verbal inflection paradigms. Controlling for other confounding factors, it will be shown that zero forms 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. I will argue that the findings of this study do not support this hypothesis. Section 7 concludes.

2 Zero forms and coding efficiency

The modern understanding of coding efficiency or form-frequency effects 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.

Markedness not always being used in this way, Haspelmath (2008b,c) convincingly argues that we do not need the notion of markedness to explain such crosslinguistically robust patterns. Instead, he showed that the length, complexity or availability of expression (e.g. markers of grammatical categories) can be accounted for by their frequency in language use itself. In a recent study, Haspelmath (2021: 2) proposes 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 seman-

> tic 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. While there is substantial evidence for such form-frequency effects between comparable grammatical expressions, they usually only concern the length of forms in terms of shorter vs. longer forms.[1] On the other hand, the participation of zero forms 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 & 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 & 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 & 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 form-frequency effects in the expected way, their results suggest that the occurrence of zero forms does not follow that of shorter forms. They note that fitting a model predicting the length of markers on the basis of their type frequency, a simple Poisson model strongly overestimates the occurrence of zero forms. In other words, based on the frequency information, many more zero forms are predicted than observed.[2]

Another example is the occurrence of zero forms for person and number indexing on verbs. For instance, Bickel et al. (2015); Cysouw (2003) and Siewierska (2010) find that zero forms for indexing on the verb are typologically uncommon; they do not find evidence for a paradigmatic

[1] 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 (2008a) for reflexive marking, Haspelmath & Karjus (2017) for number marking and Ye (2020) for (in)dependent possessor marking.

[2] 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 markers to have a very low probability of 0.02 (Guzmán Naranjo & Becker 2021a: 6).

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 & Moroz (2022) also mention zero forms in verbal indexing. Analysing the length of the indexes on verbs in a typological sample, they argue for an atractor state in which the lengths of different indexes are associated with their frequencies in language use. Seržant & 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 & Moroz (2022: 7) hypothesise 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 & 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 important point is that coding efficiency, being able to account for the length of comparable more and less frequent expression, does not seem to be applicable to the frequency distribution of zero forms in person and number indexing on verbs.

In order to shed more light on the relation between coding efficiency and the distribution of zero forms, the present study offers a first crosslinguistic and quantitative overview of the distribution of zero forms in nominal and verbal inflection morphology. With this, we can test whether zero forms are indeed preferred in certain morphosyntactic functions that occur frequently in language use, or if coding efficiency, as other previous studies suggest, is a less important factor to account for the distribution of zero forms.

3 A working definition of zero forms

Without any assumptions about the ontological status of zero markers or zero morphemes, I follow Stolz & Levkovych (2019) in using zero marking as a descriptive shorthand for the absence of material exponence of a given morphosyntactic function.[3] In other words, I do not assume the presence of a zero morpheme, but understand it as the absence of exponence that is otherwise used to encode other functions of the same inflection paradigm. Thus, zero forms can only occur in opposition to overt marking in the same inflection paradigm.

Another problematic issue of the "traditional" zero morpheme approach is that it may postulate a zero morpheme for any single morphosyntactic function that does not correspond to

[3]For discussions on the ontological status of zero markers, see Garcia & van Putte (1989); Jakobson ([1939] 1983); Lemaréchal (1997); Mel'cuk (2002). For a discussion on issues related to the use of zero morphs in morpheme-based, segmental approaches to morphology, see Anderson (1992); Blevins (2016); Pullum & Zwicky (1991).

an overt exponent is. As Anderson (1992: 30) notes, it "leads to the formal problem of assigning a place in the structure (and linear order) to all of those zeros." Therefore, I will neither use a morpheme-based approach for the purposes of the present study nor argue for zero morphemes. I follow a word and paradigm approach to morphology instead (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 marker or 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. This way, further segmentation which may require language-specific insights and which may not always be desirable or useful is avoided.

For a definition of zero forms, we also require a definition of all the material that does not serve as an exponent of any morphosyntactic function but represents the lexeme itself. This is what I will call a **stem**. For the purposes of the present paper, the stem correspond to the phonetic material that is shared by all forms of an inflectional paradigm. An exponent of a given cell of the paradigm then corresponds to the additional material of the form in that cell. If there is no phonetic material in addition to the stem in the form of a cell of the paradigm, I call this form a **zero form**.

A simple example is the singular form of most nouns in English. 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. Because exponents of functions are defined based on the relations between the forms of the different cells of the inflection paradigm, I will not distinguish between a marker 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 or markers to single abstract morphosyntactic values but to the relevant value combinations of the inflection paradigms.

This stands in contrast with the typological tradition of analysing marker length and form-

frequency effects including zero marking in inflectional morphology. Usually, markers have not been associated with cells but with single abstract values (Bybee 1985; Dahl 1985; Greenberg 1966; Haspelmath 2008b,c; Stolz & Levkovych 2019). The issue with this approach is that it does not reflect the morphological reality of many if not most languages. Values are not marked in isolation but often occur in combinations, and it is not always possible to justify a segmental analysis. Croft (2003: 93-94) notes this issue for zero forms as well, and his solution is very similar to the one proposed for the present study:

> Frequently, however, there is zero coding which involves more than one category [...] Should this be taken as evidence for the unmarked status of third person, singular, animate, or all of the above? The answer is all of the above [...] In fact, in the Ngakalan case the only opportunity for the absence of overt expressions is when all three of the categories cumulated in the morpheme have their unmarked values; this is a common phenomenon. (Croft 2003: 93-94)

Such interactions between the marking of different functions are often additionally complicated by inflection classes, which make use of different markers for the cells of the paradigms. In Sections 4.3 and 4.4, I show in more detail how the approach taken in the present paper deals with variation in the exponence due to inflection classes, with the shared exponence of different values, 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 stems 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 analysed in this study is 141.[4] 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 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]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.

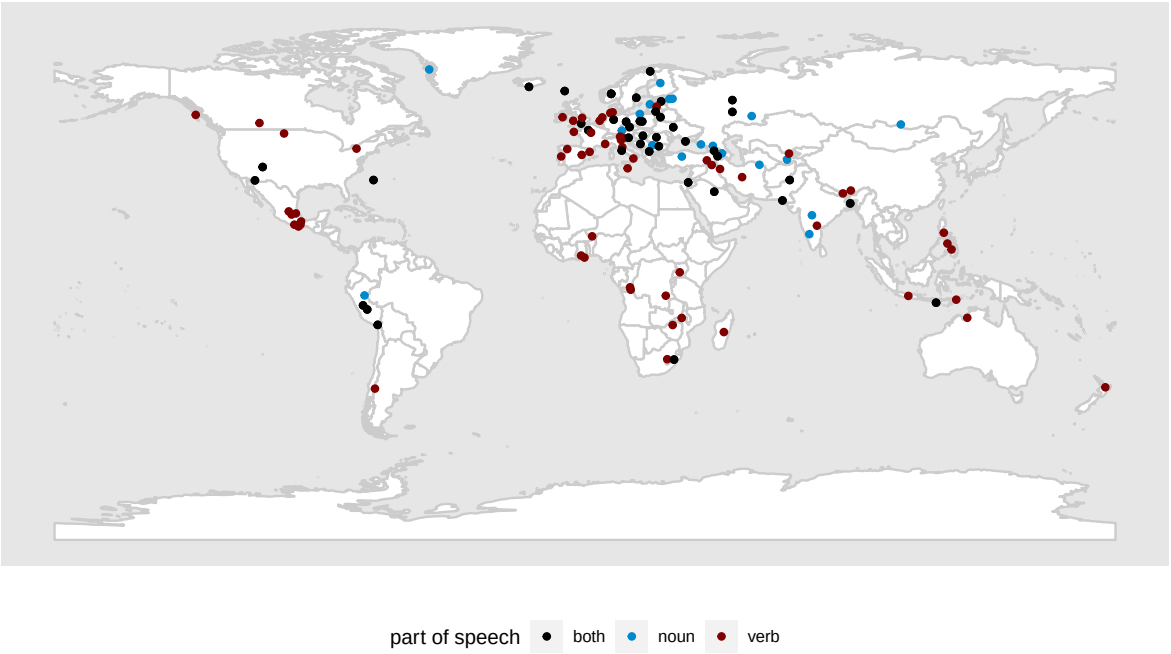

Figure 1: Map of the dataset.

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 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, Dalmia & Littell 2018). Epitran currently has modules to transcribe 28 of the languages used here.[5]

While not perfect, Epitran offers a somewhat more realistic representation of the forms

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

occupying the different cells of inflection paradigms. Table 1 illustrates this by showing the transcriptions generated 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.[6] Except for two cases, the automatically generated phonological transcription is accurate. Epitran only seems to struggle with the third person plural forms in the past imperfective (*allumaient* / *alymaj*) and present conditional (*allumeraient* / *alymɛraj*), where the orthographic segment *-aient* is transcribed as *-aj*, while is should have been transcribed as *-ɛ* in consistency with the first and third person singular forms. However, for the purposes of the present study, this is irrelevant, since the objective is to identify the stem as the longest common substring across all forms and those cells that consist of no more material than the longest common substring, i.e. a zero form.

