Zero marking in inflection: A token-based approach

Laura Becker
University of Freiburg

ABSTRACT

This study examines zero marking, i.e. the absence of an overt exponent, in adjectival, nominal, and verbal inflectional morphology across languages. The first part of the study provides an overview of the distribution of zero markers in inflection paradigms using the UniMorph dataset. The results show that there is a general preference against zero marking. The distribution of zero markers varies to a great extent across languages and lemmas, the only robust trend being that they are avoided in cells that express a high number of grammatical values. The second part of this study examines the association between marker frequencies and phonological length, using the Universal Dependencies treebanks. While token frequency is a good predictor for the length of overt markers, it does not account for the occurrence of zero markers. This is taken as evidence to support a differential non-development scenario of zero marking rather than a phonetic reduction scenario.

Keywords: token-based typology, corpus typology, zero marking, zero exponence

INTRODUCTION 1

The present study examines the distribution of zero markers in adjectival, nominal, and verbal inflectional morphology.[1] In typology,

[1] I wish to thank the participants of the Freiburg Linguistics reading group for their helpful comments on earlier versions of this study: Uta Reinöhl, Peter

zero marking plays an important role for coding efficiency or form-frequency effects in morphosyntax. The analysis of form-frequency effects goes back to the early findings by Zipf (1935) that more frequent lexical elements tend to be shorter than less frequent ones. There is cross-linguistic evidence that, in inflectional morphology as well, more frequent or predictable markers tend to be shorter or at least not longer than comparable less frequent markers (Greenberg 1966; Guzmán Naranjo and Becker 2021; Haspelmath 2008b; Haspelmath *et al.* 2014; Haspelmath 2021; Haspelmath and Karjus 2017; Stave *et al.* 2021).

Such effects can be subsumed under the term of coding efficiency. The coding of grammatical expressions is efficient, because it saves effort in the production and processing of speech but maintains the successful transfer of information (cf. Levshina 2022, for an overview of efficiency in language and communication).

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

To do so, I analyze the distribution of zero markers in the UniMorph dataset (McCarthy *et al.* 2020), a cross-linguistic database of inflectional paradigms for individual lemmas. I first provide some theoretical background on zero marking and coding efficiency and introduce a working definition of zero markers in Section 2. Section 3 describes the dataset as well as the marker extraction procedure, and discusses examples of zero markers. I then analyze the probability of zero marking using the UniMorph dataset in Section 4. As will be seen, zero marking is generally dispreferred across languages and parts-of-

Arkadiev, Matías Guzmán Naranjo, Marvin Martiny, and Naomi Peck. I also thank the three anonymous reviewers for their valuable comments on earlier versions of this paper. This paper was supported by a Junior Fellowship from the Freiburg Institute for Advanced Studies, University of Freiburg.

speech. Section 5 then zooms in on those cells and values of adjectival, nominal, and verbal inflectional paradigms that are most likely to be zero marked across languages. In Section 6, I turn to the distribution of zero markers in language use. Using corpus data from the Universal Dependencies treebanks (Zeman *et al.* 2023), I analyze the association between token frequencies of inflection markers and their phonological length, including the distribution of zero markers. As we will see, frequency does not affect zero markers in the same way as it affects overt markers. Section 7 discusses the findings of this study with a special focus on the role of coding efficiency to account for the distribution of zero marking. Section 8 concludes.

ZERO MARKING 2

This section presents the relevant theoretical notions related to zero marking. Section 2.1 introduces zero marking and its relation to coding efficiency in typology. In Section 2.2, I propose a working definition of zero markers for the purposes of the present study. Throughout the paper, I use zero marking to refer to the absence of phonetic exponence ("zero exponence") of a morphosyntactic function.

Zero marking and coding efficiency 2.1

The modern understanding of coding efficiency began with Zipf (1935), who showed that more frequent words tend to be shorter than less frequent words. Greenberg (1966, 1963) was one of the first typologists to relate the token frequencies of grammatical values 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. For instance, Greenberg (1966, 32–37) showed how the markedness of singular, plural, and dual forms of nouns, verbs, and adjectives is reflected in their distribution in corpora from various languages. He noted that the formally unmarked (no exponent) number value, singular, is substantially more frequent than the formally marked number values (overt exponent) of plural and dual in corpus data from different languages.

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

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

This hypothesis includes the assumption that zero forms pattern with shorter forms in being used to encode comparatively frequent expressions. Applied to inflectional morphology, we should thus expect zero marking for highly frequent values of morphosyntactic features. By now there is indeed much evidence for effects of coding efficiency between comparable grammatical expressions. However, examples usually only involve a difference in length, i.e. shorter vs. longer forms.[2] The participation of zero forms has not yet been the focus of any systematic cross-linguistic study. There are some indications from the literature, though, which suggests that coding efficiency and frequency may not be a suitable explanation for the distribution of zero markers. Stolz and Levkovych (2019) provide a qualitative overview of the distribution of zero marking in inflection ("absence of material exponence, AOME") from the perspective of canonical morphology. They note that "[f]rom the small number of cases discussed above it transpires that frequency might not always be the most powerful factor

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

to make a given word-form or category a candidate for AOME" (Stolz and Levkovych 2019, 396–397).

Guzmán Naranjo and Becker (2021) come to a similar conclusion based on a quantitative analysis of the association between the length of nominal inflection markers and their distribution across paradigms. They also use the UniMorph database, but focus on nominal inflection and test different distributional factors for their association with marker length. Although they find that marker length is associated with their type frequency, their results suggest that other measures such as the entropy of the marker are better predictors for their length. With their main focus being on predicting marker length from distributional measures, one detail of their analysis concerns zero marking and is highly relevant for the present study. Guzmán Naranjo and Becker (2021) note that a simple Poisson model to predict marker length strongly overestimates the occurrence of zero markers. This suggests that the distribution of zero markers does not simply follow the pattern of shorter ones.

Another area in which zero marking has been mentioned to behave differently is the occurrence of zero markers for person and number marking on verbs. Several quantitative typological studies (Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010) find that zero marking for person marking is rather uncommon across languages. In contrast to the traditional view in typology, these studies do not find evidence for a paradigmatic preference of third person (singular) being zero marked on the verb. However, all three studies show that if a person marker is zero, it more likely expresses third person (singular) than first or second person.

Seržant and Moroz (2022) also mention zeros in verbal person-number marking. Analyzing the length of person-number markers in a typological sample, they argue for an attractor state in which the lengths of different indexes are associated with their frequencies in language use. Seržant and Moroz (2022, 6) note that "[…] articulatory efficiency plays an important role here: the more expected the sign is the shorter it is. Nevertheless, zero is not preferred." They motivate the cross-linguistic avoidance of zero forms by invoking two types of efficiency: processing and planning efficiency. Seržant and Moroz (2022, 7) hypothesize that an overt exponent facilitates processing on the addressee's side. They also propose that avoiding zero marking makes planning more efficient on the speaker's side, "[...] because it provides a straightforward link from meaning to coding, while zero is inherently ambiguous by being linked to various meanings and domains" (Seržant and Moroz 2022, 7). Whether or not the avoidance of zero marking can indeed be accounted for by processing or planning efficiency requires proper psycholinguistic testing. The relevant point is that coding efficiency does not seem to be applicable to the frequency distribution of zero markers in person indexing in the same way as it is for overt markers.

2.2 *A working definition of zero markers*

The discussion and use of zero marking has a long tradition in morphology and in linguistics in general. It goes back to Pāṇini, who introduced the idea of zero morphs for morphemes that lack a phonetic representation as the outcome of morphological rules (Robins 1997, 181–182). The concept of zero morphs for linguistic analysis was also widely applied in later work by structuralists (e.g. Bloch 1947; Bloomfield 1933; Jakobson 1983[1939]; de Saussure 1916).[3] Starting with Haas (1957), linguists began to criticize the assumption of zero morphs in the structuralist tradition and argued for stricter criteria to define zero morphs in order to avoid the assumption of excessive linguistic structure (e.g. Sanders 1988; Mel'čuk 2002; McGregor 2003). This was because linguists may postulate a zero morph for any single morphosyntactic function that does not correspond to an overt exponent. 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".[4] Others, such as Arkadiev (2016), Contini-Morava (2006) and Mithun (1986), used data from typologically diverse languages to show that the absence of phonetic material can also correspond to the absence of a morphosyntactic feature rather than to

[3] For more details, see Meier 1961. See also Al-George 1967, Diehl 2008 and McGregor 2003 for more details on the history of linguistic zero.

[4] For examples and discussions of issues related to the use of zero morphs in morpheme-based, segmental approaches to morphology, see Anderson 1992, Pullum and Zwicky 1991, Blevins 2016 and Bank and Trommer 2015. For overviews of zero exponence in morphological theories, see Trommer 2012 and Dahl and Fábregas 2018.

zero marking. For instance, Lakota has overt markers for first and second person arguments on the verb, but no overt third person markers. Mithun (1986, 201–203) proposes to analyze the Lakota pattern as agreement that is restricted to first and second person arguments instead of analyzing agreement with third person arguments as zero marked.

In line with those more cautious approaches to zero morphs, this study uses the notion of "zero marker" as a descriptive shorthand for the absence of material exponence of a given morphosyntactic function (cf. Stolz and Levkovych 2019). In other words, I do not assume the presence of a zero morph. Instead, I understand zero markers as the absence of exponence which expresses a certain morphosyntactic function in addition to the lexical content of a word form. This also means that zero markers can only occur in contrast to at least one other, overtly coded morphosyntactic function of the same inflectional paradigm.

To analyze the distribution of zero markers in inflectional morphology, we need to identify the invariable, lexical parts (stems) as well as the potential exponents of a morphosyntactic function in an inflected word form. This conforms with the basic intuition that we want to separate the segments that convey the word's lexical meaning from the segments that convey morphosyntactic information (cf. Matthews 1972).[5] For the purposes of the present study, I define stems, markers, and zero markers as shown in (2), (3), and (4), respectively. These definitions are motivated by both theoretical and practical considerations regarding the dataset and annotations available.

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

(3) *Marker*
A marker encodes the morphosyntactic function of a word

[5] In reality, the identification of stems is not always this straightforward. There are many different ways in which the lexical parts of inflected words can vary in their phonological shape. Baerman and Corbett (2012) provide a number of examples and introduce a canonical approach to stems to capture that variation.

form, i.e. a value of some morphosyntactic feature defined for that word or a bundle of values of several such features. The marker corresponds to the phonetic material outside of the stem of a word form; it can be discontinuous.

(4) *Zero Marker*
A zero marker occurs when the word form does not feature any overt marker (as defined in (3)) to encode its morphosyntactic function. If the morphosyntactic function of the word consists of several morphosyntactic features, zero marking applies to the combination of feature values and not to single feature values in isolation.

Consider a simple example of stem and marker identification. The paradigm of English nouns consists of two cells: singular and plural. 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 a plural marker. In the singular cell, 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* as zero marked.

However, as will be described in detail in Section 3.3, I automatically adjusted the stems extracted according to the definition in (2) in order to account for stem allomorphy to a certain extent. This is motivated by the fact that many stem alternations are phonologically driven, which means that they do not necessarily provide meaningful insights about the inflectional properties of a system in general and about the distribution of zero marking in particular. Ignoring such alternations allocates additional material to the marker segments and runs the risk of systematically underestimating the number of zero markers. The adjusted marker$_A$ and zero marker$_A$, which take into account stem alternations, are operationalized as described in (5) and (6), respectively.[6]

[6] From a theoretical perspective, it may be desirable to adjust the definition of stems and then derive the new definition of markers from that. The definitions given in (5) and (6) reflect the data extraction process, in that I extracted the ad-

(5) *Marker$_A$*
A marker$_A$ is extracted from a marker, as defined in (3), by removing all material from the affix positions that the system does not use for inflection.

(6) *Zero Marker$_A$*
A zero marker occurs when the word form does not feature any overt marker (as defined in (5)) to encode its morphosyntactic function.

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

Besides practical considerations, this method is also based on theoretical grounds and follows the definition of stems by Beniamine and Guzmán Naranjo (2021), Bonami and Beniamine (2021), and Guzmán Naranjo and Becker (2021). Despite much theoretical work on the role and identification of stems in morphology, Bonami and Beniamine (2021) note that "there is no agreed upon method for identifying which part of an inflected word is a stem, and that the heuristics used by morphologists in that area are neither systematic nor principled enough".[7] They compare two types of stem identification based on prioritizing two different principles, namely to avoid stem allomorphy and to avoid discontinuous stems. Since those two principles are in conflict with each other many times, every approach to stem identification needs to rank them in some way to resolve such conflicts. Bonami and Beniamine (2021) compare the two methods of either adhering to the first or the second principle, resulting in what they call "unique discontinuous stems" (no stem allomorphy allowed) and "continuous stem sets" (no discontinuous stems allowed). While the

justed markers$_A$ and zero markers$_A$ from the original markers without extracting adjusted stems. I therefore omit the step of defining adjusted stems and focus directly on the alternative definitions of markers$_A$ and zero markers$_A$.

[7] For work on stem identification and stem allomorphy, see Blevins 2003, Bonami 2012, Brown 1998, Maiden 1992, Montermini and Bonami 2013, Pirrelli and Battista 2000, Spencer 2012, Stump 2001 and Stump and Finkel 2013.

first method of unique discontinuous stems allocates all the variation of word forms to the exponents, leading to more exponent allomorphy, the second method of continuous stem sets keeps exponent allomorphy minimal, but leads to a high degree of stem allomorphy, since all variation that is enclosed by stem segments has to be included in the stems. What this shows is that neither approach creates more allomorphy; they simply allocate it differently. Of course, which of the two approaches is more useful depends on the research question at hand.

One of the questions discussed by the authors is what types of stems are more helpful in addressing the 'Inflected Word Recognition Problem' (IWRP), i.e. understanding what allows speakers to draw inferences from a word's form about its content. This results in the task of separating the lexical and the inflectional parts of a word form, and Bonami and Beniamine (2021) note that "[i]n terms of the IWRP, the answer is quite simple. Sets of continuous stems are by definition less useful than a unique discontinuous stem: the unique discontinuous stem identifies exactly that part of the word that has no exponential value, while stem allomorphs blur the distinction between exponential and nonexponential material." As the identification of zero forms relies on separating lexical segments from exponents of morphosyntactic information in word forms, the IWRP is of high relevance to this study and provides the theoretical grounds for the definition of stems given in (2).

This study will largely follow a word and paradigm approach to inflection (cf. Anderson 1992; Blevins 2016; Hockett 1967; Matthews 1972; Robins 1959; Stump 2001; Zwicky 1985). This approach bases morphological analyses on the paradigmatic relation between different word forms that represent the different morphosyntactic functions a given word can have. The exponent of a cell in an inflectional paradigm is determined through the relation of that word form to the forms used for the other cells of the paradigm. The word and paradigm approach has a very important practical advantage. It allows us to refrain from further segmentation of exponents into morphemes, which may require language-specific insights and which may not always be desirable or useful (cf. Blevins 2005, 2006).

Although morphological segmentation analyses may sometimes be uncontroversial, there are many cases where a morpheme analysis is less than clear. Various examples are given in Spencer 2012, one of them being the Spanish subjunctive verb form *cantaríamos* 'we would sing'. Several theoretical motivations exist to segment this word form into morphemes in five different ways: (i) *cant-a-r-í-a-mos*, (ii) *canta-ríamos*, (iii) *cant-a-ría-mos*, (iv) *canta-r-í-a-mos* and (v) *cantar-í-amos* (Spencer 2012, 93). The fact that these profoundly varying morphological analyses are motivated in the literature suggests that such morpheme segmentations are always theoretically guided, whether explicitly or implicitly. It is likely that segmentation into morphemes in lesser-studied languages involves even more theoretical uncertainty, given that we may know much less about morphological structure and its diachrony than for languages like Spanish.

As will be shown in more detail in Sections 3.3 and 3.4, cells of paradigms are defined by (a combination of) values of morphosyntactic features. For instance, the inflectional paradigms of German nouns combine the morphosyntactic features of case and number. While nouns are inherently specified for gender, each word form in context is also specified for number and case, so that each cell of the paradigm corresponds to a number-case combination (e.g. dative plural).

For the purposes of this study, I do not distinguish between an exponent for plural number and one for dative case. Instead, I treat the material in addition to the stem in the dative plural cell as the marker of the dative-plural function. When no additional phonetic material is used, this cell is then analyzed as being zero marked (cf. Table 9). I do not assign zero markers to single abstract morphosyntactic values but to the relevant value combinations of the inflectional paradigms. The theoretical reason for this is that exponents of morphosyntactic functions are defined based on the relations between the forms of the different cells of the inflection paradigm, which combine these functions. This also reflects the morphological reality of many (if not most) languages, in that morphosyntactic functions are usually not marked in isolation but often occur in combination. As mentioned above, it is not always trivial to justify a segmental analysis. The practical reason is that there is still no language-independent and theory-independent way of segmenting distinct morphosyntactic exponents, and such segmentations are not (yet) automatable. Since automatic processing is indispensable for the purposes of the present study, no further segmentation of morphosyntactic exponents will be carried out.