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

	1SG	2SG	3SG	1PL	2PL	3PL
PRS.IND	allume alym	allumes alym	allume alym	allumons alymɔn	allumez alymɛz	allument alym
PST.IPF.IND	allumais alymɛ	allumais alymɛ	allumait alymɛ	allumions alymjɔn	allumiez alymjɛz	allumaient alymaj
PST.PFV.IND	allumai alymɛ	allumas alyma	alluma alyma	allumâmes alymɑm	allumâtes alymɑt	allumèrent alymɛr
FUT	allumerai alymɛrɛ	allumeras alymɛra	allumera alymɛra	allumerons alymɛrɔn	allumerez alymərɛz	allumeront alymɛrɔn
PRS.COND	allumerais alymɛrɛ	allumerais alymɛrɛ	allumerait alymɛrɛ	allumerions alymɛrjɔn	allumeriez alymɛrjɛz	allumeraient alymɛraj
PRS.SUBJ	allume alym	allumes alym	allume alym	allumions alymjɔn	allumiez alymjɛz	allument alym
PST.SUBJ	allumasse alymas	allumasses alymas	allumât alymɑ	allumassions alymasjɔn	allumassiez alymasjɛz	allumassent alymas

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 information provided in the source or in reference grammars. Other manual changes 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

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

annotation to a single label for a given value in all languages.[7]

4.3 Segmentation and extraction of zero forms

In order to analyse the distribution of zero forms, I automatically segmented the forms following the method developed in Beniamine & Guzmán Naranjo (2021) and Guzmán Naranjo & Becker (2021a). As mentioned in Section 3, the segmentation or analysis follows a word and paradigm approach to morphology, i.e. whole forms are paired with morphosyntactic functions according to their distribution across the inflectional paradigms. The shortest common substring of phonological material shared between all cells in the paradigm of a given lexeme is called the stem. All the remaining phonological material not shared by all cells in the paradigm of a given lexeme corresponds to the exponent of that cell. To give an example, Table 2 shows the paradigm of the French verb *allumer (alyme)* from Table 1.[8] Comparing the forms of the different cells of the paradigm, we can analyse the string of *alym* as the stem, i.e. as the longest common substring between all forms of the paradigm. 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 analysed as zero forms.

Table 2: Part of the inflectional paradigm of French *alyme* 'turn on (light)'

cell	form	stem	marker
PRS.IND.1SG	alym	alym	-
PRS.IND.2SG	alym	alym	-
PRS.IND.3SG	alym	alym	-
PRS.IND.1PL	alymɔn	alym	-ɔn
PRS.IND.2PL	alymɛz	alym	-ɛz
PRS.IND.3PL	alym	alym	-
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.COND.2PL	alymɛrjɛz	alym	-ɛrjɛz
PRS.COND.3PL	alymɛraj	alym	-ɛraj
PRS.SUBJ.1SG	alym	alym	-
PRS.SUBJ.2SG	alym	alym	-
PRS.SUBJ.3SG	alym	alym	-
PRS.SUBJ.1PL	alymjɔn	alym	-jɔn
PRS.SUBJ.2PL	alymjɛz	alym	-jɛz
PRS.SUBJ.3PL	alym	alym	-

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'

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

[8]For reasons of space, Table 2 is restricted to the present tense forms.

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 analysed to consist of the three stem consonants only, i.e. as *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 are analysed as a part of the cells' exponents, respectively. Therefore, lexemes such as *Klos* in German do not have zero forms.

Table 3: Inflectional paradigms of three German nouns

	Klos 'dumpling'			*Abend* 'evening'			*Kreuz* 'cross'		
cell	form	stem	marker	form	stem	marker	form	stem	marker
NOM.SG	klos	kls	-o-	abənt	abən	-t	kroy͡ts	kroy͡ts	-
ACC.SG	klos	kls	-o-	abənt	abən	-t	kroy͡ts	kroy͡ts	-
DAT.SG	klos	kls	-o-	abənt	abən	-t	kroy͡ts	kroy͡ts	-
GEN.SG	kloses	kls	-o-es	abəndes	abən	-des	kroy͡tsəs	kroy͡ts	-es
NOM.PL	kløsə	kls	-ø-ə	abəndə	abən	-de	kroy͡tsə	kroy͡ts	-ə
ACC.PL	kløsə	kls	-ø-ə	abəndə	abən	-de	kroy͡tsə	kroy͡ts	-ə
DAT.PL	kløsən	kls	-ø-ən	abəndən	abən	-den	kroy͡tsən	kroy͡ts	-ən
GEN.PL	kløsə	kls	-ø-ə	abəndə	abən	-de	kroy͡tsə	kroy͡ts	-ə

Table 3 also shows that the Epitran transcription reflects the final devoicing of plosives in German. As can be seen for the second noun, *Abend (abənt)* 'evening', the segment that is usually analysed as the final consonant of the stem surfaces as a voiceless *-t* in all singular cells, where it occurs in a word-final position. If the consonant does not occur word-finally, i.e. in all plural cells, it surfaces as voiced *-d-*. Because this consonant is not identical across all cells of the paradigm, it is analysed as a part of the marker instead of the stem in the present approach. Hence, the lexeme of *Abend* does not have any zero form either.

The third noun given in Table 3, *Kreuz (kroy͡ts)* '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 material exponence of their specific morphosyntactic functions.

Besides umlauting and final devoicing, Table 4 shows how the automatic segmentation into stems and markers 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.[9] Again, this leads to a situation where the stem does not include the segment undergoing metathesis; only the string *ɟomo* is analysed as the stem. This in turn leads to the nominative singular cell having the marker *-r*; usually, the nominative singular does not receive any morphological marking in Hungarian,

[9]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 marker *-ig* does not cause metathesis.

as can be seen in the second example in Table 4.[10] The noun *gép (ge:p)* '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 analysed as a zero form for the purposes of the present study.

Table 4: Inflectional paradigm of two Hungarian nouns

	gyomor 'stomach'			*gép* 'machine'		
cell	form	stem	marker	form	stem	marker
NOM.SG	ɟomor	ɟomo	-r	ge:p	ge:p	-
ACC.SG	ɟomrot	ɟomo	-r-t	ge:pɛt	ge:p	-ɛt
DAT.SG	ɟomornɒk	ɟomo	-rnɒk	ge:pnɛk	ge:p	-nɛk
INSTR.SG	ɟomorrɒl	ɟomo	-rrɒl	ge:ppɛl	ge:p	-pɛl
PRP.SG	ɟomore:rt	ɟomo	-re:rt	ge:pe:rt	ge:p	-e:rt
TRANSL.SG	ɟomorra:	ɟomo	-rra:	ge:ppe:	ge:p	-pe:
TERM.SG	ɟomorig	ɟomo	-rig	ge:pig	ge:p	-ig
FRML.SG	ɟomorke:nt	ɟomo	-rke:nt	ge:pke:nt	ge:p	-ke:nt
IN+ESS.SG	ɟomorbɒn	ɟomo	-rbɒn	ge:pbɛn	ge:p	-bɛn
ON+ESS.SG	ɟomron	ɟomo	-r-on	ge:pɛn	ge:p	-ɛn
AT+ESS.SG	ɟomorna:l	ɟomo	-rna:l	ge:pne:l	ge:p	-ne:l
IN+ALL.SG	ɟomorbɒ	ɟomo	-rbɒ	ge:pbɛ	ge:p	-bɛ
ON+ALL.SG	ɟomorrɒ	ɟomo	-rrɒ	ge:prɛ	ge:p	-rɛ
AT+ALL.SG	ɟomorhoz	ɟomo	-rhoz	ge:phɛz	ge:p	-hɛz
IN+ABL.SG	ɟomorbo:l	ɟomo	-rbo:l	ge:pbø:l	ge:p	-bø:l
ON+ABL.SG	ɟomorro:l	ɟomo	-rro:l	ge:prø:l	ge:p	-rø:l
AT+ABL.SG	ɟomorto:l	ɟomo	-rto:l	ge:ptø:l	ge:p	-tø:l

As in the examples from French, German and Hungarian, I automatically segmented all forms of the dataset into stems and markers. Whenever the form of a cell of a given lexeme corresponded to the stem, I analysed it as a zero form because of the absence of an additional exponent for that cell.

There is one more issue that needs to be mentioned regarding the segmentation into stems and markers and the extraction of zero forms. The automatic segmentation of the whole dataset resulted in 305,276 different markers (by type of cell). Out of those, more than 50% of the markers, namely 155,407, occurred only once in the entire dataset. I excluded those markers, which resulted in the total of 149,869 markers for further analysis. Excluding markers with single occurrences was important in order to deal with irregular and suppletive forms. 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 marker ends up with all the remaining material (which would usually be analysed as being part of an irregular stem), which means that the marker 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.

[10]For reasons of brevity, Table 4 only shows the singular forms.

Table 5: Inflectional paradigms of three English verbs

	know			*think*			*heal*		
cell	form	stem	marker	form	stem	marker	form	stem	marker
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

In total, the final dataset contains 149,869 different markers (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

The advantage of the current approach is that different patterns across lexemes belonging to different inflection classes can be included and quantified. This section will briefly show two examples of zero forms in inflection paradigms of the nominal domain (Faroese and Aymara), and two of the verbal domain (Georgian and Tlatepuzco Chinantec).

4.4.1 Faroese

Nouns in Faroese (Germanic, Faroese Islands) show zero forms across a number of cells of their paradigms. They can express NOM.SG, ACC.SG, DAT.SG, GEN.SG, NOM.PL and ACC.PL cells. Importantly, these cells are not necessarily expressed by a zero form; Table 6 shows the proportions of zero forms for each of these cells based on the inflection paradigms of 2136 nouns in total. As we can see, there are substantial differences regarding the proportions of zero forms. For instance, the accusative singular indefinite cell is expressed by a zero form in almost 50% of the lexemes in the dataset. The genitive singular indefinite cell can also be zero marked. This, however, is only rarely the case, namely in 4% of the nouns. Such a token-based approach thus allows for a much more fine-grained analysis of the distribution of zero forms, as it can quantify to what extent zero forms occur in each cell.