The segmentation into stems and markers is often additionally complicated by inflection classes, which use different exponents to signal grammatical functions. Sections 3.3 and 3.4 show in more detail how the present approach deals with variation in the exponents due to inflection classes, with stem alternations and with suppletive forms.

3 DATASET AND SEGMENTATION

3.1 *Dataset*

The data used in this study comes from the UniMorph database (McCarthy *et al.* 2020), a large-scale cross-linguistic database of complete inflectional paradigms of adjectives, nouns, and verbs for individual lexemes from different languages. The present study includes adjectival, nominal, and verbal paradigms for 39, 62, and 96 languages, respectively. Some languages are featured with paradigms for more than one part-of-speech; a total of 114 languages is analyzed in this study. Figure 1 shows the geographical distribution of the languages in the dataset.[8]

While the dataset is not a balanced typological sample in the strict sense, it does include languages from all six macro areas (Africa, Eurasia, Papunesia, Australia, North America and South America), which ensures that typological and areal diversity is captured at least to some degree. Table 1 provides an overview of the final dataset with the number of languages, lemmas, paradigm cells, marker types and observations by part-of-speech. The morphosyntactic annotation in the UniMorph dataset follows the guidelines described by Sylak-Glassman (2016, 3), who notes: "This paper presents the Universal Morphological Feature Schema (UniMorph Schema), which is a set of morphological features that functions as an interlingua for inflectional morphology by defining the meaning it conveys in language-independent

[8] More details about the languages, the parts-of-speech, and the number of lexemes is provided in the files `affixation.csv` and `lemmas.csv` in the supplementary materials. All supplementary materials referred to in this paper can be found here: `https://osf.io/p4mkc/?view_only=5238ace9cb1d4f4d998486ebb28f4fd8`

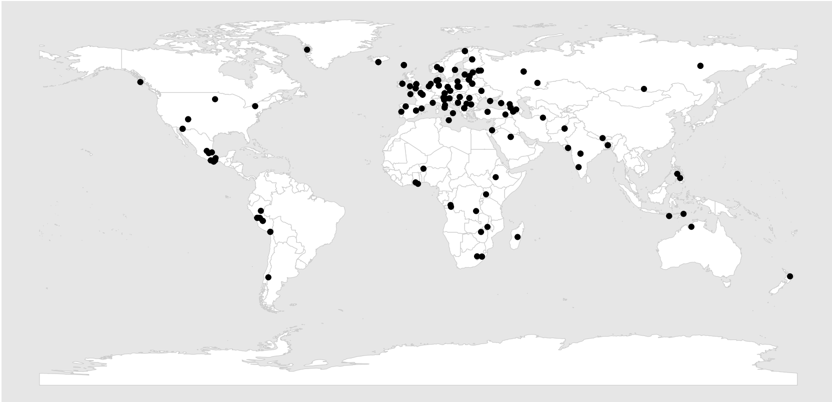

Figure 1: Location of the languages in the dataset

Table 1: Dataset overview

	N langs	N lemmas	N cells	N markers	N obs
adjectives	39	157355	961	5552	6348198
nouns	62	610242	727	19537	6261881
verbs	96	129377	2753	47457	4407743

terms. The features of the Universal Morphological Feature Schema have precise definitions based on attested cross-linguistic patterns and descriptively-oriented linguistic theory, and can capture the maximal level of semantic differentiation within each inflectional morphological category.” Annotations thus do not necessarily follow the linguistic traditions of particular languages but are defined and used in the sense of comparative concepts in typology (cf. Haspelmath 2018).

3.2 *Data pre-processing*

I excluded a number of languages available in UniMorph from the final analysis on the basis of unclear or insufficient annotations in the original datasets, some of which were annotated only automatically with no manual checks. Since the database is somewhat biased towards languages spoken in Eurasia (mostly Indo-European languages), I only included languages from this area with paradigms for more than 30 lemmas. For languages from other macro areas, especially from Africa or the Americas, I did not apply this threshold of 30 lemmas in order to include more non-Indo-European languages and to keep the dataset as diverse as possible.[9]

The next step was to pre-process the data to remove errors and to make annotations more consistent across languages.[10] Pre-processing consisted of different global and dataset-specific corrections. Global corrections included resolving annotation inconsistencies across languages. For example, the value "indefinite" was coded as "INDF" in some languages and as "NDEF" in others. Similarly, the annotation of person-number combinations in verbs varied, e.g. between "SG;1", "1;SG", "1SG" for first person singular. In such cases, I adjusted the annotation to a single label across all languages. I also removed complex lemmas containing a space or "-". This removed some erroneous lemmas that were complex expressions rather than nouns, adjectives, or verbs. In some languages, both parts of a complex noun or adjective are inflected. Keeping such lemmas would have caused the marker extraction to detect infixation for complex lemmas with suffixes on two or more parts. Removing them avoided the artificial creation of more complex inflection patterns. Similarly, periphrastic forms were removed in the case of inflected auxiliaries, which would also have led to the erroneous analysis of infixation. This conservative approach of removing such forms was chosen over, e.g., splitting them or analyzing the inflected auxiliaries only. This alternative would have involved many additional case-by-case modifications of the original data, which in turn would have made it more prone to additional errors. Moreover, it would have increased the number of inflected forms from single auxiliaries, potentially misrepresenting the distribution of markers across lemmas. Complex forms were also removed if they contained a separate marker that occurred before or after the inflected verb form, depending on the cell of the paradigm. This was especially common with verbal paradigms, e.g.

[9] For adjectives, only Zulu has fewer than 30 lemmas (17); for nouns, this is the case only for Kalaalisut (23). For verbal paradigms, the languages with fewer than 30 lemmas are Sotho (26), Mapudungun (26), Murrinpatha (29), and Zarma (27).

[10] Detailed documentation of all pre-processing steps can be found in `preprocessing.txt` in the supplementary materials. For the implementation, see `code-preprocessing.R`.

verbal particles in German, or reflexive markers in Italian and Macedonian.

Dataset-specific cleaning steps included deleting "?" following interrogative verb forms in the Turkish data or deleting the indefinite article from Romanian nominal forms. Other cleaning steps were related to the alphabetic scripts used. For example, the Serbian-Bosnian-Croatian dataset contained forms in the Latin script with a handful of forms in Cyrillic. The latter were removed to allow for consistent processing. Some datasets (e.g. Old French or Yoloxochitl Mixtec) contained alternative forms for certain cells. In such cases, I systematically kept the first form and removed the other(s).[11] Other dataset-specific operations included deleting single forms containing obvious errors (e.g. misalignment, cells with missing data).

Following data cleaning, I added phonological transcriptions to the inflected forms whenever possible. For some languages (e.g. Palantla Chinantec), the UniMorph database already provided the inflected forms in a phonological transcription. For most other languages, however, forms were given in the standard orthographic representation. This may well be problematic, especially for 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 replaced the orthographic forms by a phonological transcription using Epitran (Mortensen *et al.* 2018). Epitran currently has modules to transcribe 31 of the languages used here.[12]

While not perfect, Epitran offers a more realistic representation of the forms occupying the different cells of inflectional paradigms. Table 2 illustrates this by showing the transcriptions generated with

[11] It would have been insightful to include overabundance in a systematic way. Overabundance refers to the phenomenon of two distinct forms being available to express a single cell in a paradigm (cf. Thornton 2012). However, alternative forms are not systematically annotated in the UniMorph datasets. If provided, their relation differs greatly across datasets and is not usually documented in the dataset descriptions. Alternatives can represent diachronic, dialectal, or stylistic variants; in other cases, their alternation behavior remains unclear. It is also unclear how many overabundant forms are not provided in UniMorph. Including overabundance is thus not possible with the approach used in this study.

[12] For details, see `epitran.py` in the supplementary materials.

Table 2: Phonological transcription of the French verb *allumer* 'turn on (light)'

	1SG	2SG	3SG	1PL	…
PRS.IND	*allume* alym	*allumes* alym	*allume* alym	*allumons* alymɔ̃	
PST.IPFV.IND	*allumais* alymɛ	*allumais* alymɛ	*allumait* alymɛ	*allumions* alymiɔ̃	
PST.PFV.IND	*allumai* alymɛ	*allumas* alyma	*allumat* alyma	*allumâmes* alymɑm	
FUT	*allumerai* alymrɛ	*allumeras* alymra	*allumera* alymra	*allumerons* alymɛrɔ̃	
PRS.COND	*allumerais* alymrɛ	*allumerais* alymrɛ	*allumerait* alymrɛ	*allumerions* alymriɔ̃	
PRS.SBJV	*allume* alym	*allumes* alym	*allume* alym	*allumions* alymiɔ̃	
PST.SBJV	*allumasse* alymas	*allumasses* alymas	*allumât* alymɑ	*allumassions* alymasiɔ̃	
…					

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 orthographic representation, and the second row shows the phonological transcription generated with Epitran. For the remaining 81 languages, the forms in UniMorph are given in their orthographic representation, which reflect the phonological shapes to a varying degree. To consider the potential influence that the type of phonological representation may have on the detection of zero forms, I manually coded whether or not the representation was phonological.[13] Orthographic representations that systematically reflected phonology were treated as phonological representations. This led to 31 languages with a transcription generated using Epitran, 63 languages with original representations that systematically reflect phonological shape, and 20 languages with orthographies that do not always reflect phonological shape. The type of phonological representation was then added as a control variable in the analysis.

[13] For details by language, see `affixation.csv` in the supplementary files.

Extracting stems and zero markers 3.3

In order to analyze the distribution of zero markers, I automatically segmented the inflected word forms following the method developed in Beniamine and Guzmán Naranjo 2021 and Guzmán Naranjo and Becker 2021. As mentioned in Section 2.2, the segmentation follows a word and paradigm approach to morphology, in that whole forms are paired with morphosyntactic functions according to their distribution across the inflectional paradigms. This means that the subsequence shared by all cells of the paradigm is automatically extracted and taken as the stem according the working definition given in (2). All material not included in this subsequence is analyzed as the marker of a given cell, as defined in (3). If the form corresponds to the longest common subsequence (i.e. the stem), the marker is analyzed as zero according to the definition in (4). This automated detection of stems and markers is necessary for two reasons. First, it is not feasible to apply manual, language-specific segmentations to this dataset. Second, this method allows for a single, consistent way of detecting zero marking across languages, which is necessary for the cross-linguistic comparisons made in this study.[14]

To give a simple example of the segmentation into stems and markers and of the detection of zero markers, Table 3 shows parts of the present tense paradigm of the French verb *allumer* from Table 2.[15] Comparing the forms of the different cells of the paradigm, the string *alym* is identified as the longest common subsequence between all forms of the paradigm. For the purposes of the present paper, this subsequence is analyzed as the stem. All remaining material is analyzed as the marker of a particular cell. In cells where the form corresponds to the stem, markers are analyzed as zero. This is the case for some of the present tense forms; such cells are shaded in grey in Table 3.

In the remainder of this section, I discuss the extraction of stems and markers using examples that may appear less straightforward, in

[14] Stem alternations are not accounted for by this extraction method; Section 3.4 shows how they are included in the present study.

[15] This example involves a continuous stem as well as continuous markers. Examples of discontinuous stems are shown later in this section and in Section 3.4.

Table 3:
Marker extraction for the French verb *allumer* '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.COND.1SG	alymɛrɛ	alym	-ɛrɛ
PRS.COND.2SG	alymɛrɛ	alym	-ɛrɛ
PRS.COND.3SG	alymɛrɛ	alym	-ɛrɛ
PRS.COND.1PL	alymɛrjɔn	alym	-ɛrjɔn
PRS.SBJV.1SG	alym	alym	-
PRS.SBJV.2SG	alym	alym	-
PRS.SBJV.3SG	alym	alym	-
PRS.SBJV.1PL	alymjɔn	alym	-jɔn
…	…	…	…

that the identified stems (and thus also markers) do not correspond to stems as traditionally analyzed in the literature, or in that they are discontinuous.

One example comes from Ayamara (Aymaran), a language with nominal inflection known for its subtractive morphology. The accusative singular cell is usually analyzed as being expressed by the subtraction of the final vowel of the nominative singular form (cf. Coler 2015). Table 4 illustrates this with parts of the paradigms of two Aymara nouns. For the purposes of this study, the accusative singular form corresponds to the stem, because it equals the longest common subsequence of all forms of the lexeme. Compared to the accusative form, the nominative form has an additional final vowel, which is also found in all other forms of the paradigm, except for the inessive (INESS) and equative (EQTV) forms.

Traditionally, the nominative form with the final vowel is analyzed as the stem of the noun, while the accusative is argued to be a subtractive form, i.e. consisting of less material than the stem of the lexeme (Baerman *et al.* 2017; Coler 2015, 2018). There are valid diachronic 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

Cell	Form	Stem	Marker	Form	Stem	Marker
NOM.SG	anu	an	-u	chaski	chask	-i
ACC.SG	an	an	-	chask	chask	-
GEN.SG	anuna	an	-una	chaskina	chask	-ina
COM.SG	anumpi	an	-umpi	chaskimpi	chask	-impi
ABL.SG	anuta	an	-uta	chaskita	chask	-ita
ALL.SG	anuru	an	-uru	chaskiru	chask	-iru
INESS.SG	anpacha	an	-pacha	chaskpacha	chask	-pacha
EQTV.SG	anjama	an	-jama	chaskjama	chask	-jama
…	…	…	…	…	…	…

Table 4: Marker extraction for the Aymara nouns *anu* 'dog' and *chaski* 'messenger'

process in Aymara. Nevertheless, aiming at a synchronic, comparable analysis across languages here, I treat the accusative form as the stem of the lexeme. In the Aymara data, the accusative is zero marked in all 1,522 nouns of the dataset, without exception.

Another rather unusual case of zero marking can be found in Georgian (Kartvelian) verbs. Besides a number of other theoretically interesting patterns, Georgian verbs have been cited in the typological and morphological literature for their cross-linguistically unusual 2SG zero marker (e.g. Anderson 1992; Blevins 2016; Stolz and Levkovych 2019). However, not all lexemes express the 2SG form with a zero marker in the sense of the present study. Only one out of 118 verbal lexemes in the dataset features a zero marker in the 2SG present tense cell. Table 5 shows this for the verb *ts'ers* 'write', in opposition to *ak'etebs* 'make'.[16]

In general, Georgian verbs take a so-called preverb in some but not all of the tenses (Hewitt 1995, 148–169). When present, it precedes the prefixal part of agreement marking on the verb. As we can see in Table 5, present and imperfect forms occur without the verbal prefix, while the future, aorist, and perfect forms all make use of the prefix (*da-* and *ga-* in the examples in Table 5). In most TAM series, many Georgian verbs also have so-called thematic suffixes (Hewitt 1995, 143–147), such as *-eb* in *ak'etebs* 'make'. The presence of such thematic suffixes in the present tense results in the absence of

[16] The segment *-a-* is not part of the verb stem of *ak'etebs* 'make', as it does not occur in all forms of the paradigm, e.g. the imperfective masdar form *k'etebi*.

Table 5: Marker extraction for the Georgian verbs *ts'ers* 'write' and *ak'etebs* 'make'

Cell	Form	Stem	Marker	Form	Stem	Marker
PRS.1SG	vts'er	ts'er	v-	vak'eteb	k'et	va-eb
PRS.2SG	ts'er	ts'er	-	ak'eteb	k'et	a-eb
PRS.1PL	vts'ert	ts'er	v-t	vak'etebt	k'et	va-ebt
IMPF.1SG	vts'erde	ts'er	v-de	vak'etebdi	k'et	va-ebdi
IMPF.2SG	ts'erde	ts'er	-de	ak'etebdi	k'et	a-ebdi
IMPF.1PL	vts'erdet	ts'er	v-det	vak'etebdit	k'et	va-ebdit
FUT.1SG	davts'er	ts'er	dav-	gavak'eteb	k'et	gava-eb
FUT.2SG	dats'er	ts'er	da-	gaak'eteb	k'et	gaa-eb
FUT.1PL	davts'ert	ts'er	dav-t	gavak'etebt	k'et	gava-ebt
AOR.1SG	davts'ere	ts'er	dav-e	gavak'ete	k'et	gava-e
AOR.2SG	dats'ere	ts'er	da-e	gaak'ete	k'et	gaa-e
AOR.1PL	davts'eret	ts'er	dav-et	gavak'etet	k'et	gava-et
…	…	…	…	…	…	…

zero marking for most of the verbs. The thematic suffix *-eb/-ob* is part of the second person singular present form; as it is not used in the aorist forms, the former does not correspond to the longest common subsequence 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, like the verb *ts'ers* 'write' in Table 5.

Arabic (Semitic) is well known to have roots that consist of discontinuous consonants, with prefixed, infixed, and suffixed vowels, and other consonants to mark the grammatical values of a given form in the paradigm (e.g. Boudelaa and Marslen-Wilson 2001; Ratcliffe 1998; Schramm 1962; Yip 1988). The automatic extraction of the longest common subsequence detects these consonants and assigns all additional material to the markers. This is shown for two verbs, *ʔarsala* 'send' and *iktašafa* 'discover' in Table 6.

Another language that is interesting from the point of view of marker extraction is Tohono O'odham (Uto-Aztecan, Mexico, USA). Some nouns in Tohono O'odham mark the plural using partial reduplication of the stem (Hill and Zepeda 1998). Table 7 shows this for the two nouns *ban* 'coyote' and *ceoj* 'boy', using the phonological transcription generated by Epitran.