Table 6: Proportions of zero forms in Faroese noun inflections

cell	N zero	prop zero
NOM.SG.INDF	662	0.31
ACC.SG.INDF	996	0.47
DAT.SG.INDF	329	0.15
GEN.SG.INDF	90	0.04
NOM.PL.INDF	246	0.12
ACC.PL.INDF	246	0.12

Table 7 shows five examples of nouns belonging to different inflection classes to illustrate different patterns in terms of zero forms.

Table 7: Inflectional paradigms of five Faroese nouns

cell	*arbeiði* 'work'	*álvur* 'elf'	*havskip* 'seagoing ship'	*ísur* 'ice'	*hugsan* 'thought'
NOM.SG.INDF	**arbeið**i	**álv**ur	**havskip**	ísur	**hugsan**
ACC.SG.INDF	**arbeið**i	**álv**	**havskip**	**ís**	**hugsan**
DAT.SG.INDF	**arbeið**i	**álv**i	**havskip**i	**ís**i	**hugsan**
GEN.SG.INDF	**arbeið**is	**álv**s	**havskip**s	**ís**	**hugsan**ar
NOM.SG.DEF	**arbeið**ið	**álv**urin	**havskip**ið	**ís**urin	**hugsan**in
ACC.SG.DEF	**arbeið**ið	**álv**in	**havskip**ið	**ís**in	**hugsan**ina
DAT.SG.DEF	**arbeið**inum	**álv**inum	**havskip**inum	**ís**inum	**hugsan**ini
GEN.SG.DEF	**arbeið**isins	**álv**sins	**havskip**sins	**ís**ins	**hugsan**arinnar
NOM.PL.INDF	**arbeið**ir	**álv**ar	**havskip**	**ís**ar	**hugsan**ir
ACC.PL.INDF	**arbeið**ir	**álv**ar	**havskip**	**ís**ar	**hugsan**ir
DAT.PL.INDF	**arbeið**um	**álv**um	**havskip**um	**ís**um	**hugsan**um
GEN.PL.INDF	**arbeið**a	**álv**a	**havskip**a	**ís**a	**hugsan**a
NOM.PL.DEF	**arbeið**ini	**álv**arnir	**havskip**ini	**ís**arnir	**hugsan**irnar
ACC.PL.DEF	**arbeið**ini	**álv**arnar	**havskip**ini	**ís**arnar	**hugsan**irnar
DAT.PL.DEF	**arbeið**unum	**álv**unum	**havskip**unum	**ís**unum	**hugsan**unum
GEN.PL.DEF	**arbeið**anna	**álv**anna	**havskip**anna	**ís**anna	**hugsan**anna

The stem of the first noun, *arbeiði* 'work', is *arbeið*, which does not correspond to any inflected form. Therefore, this noun cannot be expressed by a zero form. This is different in the case of the other nouns shown in Table 7. The cells that are expressed by zero forms are shaded in gray. The stem of the noun *álvur* 'elf' is *álv*, which coincides with the form of the accusative singular indefinite cell. For *havskip* 'seagoing ship', zero forms occur in the nominative and accusative indefinite cells in both singular and plural number. The next noun, *ísur* 'ice' is an example of nouns that use a zero form in the genitive singular indefinite cell (besides the accusative singular indefinite). Finally, the noun *hugsan* 'thought' shows a zero form in the nominative, accusative and dative singular indefinite cells.

4.4.2 Aymara

Aymara (Aymaran, Argentina, Bolivia, Chile, Peru) is a language with nominal inflection known for its subtractive morphology. The accusative singular cell is usually analysed as being expressed by the subtraction of the final vowel of the nominative singular form (cf. Coler 2015). Table 8 illustrates this with the paradigms of three Aymara nouns; for simplicity, only the singular and non-possessive forms are shown.

Table 8: Inflectional paradigm of 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

As can be seen in Table 8, 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 analysed 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 the lexeme (Baerman, Brown & Corbett 2017; Coler 2015, 2018). 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.

4.4.3 Georgian

Another rather unusual case of zero markers 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. Anderson 1992; Blevins 2016; Stolz & Levkovych 2019). 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 9 shows four examples of verb paradigms in

Georgian.[11]

Table 9: Parts of the inflectional paradigm of 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	va**k'et**eb
PRS.2SG	**t'ex**	**k'vec**	**cn**ob	a**k'et**eb
PRS.3SG	**t'exs**	**k'vecs**	**cn**obs	a**k'et**ebs
PRS.1PL	v**t'ex**t	v**k'vec**t	v**cn**obt	va**k'et**ebt
PRS.2PL	**t'ex**t	**k'vec**t	**cn**obt	a**k'et**ebt
PRS.3PL	**t'ex**en	**k'vec**en	**cn**oben	a**k'et**eben
IMPERF.1SG	v**t'ex**di	v**k'vec**di	v**cn**obdi	va**k'et**ebdi
IMPERF.2SG	**t'ex**di	**k'vec**di	**cn**obdi	a**k'et**ebdi
IMPERF.3SG	**t'ex**da	**k'vec**da	**cn**obda	a**k'et**ebda
IMPERF.1PL	v**t'ex**dit	v**k'vec**dit	v**cn**obdit	va**k'et**ebdit
IMPERF.2PL	**t'ex**dit	**k'vec**dit	**cn**obdit	a**k'et**ebdit
IMPERF.3PL	**t'ex**dnen	**k'vec**dnen	**cn**obden	a**k'et**ebdnen
FUT.1SG	gav**t'ex**	šev**k'vec**	gav**cn**ob	gava**k'et**eb
FUT.2SG	ga**t'ex**	še**k'vec**	ga**cn**ob	gaa**k'et**eb
FUT.3SG	ga**t'ex**s	še**k'vec**s	ga**cn**obs	gaa**k'et**ebs
FUT.1PL	gav**t'ex**t	šev**k'vec**t	gava**cn**obt	gava**k'et**ebt
FUT.2PL	ga**t'ex**t	še**k'vec**t	ga**cn**obt	gaa**k'et**ebt
FUT.3PL	ga**t'ex**en	še**k'vec**en	ga**cn**oben	gaa**k'et**eben
AOR.1SG	gav**t'ex**e	šev**k'vec**e	gav**cn**e	gava**k'et**e
AOR.2SG	ga**t'ex**e	še**k'vec**e	ga**cn**e	gaa**k'et**e
AOR.3SG	ga**t'ex**a	še**k'vec**a	ga**cn**a	gaa**k'et**a
AOR.1PL	gav**t'ex**et	šev**k'vec**et	gav**cn**et	gava**k'et**et
AOR.2PL	ga**t'ex**et	še**k'vec**et	ga**cn**et	gaa**k'et**et
AOR.3PL	ga**t'ex**es	še**k'vec**es	ga**cn**es	gaa**k'et**es

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 9, 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 9). In most TAM series, many Georgian verbs also have so-called thematic suffixes (Hewitt 1995: 143-147), as 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 9, *t'exs* 'break' and *kv'ecs* 'cut off'.

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

4.4.4 Tlatepuzco Chinantec

Another language with a noteworthy verbal inflectional paradigm is Tlatepuzco Chinantec (Otomanguean, Mexico). Tlatepuzco Chinantec has a complex inflectional paradigm because it combines various patterns of stem and tone changes.[12] Yet, some inflection classes involve neither stem nor tone changes. This leads to zero forms in the irrealis or future cells of the paradigm, which is typologically uncommon (cf. Bybee, Perkins & Pagliuca 1994). Table 10 shows the inflectional paradigms of three verbs. The first two verbs *køgʔ²* 'eat' and *ʔlúg²* 'heal' both have stem changes across and within present, past, and future tense (irrealis) forms. The forms of *køgʔ²* 'eat' have different tones for first vs. second and third person forms in all three tenses. The forms of *ʔlúg²* 'heal' all share the tone pattern but have distinct segmental realizations across tenses. The third verb, *ʔian¹²* 'finish', only has the additional segmental marker *mi³-* for the past tense forms. Because the paradigm does not include any other tonal or segmental changes, both present and future forms correspond to the stem, i.e. the longest common substring of all forms. This leads to the typologically uncommon situation that the verb *ʔian¹²* 'finish' uses a zero form to express future.

Table 10: Inflectional paradigm of three Tlatepuzco Chinantec verbs

cell	*køgʔ²* 'eat'	*ʔlúg²* 'heal'	*ʔian¹²* 'finish'
PRS.1SG	køgʔ¹²	ʔlug²	ʔian¹²
PRS.1PL	køgʔ¹²	ʔlug²	ʔian¹²
PRS.2	køgʔ²	ʔlug²	ʔian¹²
PRS.3	køgʔ²	ʔlúg²	ʔian¹²
PST.1SG	mi³-køgʔ¹²	mi³-ʔlug²	mi³-ʔian¹²
PST.1PL	mi³-køgʔ¹²	mi³-ʔlug²	mi³-ʔian¹²
PST.2	mi³-køgʔ²	mi³-ʔlug²	mi³-ʔian¹²
PST.3	mi³-køgʔ²	mi³-ʔlug²	mi³-ʔian¹²
FUT.1SG	køgʔ¹³	ʔliug²	ʔian¹²
FUT.1PL	køgʔ¹³	ʔliug²	ʔian¹²
FUT.2	køgʔ³	ʔliug²	ʔian¹²
FUT.3	køgʔ¹	ʔliug²	ʔian¹²

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.

[12] For more details and an analysis of the verbal inflectional system, see the Otomanguean Inflection database (`https://oto-manguean.surrey.ac.uk/Info/CPA`).

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 extremely rarely 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.[13] 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.

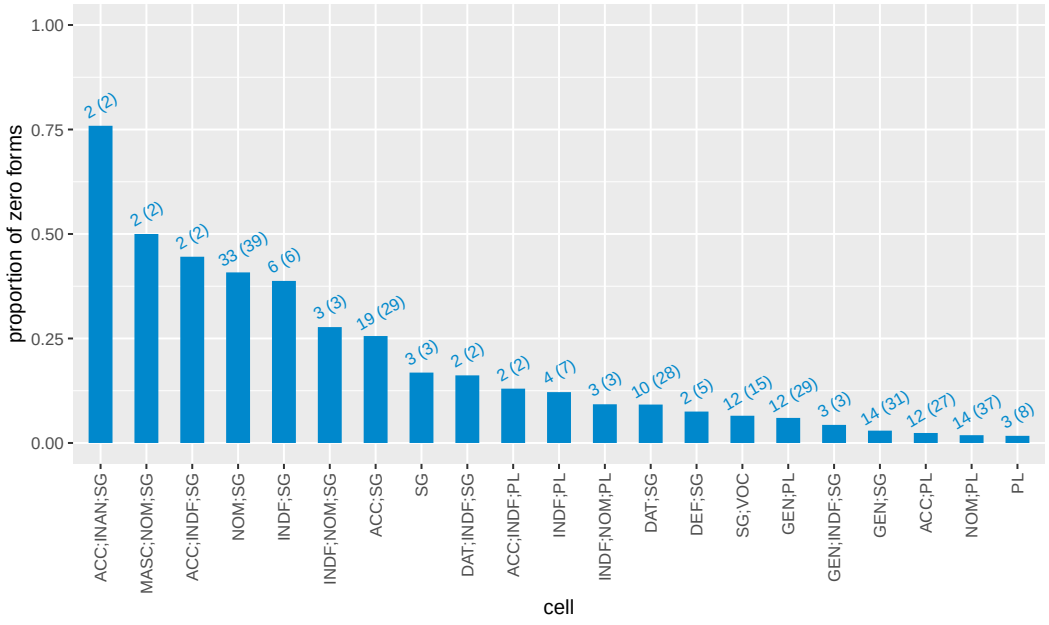

Figure 2: Nominal cells with highest proportions of zero forms

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 (Russian and Czech). The other cells with

[13] Table 13 in Appendix A provides the exact occurrences and proportions for the cells shown in Figure 2.

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 token-based approach is applied. This means that the common (implicit) assumption in the typological literature that zero forms are the default for certain morphosyntactic values (e.g. Bybee 2011; Croft 2003: Ch. 4; Diessel 2019: 224-228; Greenberg 1966: 32-37; Haspelmath 2021, Song 2018: Ch. 7) is not tenable. The first reason for this is that different values are usually not encoded in isolation; the second reason is that once we consider the expression of cells, i.e. value combinations, in single lexemes, the variation across lexemes introduces a higher level of complexity and leads to overall much less strong tendencies.

However, the proportions in Figure 2 may be 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 & 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 in-

[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.

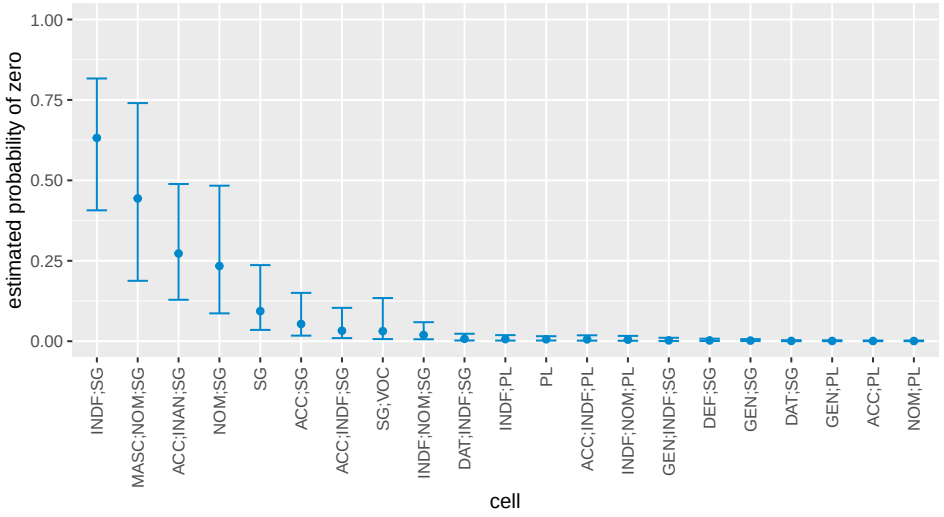

Figure 3: Conditional effects for the nominal cells with the highest proportions of zero forms

definite 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 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 nominative 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 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

[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).

[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.

0.03. They can thus hardly be viewed as being prone to zero marking, even though zero forms are occasionally used to express those cells in different 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 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 detect forms in which one could argue that the singular value is zero-coded while e.g. the dative case value of the same cell is 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.

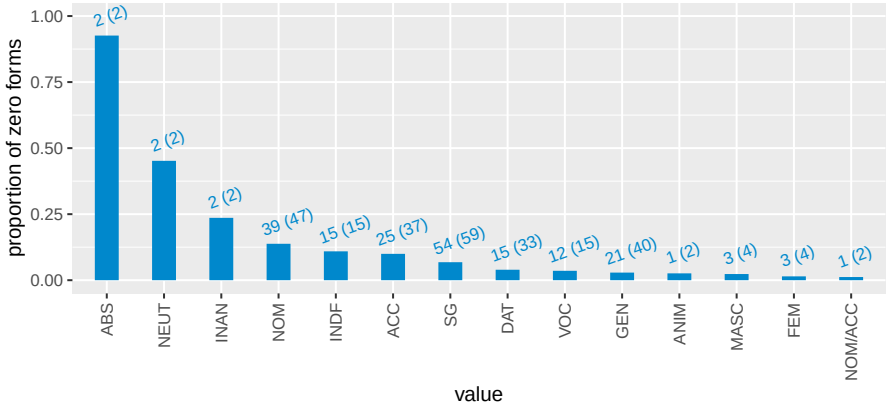

Figure 4: Nominal values with highest proportions of zero forms

[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.

The distributions in Figure 4 confirm some of the tendencies seen in the previous section but also provide new insights. Again, we find the values of inanimate, indefinite, nominative, accusative, singular, dative, and vocative amongst 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.

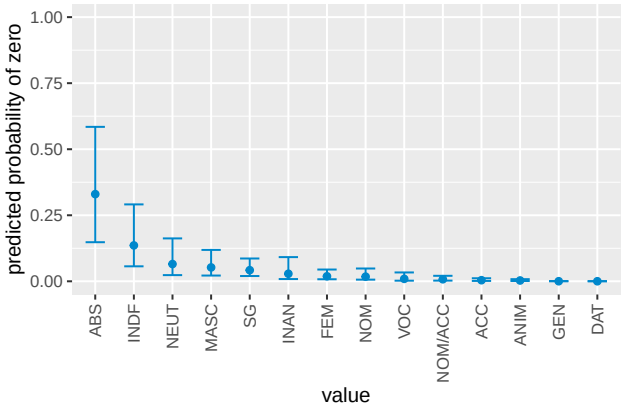

Figure 5: Conditional effects for the nominal values with the highest proportions of zero forms

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 0.10 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 a morphological singular marker in the traditional sense, but rather that they do not have surface forms in their nominal paradigms that correspond to the longest common substring of all forms, which results in the analysis of all cells having a dedicated marker, including cells with the singular value.

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 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 ocasionally in different languages but they clearly have no strong association with zero forms.

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.[22]

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

[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).

[22] Table 14 in Appendix B provides the exact occurrences and proportions for the cells shown in Figure 6.

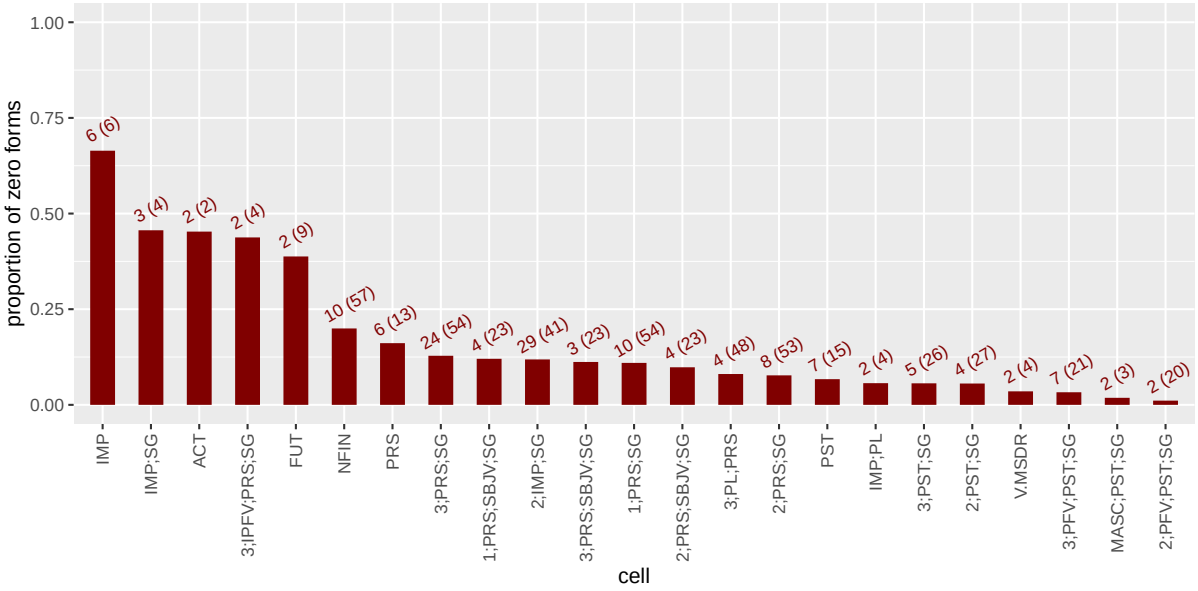

Figure 6: Verbal cells with the highest proportions of zero forms

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, Perkins & Pagliuca 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, Hili-

gaynon 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 the base or the stem of other verb forms.