Cell	Form	Stem	Marker	Form	Stem	Marker
IPFV.1SG	ʔursilu	rsl	ʔu-i-u	ʔaktašifu	ktšf	ʔa-a-i-u
IPFV.2SG.F	tursilīna	rsl	tu-i-īna	taktašifīna	ktšf	ta-a-i-īna
IPFV.3PL.M	yursilūna	rsl	yu-i-ūna	yaktašifūna	ktšf	ya-a-i-ūna
PFV.1SG	ʔarsaltu	rsl	ʔa-a-tu	iktašaftu	ktšf	i-a-a-tu
PFV.2SG.F	ʔarsalti	rsl	ʔa-a-ti	iktašafti	ktšf	i-a-a-ti
PFV.3PL.M	ʔarsalū	rsl	ʔa-a-ū	iktašafū	ktšf	i-a-a-ū
…	…	…	…	…	…	…

Table 6: Marker extraction for Arabic verbs *ʔarsala* ‘send’ and *iktašafa* ‘discover’

Cell	Form	Stem	Marker	Form	Stem	Marker
SG	ban	ban	-	t͡ʃɨɒd͡ʒ	t͡ʃɨɒd͡ʒ	-
PL	ba:ban	ban	-:ba-	t͡ʃɨt͡ʃɒd͡ʒ	t͡ʃɨɒd͡ʒ	-t͡ʃ-

Table 7: Tohono O’odham nouns *ban* ‘coyote’ and *ceoj* ‘boy’

Applying the automatic stem extraction for the purposes of this study, the reduplicated stem is analyzed as infixation, i.e. the marker of the plural cell occurs within the sequence shared by both cells.

3.4 *Stem alternations and suppletion*

The previous examples showed that stems correspond to continuous strings to a differing degree; in fact, alternations within stems are common across languages. Stem alternations can be defined as phonological changes within the material expressing the lexical meaning of a word across the cells of a paradigm (cf. Paster 2016; Baerman and Corbett 2012). As was mentioned in Section 2.2, such alternations do not necessarily provide meaningful insights about the inflectional properties of a system. For inflected forms with stem alternations, the stem and marker extraction method shown in Section 3.3 would result in material being analyzed as part of the marker that could otherwise be considered as belonging to the stem. Therefore, this method runs the risk of detecting fewer zero markers than potentially there could be.

To gauge the effect of marker material resulting from stem alternations, I extracted another set of zero markers$_A$, as defined in (5), by removing material that could be analyzed as a stem alternation. To do so, I determined the position(s) of inflectional affixation for all language and part-of-speech combinations in the dataset. This was done

Table 8: Marker_A extraction

Affix position	Removal	Marker	Marker_A
pfx	remove infixes and suffixes	pfx-ifx-sfx	pfx-
sfx	remove prefixes and infixes	pfx-ifx-sfx	-sfx
pfx + sfx	remove infixes	pfx-ifx-sfx	pfx-sfx
ifx + sfx	remove prefixes	pfx-ifx-sfx	-ifx-sfx
pfx + ifx + sfx	/	pfx-ifx-sfx	pfx-ifx-sfx

based on language descriptions and on the extracted stems and markers used in this study. Given the observed patterns, I distinguished between the following five categories of affix position: prefix, suffix, prefix + suffix, infix + suffix, prefix + infix + suffix.[17] Using this classification, all material that had originally been assigned to the marker but did not occur in a regular affix position for a given language and part-of-speech was removed. A schematic overview of this step is shown in Table 8. For instance, if a language and part-of-speech combination is classified as having prefixes only, all additional material that would be classified as an infix or suffix was removed. Similarly to the first step of stem and zero marker extraction, these marker adjustments were automated so that they could be applied systematically for all the languages in the dataset without any additional manual annotations. For the type prefix + infix + suffix only, no additional material could be removed from markers, because all available affix positions were already used by inflectional morphology. The three languages in this category are Arabic, Hebrew, and Maltese; I applied no further changes to the markers in these cases.

The following paragraphs provide a few examples of how markers_A, as defined in (5) and (6) (cf. Section 2.2), were extracted in the presence of stem alternations. One example is a vowel change in German nouns, where a back stem vowel in the singular cells is opposed to a front stem vowel in the plural cells. This is shown for the German noun *Kloß* 'dumpling' in Table 9. All forms are given in the phonological transcription generated with Epitran.

In the case of *Kloß*, the longest common subsequence is not continuous. Due to the umlaut in the plural forms, the automatically ex-

[17] The list of languages and affix position values can be found in `affixation.csv` in the supplementary materials.

Cell	Form	Stem	Marker	$Marker_A$
NOM.SG	klos	kls	-o-	-
ACC.SG	klos	kls	-o-	-
DAT.SG	klos	kls	-o-	-
GEN.SG	kloses	kls	-o-es	-es
NOM.PL	kløsə	kls	-ø-ə	-ə
ACC.PL	kløsə	kls	-ø-ə	-ə
DAT.PL	kløsən	kls	-ø-ən	-ən
GEN.PL	kløsə	kls	-ø-ə	-ə

Table 9:
Marker extraction
of the German noun
Kloß 'dumpling'

tracted stem of *Kloß* consists of the three consonants *kls*. The vowel change from /o/ in the singular to /ø/ in the plural is analyzed as a part of the cells' markers, respectively. Therefore lemmas such as *Kloß* in German do not have zero marking according to the first method of marker extraction. However, German nouns are classified as using suffixes only for inflection. Adjusting the markers by removing all material that is not a suffix takes into account that the alternation between /o/ and /ø/ is a stem alternation. The $markers_A$ no longer contain infixal material and are analyzed as zero for the nominative, accusative, and dative singular cells. Another process of stem alternation is metathesis. Table 10 shows how this is dealt with in the case of the Hungarian noun *gyomor* 'stomach'. In this example, the final segment *-or* is metathesized when certain affixes are added to the stem. Again, this leads to a situation where the stem does not include the segment undergoing metathesis, and the discontinuous string *ɟomr*

Cell	Form	Stem	Marker	$Marker_A$
NOM.SG	ɟomor	ɟomr	-o-	-
ACC.SG	ɟomrot	ɟomr	-ot	-ot
DAT.SG	ɟomornɒk	ɟomr	-o-nɒk	-nɒk
INSTR.SG	ɟomorːɒl	ɟomr	-o-ːɒl	-ːɒl
TERM.SG	ɟomorig	ɟomr	-o-ig	-ig
ON.ESS.SG	ɟomron	ɟomr	-on	-on
ON.ALL.SG	ɟomorːɒ	ɟomr	-o-ːɒ	-ːɒ
ON.ABL.SG	ɟomorːoːl	ɟomr	-o-ːoːl	-ːoːl
...	...	...	...	...

Table 10:
Marker extraction
for the Hungarian noun
gyomor 'stomach'

Table 11: Marker extraction for the Slovenian adjective *absúrden* ‘absurd’

Cell	Form	Stem	Marker	Marker$_A$
NOM.SG.M.INDF	absúrden	absúrdn	-e-	-
NOM.SG.N	absúrdno	absúrdn	-o	-o
NOM.SG.F	absúrdna	absúrdn	-a	-a
DAT.SG.M	absúrdnemu	absúrdn	-emu	-emu
DAT.SG.N	absúrdnemu	absúrdn	-emu	-emu
DAT.SG.F	absúrdni	absúrdn	-i	-i
…	…	…	…	…

is analyzed as the stem. This in turn leads to the infixal marker *-o-* in the NOM.SG cell, for instance. However, Hungarian only uses suffixation for nominal inflection, and the NOM.SG is usually (81% in this dataset) not overtly marked in Hungarian. Therefore the adjusted markers$_A$ no longer feature material that is infixal, and the NOM.SG is zero marked for the noun *gyomor* as well.

Another example of stem-internal alternations is epenthesis, the addition of phonological material in the stem in some but not all of the cells in the paradigm. One example of epenthesis is found with certain types of adjectives in Slovenian, which feature stem-final consonant clusters. This can be seen with the adjective *absúrden* ‘absurd’ in Table 11. Similarly to the previous examples, Slovenian adectives only use suffixation to mark inflection. In Table 11, in all but one inflected form, the stem ends in the cluster */rdn/*, and an overt suffix is added to the stem. The NOM.SG.M.INDF cell, however, is not marked by an additional suffix. Instead, the epenthetic vowel */-e-/* is inserted between the stem-final consonants to break up the consonant cluster. The adjusted markers$_A$ remove all infixal material for Slovenian adjective markers, which results in the NOM.SG.M.INDF cell being analyzed as zero marked.

Stem alternations are relevant in yet another way in Tlatepuzco Chinantec (Otomanguean). This language has a complex inflectional paradigm, combining various patterns of stem and tone changes. Table 12 shows the inflectional paradigm of the verb *køgʔ*^{2} ‘eat’. The forms of *køgʔ*^{2} have different tones for first vs. second and third person forms in all three tenses. Given that the tones are represented by superscript numbers following the tone-bearing unit, they are taken into account by the extraction and detection of zero markers. While

Cell	Form	Stem	Marker	Marker$_A$
PRS.1SG	køgʔ^{12}	køgʔ	-^{12}	-
PRS.1PL	køgʔ^{12}	køgʔ	-^{12}	-
PRS.2	køgʔ^{2}	køgʔ	-^{2}	-
PRS.3	køgʔ^{2}	køgʔ	-^{2}	-
PST.1SG	mi^{3}-køgʔ^{12}	køgʔ	mi^{3}-^{12}	mi^{3}-
PST.1PL	mi^{3}-køgʔ^{12}	køgʔ	mi^{3}-^{12}	mi^{3}-
PST.2	mi^{3}-køgʔ^{2}	køgʔ	mi^{3}-^{2}	mi^{3}-
PST.3	mi^{3}-køgʔ^{2}	køgʔ	mi^{3}-^{2}	mi^{3}-
FUT.1SG	køgʔ^{13}	køgʔ	-^{13}	-
FUT.1PL	køgʔ^{13}	køgʔ	-^{13}	-
FUT.2	køgʔ^{3}	køgʔ	-^{3}	-
FUT.3	køgʔ^{1}	køgʔ	-^{1}	-

Table 12: Marker extraction for the Tlatepuzco Chinantec verb *køgʔ*^{2} 'eat'

present and future tense forms do not make use of an additional segmental marker, the tone annotations are extracted as marker material. Given that otherwise Tlatepuzco Chinantec verbs only use prefixation, I removed all infixal and suffixal material for the adjusted markers$_A$. As can be seen in Table 12, the adjusted markers$_A$ now capture tonal changes as changes to the stem, and the present and future tense cells are now taken to be zero marked. Although this automated way of accounting for stem alternations is able to deal with almost all of the relevant cases, there is one type of alternation that this method cannot capture. If a stem alternation occurs at the edge between stem and affix, then the extraction methods used for this study are not able to detect that the boundary between marker and stem should occur in a different position.

One example is the so-called consonant gradation in Northern Saami (Uralic). It can be described as an alternation of the final stem consonants across the cells of the paradigm, leading to their weakening or strengthening (cf. Bakró-Nagy 2022). An example of Northern Saami adjectives is shown in Table 13. We see that the final stem consonant of the adjective *aiddolaš* 'exact' alternates between */-š/*, */-čč/* and */-žž/*. The extraction process used here analyzes this alternation as part of the marker. By contrast, the adjective *bahá* 'angry' shows the marker extraction for adjectives with no stem alternations. For such adjectives, the NOM.SG, ACC.SG, and GEN.SG cells are zero marked.

Table 13: Marker extraction for the Northern Saami adjectives *aiddolaš* 'exact' and *bahá* 'angry'

Cell	Form	Stem	Marker$_{(A)}$	Form	Stem	Marker$_{(A)}$
NOM.SG	aiddolaš	aiddola	-š	bahá	bahá	-
ACC.SG	aiddolačča	aiddola	-čča	bahá	bahá	-
GEN.SG	aiddolačča	aiddola	-čča	bahá	bahá	-
ILL.SG	aiddolažžii	aiddola	-žžii	bahái	bahá	-i
COM.SG	aiddolaččain	aiddola	-ččain	baháin	bahá	-in
FRML.SG	aiddolažžan	aiddola	-žžan	bahán	bahá	-n
PRP.SG	aiddolaččas	aiddola	-ččas	bahás	bahá	-s

Thus, in cases of alternation at the edge between the stem and the inflectional affix, this method of marker extraction is unable to detect zero marking.

In its most extreme form, a stem alternation that includes the edge segments of stems is suppletion. Suppletion refers to stem alternations where maximally different phonological forms are used to express the same lexical component of an inflected word form across different cells of the paradigm (cf. Mel'čuk 1994; Corbett 2007). Suppletive forms go beyond alternations that can be described in terms of phonological or prosodic relations between forms (at least synchronically). Consider the English examples given in Table 14, where we see the verbs *think* and *go*, both with suppletive stems. In the case of *think*, the suppletion does not affect the entire stem, as the initial segment *θ-* is found in all cells of the paradigm. As a consequence, the extracted marker ends up with all the remaining material (which would usually be analyzed as being part of a suppletive stem). In the case of *go*, suppletion is complete in that no segment is shared between all cells of the paradigm. The complete phonological strings of each form are thus extracted as markers of their respective cells. As the examples from Northern Saami and English showed, neither marker extraction method used

Table 14: Marker extraction for the English verbs *think* and *go*

Cell	Form	Stem	Marker$_{(A)}$	Form	Stem	Marker$_{(A)}$
NFIN	θɪŋk	θ	-ɪŋk	gow	-	gow
PRS.3SG	θɪŋks	θ	-ɪŋks	gowz	-	gowz
PTCP.PRS	θɪŋkɪŋ	θ	-ɪŋkɪŋ	gowɪŋ	-	gowɪŋ
PST	θɔt	θ	-ɔt	wɛnt	-	wɛnt
PTCP.PST	θɔt	θ	-ɔt	gɔn	-	gɔn

for this study has a principled way of removing alternating stem segments that are adjacent to affixal material from the marker. Therefore neither method detects potential zero marking with suppletive forms, as they will always assign phonological material to the marker. While it is possible to exclude markers that occur only once per cell (cf. Section 3.5), many suppletive forms do not correspond to such hapax legomenon markers. Especially larger datasets often include complex lemmas such as *overthink* or *undergo* in English, for example. The extracted markers *-gow* and *-ɪŋk* from Table 14 occur 11 times in the verbal paradigms of English. The stem alternation pattern shown for Northern Saami in Table 13 occurs systematically (26 times) in the dataset. In such cases, I do not have any principled way of excluding the markers from the analysis.

To remain agnostic about the effect of stem alternations and to apply a systematic approach to all languages, I performed the analyses in Sections 4 and 5 for both sets, markers and markers_A. Since the results were very similar with no substantial differences, I only report the results of markers_A, for reasons of brevity. Details about the results based on the originally extracted markers can be found in the supplementary materials as indicated in the relevant sections. Given that no substantial differences were found for the distribution of zero markers in inflection paradigms, I only analyze the distribution of markers_A in the corpus data in Section 6. Whenever markers are mentioned in the following sections, I refer to markers_A, if not stated otherwise.

Hapax legomenon markers 3.5

The dataset includes a number of markers that occur only once per cell for a given language and part-of-speech combination. Some of these hapax legomenon markers are the result of stem alternations, but most of them result from the remaining errors in the dataset. In total, I identified the following number of hapax legomenon markers: 9,223 for adjectives, 23,539 for nouns, and 54,768 for verbs. In terms of marker types, hapax legomenon markers make up a large proportion, namely 0.45, 0.46, and 0.42 for adjectives, nouns, and verbs, respectively. In terms of the total number of occurrences, however, they only amount to a proportion of 0.003 for adjectives, 0.008 for nouns, and 0.03 for verbs.

One example of a hapax legomenon marker as the result of stem alternation comes from Northern Saami. The adjective *čáppat* 'pretty' features gradation similarly to the example shown in Table 13. In this case, stem-final *-pp* alternates with *-bb* across cells of the paradigm. This type of alternation is only attested once in the dataset, making all markers extracted from the lemma *čáppat* hapax legomenon markers.

Most hapax legomenon markers, however, result from remaining material that is not part of the inflected word forms or from errors in the automatic phonological transcription performed by Epitran. To give one example, in the Hungarian dataset, the impersonal verb *fái* 'hurt' features the string "only3rdpersonforms" as the verb form in a number of cells. This string is of course not a Hungarian verb form, but an additional linguistic annotation, which causes the extraction of the longest common substring to find nonsensical strings and hence hapax legomenon markers.

Visual inspection of the hapax legomenon markers suggests that most result from the automatic phonological transcription using Epitran. For instance, the German adjective *maːkabeʁə* 'macabre' shows an alternation between stem-final *-b* and *-p* in the phonological transcription. All forms except the comparative form have *-b*, while the comparative form *maːkapʁɒ* has *-p*, which leads to hapax legomenon markers.

In order to exclude such markers, as they do not provide much insight into the distribution of zero marking, I removed all hapax legomenon markers from the dataset. Given that their proportion of the total number of observations is very low, it is safe to assume that their removal will not artificially distort the distribution of zero markers.