Another value that appears in a number of cells in Figure 6 is the present tense (PRS), i.e. 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 section, 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.[23] The conditional effects for all 23 cells are shown in Figure 7.

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

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

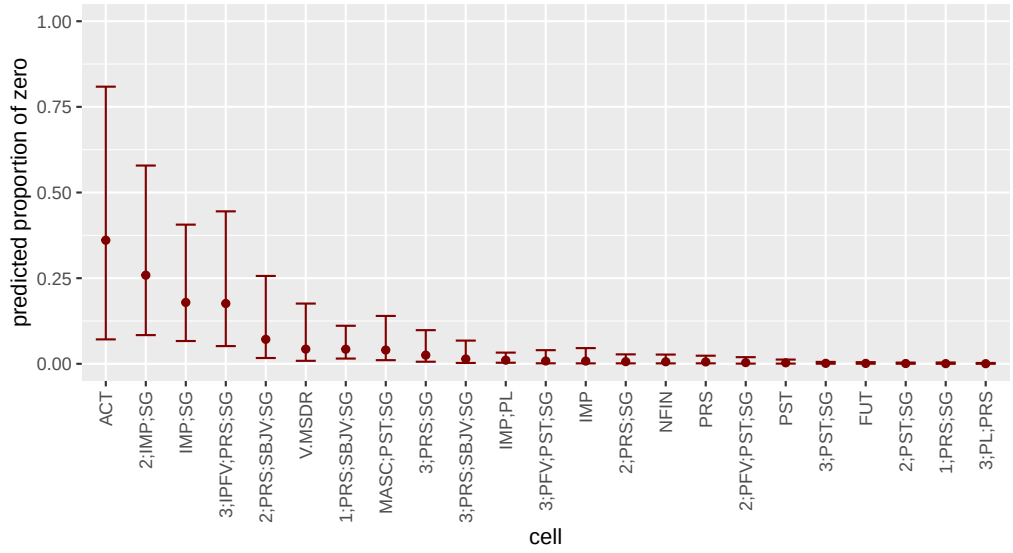

Figure 7: Conditional effects for the verbal cells with the highest proportions of zero forms

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.

The only other cell with an estimated probability of zero forms above 0.1 is the third person singular imperfective present (3.IPFV.PRS.SG, 0.18) cell, also with a very large uncertainty interval. This indicates that third person present tense forms are comparatively more likely to be zero marked than most other cells of verbal inflection paradigms. With the current approach and dataset, it is difficult to say that this generally holds for present tense forms, as has been suggested in the typological literature (Bybee & Dahl 1989: 55; Bybee 1994: 248), simply because present tense values usually occur in combination with other values in the verbal paradigms. Thus, it may be the combination of third person and present tense (and 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 for nouns, 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.

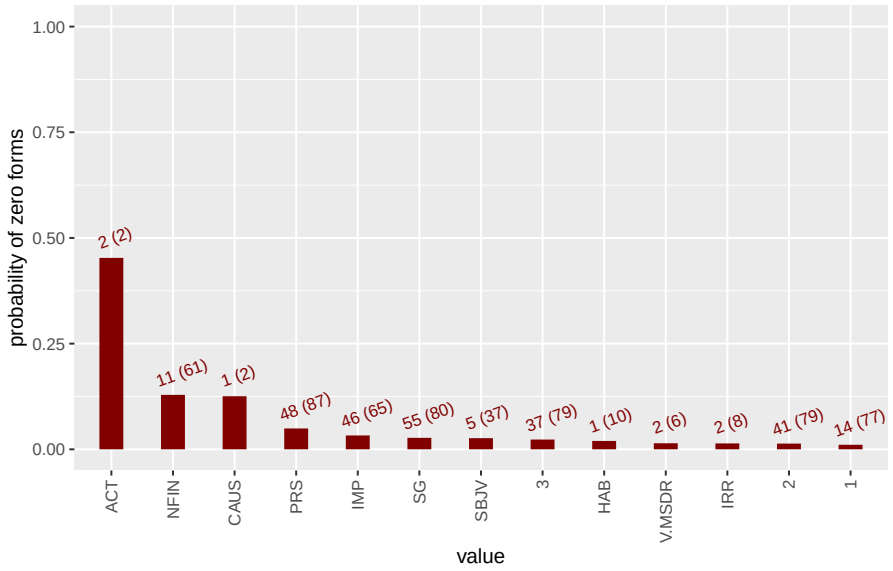

Figure 8: Verbal cells with the highest proportions of zero forms

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.

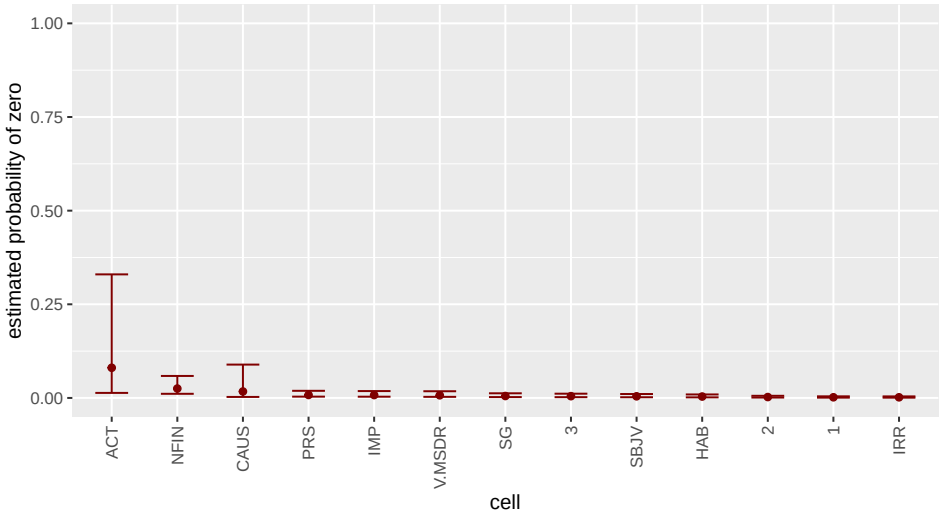

Figure 9: Conditional effects for the verbal values with the highest proportions of zero forms

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 in general, no value of verbal paradigms shows a very strong association with zero forms, let alone a preference for being expressed by a zero form.

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 these 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.[24]

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 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.

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

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, Gelman & Gabry (2017). Appendix C describes the different models and the process of model comparison and selection in more detail.[25] 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.

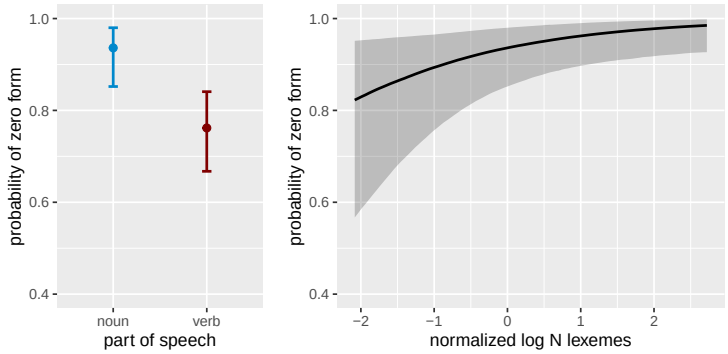

Figure 10: Conditional effects for parts of speech and number of lexemes

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.[26] The left plot in Figure 10 shows the effects of the part of speech. The points represent the means of the posterior 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 very wide, we can expect to find slightly more zero forms if more data is available.

[25] See the file "code.R" in the supplementary materials for more information on the model specifications.

[26] 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 smallest 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.

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. I will briefly sketch two factors that potentially play a role in the higher probability of zero forms in nominal over verbal inflectional paradigms.

The first factor is the place of inflection marking. Already Bybee, Pagliuca & Perkins (1990) have shown 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) shows 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 three domains, at least 80% of all markers in his dataset are suffixes (Cysouw 2009: 2). He further notes that the suffixation preference is rather weak for person marking on the verb. Moreover, distinguishing between person marking paradigms based on the number of markers, Cysouw (2009: 3) shows that the suffixation preference only holds for systems with a larger number of markers (six or more). For systems with fewer markers, his data shows 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, Pagliuca & Perkins 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; Cho, McQueen & Cox 2007; Keating et al. 2003; Kim 2004; Smith 2005). Especially word-initial consonants tend to be strengthened and lengthened (e.g. Cho & Keating 2001, 2009; Fougeron 2001; White, Benavides-Varela & Mády 2020), which is relevant here, since Bybee, Pagliuca & Perkins (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 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, Pagliuca & Perkins (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 equally account for the higher proportion of zero markers in the nominal domain.