3.6 *Morphomic paradigms*

Another potential factor influencing the distribution of zero marking is the distribution of inflected word forms across the paradigm. Many paradigms have syncretic cells, where a single form expresses more than one cell. Taking this into account and considering only the different forms that are found in a paradigm may thus lead to different probabilities of zero markers. To examine how much the results change if proportions of zero marking are established using distinct forms only, I collapsed the data into morphomic paradigms (cf. Boyé and Schalchi 2016). Morphomic paradigms consist of all the different forms that a given word can have without taking into account their meaning. Syncretic forms are counted in only once in morphomic paradigms. Section 4 therefore analyzes the distributions of markers in morphomic paradigms in addition to paradigms that include information on cells. The analysis of the effect of token frequency in language use on the distribution of zero marking in Section 6 is also based exclusively on forms, i.e. morphomic paradigms.

ESTIMATING THE PROBABILITY OF ZERO MARKERS 4

Observed distributions 4.1

In order to examine the probability of zero markers in adjectival, nominal, and verbal inflection, Table 15 and Figure 2 provide an overview of the observed distribution of zero marking in inflection. The second column of Table 15, "N forms zero", shows the number of inflected word forms across parts-of-speech that are zero marked. The third column, "prop forms zero", indicates the proportion of zero-marked word forms in the entire dataset. We see that the proportions of zero markers are very low for adjectives; verbs show a somewhat higher proportion, and nouns have the highest proportions of zero marking at about 0.1. Zero marking is clearly not common in inflection of any of the parts-of-speech. The last two columns of Table 15 show the number of cells where zero marking is absent and the number where zero marking is used for all lemmas. Unsurprisingly, we find a high number of cells

pos	N forms zero	prop forms zero	N cells no zero marking	N cells all zero marking
adj	45,859	0.007	1,439	12
noun	648,859	0.104	1,227	5
verb	141,268	0.032	3,771	26

Table 15: Observed proportions of zero markers

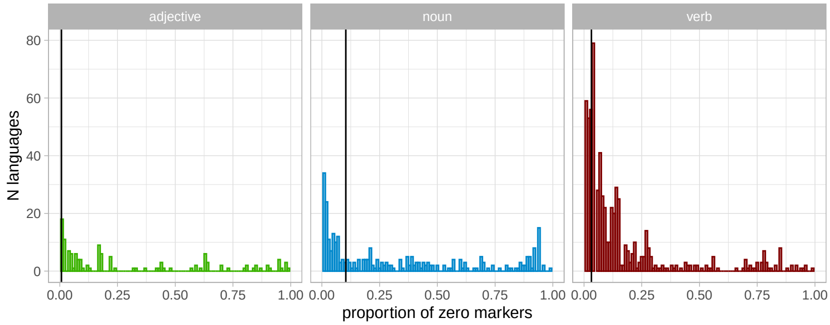

Figure 2: Number of languages by part-of-speech and proportion of zero markers (solid lines correspond to the overall proportions of zero markers)

with no zero marking at all, and only a very small number of cells that feature zero marking consistently across all lemmas.[18] For the last two columns, we find an increasing number of cells from nouns to adjectives to verbs. This reflects the number of cells that those three parts-of-speech distinguish in the dataset, with 727, 961, and 2,753 cells for nouns, adjectives, and verbs, respectively. Figure 2 shows a histogram of the proportions of zero marking in adjectival, nominal, and verbal inflection. The overall proportions are indicated by a vertical line. We can see that they vary to a great extent across languages and parts-of-speech. All three parts-of-speech exhibit a preference for proportions of 0 or close to 0. This preference is most pronounced for adjectives and verbs. For nouns, we find a more balanced distribution, with more proportions above 0.5 for zero marking.

There are five additional factors that are relevant for estimating the probability of zero markers in inflection: the number of cells in a paradigm, the number of morphosyntactic values expressed per cell, the number of lemmas for which paradigms are available, the usual affix position, and the type of phonological representation. The number of cells in a paradigm can be taken as a measure of paradigm size. It is an important factor to include, since it is possible that zero markers are less likely to occur in a larger paradigm that makes more morphosyntactic distinctions. Table 16 gives an overview of the number of cells per paradigm in the dataset, showing the minimum, maximum,

[18] The figure of 26 cells that are expressed by zero markers exclusively is rather high; this can in part be explained by many cells in the verbal paradigm that only occur in single languages.

	min	max	Q1	median	Q3
adjective	3	256	13.5	26	51
noun	2	256	8.5	14	23.5
verb	2	432	15	30	50.5

Table 16: Number of cells

median, the first and the third quartile. As the number of cells spans several magnitudes, I use log-transformed values for the analysis.

Another important factor for estimating the probability of zero marking is the number of morphosyntactic values expressed per cell.[19] For the purposes of this study, we can take the number of values per cell to represent the semantic complexity of the inflectional markers. A summary of the number of values per cell is shown in Table 17.

	min	max	Q1	median	Q3
adjective	1	5	2	3	3
noun	1	4	2	2	2
verb	1	7	1.75	2	2.25

Table 17: Number of morphosyntactic values per cell

Including this factor in the analysis is important, since one could expect more complex markers (which express more complex meanings) to be encoded by more material. The average number of lemmas for which inflectional paradigms are available is not inherently related to the probability of zero marking, but may influence it. As can be seen in Table 18, the median number of lemmas differs greatly across

	min	max	Q1	median	Q3
adjective	17	98464	131	507	1994
noun	23	235294	248	1240	4591
verb	26	30032	109	374	910

Table 18: Number of lemmas

languages. It is therefore an important factor to be controlled for. Another factor that is included in the analysis for its potential effect on the probability of zero marking is the position of the marker regarding the stem. As described in Section 3.3, I distinguish between five affix

[19] For the remainder of this study, I will use "values" to refer to "morphosyntactic values".

Table 19: Affix position

	pfx	pfx + sfx	pfx + sfx + ifx	sfx	sfx + ifx
adjective	36	259	48	1365	0
noun	8	84	62	1436	2
verb	407	889	164	3093	8

positions found in the dataset. Table 19 shows the number of cells per part-of-speech expressed by markers in the five positions. For the analysis, I merged the two positions that include infixes, because the sfx + ifx category on its own has too few observations to allow for any meaningful insights. This leaves the following four values of affix position that are considered in the analysis: `pfx`, `pfx+sfx`, `sfx`, and `has_ifx`.

4.2 *Modelling the probability of zero marking*

To estimate the probability of zero marking in inflection, I aggregated the data by type of cell, language, and part-of-speech. This means that each datapoint corresponds to a proportion of zero marking (0.81) for a given type of cell (NOM.SG) in a given language (Hungarian) for a given part-of-speech (noun). As shown in Table 15, the dataset contains cells with proportions of zero marking that equal 0 or 1. Therefore I fitted a Bayesian zero-one-inflated beta regression model. Zero-one-inflated beta regression models consist of two components. The first component is the regular beta regression model, which deals with proportion values within the interval (0,1). The second component is a logistic regression component that estimates the probability of either of the extremes 0 or 1 as opposed to the proportion data within (0,1).

The models were fitted 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 languages using a phylogenetic regression term, following the method described in Guzmán Naranjo and Becker 2022. 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 individual languages to co-vary according to the tree.[20] In other words,

[20] The phyologenetic tree is taken from Glottolog (Hammarström *et al.* 2021). For details, see `code-phylogeny.R` in the supplementary materials.

if two languages share many nodes on the tree, the model forces their coefficients to be very similar. If two languages are not related at all, the model allows their estimates to vary freely. For instance, if five closely related languages have very high observed proportions of zero markers in a given cell, the model does not take those five observations as independent data points, but assigns much less confidence and/or lowers the predicted probability of zero marking in that cell.

The final model predicts the probability of zero marking from the part-of-speech, affix position, number of values per cell, number of lemmas, and orthographic representation. In addition, I used type of cell and phylogenetic relations between languages as group-level effects.[21] Figures 3 and 4 show the conditional effects for the different predictors for the beta and the zero-one-inflation components, respectively.[22] The points and solid lines correspond to the mean values of the posterior distributions; the error bars and error bands show the 95% credible interval. This approach allows a straightforward interpretation: given the data and the model, we can be 95% certain that the estimated values lie within that interval. Note that the three numerical predictors are all standardized, so that they have a mean of zero and a standard deviation of 1.

From Figure 3, we see that none of the predictors has a clear impact on the probability of zero marking within the interval (0,1). Across all predictors, the mean predictions lie between 0.15 and 0.3. The results thus show that the probability of zero marking to occur, excluding systematic absence or presence thereof, does not depend much on the predictors explored here. This does not necessarily mean that a better model is needed. It suggests that there is a high degree of idiosyncratic variation across languages, and that no clear association

[21] To select a reasonable combination of predictors, I fitted several models and compared their performance using approximated leave-one-out cross-validation as described by Vehtari *et al.* (2017). Due to the low number of proportions of 1, I modelled conditional-one-inflation with an intercept-only model. See `code-prob.R` in the supplementary files for details on the conditional-one-inflation.

[22] I only report the results of the model based on markers$_A$ which allow for stem alternations. All conditional effects of the model based on markers without stem alternations can be found in `ce-probcheck-mu-<predictor>.pdf` and `ce-probcheck-zoi-<predictor>.pdf` in the supplementary materials.

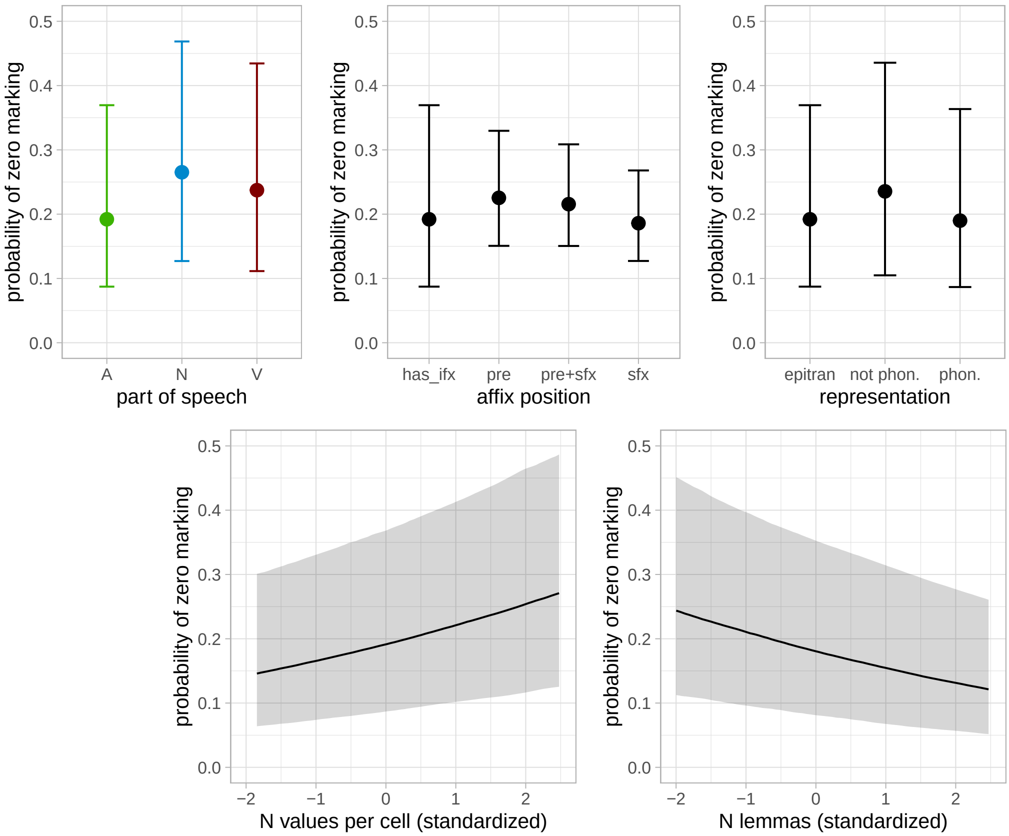

Figure 3: Conditional effects for the beta regression component

can be established with other relevant grammatical properties of the inflectional systems, at least not the ones tested here.

Figure 4 shows the model results for the zero-one-inflation component. It predicts the probability of a cell being exclusively zero marked or never zero marked, as opposed to probability values between those two extremes. As was shown in Table 15, no zero marking per cell is common in the data (6,437 markers out of 7,861), while exclusively zero marked cells are very rare (43 markers out of 7,861). This means that zero-one-inflation predictions largely correspond to the probability of no zero marking for a given cell. We can thus interpret the conditional effects shown in Figure 4 as the probability of the absence of zero marking. For the predictors part-of-speech, affix po-

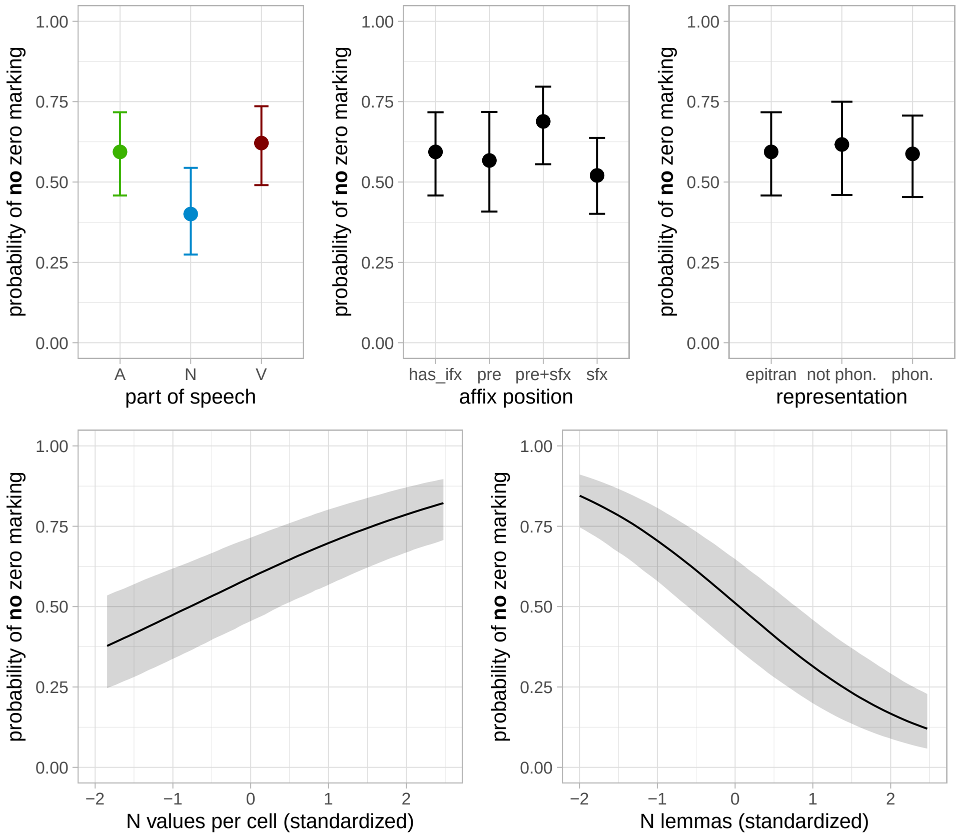

Figure 4: Conditional effects for the zero-one-inflation component

sition, and phonological representation, we find no substantial trends regarding a preference against zero marking. For part-of-speech, adjectives and verbs appear to have a slightly higher probability than nouns of avoiding zero marking altogether, but we have little certainty about this difference. The same can be said about the affix order pfx + sfx; it has a slightly higher tendency to avoid all zero marking than the other positions, but no clear picture emerges.

In contrast to the predictions from the beta component, we do find clear effects of the number of values per cell and the number of lemmas. The more lemmas are available, the lower the probability of encountering not a single case of zero marking. This is expected and shows that the number of lemmas needs to be controlled for. The number of values per cell has a positive effect on the probability of avoiding zero marking altogether. While cells expressing fewer values show no strong preference for or against zero marking, the model predicts a strong preference against all zero marking for cells with many values. This does not restrict where zero marking is likely to occur, but it predicts the total absence of zero marking for complex cells, with a high probability of 0.8.