Another factor that may be relevant for explaining the difference in zero marking between the nominal and the verbal domain is the proportion of loan words. Crosslinguistic studies have shown that, independently of their morphological properties, languages borrow proportionally more nouns than verbs (Tadmor & Haspelmath 2009: 61). Loan words may not take inflectional morphology, and given how zero forms were detected in the present study, the forms of lexemes lacking inflection were analysed as zero forms, since they correspond to the longest common substring of all cells of the paradigm. Thus, a higher proportion of borrowed lexemes that do not exhibit inflectional marking could have resulted in a slightly higher number of zero forms in nominal inflection paradigms. For the data used in the present study, there is no good way to exclude such lexemes without careful, case-by-case manual checks. However, this effect, if present, should be very weak. A lexeme being borrowed from another language does not necessarily lead to the absence of inflection morphology, and we can assume that the number of borrowed lexemes in the UniMorph database is low to begin with.

Related to the borrowability of nouns and verbs, Arkhangelskiy (2020) makes an important point that could also be relevant for the different probabilities of zero forms in nominal and verbal paradigms. He shows that the higher rate of nominal compared to verbal borrowings may in fact be due to a generally longer life span of verbs. In other words, Arkhangelskiy (2020) shows that borrowed verbs survive longer in a language than borrowed nouns after the contact between languages is disrupted. If and how this could influence the development and persistence of zero forms would have to be tested in a separate study. Nevertheless, those results point to more substantial differences between nouns and verbs, which could in turn account for the observed differences in zero forms.

6.2 Coding efficiency does not explain the distribution of zero forms

As was introduced in Section 2, the occurrence and distribution of zero forms was related to coding efficiency and form-frequency effects in previous studies. To be precise, the grammatical form-frequency correspondence hypothesis (Haspelmath 2021: 2) predicts more frequent morphosyntactic functions to prefer zero coding, i.e. the use of zero forms. However, Section 2 also mentioned that the results from other previous studies (Bickel et al. 2015; Cysouw 2003; Guzmán Naranjo & Becker 2021a; Seržant & Moroz 2022; Siewierska 2010; Stolz & Levkovych 2019) point against coding efficiency and frequency as factors that shape the distribution of zero forms.

The findings from the present study are in line with coding efficiency playing a minor role, if at all, in the distribution of zero forms. While the present study did not examine frequency effects of zero forms directly, we saw that there is no morphosyntactic function for which zero forms would be a preferred or most likely coding strategy. Form-frequency effects develop through coding efficiency driving language change. If coding efficiency is supposed to account for the distribution of zero forms, we should find that they are systematically preferred in

highly frequent functions. However, as we saw in Sections 5.1 and 5.2, no nominal or verbal cells or values stood out as having such a crosslinguistic preference for zero forms.

In addition, 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 many lexemes of many 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 (we will return to this point in Section 6.3).

Taking together the lack of a clear preference for zero forms and their very inconsistent distribution, it appears very unlikely that coding efficiency would directly lead to zero forms in inflectional morphology. However, once we consider the morphosyntactic values that have a higher probability of being expressed by a zero form than the other morphosyntactic values, we find that at least some of them correspond to the values that already Greenberg (1966) showed to be the more or most frequent value of the morphosyntactic 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 give a rough impression of the distribution of the relevant morphosyntactic values in language use, Table 11 shows their token frequencies in the Universal dependencies treebanks (Zeman et al. 2021). Table 12 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

Table 11: Distribution of nominal features in the Universal Dependencies

value	N value	N total	N langs	prop value
number				
singular	3789088	5144632	80	0.7365129
plural	1347473	5152931	88	0.2614964
case				
nominative	887053	3166314	57	0.2801532
accusative	708587	2868514	46	0.2470223
absolutive	9801	32220	5	0.3041899
ergative	1984	22465	3	0.08782258
dative	234759	3182335	44	0.08831516
genitive	862330	3070696	56	0.2808256
vocative	6205	1849520	24	0.003354925
definiteness				
definite	384650	971411	17	0.3959704
indefinite	458271	957243	16	0.4787405

of occurrences difficult. Therefore, I also extracted the total number of lexemes in all languages with a given feature value, e.g. singular (on nouns), for which the relevant morphosyntactic feature, i.e. number, is annotated. The "N total" column shows that there are 5144632 nouns with a number annotation in the Universal dependencies. The last column of Tables 11 and 12 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 11 and 12 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.[27] As for the indefinite value, which also showed a comparatively high probability 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 12 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.

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

[27] Table 11 also includes numbers for the ergative, dative, genitive and vocative case values for reasons of comparison.

Table 12: Distribution of verbal features in the Universal Dependencies

value	N value	N total	N langs	prop value
tense				
past	739008	1666938	80	0.4433326
present	857398	1667680	77	0.5141262
future	47114	1049995	44	0.04487069
mood				
Indicative	1216937	1329281	75	0.9154851
Imperative	37822	1331878	77	0.0283975
person				
1	132245	1232404	78	0.1073065
2	68328	1232234	73	0.05545051
3	1021191	1231610	76	0.8291513
voice				
active	1032873	1227679	40	0.8413217
passive	144708	1011323	52	0.1430878

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. It is plausible that frequency plays an indirect role in the development of zero forms in such functions. As will be mentioned in Section 6.3, one possible process leading to the development of zero forms is the differential non-development of an overt exponent. In that case, an exponent for a different value of the same grammatical category develops, while the zero form develops in relation to that exponent, as it grammaticalizes. Frequency could play a role in this type of processes in that the functions for which exponents develop are likely to be less frequent, so that speakers start expressing those overtly for successful communication. The function that is eventually encoded by a zero form in opposition then is simply a function that was not sufficiently infrequent to motivate the development of an overt marker. Importantly, frequency does not play a direct role in such cases. Moreover, Table 12 shows that this explanation does not hold for imperative forms. Cells with the imperative value were shown to have a comparatively high probability of zero forms, but we see that they make up a much smaller proportion of verbs than e.g. indicative forms. We will return to a possible account for the development of zero forms with imperatives in Section 6.3. Crucially, we can conclude that there is no strong evidence for coding efficiency being involved in the development of zero forms in a direct way.

6.3 Zero forms as a diachronic by-product

Similarly to the argumentation by Cristofaro (2019, 2021), I take the results of the present study to suggest that the occurrence of zero forms is a by-product of other diachronic processes rather than an attractor state of its own. This means that using zero forms for certain

grammatical functions is not a preferred state as such that would motivate the development of zero forms. Rather, various other processes, which may be motivated by efficiency or other functional factors, can lead to developments in which zero markers form. Importantly, these processes can be independent of each other and they need not be coherent or caused by the same factors.

In the remainder of this section, I will give a brief overview of different processes that have been related to the development of zero marking in the literature: differential phonetic reduction, differential non-development, differential morphosyntactic reduction, and reanalysis (Bybee 1985, 1994; Cristofaro 2019, 2021; Haspelmath 2008b; Koch 1995). Although all processes have at least been mentioned in the previous literature, they have not yet been discussed together as different ways of leading to the development of zero forms. A more detailed analysis of those processes would go beyond the scope of the present paper. The purpose of discussing them here is to show that the development of zero forms is not a homogeneous process motivated by an efficient end-state that languages adapt to. Instead, we should understand the development and distribution of zero forms as a by-product of other, independent diachronic processes.

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 2008b). 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, she does not explicitly relate phonetic reduction to the creation of zero forms (i.e. the total reduction of phonetic material associated with a function), and none of her examples of phonetic reduction show cases in which this total reduction to a zero form would have taken place. We only find very few examples of phonetic reduction leading to zero forms in the literature. Haspelmath (2008b: 206) presents the following two examples: the third person agreement marker in Polish and the English singular marker of the singular noun *day* as opposed to its plural form *days* (from Old English *dæg* 'day.SG', *dagas* 'day.PL').[28] Yet, he 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" (Haspelmath 2008b: 207). Given that we find very few examples in the literature for phonetic reduction being involved in the creation of zero forms, 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 zero marking 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

[28]Based on the historical sources, it is not entirely clear whether phonetic reduction is responsible for the Polish pattern.

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 as an inflectional marker, 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 marker developing for another cell in the paradigm. Although convincing examples with diachronic data tracing the details of e.g. a developing plural marker (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).

Another commonly cited type of examples for differential non-development is the development of zero third person agreement markers on verbs (Bickel et al. 2015; Cristofaro 2021). A common source for person agreement markers are personal pronouns (Heine & Song 2011: 595). Third persons, however, are often either referred to by lexical expressions or left unexpressed if they are accessible enough in a given discourse context. Also, languages may not have dedicated free pronouns for the third person and use demonstratives if needed. In such a situation, already the use of the first and second person pronouns differs greatly from the use of elements referring to third persons. Such a situation can lead to the development of first and second person agreement markers from free pronouns, while no parallel development of a third person agreement marker takes place, given that there is no parallel source element. Examples of this have been described in the literature for, e.g., Tabasaran (Bogomolova 2018; Helmbrecht 1996) and Plains Cree (Mithun 1991).

Importantly, in both examples, i.e. the zero singular and the zero third person form, it is not the final state of having a zero-expressed value that drives the diachronic process. Rather, the developing zero form is simply 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 other values of that category, and with no overt exponent for the value in question.

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 amongst 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 (Sadock & Zwicky 1985: 173; Levshina 2018: 25; Aikhenvald 2010: 147; Nikolaeva 2007: 163).

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. However, more diachronic, language-specific work with such examples is needed to determine if this is really the case.

Imperatives offer another interesting piece of evidence for the question of how zero forms develop. Theoretically, zero forms can develop by differential morphosyntactic reduction. Differential morphosyntactic reduction means that the use of a morphosyntactic, e.g. imperative, marker is omitted in certain contexts, leading to the use of a zero form for a function that used to be overtly coded before. While Haspelmath (2008b: 210) argues that examples of this process are not attested, the optional omission of imperative markers with certain verbs or in certain discourse contexts may be an example of differential morphosyntactic reduction. For instance, Nikolaeva & Tolskaya (2011: 221-222) note for the imperative in Udihe (Tungusic) that in the "[s]ingular it is also possible to use the bare Present stem without any personal inflection [...] Such forms are particularly expressive and are used to give a categorical order." Again, a more detailed diachronic analysis would have to clarify if we really deal with a formerly used marker that becomes omitted under certain conditions, leading to the imperative function expressed by a zero form. Notwithstanding, such examples are highly relevant for understanding yet another type of processes that may lead to zero forms in inflectional morphology.

The last process that can lead to the development of zero forms in morphology is the reanalysis of a marker as part of the stem, resulting in the absence of exponence of a morphosyntactic feature (combination), i.e. a zero form. This phenomenon is well known from historical linguistics as "Watkin's law" (cf. Bybee 1985; Koch 1995; Watkins 1962). 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 marker is reanalysed as part of the stem. At the same time, the former marker is added to the other forms of the paradigm as well, restructuring the entire paradigm. We are far from understanding the details of such processes, e.g. what the frequencies of the various cells of the paradigms are. The main point here is again that if zero forms can develop from the reanalysis of a former marker as part of the stem as suggested in the literature, it is yet another piece of evidence for zero markers being a by-product of different, independent diachronic processes.

7 Conclusion

This paper offered a first quantitative crosslinguistic study of the occurrence and distribution of zero forms in nominal and verbal inflectional morphology. Because of its token-based approach, it could take into account the behavior of single lexemes and capture the variation across inflection classes and irregular forms. In addition, a more realistic picture of the distribution of zero forms emerged through the word and paradigm approach; the analysis of

exponence was based on forms as they occur in different cells of the paradigms, without the additional abstraction towards single values of morphosyntactic features that are commonly not expressed in isolation.

The results of this study showed that no cells, neither in nominal nor in verbal paradigms, have a strong association with 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. Nevertheless, the results clearly showed that we deal with a very high degree of variation across languages and lexemes in the distribution of zero forms. The findings of the present study therefore do not support the hypothesis of coding efficiency driving the development of zero forms. If that were the case, we would expect a stronger crosslinguistic preference towards zero forms and a more consistent pattern. Rather, the results are supporting evidence for the hypothesis that zero forms develop through many different, unrelated diachronic processes, which may be motivated for reasons of efficiency themselves.

References

Aikhenvald, Alexandra. 2010. *Imperatives and commands*. Oxford: Oxford University Press.

Anderson, Stephen R. 1992. *A-morphous morphology*. Cambridge: Cambridge University Press.

Arkhangelskiy, Timofey. 2020. Verbal borrowability and turnover rates. *Diachronica* 37(4). 451–473.

Baerman, Matthew, Dunstan Brown & Greville G. Corbett. 2017. *Morphological complexity*. Cambridge: Cambridge University Press.

Beckman, Jill. 1998. *Positional faithfulness*. Amherst: University of Massachusetts dissertation.

Beniamine, Sacha & Matías Guzmán Naranjo. 2021. Multiple alignments of inflectional paradigms. *Proceedings of the Society for Computation in Linguistics* 4. Article 21.

Bickel, Balthasar et al. 2015. Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories. In Jürg Fleischer, Elisabeth Rieken & Paul Widmer (eds.), *Agreement from a diachronic perspective*, 29–52. Berlin: De Gruyter.

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

Bogomolova, Natalia. 2018. The rise of person agreement in East Lezgic: Assessing the role of frequency. *Linguistics* 56(4). 819–844.

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

Bybee, Joan. 1985. *Morphology. A Study of the Relation of Meaning and Form*. Vol. 9 (Typological Studies in Language). Amsterdam and Philadelphia: John Benjamins.

Bybee, Joan. 1994. The grammaticization of zero: Asymmetries in tense and aspect systems. In William Pagliuca (ed.), *Perspectives on grammaticalization*, 235–254. Amsterdam: Benjamins.

Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In Brian Joseph & Richard Janda (eds.), *Handbook of historical linguistics*, 602–623. Oxford: Blackwell.

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

Bybee, Joan. 2011. Markedness. In Jae Jung Song (ed.), *The Oxford handbook of typology*, 1–11. Oxford: Oxford University Press.

Bybee, Joan. 2015. *Language change*. Cambridge: Cambridge University Press.

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

Bybee, Joan, William Pagliuca & Revere Perkins. 1990. On the asymmetries in the affixation of grammatical material. In William Croft, Suzanne Kemmer & Keith Denning (eds.), *Studies in Typology and Diachrony. Papers presented to Joseph H. Greenberg on his 75th birthday*, 1–42. Amsterdam: Benjamins.

Bybee, Joan, Revere Perkins & William Pagliuca. 1994. *The evolution of grammar. Tense, aspect, and modality in the languages of the world*. Chicago: The University of Chicago Press.

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

Cho, Taehong & Patricia Keating. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics* 29(2). 155–190.

Cho, Taehong & Patricia Keating. 2009. Effects of initial position versus prominence in English. *Journal of Phonetics* 37(4). 466–485.

Cho, Taehong, James McQueen & Ethan Cox. 2007. Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics* 35. 210–243.

Coler, Matt. 2015. Aymara inflection. In Matthew Baerman (ed.), *The Oxford Handbook of Inflection*, 1–30. Oxford: Oxford University Press.

Coler, Matt. 2018. *Subtractive Morphology & Disfixation in Aymara Case*. Groningen.

Cristofaro, Sonia. 2019. Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals. In Karsten Schmidtke-Bode et al. (eds.), *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence* (Conceptual Foundations of Language Science), 25–46. Berlin: Language Science Press.

Cristofaro, Sonia. 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.

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

Cysouw, Michael. 2003. *The paradigmatic structure of person marking*. Oxford: Oxford University Press.

Cysouw, Michael. 2009. The asymmetry of affixation. *Snippets* 20. 10–144.

Dahl, Östen. 1985. *Tense and aspect systems*. Oxford: Blackwell.

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

Fougeron, Cécile. 2001. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics* 29(2). 109–135.

Garcia, Erica C. & Florimon van Putte. 1989. Forms are silver, nothing is gold. *Folia Linguistica Historica* 21. 365–384.

Givón, Talmy. 2018. *On understanding grammar*. Amsterdam: Benjamins.

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

Guzmán Naranjo, Matías & Laura Becker. 2021a. Coding efficiency in nominal inflection: expectedness and type frequency effects. *Linguistics Vanguard* 7(s3).

Guzmán Naranjo, Matías & Laura Becker. 2021b. Statistical bias control in typology. *Linguistic Typology*.

Hall, Christopher. 1988. Integrating diachronic and processing principles in explaining the suffixing preference. In John Hawkins (ed.), *Explaining language universals*, 321–349. London: Basil Blackwell.

Hammarström, Harald et al. 2021. *Glottolog 4.4*. Leipzig: Max Planck Institute for the Science of Human History.

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

Haspelmath, Martin. 2008b. Creating economical morphosyntactic patterns in language change. In Jeff Good (ed.), *Linguistic universals and language change*, 185–214. Oxford: Oxford University Press.

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

Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*. 1–29.

Haspelmath, Martin & Andres Karjus. 2017. Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency. *Linguistics* 55(6). 1213–1235.

Haspelmath, Martin et al. 2014. Coding causal–noncausal verb alternations: A form–frequency correspondence explanation. *Journal of Linguistics* 50(3). 587–625.

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

Heine, Bernd & Kyung-An Song. 2011. On the grammaticalization of personal pronouns. *Journal of Linguistics* 47(3). 587–630.

Helmbrecht, Johannes. 1996. The syntax of personal agreement in East Caucasian languages. *STUF - Language Typology and Universals* 49(2). 127–148.

Hewitt, George. 1995. *Georgian: A structural reference grammar*. Amsterdam: Benjamins.

Jakobson, Roman. [1939] 1983. Zero sign. In Linda Waugh & Morris Halle (eds.), *Russian and Slavic grammar: Studies 1931-1981*, 1–14. New York: De Gruyter.

Keating, Patricia et al. 2003. Domain-initial strengthening in four languages. *Phonetic Interpretation: Papers in Laboratory Phonology* 6.

Kim, Sahyang. 2004. *The role of prosodic phrasing in Korean word segmentation*. University of California, Los Angeles dissertation.

Koch, Harold. 1995. The creation of morphological zeros. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1994*, 31–731. Springer.

Lemaréchal, Alain. 1997. *Zéro(s)*. Paris: Presses Universitaires de France.

Levshina, Natalia. 2018. *Towards a theory of communicative efficiency in human languages*. Leipzig: Leipzig University Habilitationsschrift.

Matthews, P.H. 1972. *Inflectional Morphology: A Theoretical Study Based on Aspects of Latin Verb Conjugation*. Cambridge University Press.

McCarthy, Arya D. et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 3922–3931. Marseille, France: European Language Resources Association.

Mel'cuk, Igor. 2002. Towards a formal concept zero linguistic sign: Applications in typology. In Sabrina Bendjaballah et al. (eds.), *Morphology 2000: Selected papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000*, 241–258. Amsterdam: Benjamins.