As mentioned in Section 3.6, it is important to consider the distribution of zero marking in morphomic paradigms as well. I fitted another Bayesian zero-one-inflated beta regression model using morphomic paradigms with the same predictors as described above. Only the predictors including information on cells (cell, number of values per cell) were no longer included. The predictions from the beta regression component are similar to those of the full paradigms, which is why I do not discuss them here in detail.[23] The overall predicted probability of zero marking is just below 0.2, which is slightly lower than in full paradigms. This suggests that zero marking is syncretic in a portion of the dataset. As the credible intervals are very wide in both models and overlap, we cannot be very certain about this finding. For the zero-one-inflation component of the model, the conditional effects of part-of-speech and affix position allow for additional insights. The model predictions for these two variables are shown in Figure 5. We

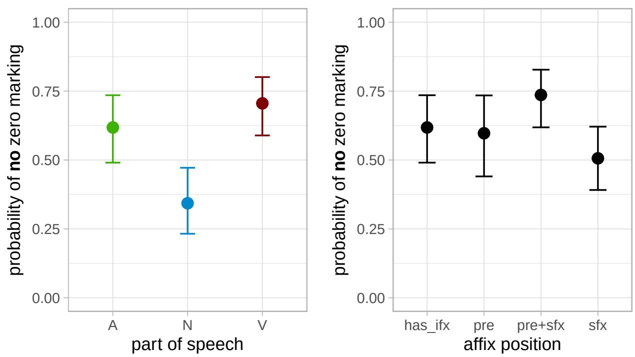

Figure 5: Conditional effects for the zero-one-inflation component of morphomic paradigms

[23] See the file `code-morphomic.R` for details. The conditional effects for all predictors of the model using morphomic paradigms are found in the supplementary materials as `ce-probmorph-mu-<predictor>.pdf` and `ce-probmorph-zoi-<predictor>.pdf`.

see that the patterns are similar, only the differences between parts-of-speech are much stronger now. With morphomic paradigms, we can be certain that verbs and adjectives have a stronger tendency than nouns to avoid zero marking altogether. The same holds for the affix position. Figure 5 shows that systems with prefixes and suffixes are more likely to avoid zero marking altogether than systems with suffixes only.

FUNCTIONS ASSOCIATED WITH ZERO MARKING 5

Cells with the highest probability of zero marking 5.1

To explore which cells are most likely to be zero marked, I subsetted the dataset to include only those cells with a proportion of zero forms ≥ 0.1 in at least 10% of the languages. Subsetting the data in such a way was necessary because of the high number of cell types. The threshold is a heuristic chosen to restrict the following analysis only to cells with a reasonable cross-linguistic probability of being expressed by zero markers. This steps retains the 18 types of cells that show the strongest association with zero marking in the observed distributions.[24]

In order to estimate the probability of zero marking in these cells, I fitted a Bayesian beta regression model that predicts the probability of zero marking from the type of cell.[25] In addition, I added the number of values per cell and lemmas as group-level intercepts as well as phylogenetic controls to account for phylogenetic biases in the data.

[24] The exact figures, including the number of languages per cell, are found in `cells-merged.csv` in the supplementary materials.

[25] In this case, I used beta regression instead of zero-one-inflated beta regression for a combined prediction from both processes. To do so, I converted proportions of zero to 0.0000001 and proportions of 1 to 0.9999999. Again, I compared several models using approximated leave-one-out-cross-validation. See `code-cells.R` in the supplementary materials for details.

Figure 6 shows the observed proportions of zero forms (black triangles) together with the model predictions (dots, error bars, and error bands).[26] Again, the dots represent the mean values of the posterior distribution of the zero probabilities, and the error bars and bands show the 95% credible intervals. The observed proportions of zero forms still differ across cells and parts-of-speech, ranging from 0.1 (2SG.PRS verb forms and DAT.SG adjectives) to above 0.7 (INDF.SG nouns). Although adjectives have fewer cells that met the threshold criteria than nouns and verbs, Figure 6 shows that the cells that do meet them have comparatively high proportions of zero marking. In nominal cells, we find a wider range, including the highest overall proportions of zero marking. Verbs show the lowest proportions of zero marking compared to the other parts-of-speech.

When comparing the results of the model with the observed proportions, the predicted probabilities of zero markers reflect the observed proportions, for the most part. The top plot in Figure 6 shows a few differences, though. For some cells, the predicted probability is much lower than their observed proportions, namely for PL.VOC in adjectives, as well as ACC.SG and INDF.SG in nouns. This points to a bias in the observed distributions, which is also reflected in the large credible intervals of the predictions. The PL.VOC cell is featured in four languages of the dataset, namely in Czech, Georgian, Irish, and Sanskrit. In this case, the high proportion of zero marking is mainly an artefact of the data. The PL.VOC cell is exclusively zero marked in the Czech data. Irish has a low proportion of zero marked PL.VOC cells (0.22), and Georgian as well as Sanskrit do not feature zero marking for the PL.VOC cells of adjectives. Thus, in this case, the high overall proportion largely comes from a single language, which is then adjusted to a much lower prediction in the model, together with large credible intervals to indicate the high level of uncertainty. A similar explanation applies to the ACC.SG cell in nouns. It is featured in 26 languages in the dataset, including phylogenetically unrelated languages. However, the higher observed proportion of zero marking is due to high proportions in a few, mostly related, languages with large

[26] All conditional effects of the model based on markers without stem alternations can be found in `ce-cells-check-<predictor>.pdf` in the supplementary materials.

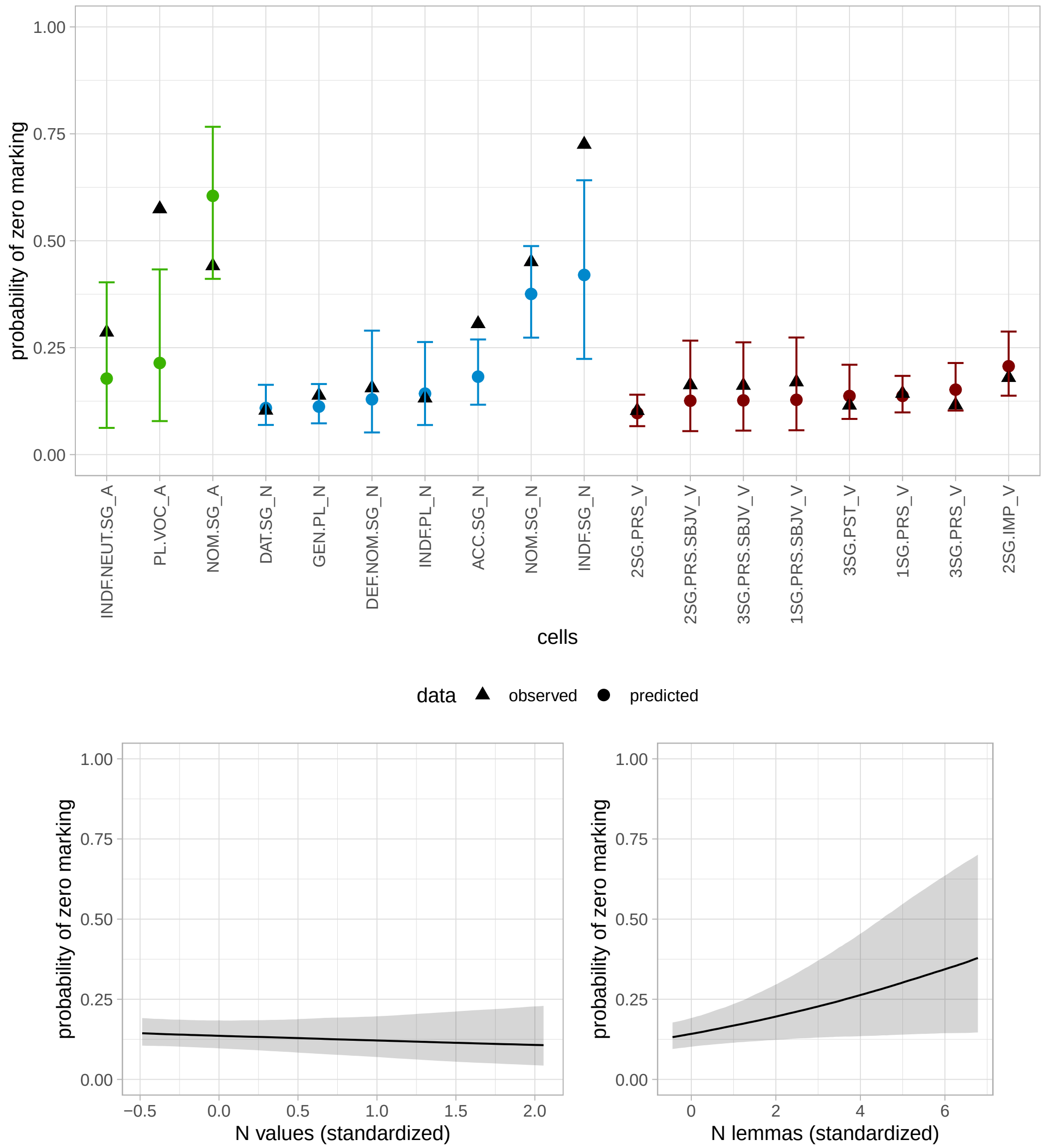

Figure 6: Conditional effects for cells most strongly associated with zero marking

datasets.[27] For the INDF.SG cell, the lower predicted probability of zero marking is also the consequence of a bias in the observed proportions. Here, the bias comes from Norwegian Bokmål, which makes up more than 50% of all observations for this cell, and which has a very high proportion (0.88) of zero marking.

Comparing the predictions across cells and parts-of-speech, we see that adjectival cells have a very high probability of being zero marked. This is noteworthy, as adjectives had only very few cells that met the threshold to begin with. While generally not associated with zero marking, the adjectival cells that are zero marked appear to be those with the strongest association with zero marking across parts-of-speech. Nominal cells are generally predicted to have lower probabilities of zero marking, except for the NOM.SG and the INDF.SG cells, which rank second and third for the predicted probability of zero marking. All verbal cells range between 0.1 and 0.25 for the probability of zero marking. The cell that stands out for having the highest probability of zero marking is the 2SG.IMP cell, which will be further discussed in Section 7.2.

5.2 *Values with the highest probability of zero marking*

The fact that the languages in the dataset differ with respect to the combinations of values in single cells makes it somewhat difficult to assess the association between zero marking and cells that are less common in the dataset. It is therefore important to consider the association of single grammatical values and zero marking as well. Note that, due to the way in which zero markers were extracted, pulling apart the values of cells and analyzing their association with zero marking does not translate directly into the traditional analysis of an abstract feature value, e.g. singular, as being zero marked. Rather, the singular value being expressed by a zero marker refers to all cells in the dataset that encode singular (potentially besides other feature values) and that are zero marked.

[27] This includes German (0.77), Old English (0.50), Finnish (0.37), Russian (0.35), Ukrainian (0.23), Polish (0.22), and Serbian-Croatian-Bosnian (0.30).

In order to examine the association of single values with zero marking, I applied a similar threshold heuristic as in Section 5.1 to select those values that show the strongest association with zero marking. I only included values with an overall proportion of zero marking ≥ 0.02 that are featured in 10% of the languages per part-of-speech. This led to the selection of 21 values in total.[28] To assess how robust the observed proportions of zero marking are, I fitted a Bayesian beta regression model, adding a phylogenetic control and the number of cells and lemmas as group-level effects.[29]

Figure 7 shows the observed proportions (triangles) together with the model predictions (dots, lines).[30] The dots represent the mean values of the posterior distribution of the zero marker probabilities; error bars and bands indicate the 95% uncertainty intervals. The distributions in Figure 7 mostly mirror the tendencies seen in Figure 6 in the previous section. Almost all values that meet the threshold (and are thus the values with the highest proportions of zero marking) have also been part of the cells most likely to be zero marked. Only the nominal value VOC, and the verbal values PROG, PL, and NFIN have not been part of the cells most associated with zero marking. Compared to cells, values show much lower absolute proportions of zero marking. This is expected, since single values potentially occur in many different contexts, not all of which are necessarily zero marked. As for the three parts-of-speech, we now see the highest proportions for nominal values. Adjectival and verbal values show lower proportions of zero marking.

Turning to the model predictions, we see that in the case of values, the probability of zero marking is generally estimated by the model to be higher than the observed proportions. This can be explained by the fact that the model takes into account information on the affix position, the number of cells, and the number of lemmas. The effects of

[28] The exact figures, including the number of languages per value, are found in `values-merged.csv` in the supplementary materials.

[29] I used the same method as for the model described in Section 5.1. See `code-values.R` in the supplementary materials for details.

[30] All conditional effects of the model based on markers without stem alternations can be found in `ce-values-check-<predictor>.pdf` in the supplementary materials.

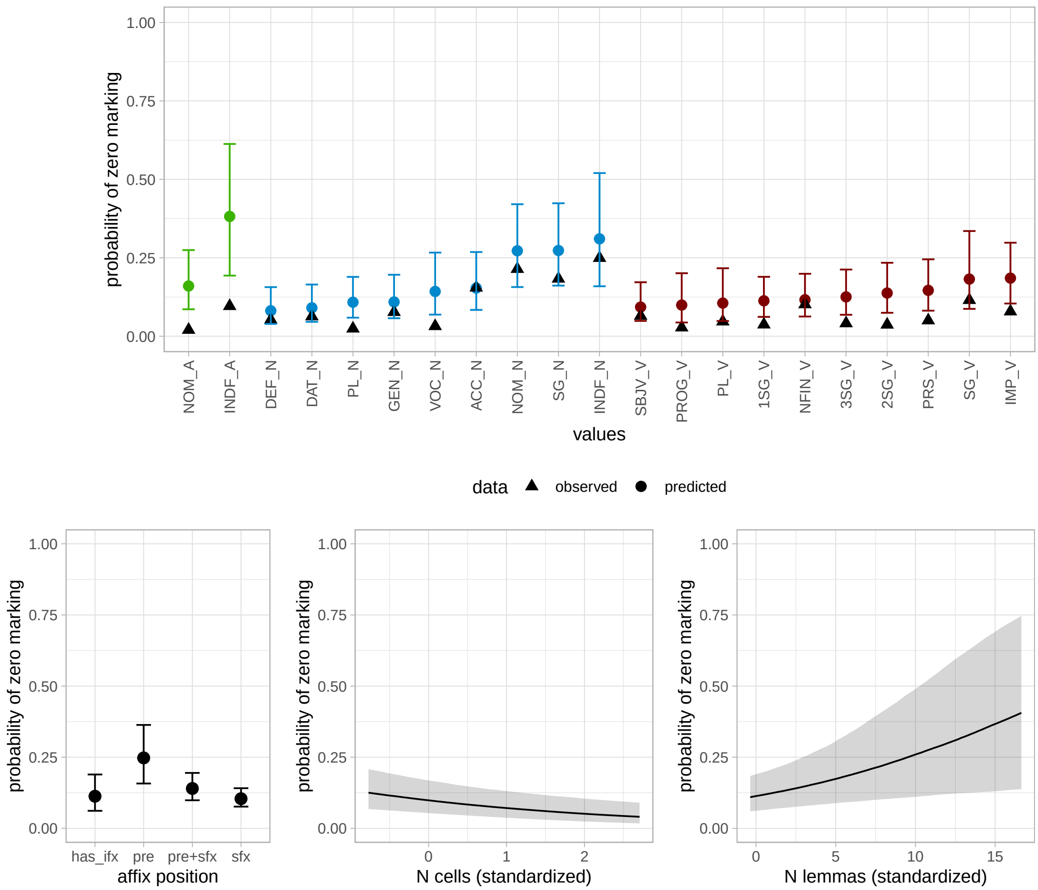

Figure 7: Conditional effects for the values most associated with zero marking.

single values thus correspond to their effects once all the other predictors are controlled for. Interestingly, the affix position is also relevant in this case. The model predicts a higher probability of zero marking for systems with prefixes as opposed to those with suffixes.

The highest predicted probabilities of zero marking are found for the indefinite value in adjectival and nominal inflection. This mirrors the model results of cells shown in Figure 6. Other values with a comparatively high probability of zero marking are SG and NOM for nouns, as well as IMP and SG for verbs. These results also reflect the tendencies seen with cells in Section 5.1.

6 THE FREQUENCY OF ZERO MARKERS IN LANGUAGE USE

To assess the usage frequencies of inflection markers and their phonological length including zero, I analyzed the distribution of zero markers in the Universal Dependencies treebanks (UD) (Zeman *et al.* 2023). To do this, I merged the adjective, noun, and verb forms in UniMorph identified as zero forms with the Universal Dependencies data. I only included the languages for which a phonological transcription was available, so that marker length could be approximated in a more realistic way. From the original dataset, 20 languages have phonological transcriptions and are represented in UD. When merging UniMorph forms with forms in UD, I did not include cell information, but merged the forms purely based on their orthographic representation. The identification of zero markers, however, was based on the phonological transcriptions and the marker$_A$ extraction, as described in Section 3. The resulting dataset contains 9,975 types of markers, which are made up of 51 types of zero markers (across different language and part-of-speech combinations) and 9,924 distinct types of overt markers. In terms of token frequencies, zero markers make up 23% of all the marker occurrences (7,382,497 tokens in total). For the purposes of this study, the distribution of zero and overt markers in UD

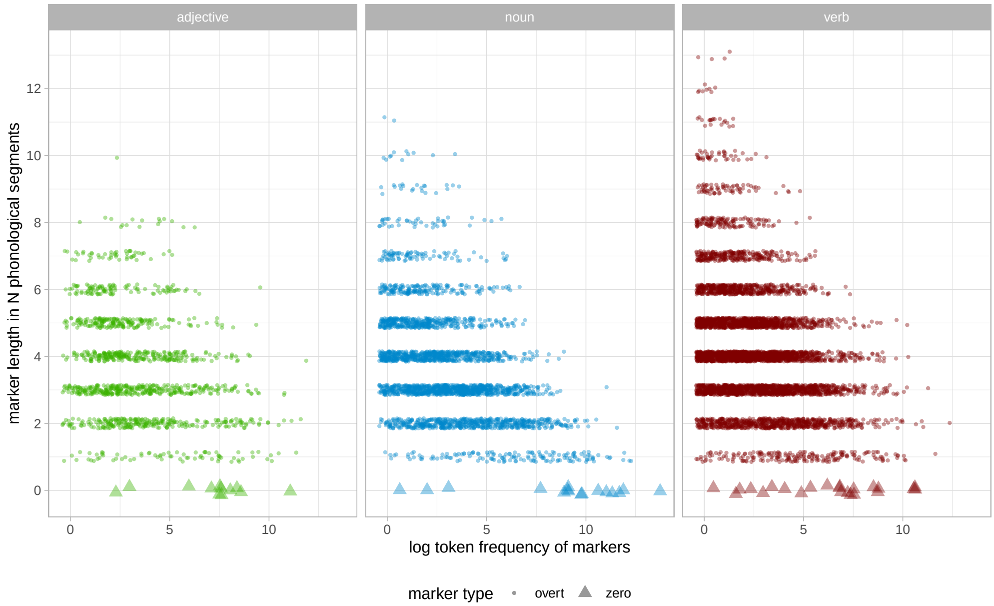

Figure 8: Association between marker token frequency and length

is measured by their log-transformed token frequencies.[31] The length of the markers corresponds to the number of phonological segments identified with the UniMorph dataset. Figure 8 shows the relation between log token frequencies and marker length for adjectives, nouns and verbs. Overt markers are shown as dots, and zero markers are indicated by triangles. As expected, Figure 8 shows a consistent tendency across the three parts-of-speech for more frequent markers to

[31] Frequency is but one of several possible measures of the distribution of linguistic expressions. Common alternatives are contextual probability and informativity (average contextual probability). Some studies suggest that these measures are more strongly associated with the length of an expression (e.g. Barth 2019; Cohen Priva 2015; Jurafsky *et al.* 2001; Levshina 2018; Piantadosi *et al.* 2011). However, which measure is "best" seems to depend on the corpus size and the phenomenon at hand. Given that there is no good suggestion from the literature as to which measure is most strongly associated with expression length in inflectional morphology, the present study uses frequency as a first, straightforward approach. Future research will be necessary to assess efficiency effects using other distribution measures.

be shorter. For less frequent markers, however, there does not seem to be a strong tendency to be longer; we also find many infrequent markers that are short. Figure 8 does not show any clear tendencies for zero markers either. For adjectives and nouns, they appear to have comparatively high frequencies, whereas no such trend is apparent for verbs.

To test the association shown in Figure 8, I fitted a Bayesian hurdle Poisson model, predicting the marker length from their frequencies. Similarly to the zero-one-inflated beta models, a hurdle Poisson model consists of two components. The Poisson component predicts count data, and the hurdle consists of a logistic regression component that predicts the probability of markers of length zero. This allows us to compare the effect of frequency on marker length between zero and overt markers.

In order to determine which predictors other than token frequency should be included, I fitted a series of 9 models that included different combinations of token frequency with part-of-speech, affix position, and number of cells. The performance of these models was then compared to select the final model. I used approximated leave-one-out cross-validation for the comparison, following the method described by Vehtari *et al.* (2017).[32] The final model includes token frequency and affix position as well as their interaction and the phylogenetic control.

Figure 9 shows the conditional effects for the Poisson component, i.e. the part of the model that predicts the length of overt markers. We find a clear negative effect of marker frequency, confirming previous results from the literature. On average, low frequency markers are predicted to be about 0.15 phonological segments longer than high frequency markers. The position of the affix also proves relevant for marker length. Despite the effect being smaller, the model predicts a substantial difference in marker length between systems only using suffixes and all other systems. This becomes more evident when considering the interaction between token frequency and affix position. The effect of frequency is greater for systems using only suffixes than for all other systems, reaching an average difference of 0.25 phonological segments between low-frequency and high-frequency markers.

[32] See `code-ud.R` in the supplementary materials for details.

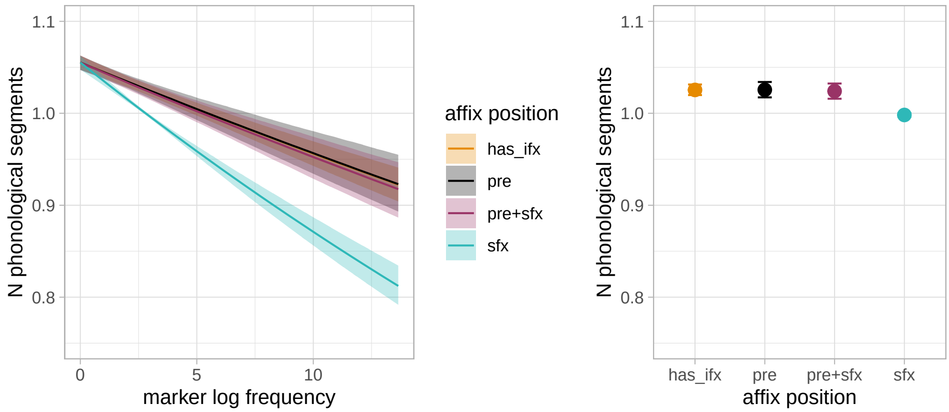

Figure 9: Conditional effects for the Poisson component

We can thus conclude that suffixes are more sensitive to the effect of marker frequency than the other types of affixes. We see the conditional effects for the hurdle component in Figure 10. They represent the effect that the predictors have on the probability of a zero marker occurring.

In stark contrast with the effects predicted for the phonological length of markers, neither token frequency nor affix position affect the probability of a zero marker. The small credible intervals show that this is not an issue of uncertainty or too few observations. We can be confident in the model results that, given the data, the probability of zero marking occurring is not associated with the token frequency of that marker or the affix position that the system uses. This means that there is indeed a clear difference between the effect of frequency on

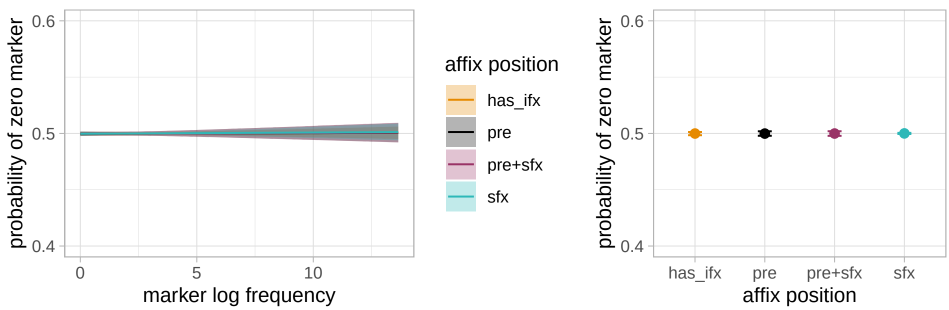

Figure 10: Conditional effects for the hurdle component

marker length in general and the occurrence of zero marking. Zero marking does not simply follow the general trend of marker length being associated with marker frequency.

DISCUSSION 7

The probability of zero marking 7.1

The results of this study allow for a number of important insights into cross-linguistic trends of zero marking in inflection. The model results predicting the probability of zero marking in inflectional paradigms (Section 4) showed three important points. First, zero marking generally affects adjectives, nouns, and verbs fairly equally, and the occurrence of zero marking is not sensitive to the affix position(s) used for inflection. The only notable difference across parts-of-speech and affix positions was found with the total absence of zero marking (zero-one-inflation component). Adjectives and verbs were more likely than nouns to avoid zero marking altogether. The same was seen for systems with prefixes and suffixes as opposed to systems with suffixes only. This effect was shown to be more pronounced when analyzing morphomic paradigms (cf. Figure 5), which are based on forms alone and where syncretic forms are counted only once per paradigm. As the overall probability of zero marking is rather low (0.1–0.3), zero marking is not a default strategy for inflection. This finding provides quantitative support for the proposal by Stolz and Levkovych (2019, 373), who argue that zero marking in inflection should be treated as a “morphological mismatch on a par with established categories such as suppletion and syncretism”. Zero marking is not a common strategy to encode inflection.

Second, we saw an effect of part-of-speech and affix position when analyzing zero marking in morphomic paradigms. Based on forms only, with no information about cells, zero marking was more likely to be absent altogether in adjectives and verbs as opposed to nouns. The same applied to systems with prefixes and suffixes as opposed to suffixes only. This does not mean that nouns and systems with suffixes have a stronger preference for zero marking. It rather suggests that the complete absence of zero marking is less likely in those cases.

Third, an increasing number of values per cell was shown to be a strong predictor for a high probability of zero marking being avoided altogether. The predictor number of values per cell can be taken to quantify how semantically complex a marker is. The fact that more complex cells strictly avoid zero marking is reminiscent of what has been discussed as isomorphism or iconicity in the literature (cf. Haspelmath 2008b; Lehmann 1974; Downing and Stiebels 2012; Givón 1991). While approaches differ in their details, the general idea is that the complexity or amount of linguistic structure reflects the complexity or amount of functional structure (meaning). It remains an open question, however, whether the number of morphosyntactic values per cell reflects functional complexity in the first place, and what the functional motivation for any such effect might be. It is likely that usage distributions and frequencies are a confounding factor, in that cells expressing more values may also be cells that are used less frequently. Their preference for longer markers could thus be a consequence of frequency rather than some iconicity principle.

7.2 *Cells and values associated with zero marking*

Sections 5.1 and 5.2 focused on a selection of cells and morphosyntactic values and their association with zero marking. The results showed that even though zero marking exhibits a high degree of variation across lemmas and languages, it is not distributed randomly across inflectional paradigms. Some cells and values are comparatively likely to be zero marked across languages. For adjectives and nouns, INDF, NOM, and SG (and cell combinations thereof) were the values with the highest predicted probability of zero marking. For verbs, the probabilities of zero marking tended to be lower. The values of IMP, SG, 3, and PRS (and cell combinations thereof) stood out as those with the highest probability of zero marking. The NOM.SG cell for adjectives was the only cell for which the probability of zero marking was predicted to be above 0.5. In other words, this is the only cell for which we can expect zero marking to be more likely than overt marking. In all other cases, predicted probabilities were well below 0.5. This means that the vast majority of inflectional marking is in fact overt, and zero marking is more of an exception.

The values of NOM and SG, as well as their combination, have long been associated with zero marking in the typological literature (e.g. Croft 2003; Greenberg 1963, 1966; Haspelmath and Karjus 2017; Haspelmath 2021; Jakobson 1983[1939]; Koch 1995). Interestingly, there is less discussion in the literature about zero marking of the INDF value, which showed the strongest trend towards zero marking in this study. Two verbal values that have been related to zero marking in the literature are third person (Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010) and present tense (Bybee and Dahl 1989, 55; Bybee 1994, 248). The results of this study confirm the association. Although neither values show a cross-linguistic preference towards being zero marked, they are part of the values with the highest probabilities of zero marking.

Imperatives, especially 2SG forms, have also been mentioned in the literature as being prone to zero marking (e.g. Aikhenvald 2010; Croft 2003; Greenberg 1966; Haspelmath 2021; Koch 1995; Siewierska 2010). The results of the present study thus fit well with the expectations from the literature. Instead of phonetic reduction, previous studies have argued for a functionally motivated non-development scenario for zero marking in (2SG) imperatives. The idea is that the second person is highly recoverable in imperative contexts, e.g. as opposed to contexts of indicative verb forms. Thus, on the level of syntax, many languages allow or require the use of imperatives with no overt second person subject pronoun. This in turn means that the source construction of a verbal person marker is often not available for imperative forms (Aikhenvald 2010, 147; Nikolaeva 2007, 163; Sadock and Zwicky 1985, 173). The cross-linguistically common absence of a suitable source construction for person markers in imperative contexts may thus ultimately account for the high probability of zero marking, especially for person-number agreement values. In addition, the use of bare verb forms for imperatives has been motivated by iconicity (Aikhenvald 2010, 46). According to her, 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.

7.3 *Frequency effects and affix position*

Section 6 examined the association between the token frequency of inflection markers and their length, including zero marking. For overt inflectional markers, the present study provided further evidence of Zipfian effects. Markers with a higher log frequency were predicted to have shorter forms (i.e. number of phonological segments). This corroborates previous findings about form-frequency effects for inflectional markers (cf. Haspelmath and Karjus 2017; Stave *et al.* 2021).

An aspect that has not so far been addressed in quantitative corpus studies is the effect that the position of the inflection marker has. The results from this study showed a clear difference between inflectional systems using only suffixes and those that use different combinations of prefixes, suffixes, and infixes. If inflectional markers are strictly suffixes, their length is predicted to be shorter than if the system uses a combination of affix positions. The effect of token frequency on marker length was also shown to be stronger for suffixes than for other combinations of marker positions. This means that suffixes are more susceptible to frequency effects on marker length than other affix positions are.

A potential explanation for this difference across affix positions is phonetic reduction over time. We know from the literature that phonetic material at the end of words is reduced at higher rates than material at the beginning of words (Bybee *et al.* 1990, 19; Hall 1988). There is also evidence for word-initial (or domain-initial) syllables to be more prominent than other syllables (e.g. Beckman 1998; Smith 2005; Cho *et al.* 2007; Kim 2004; Keating *et al.* 2004). Especially word-initial consonants tend to be strengthened and lengthened (e.g. White *et al.* 2020; Cho and Keating 2009; Fougeron 2001; Cho and Keating 2001). This is relevant, since Bybee *et al.* (1990, 26) find that inflectional prefixes are cross-linguistically significantly more likely to have initial consonants than inflectional suffixes. Taken together, it is plausible that these properties contribute to suffixes being more likely candidates for phonetic reduction over time than affixes in other positions.

Support for the non-development scenario of zero markers 7.4

The other major finding from Section 6 is that the association between token frequency and marker length did not hold for zero markers. Their distributions in the Universal Dependencies treebanks showed that neither token frequency nor affix position were associated with the occurrence of a zero marker. This is evidence against the traditional (implicit) assumption in typology that zero markers behave like short markers in terms of their distribution in language use (e.g. Bybee 2011; Croft 2003; Greenberg 1966; Haspelmath 2021). At the same time, the results from this study confirm previous studies, arguing that coding efficiency and frequency may not be suitable or a sufficient explanation for zero marking in inflectional morphology (Stolz and Levkovych 2019; Guzmán Naranjo and Becker 2021; Bickel *et al.* 2015; Cysouw 2003; Siewierska 2010; Seržant and Moroz 2022).

The difference between overt and zero markers in terms of their association with token frequencies also provides evidence for the non-development scenario leading to zero markers. The other potential mechanism leading to zero marking is phonetic reduction. Phonetic reduction is commonly invoked as the mechanism responsible for the shortening of forms and the development of zero forms (Bybee 2003, 2007, 2015; Givón 2018; Haspelmath 2008a; Lehmann 2015). Bybee (2003, 2015) in particular has argued for phonetic reduction being a consequence of the repetition and automatization in production in the course of grammaticalization.

The main alternative to phonetic reduction is the differential non-development of a marker (cf. Bybee 1994; Cristofaro 2019, 2021; Haspelmath 2008a). For instance, we can imagine a scenario in which number is not marked on nouns at a given point in time. For independent reasons, plural marking could be developed. At the same time that the plural marker develops into an inflectional exponent, its absence becomes more systematically associated with the singular. Then, at some point, the singular is expressed by a zero form. In such a scenario, the zero marker results from the opposition to another new exponent in a different cell of the paradigm.

We can assume that phonetic reduction is at least in part responsible for the patterns found with overt markers, since we found a strong association between token frequency and marker length. Given that such an effect was not found for zero markers, the role of phonetic reduction as the main factor driving their development is questionable. As was mentioned above, the other main mechanism that can lead to the development of zero marking is the differential non-development of an inflection marker. For such a scenario, usage token frequencies may still play a role, but much more indirectly. In a non-development scenario, the zero marker is merely a consequence of the development of a different marker. The development process thus depends on a number of factors that are not directly related to the zero marker itself. The results from Section 6 cannot offer direct evidence in favor of the non-development scenario, but they are more compatible with this scenario than with the phonetic reduction scenario. There is certainly no single answer as to which mechanism leads to zero marking; it is likely that both these mechanisms and others are involved, although probably to differing degrees. Diachronic corpus work is needed to shed more light on the development of zero marking and its cross-linguistic tendencies.

8 CONCLUSION

This study offers the first token-based overview of zero marking in adjectival, nominal, and verbal inflectional morphology across languages. Using the UniMorph dataset, it takes into account the behavior of single lemmas to capture variation across inflection classes and irregular forms. Regarding the probability of zero marking in inflection, the results showed that zero marking is generally not a preferred marking strategy, as it is predicted to occur in only 10–30% of inflected forms. No single cells or values showed a strong association with zero marking. Nevertheless, the values with the highest probability of zero marking (NOM, SG, INDF, 3, PRS, IMP) confirmed earlier observations from the typological literature. The findings further evidenced a high degree of idiosyncratic variation across languages and lemmas in the distribution of zero markers.