Mithun, Marianne. 1991. The development of bound pronominal paradigms. In Winfred Lehmann & Helen-Jo Jakusz Hewitt (eds.), *Language typology 1988: Typological models in the service of reconstruction*, 85–104. Amsterdam: Benjamins.

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

Nikolaeva, Irina. 2007. *Finiteness: Theoretical and empirical foundations*. Oxford: Oxford University Press.

Nikolaeva, Irina & Maria Tolskaya. 2011. *A Grammar of Udihe*. De Gruyter Mouton.

Pullum, Geoffrey & 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*. Manual. R Foundation for Statistical Computing. Vienna, Austria.

Sadock, Jerrold & Arnold Zwicky. 1985. Speech act distinctions in grammar. In Timothy Shopen (ed.), *Language typology and syntactic description. Volume 1*, 155–196. Cambridge: Cambridge University Press.

Seržant, Ilja A. & 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.

Siewierska, Anna. 2010. Person asymmetries in zero expression and grammatical functions. In Franck Floricic (ed.), *Essais de linguistique generale et de typologie linguistique offerts au professeur Denis Creissels à l'occasion de ses 65 ans*, 425–438. Paris: Presses de l'École Normale Supérieure.

Smith, Jennifer. 2005. *Phonological augmentation in prominent positions*. Oxfordshire: Taylor & Francis.

Song, Jae Jung. 2018. *Linguistic typology*. Oxford: Oxford University Press.

Stave, Matthew et al. 2021. Optimization of morpheme length: a cross-linguistic assessment of Zipf's and Menzerath's laws. *Linguistics Vanguard* 7(s3).

Stolz, Thomas & Nataliya Levkovych. 2019. Absence of material exponence. *Language Typology and Universals* 72(3). 373–400.

Stump, Gregory. 2001. *Inflectional morphology: A theory of paradigm structure*. Cambridge: Cambridge University Press.

Tadmor, Uri & Martin Haspelmath (eds.). 2009. *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton.

Vehtari, Aki, Andrew Gelman & Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5). 1413–1432.

Watkins, Calvert. 1962. *Indo-European origins of the Celtic verb*. Dublin: Dublin Institute for Advanced Studies.

White, Laurence, Silvia Benavides-Varela & Katalin Mády. 2020. Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics* 81. 100982.

Ye, Jingting. 2020. Independent and dependent possessive person forms: Three universals. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 44(2). 363–406.

Zeman, Daniel et al. 2021. *Universal Dependencies 2.9*.

Zipf, George Kingsley. 1935. *The psychobiology of language: An introduction to dynamic philology*. Cambridge, MA: MIT Press.

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort*. Cambridge, MA: Addison-Wesley Press.

Zwicky, Arnold. 1985. How to describe inflection. In Mary Niepokuj et al. (eds.), *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, 372–386. Berkeley, CA: Berkeley Linguistics Society.

Appendix A

Table 13: Nominal cells with higher proportions of zero forms (≥ 0.01) in at least 2 languages

	cell	N zero forms	N lexemes	prop zero	N lang
1	ACC;INAN;SG	107	141	0.76	2 (2)
2	MASC;NOM;SG	49	98	0.50	2 (2)
3	ACC;INDF;SG	2788	6257	0.45	2 (2)
4	NOM;SG	57640	141242	0.41	33 (39)
5	INDF;SG	3264	8418	0.39	6 (6)
6	INDF;NOM;SG	3673	13248	0.28	3 (3)
7	ACC;SG	32922	128648	0.26	19 (29)
8	SG	292	1734	0.17	3 (3)
9	DAT;INDF;SG	1012	6257	0.16	2 (2)
10	ACC;INDF;PL	692	5324	0.13	2 (2)
11	INDF;PL	1787	14686	0.12	4 (7)
12	INDF;NOM;PL	981	10593	0.09	3 (3)
13	DAT;SG	8526	92742	0.09	10 (28)
14	DEF;SG	1041	13849	0.08	2 (5)
15	SG;VOC	2950	45281	0.07	12 (15)
16	GEN;PL	6760	112713	0.06	12 (29)
17	GEN;INDF;SG	572	13228	0.04	3 (3)
18	GEN;SG	3527	119516	0.03	14 (31)
19	ACC;PL	2829	118357	0.02	12 (27)
20	NOM;PL	2462	132140	0.02	14 (37)
21	PL	67	3918	0.02	3 (8)

Appendix B

Table 14: Verbal cells with highest proportions of zero forms (≥ 0.01) in at least 2 languages

	cell	N zero	N lexemes	prop zero	N lang
1	IMP	2455	3696	0.66	6 (6)
2	IMP;SG	1305	2860	0.46	3 (4)
3	ACT	960	2120	0.45	2 (2)
4	3;IPFV;PRS;SG	1865	4262	0.44	2 (4)
5	FUT	809	2086	0.39	2 (9)
6	NFIN	16009	80178	0.20	10 (57)
7	PRS	816	5061	0.16	6 (13)
8	3;PRS;SG	9906	77204	0.13	24 (53)
9	1;PRS;SBJV;SG	4747	39439	0.12	4 (23)
10	2;IMP;SG	5896	49668	0.12	29 (41)
11	3;PRS;SBJV;SG	4476	39920	0.11	3 (23)
12	1;PRS;SG	5951	55833	0.11	9 (52)
13	2;PRS;SBJV;SG	3876	39476	0.10	4 (23)
14	3;PL;PRS	4199	51818	0.08	4 (47)
15	2;PRS;SG	4236	54920	0.08	8 (51)
16	PST	1813	27001	0.07	7 (15)
17	IMP;PL	162	2859	0.06	2 (4)
18	3;PST;SG	1273	22624	0.06	5 (26)
19	2;PST;SG	1261	22639	0.06	4 (27)
20	V.MSDR	40	1134	0.04	2 (4)
21	3;PFV;PST;SG	1015	30844	0.03	7 (21)
22	MASC;PST;SG	122	6638	0.02	2 (3)
23	2;PFV;PST;SG	330	29977	0.01	2 (20)

Appendix C

The model formulas are given in Table 15. Besides the paradigm size (`n_cells`), the number of values per cell (`n_values`) and the number of lexemes for which inflectional paradigms are available (`n_lexemes`) may influence the probability of zero forms. As can be seen in Table 15, besides the main model m01, I included additional models to test for the effect of paradigm size (m02-m04), of number of values per cell (m05-m07) and of number of lexemes (m08-m10). For each of the three additional variables, I fitted a model using them as the single population-level effect, using it in addition to the part of speech, and including the interaction between

the two population-level effects. For details about the models and their comparison, see the file "code.R" in the supplementary materials.

Table 15: Model series to examine the probability of zero forms

main model	
m01	pos + (1|phylo)
check for paradigm size	
m02	n_cells + (1|phylo)
m03	pos + n_cells + (1|phylo)
m04	pos * n_cells + (1|phylo)
check for number of values per cell	
m05	n_values + (1|phylo)
m06	pos + n_values + (1|phylo)
m07	pos * n_values + (1|phylo)
check for number of lexemes	
m08	n_lexemes + (1|phylo)
m09	pos + n_lexemes + (1|phylo)
m10	pos * n_lexemes + (1|phylo)

In order to select the optimal and final model, I compared their performance using approximated leave-one-out cross-validation (LOO-CV) following (Vehtari, Gelman & Gabry 2017). The basic principle behind leave-one-out cross-validation is to re-fit the model leaving out one datapoint at a time and then predict this data point. By doing so, the overall model performance can be evaluated against data which has not been used to train the model. Approximate LOO is an efficient approximation of LOO-CV, and it estimates the model's ELPD value (theoretical expected log pointwise predictive density) for a new dataset. The absolute value itself is difficult to interpret and not relevant here; it is rather the relative difference of ELPD values that can be used to compare models in terms of their predictive power.

The relative ELPD values of the models m01-m10 are given in Table 16. The model with the highest ELPD value, i.e. the best-performing model is shown on top with its ELPD value set to 0; the ELPD differences between the best and the other models are shown by negative values. To assess the ELPD differences, the rightmost column in Table 16 provides the estimated standard errors for the ELPD differences.

We can see that m09, including information on the number of lexemes in addition to the part of speech is the best model in terms of predictive power. Therefore, it was selected as the final model for the analysis in Section 5.3. However, m01, using only part of speech is only marginally worse in terms of its fit than m09. The ELPD difference between these two models is 0.7, which is less than half of the standard error of the ELPD difference. The comparison also shows for the remaining models that their predictive power is not much worse than the one of m09; in almost all cases, the ELPD difference is smaller than its standard error. Interestingly,

Table 16: Approximated LOO

	predictors	ELPD difference	SE difference
m09	pos + n_lexemes + (1|phylo)	0	0
m01	pos + (1|phylo)	-0.7	2.5
m06	pos + n_values + (1|phylo)	-1.0	3.3
m10	pos * n_lexemes + (1|phylo)	-1.2	0.5
m04	pos * n_cells + (1|phylo)	-1.3	3.2
m07	pos * n_values + (1|phylo)	-1.5	3.5
m03	pos + n_cells + (1|phylo)	-1.6	2.8
m05	n_values + (1|phylo)	-3.2	4.5
m08	n_lexemes + (1|phylo)	-3.6	3.1
m02	n_cells + (1|phylo)	-3.9	4.2

the three worst performing models m05, m08 and m02 are the three models that do not include information about the part of speech. This, taken together with the fact that m02, using only part of speech as a population-level effect, is the second best-performing model suggests that the part of speech includes the most important information for predicting the probability of zero forms.