In addition, the study analyzed the token frequencies of zero markers together with those of overt markers in several corpora from the Universal Dependencies treebanks. For overt markers, the results showed that the token frequency has a stronger effect on the phonological length of suffixes compared to other affixes. This fits into a broader picture of phonetic differences between suffixes and other positions. For the probability of zero markers, however, no association with their frequency was found. This is new evidence for a fundamental difference between the distribution of overt and zero markers. Zero markers do not simply follow the distributional patterns of short markers. This difference supports a differential non-development scenario of zero marking, rather than a phonetic reduction scenario.

ABBREVIATIONS

1 – first person, 2 – second person, 3 – third person, ABL – ablative, ACC – accusative, ALL – allative, AOR – aorist, COM – comitative, COND – conditional, DAT – dative, DEF – definite, EQTV – equative, ESS – essive, F – feminine, FRML – formal case, FUT – future, GEN – genitive, ILL – illative, IMP – imperative, IMPF – imperfect, INESS – inessive, IND – indicative, INDF – indefinite, INSTR – instrumental, IPFV – imperfective, M – masculine, N – neuter, NFIN – non-finite, NOM – nominative, ON – surface, PFV – perfective, PL – plural, PROG – progressive, PRP – purposive, PRS – present, PST – past, PTCP – participle, SG – singular, SBJV – subjunctive, TERM – terminative, VOC – vocative

REFERENCES

Alexandra AIKHENVALD (2010), *Imperatives and commands*, Oxford University Press, Oxford.

Sergiu AL-GEORGE (1967), The semiosis of zero according to Pāṇini, *East and West*, 17(1):115–124.

Stephen R. ANDERSON (1992), *A-morphous morphology*, Cambridge University Press, Cambridge.

Peter ARKADIEV (2016), Возможны ли однопадежные системы? [Are monocasual systems possible?], in Józefina PIĄTKOWSKA and Gennadij ZELDOWICZ, editors, *Znaki czy nie znaki? – II. zbiór prac lingwistycznych*, pp. 9–37, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa.

Matthew BAERMAN, Dunstan BROWN, and Greville G. CORBETT (2017), *Morphological complexity*, Cambridge University Press, Cambridge.

Matthew BAERMAN and Greville CORBETT (2012), Stem alternations and multiple exponence, *Word Structure*, 5(1):52–68, doi:10.3366/word.2012.0019.

Marianne BAKRÓ-NAGY (2022), Consonant gradation, in Marianne BAKRÓ-NAGY, Johanna LAAKSO, and Elena SKRIBNIK, editors, *The Oxford guide to Uralic languages*, pp. 859–867, Oxford University Press, Oxford.

Sebastian BANK and Jochen TROMMER (2015), Learning and the complexity of Ø-marking, in Matthew BAERMAN, Dunstan BROWN, and Greville CORBETT, editors, *Understanding and measuring morphological complexity*, Oxford University Press, Oxford.

Danielle BARTH (2019), Effects of average and specific context probability on reduction of function words BE and HAVE, *Linguistics Vanguard*, 5(1):20180055, doi:10.1515/lingvan-2018-0055.

Jill BECKMAN (1998), *Positional faithfulness*, Ph.D. thesis, University of Massachusetts, Amherst.

Sacha BENIAMINE and Matías GUZMÁN NARANJO (2021), Multiple alignments of inflectional paradigms, *Proceedings of the Society for Computation in Linguistics*, 4:216–227, doi:10.7275/ymc0-p491.

Balthasar BICKEL, Alena WITZLACK-MAKAREVICH, Taras ZAKHARKO, and Giorgio IEMMOLO (2015), Exploring diachronic universals of agreement: Alignment patterns and zero marking across person categories, in Jürg FLEISCHER, Elisabeth RIEKEN, and Paul WIDMER, editors, *Agreement from a diachronic perspective*, pp. 29–52, De Gruyter, Berlin.

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

James BLEVINS (2005), Word-based declensions in Estonian, in Geert BOOIJ and Jaap VAN MARLE, editors, *Yearbook of Morphology 2005*, pp. 1–25, Springer, Dordrecht, doi:10.1007/1-4020-4066-0_1.

James BLEVINS (2006), Word-based morphology, *Journal of Linguistics*, 42(3):531–573, doi:10.1017/S0022226706004191.

James BLEVINS (2016), *Word and paradigm morphology*, Oxford University Press, Oxford.

Bernard BLOCH (1947), English verb inflection, *Language*, 23(4):399–418, doi:10.2307/410300.

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

Olivier BONAMI (2012), Stems in inflection and lexeme formation, *Word Structure*, 5(1), doi:10.3366/word.2012.0016.

Olivier BONAMI and Sacha BENIAMINE (2021), Leaving the stem by itself, in Sedigheh MORADI, Marcia HAAG, Janie REES-MILLER, and Andrija PETROVIC, editors, *All things morphology: Its independence and its interfaces*, pp. 81–98, Benjamins, Amsterdam, doi:10.1075/cilt.353.05bon.

Sami BOUDELAA and William D MARSLEN-WILSON (2001), Morphological units in the Arabic mental lexicon, *Cognition*, 81(1):65–92, doi:10.1016/S0010-0277(01)00119-6.

Gilles BOYÉ and Gauvin SCHALCHI (2016), The status of paradigms, in Andrew HIPPISLEY and Gregory STUMP, editors, *The Cambridge handbook of morphology*, pp. 206–234, Cambridge University Press, Cambridge.

Dunstan BROWN (1998), Stem indexing and morphonological selection in the Russian verb: A network morphology account, in Ray FABRI, Albert ORTMANN, and Teresa PARODI, editors, *Models of inflection*, pp. 196–224, Niemeyer.

Joan BYBEE (1994), The grammaticization of zero: Asymmetries in tense and aspect systems, in William PAGLIUCA, editor, *Perspectives on grammaticalization*, pp. 235–254, Benjamins, Amsterdam.

Joan BYBEE (2003), Mechanisms of change in grammaticization: The role of frequency, in Brian JOSEPH and Richard JANDA, editors, *Handbook of historical linguistics*, pp. 602–623, Blackwell, Oxford.

Joan BYBEE (2007), *Frequency of use and the organization of language*, Oxford University Press, Oxford.

Joan BYBEE (2011), Markedness, in Jae Jung SONG, editor, *The Oxford handbook of typology*, pp. 1–11, Oxford University Press, Oxford.

Joan BYBEE (2015), *Language change*, Cambridge University Press, Cambridge.

Joan BYBEE and Östen DAHL (1989), The creation of tense and aspect systems in the languages of the world, *Studies in Language*, 13(1):51–103, doi:10.1075/sl.13.1.03byb.

Joan BYBEE, William PAGLIUCA, and Revere PERKINS (1990), On the asymmetries in the affixation of grammatical material, in William CROFT, Suzanne KEMMER, and Keith DENNING, editors, *Studies in typology and diachrony. Papers presented to Joseph H. Greenberg on his 75th birthday*, pp. 1–42, Benjamins, Amsterdam.

Paul-Christian BÜRKNER (2017), Brms: An R package for Bayesian multilevel models using Stan, *Journal of Statistical Software*, 80(1):1–28, doi:10.18637/jss.v080.i01.

Bob CARPENTER, Andrew GELMAN, Matthew HOFFMAN, Daniel LEE, Ben GOODRICH, Michael BETANCOURT, Marcus BRUBAKER, Jiqiang GUO, Peter LI, and Allen RIDDELL (2017), Stan: A probabilistic programming language, *Journal of Statistical Software*, 76(1):1–32, doi:10.18637/jss.v076.i01.

Taehong CHO and Patricia KEATING (2001), Articulatory and acoustic studies on domain-initial strengthening in Korean, *Journal of Phonetics*, 29(2):155–190, doi:10.1006/jpho.2001.0131.

Taehong CHO and Patricia KEATING (2009), Effects of initial position versus prominence in English, *Journal of Phonetics*, 37(4):466–485, doi:10.1016/j.wocn.2009.08.001.

Taehong CHO, James MCQUEEN, and Ethan COX (2007), Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English, *Journal of Phonetics*, 35(2):210–243, doi:10.1016/j.wocn.2006.03.003.

Uriel COHEN PRIVA (2015), Informativity affects consonant duration and deletion rates, *Laboratory Phonology*, 6(2):243–278, doi:10.1515/lp-2015-0008.

Matt COLER (2015), Aymara inflection, in Matthew BAERMAN, editor, *The Oxford handbook of inflection*, pp. 1–30, Oxford University Press, Oxford.

Matt COLER (2018), Subtractive morphology & disfixation in Aymara case, in *Case and Agreement in Panará (... and Beyond)*, pp. 1–9, University of Groningen, Groningen, doi:10.13140/RG.2.2.26153.03682.

Ellen CONTINI-MORAVA (2006), The difference between zero and nothing: Swahili noun class prefixes 5 and 9/10, in Joseph DAVIS, Radmila GORUP, and Nancy STERN, editors, *Advances in Functional Linguistics*, pp. 211–222, Benjamins, Amsterdam.

Greville CORBETT (2007), Canonical typology, suppletion, and possible words, *Language*, 83(1):8–42, doi:10.1353/lan.2007.0006.

Sonia CRISTOFARO (2019), Taking diachronic evidence seriously: Result-oriented vs. source-oriented explanations of typological universals, in Karsten SCHMIDTKE-BODE, Natalia LEVSHINA, Susanne Maria MICHAELIS, and Ilya SERŽANT, editors, *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*, pp. 25–46, Language Science Press, Berlin.

Sonia CRISTOFARO (2021), Typological explanations in synchrony and diachrony: On the origins of third person zeroes in bound person paradigms, *Folia Linguistica*, 55(s42-s1):25–48, doi:10.1515/flin-2021-2013.

William CROFT (2003), *Typology and universals*, Cambridge University Press, Cambridge, 2nd edition.

Michael CYSOUW (2003), *The paradigmatic structure of person marking*, Oxford University Press, Oxford.

Eystein DAHL and Antonio FÁBREGAS (2018), Zero morphemes, in Rochelle LIEBER, editor, *Oxford research encyclopedia of linguistics*, pp. 1–30, Oxford University Press, Oxford, doi:10.1093/acrefore/9780199384655.013.592.

Ferdinand DE SAUSSURE (1916), *Cours de linguistique générale*, in Bally, Charles and Albert Sechehaye, editors, Payot, Lausanne.

Catharine DIEHL (2008), The empty space in structure: Theories of the zero from Gauthiot to Deleuze, *Diacritics*, 38(3):93–119, `https://www.jstor.org/stable/20616535`.

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

Laura J. DOWNING and Barbara STIEBELS (2012), Iconicity, in Jochen TROMMER, editor, *The morphology and phonology of exponence*, Oxford Studies in Theoretical Linguistics, Oxford University Press, Oxford.

Cécile FOUGERON (2001), Articulatory properties of initial segments in several prosodic constituents in French, *Journal of Phonetics*, 29(2):109–135, doi:10.1006/jpho.2000.0114.

Talmy GIVÓN (1991), Isomorphism in the grammatical code: Cognitive and biological considerations, *Studies in Language*, 15(1):85–114, doi:10.1075/sl.15.1.04giv.

Talmy GIVÓN (2018), *On understanding grammar*, Benjamins, Amsterdam.

Joseph GREENBERG (1963), Some universals of grammar with particular reference to the order of meaningful elements, in Joseph GREENBERG, editor, *Universals of language*, pp. 73–113, MIT Press, Cambridge, MA.

Joseph GREENBERG (1966), *Language universals: With special reference to feature hierarchies*, Mouton, The Hague.

Matías GUZMÁN NARANJO and Laura BECKER (2021), Coding efficiency in nominal inflection: Expectedness and type frequency effects, *Linguistics Vanguard*, 7(s3):20190075, doi:10.1515/lingvan-2019-0075.

Matías GUZMÁN NARANJO and Laura BECKER (2022), Statistical bias control in typology, *Linguistic Typology*, 26(3):605–670, doi:10.1515/lingty-2021-0002.

William HAAS (1957), Zero in linguistics description, in John Rupert FIRTH, editor, *Studies in linguistic analysis*, pp. 33–53, Blackwell, Oxford.

Christopher HALL (1988), Integrating diachronic and processing principles in explaining the suffixing preference, in John HAWKINS, editor, *Explaining language universals*, pp. 321–349, Basil Blackwell, London.

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

Martin HASPELMATH (2008a), Creating economical morphosyntactic patterns in language change, in Jeff GOOD, editor, *Linguistic universals and language change*, pp. 185–214, Oxford University Press, Oxford.

Martin HASPELMATH (2008b), Frequency vs. iconicity in explaining grammatical asymmetries, *Cognitive Linguistics*, 19(1):1–33, doi:10.1515/COG.2008.001.

Martin HASPELMATH (2008c), A frequentist explanation of some universals of reflexive marking, *Linguistic Discovery*, 6(1):40–63, doi:10.1349/PS1.1537-0852.A.331.

Martin HASPELMATH (2018), How comparative concepts and descriptive linguistic categories are different, in Daniël VAN OLMEN, Tanja MORTELMANS, and Frank BRISARD, editors, *Aspects of linguistic variation*, pp. 83–114, De Gruyter, Berlin, doi:10.1515/9783110607963-004.

Martin HASPELMATH (2021), Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability, *Journal of Linguistics*, pp. 1–29, doi:10.1017/S0022226720000535.

Martin HASPELMATH, Andreea CALUDE, Michael SPAGNOL, Heiko NARROG, and Elif BAMYACI (2014), Coding causal–noncausal verb alternations: A form–frequency correspondence explanation, *Journal of Linguistics*, 50(3):587–625, doi:10.1017/S0022226714000255.

Martin HASPELMATH and Andres KARJUS (2017), Explaining asymmetries in number marking: Singulatives, pluratives, and usage frequency, *Linguistics*, 55(6):1213–1235, doi:10.1515/ling-2017-0026.

George HEWITT (1995), *Georgian: A structural reference grammar*, Benjamins, Amsterdam.

Jane HILL and Ofelia ZEPEDA (1998), Tohono O'odham (Papago) Plurals, *Anthropological Linguistics*, 40(1):1–42.

Charles HOCKETT (1967), The Yawelmani basic verb, *Language*, 43(1):208–222, doi:10.2307/411395.

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

Daniel JURAFSKY, Alan BELL, Michelle GREGORY, and William RAMOND (2001), Probabilistic relations between words: Evidence from reduction in lexical production, in Joan BYBEE and Paul HOPPER, editors, *Frequency and the emergence of linguistic structure*, pp. 229–254, Benjamins, Amsterdam.

Patricia KEATING, Taehong CHO, Fougeron CECILE, and Chai-Shune HSU (2004), Domain-initial strengthening in four languages, in John LOCAL, Richard ODGEN, and Rosalind TEMPLE, editors, *Phonetic interpretation. Papers in laboratory phonology VI*, pp. 145–163, Cambridge University Press, Cambridge.

Sahyang KIM (2004), *The role of prosodic phrasing in Korean word segmentation*, Ph.D. thesis, University of California, Los Angeles, `https://linguistics.ucla.edu/wp-content/uploads/2021/11/SahyangKim_dissertation.pdf`.

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

Christian LEHMANN (1974), Isomorphismus im sprachlichen Zeichen [Isomorphism in the lingusitic sign], in *Linguistic Workshop II: Arbeiten Des Kölner Universalienprojekts 1973/4*, pp. 98–123, Fink, München.

Christian LEHMANN (2015), *Thoughts on grammaticalization*, Language Science Press, Berlin.

Natalia LEVSHINA (2018), Probabilistic grammar and constructional predictability: Bayesian generalized additive models of help + (to) Infinitive in varieties of web-based English, *Glossa*, 3(1):1–22, doi:10.5334/gjgl.294.

Natalia LEVSHINA (2022), *Communicative efficiency: Language structure and use*, Cambridge University Press, Cambridge.

Martin MAIDEN (1992), Irregularity as a determinant of morphological change, *Journal of Linguistics*, 28(2):285–312, doi:10.1017/S0022226700015231.

Peter Hugoe MATTHEWS (1972), *Inflectional morphology: A theoretical study based on aspects of Latin verb conjugation*, Cambridge University Press.

Arya D. MCCARTHY, Christo KIROV, Matteo GRELLA, Amrit NIDHI, Patrick XIA, Kyle GORMAN, Ekaterina VYLOMOVA, Sabrina J. MIELKE, Garrett NICOLAI, Miikka SILFVERBERG, Timofey ARKHANGELSKIY, Nataly KRIZHANOVSKY, Andrew KRIZHANOVSKY, Elena KLYACHKO, Alexey SOROKIN, John MANSFIELD, Valts ERNŠTREITS, Yuval PINTER, Cassandra L. JACOBS, Ryan COTTERELL, Mans HULDEN, and David YAROWSKY (2020), UniMorph 3.0: Universal Morphology, in Nicoletta CALZOLARI, Frédéric BÉCHET, Philippe BLACHE, Khalid CHOUKRI, Christopher CIERI, Thierry DECLERCK, Sara GOGGI, Hitoshi ISAHARA, Bente MAEGAARD, Joseph MARIANI, Hélène MAZO, Asuncion MORENO, Jan ODIJK, and Stelios PIPERIDIS, editors, *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 3922–3931, European Language Resources Association, Marseille, France, https://aclanthology.org/2020.lrec-1.483.

William MCGREGOR (2003), The nothing that is, the zero that isn't, *Studia Linguistica*, 57(2):75–119, doi:10.1111/1467-9582.00100.

Georg Friedrich MEIER (1961), *Das Zéro-Problem in der Linguistik. Kritische Untersuchungen zur strukturalistischen Analyse der Relevanz sprachlicher Form*, Akademie Verlag, Berlin.

Igor MEL'ČUK (1994), Suppletion: Toward a logical analysis of the concept, *Studies in Language*, 18(2):339–410, doi:10.1075/sl.18.2.03mel.

Igor MEL'ČUK (2002), Towards a formal concept zero linguistic sign: Applications in typology, in Sabrina BENDJABALLAH, Wolfgang DRESSLER, Oskar PFEIFFER, and Maria VOEIKOVA, editors, *Morphology 2000: Selected* *papers from the 9th Morphology Meeting, Vienna, 24–28 February 2000*, pp. 241–258, Benjamins, Amsterdam.

Marianne MITHUN (1986), When zero isn't there, *Annual Meeting of the Berkeley Linguistics Society*, 12(0):195–211, doi:10.3765/bls.v12i0.1882.

Fabio MONTERMINI and Olivier BONAMI (2013), Stem spaces and predictability in verbal inflection, *Lingue e linguaggio*, 2:171–190, doi:10.1418/75040.

David R. MORTENSEN, Siddharth DALMIA, and Patrick LITTELL (2018), Epitran: Precision G2P for many languages, in Nicoletta CALZOLARI, Khalid CHOUKRI, Christopher CIERI, Thierry DECLERCK, Sara GOGGI, Koiti HASIDA, Hitoshi ISAHARA, Bente MAEGAARD, Joseph MARIANI, Hélène MAZO, Asuncion MORENO, Jan ODIJK, Stelios PIPERIDIS, and Takenobu TOKUNAGA, editors, *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 2710–2714, European Language Resources Association, Miyazaki, Japan, `https://aclanthology.org/L18-1429`.

Irina NIKOLAEVA (2007), *Finiteness: Theoretical and empirical foundations*, Oxford University Press, Oxford.

Mary PASTER (2016), Alternations: Stems and allomorphy, in Andrew HIPPISLEY and Gregory STUMP, editors, *The Cambridge handbook of morphology*, pp. 93–116, Cambridge University Press, Cambridge.

Steven PIANTADOSI, Harry TILY, and Edward GIBSON (2011), Word lengths are optimized for efficient communication, *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, doi:10.1073/pnas.1012551108.

Vito PIRRELLI and Marco BATTISTA (2000), The paradigmatic dimension of stem allomorphy in Italian verb inflection, *Rivista di Linguistica*, 12(2):307–380.

Geoffrey PULLUM and Arnold ZWICKY (1991), A misconceived approach to morphology, *Proceedings of the West Coast Conference on Formal Linguistics*, 10.

R CORE TEAM (2021), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, `https://www.R-project.org/`.

Robert RATCLIFFE (1998), *The "broken" plural problem in Arabic and comparative Semitic*, Benjamins, Amsterdam.

Robert Henry ROBINS (1959), In defence of WP, *Transactions of the Philological Society*, 58(1):116–144, doi:10.1111/j.1467-968X.1959.tb00301.x.

Robert Henry ROBINS (1997), *A short history of linguistics*, Routledge, New York, 4th edition.

Jerrold SADOCK and Arnold ZWICKY (1985), Speech act distinctions in grammar, in Timothy SHOPEN, editor, *Language typology and syntactic description. Volume 1*, pp. 155–196, Cambridge University Press, Cambridge.

Gerald SANDERS (1988), Zero derivation and the overt analogue criterion, in Michael HAMMOND and Michael NOONAN, editors, *Theoretical Morphology*, pp. 155–175, Academic Press, San Diego, CA.

Gene SCHRAMM (1962), An outline of Classical Arabic verb structure, *Language*, 38(4):360–375, doi:10.2307/410672.

Ilja SERŽANT and George MOROZ (2022), Universal attractors in language evolution provide evidence for the kinds of efficiency pressures involved, *Humanities and Social Sciences Communications*, 9(1):1–9, doi:10.1057/s41599-022-01072-0.

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

Jennifer SMITH (2005), *Phonological augmentation in prominent positions*, Taylor & Francis, Oxfordshire, doi:10.4324/9780203506394.

Jae Jung SONG (2018), *Linguistic typology*, Oxford University Press, Oxford.

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

Matthew STAVE, Ludger PASCHEN, François PELLEGRINO, and Frank SEIFART (2021), Optimization of morpheme length: A cross-linguistic assessment of Zipf's and Menzerath's laws, *Linguistics Vanguard*, 7(s3), doi:10.1515/lingvan-2019-0076.

Thomas STOLZ and Nataliya LEVKOVYCH (2019), Absence of material exponence, *Language Typology and Universals*, 72(3):373–400, doi:10.1515/stuf-2019-0015.

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

Gregory STUMP and Rafael FINKEL (2013), *Morphological typology: From word to paradigm*, Cambridge University Press, Cambridge.

John SYLAK-GLASSMAN (2016), The composition and use of the Universal Morphological Feature Schema (UniMorph Schema), `https://unimorph.github.io/doc/unimorph-schema.pdf`.

Anna THORNTON (2012), Reduction and maintenance of overabundance. A case study on Italian verb paradigms, *Word Structure*, 5(2):183–207, doi:10.3366/word.2012.0026.

Jochen TROMMER (2012), Ø-exponence, in Jochen TROMMER, editor, *The morphology and phonology of exponence*, pp. 326–354, Oxford University Press, Oxford.

Aki Vehtari, Andrew Gelman, and Jonah Gabry (2017), Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing*, 27(5):1413–1432, doi:10.1007/s11222-016-9696-4.

Laurence White, Silvia Benavides-Varela, and Katalin Mády (2020), Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues?, *Journal of Phonetics*, 81:100982, doi:10.1016/J.WOCN.2020.100982.

Jingting Ye (2020), Independent and dependent possessive person forms: Three universals, *Studies in Language*, 44(2):363–406, doi:10.1075/sl.19020.ye.

Moira Yip (1988), Template Morphology and the direction of association, *Natural Language & Linguistic Theory*, 6(4):551–577, doi:10.1007/BF00134493.

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

Kazunori FUJITA, Katarína GAJDOŠOVÁ, Daniel GALBRAITH, Marcos GARCIA, Moa GÄRDENFORS, Sebastian GARZA, Fabrício Ferraz GERARDI, Kim GERDES, Filip GINTER, Gustavo GODOY, Iakes GOENAGA, Koldo GOJENOLA, Memduh GÖKIRMAK, Yoav GOLDBERG, Xavier GÓMEZ GUINOVART, Berta GONZÁLEZ SAAVEDRA, Bernadeta GRICIŪTĖ, Matias GRIONI, Loïc GROBOL, Normunds GRŪZITIS, Bruno GUILLAUME, Céline GUILLOT-BARBANCE, Tunga GÜNGÖR, Nizar HABASH, Hinrik HAFSTEINSSON, Jan HAJIČ, Jan HAJIČ JR., Mika HÄMÄLÄINEN, Linh HÀ MỸ, Na-Rae HAN, Muhammad Yudistira HANIFMUTI, Sam HARDWICK, Kim HARRIS, Dag HAUG, Johannes HEINECKE, Oliver HELLWIG, Felix HENNIG, Barbora HLADKÁ, Jaroslava HLAVÁČOVÁ, Florinel HOCIUNG, Petter HOHLE, Eva HUBER, Jena HWANG, Takumi IKEDA, Anton Karl INGASON, Radu ION, Elena IRIMIA, Ọlájídé ISHOLA, Kaoru ITO, Siratun JANNAT, Tomáš JELÍNEK, Apoorva JHA, Anders JOHANNSEN, Hildur JÓNSDÓTTIR, Fredrik JØRGENSEN, Markus JUUTINEN, Sarveswaran K, Hüner KAŞIKARA, Andre KAASEN, Nadezhda KABAEVA, Sylvain KAHANE, Hiroshi KANAYAMA, Jenna KANERVA, Neslihan KARA, Boris KATZ, Tolga KAYADELEN, Jessica KENNEY, Václava KETTNEROVÁ, Jesse KIRCHNER, Elena KLEMENTIEVA, Elena KLYACHKO, Arne KÖHN, Abdullatif KÖKSAL, Kamil KOPACEWICZ, Timo KORKIAKANGAS, Mehmet KÖSE, Natalia KOTSYBA, Jolanta KOVALEVSKAITĖ, Simon KREK, Parameswari KRISHNAMURTHY, Sandra KÜBLER, Oğuzhan KUYRUKÇU, Aslı KUZGUN, Sookyoung KWAK, Veronika LAIPPALA, Lucia LAM, Lorenzo LAMBERTINO, Tatiana LANDO, Septina Dian LARASATI, Alexei LAVRENTIEV, John LEE, Phương LÊ HỒNG, Alessandro LENCI, Saran LERTPRADIT, Herman LEUNG, Maria LEVINA, Cheuk Ying LI, Josie LI, Keying LI, Yuan LI, KyungTae LIM, Bruna LIMA PADOVANI, Krister LINDÉN, Nikola LJUBEŠIĆ, Olga LOGINOVA, Stefano LUSITO, Andry LUTHFI, Mikko LUUKKO, Olga LYASHEVSKAYA, Teresa LYNN, Vivien MACKETANZ, Menel MAHAMDI, Jean MAILLARD, Aibek MAKAZHANOV, Michael MANDL, Christopher MANNING, Ruli MANURUNG, Büşra MARŞAN, Cătălina MĂRĂNDUC, David MAREČEK, Katrin MARHEINECKE, Héctor MARTÍNEZ ALONSO, Lorena MARTÍN-RODRÍGUEZ, André MARTINS, Jan MAŠEK, Hiroshi MATSUDA, Yuji MATSUMOTO, Alessandro MAZZEI, Ryan MCDONALD, Sarah MCGUINNESS, Gustavo MENDONÇA, Tatiana MERZHEVICH, Niko MIEKKA, Karina MISCHENKOVA, Margarita MISIRPASHAYEVA, Anna MISSILÄ, Cătălin MITITELU, Maria MITROFAN, Yusuke MIYAO, AmirHossein MOJIRI FOROUSHANI, Judit MOLNÁR, Amirsaeid MOLOODI, Simonetta MONTEMAGNI, Amir MORE, Laura MORENO ROMERO, Giovanni MORETTI, Keiko Sophie MORI, Shinsuke MORI, Tomohiko MORIOKA, Shigeki MORO, Bjartur MORTENSEN, Bohdan MOSKALEVSKYI, Kadri MUISCHNEK, Robert MUNRO, Yugo MURAWAKI, Kaili MÜÜRISEP, Pinkey NAINWANI, Mariam NAKHLÉ, Juan Ignacio NAVARRO HORÑIACEK, Anna NEDOLUZHKO, Gunta NEŠPORE-BĒRZKALNE, Manuela NEVACI, Lương NGUYỄN THỊ, Huyền NGUYỄN THỊ MINH, Yoshihiro NIKAIDO, Vitaly

NIKOLAEV, Rattima NITISAROJ, Alireza NOURIAN, Hanna NURMI, Stina OJALA, Atul Kr. OJHA, Adédayọ OLÚÒKUN, Mai OMURA, Emeka ONWUEGBUZIA, Petya OSENOVA, Robert ÖSTLING, Lilja ØVRELID, Şaziye Betül ÖZATEŞ, Merve ÖZÇELIK, Arzucan ÖZGÜR, Balkız ÖZTÜRK BAŞARAN, Hyunji Hayley PARK, Niko PARTANEN, Elena PASCUAL, Marco PASSAROTTI, Agnieszka PATEJUK, Guilherme PAULINO-PASSOS, Angelika PELJAK-ŁAPIŃSKA, Siyao PENG, Cenel-Augusto PEREZ, Natalia PERKOVA, Guy PERRIER, Slav PETROV, Daria PETROVA, Jason PHELAN, Jussi PIITULAINEN, Tommi A. PIRINEN, Emily PITLER, Barbara PLANK, Thierry POIBEAU, Larisa PONOMAREVA, Martin POPEL, Lauma PRETKALNIŅA, Sophie PRÉVOST, Prokopis PROKOPIDIS, Adam PRZEPIÓRKOWSKI, Tiina PUOLAKAINEN, Sampo PYYSALO, Peng QI, Andriela RÄÄBIS, Alexandre RADEMAKER, Mizanur RAHOMAN, Taraka RAMA, Loganathan RAMASAMY, Carlos RAMISCH, Fam RASHEL, Mohammad Sadegh RASOOLI, Vinit RAVISHANKAR, Livy REAL, Petru REBEJA, Siva REDDY, Mathilde REGNAULT, Georg REHM, Ivan RIABOV, Michael RIESSLER, Erika RIMKUTĖ, Larissa RINALDI, Laura RITUMA, Putri RIZQIYAH, Luisa ROCHA, Eiríkur RÖGNVALDSSON, Mykhailo ROMANENKO, Rudolf ROSA, Valentin ROȘCA, Davide ROVATI, Olga RUDINA, Jack RUETER, Kristján RÚNARSSON, Shoval SADDE, Pegah SAFARI, Benoît SAGOT, Aleksi SAHALA, Shadi SALEH, Alessio SALOMONI, Tanja SAMARDŽIĆ, Stephanie SAMSON, Manuela SANGUINETTI, Ezgi SANIYAR, Dage SÄRG, Baiba SAULTE, Yanin SAWANAKUNANON, Shefali SAXENA, Kevin SCANNELL, Salvatore SCARLATA, Nathan SCHNEIDER, Sebastian SCHUSTER, Lane SCHWARTZ, Djamé SEDDAH, Wolfgang SEEKER, Mojgan SERAJI, Syeda SHAHZADI, Mo SHEN, Atsuko SHIMADA, Hiroyuki SHIRASU, Yana SHISHKINA, Muh SHOHIBUSSIRRI, Dmitry SICHINAVA, Janine SIEWERT, Einar Freyr SIGURÐSSON, Aline SILVEIRA, Natalia SILVEIRA, Maria SIMI, Radu SIMIONESCU, Katalin SIMKÓ, Mária ŠIMKOVÁ, Kiril SIMOV, Maria SKACHEDUBOVA, Aaron SMITH, Isabela SOARES-BASTOS, Shafi SOUROV, Carolyn SPADINE, Rachele SPRUGNOLI, Steinór STEINGRÍMSSON, Antonio STELLA, Milan STRAKA, Emmett STRICKLAND, Jana STRNADOVÁ, Alane SUHR, Yogi Lesmana SULESTIO, Umut SULUBACAK, Shingo SUZUKI, Zsolt SZÁNTÓ, Chihiro TAGUCHI, Dima TAJI, Yuta TAKAHASHI, Fabio TAMBURINI, Mary Ann C. TAN, Takaaki TANAKA, Dipta TANAYA, Samson TELLA, Isabelle TELLIER, Marinella TESTORI, Guillaume THOMAS, Liisi TORGA, Marsida TOSKA, Trond TROSTERUD, Anna TRUKHINA, Reut TSARFATY, Utku TÜRK, Francis TYERS, Sumire UEMATSU, Roman UNTILOV, Zdeňka UREŠOVÁ, Larraitz URIA, Hans USZKOREIT, Andrius UTKA, Sowmya VAJJALA, Rob VAN DER GOOT, Martine VANHOVE, Daniel VAN NIEKERK, Gertjan VAN NOORD, Viktor VARGA, Eric VILLEMONTE DE LA CLERGERIE, Veronika VINCZE, Natalia VLASOVA, Aya WAKASA, Joel C. WALLENBERG, Lars WALLIN, Abigail WALSH, Jing Xian WANG, Jonathan North WASHINGTON, Maximilan WENDT, Paul WIDMER, Sri Hartati WIJONO, Seyi WILLIAMS, Mats WIRÉN, Christian WITTERN, Tsegay

WOLDEMARIAM, Tak-sum WONG, Alina WRÓBLEWSKA, Mary YAKO, Kayo YAMASHITA, Naoki YAMAZAKI, Chunxiao YAN, Koichi YASUOKA, Marat M. YAVRUMYAN, Arife Betül YENICE, Olcay Taner YILDIZ, Zhuoran YU, Arlisa YULIAWATI, Zdeněk ŽABOKRTSKÝ, Shorouq ZAHRA, Amir ZELDES, He ZHOU, Hanzhi ZHU, Anna ZHURAVLEVA, and Rayan ZIANE (2023), Universal dependencies 2.13, http://hdl.handle.net/11234/1-5287.

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

Arnold ZWICKY (1985), How to describe inflection, in Mary NIEPOKUJ, Mary VAN CLAY, Vassiliki NIKIFORIDOU, and Deborah FEDER, editors, *Proceedings of the Eleventh Annual Meeting of the Berkeley Linguistics Society*, pp. 372–386, Berkeley Linguistics Society, Berkeley, CA.

Laura Becker
0000-0002-1835-9404

Department of Linguistics
University of Freiburg
Belfortstraße 18,
79098 Freiburg im Breisgau, Germany

Laura Becker (2024), *Zero marking in inflection: A token-based approach*, Journal of Language Modelling, 12(2):349–413
https://dx.doi.org/10.15398/jlm.v12i2.361