Hedvig Skirgård* and Stephen Francis Mann*

An improved metric for estimating morphological information in corpora

https://doi.org/10.1515/lingvan-2025-0142
Received May 23, 2025; accepted May 10, 2026; published online August 26, 2026

Abstract: The emergence of large, consistently annotated corpora in many languages opens new avenues for linguistic typology by enabling the incorporation of usage-based evidence, including frequency information, into the study of language complexity. Morphological information is one of several facets of language complexity. In this paper, we use morphological feature annotations from the Universal Dependencies (UD) corpora to quantify information carried by morphology across 154 different datasets spanning 72 language varieties. We propose an information-theoretic approach that measures how surprising morphological feature values are given a token's part of speech or lemma in the corpus. These token-level quantities are aggregated to dataset-level averages, yielding a usage-weighted estimate of morphological information load. We find substantial cross-linguistic variation in morphological information, and observe moderate to strong correlations with a related information-theoretic metric proposed by Çöltekin and Rama. By contrast, we find little correspondence with questionnaire-based typological metrics derived from Grambank, which represents an alternative approach to cross-linguistic comparison based on grammatical inventories rather than usage. This illustrates the difference between studying the possible extent of the grammatical system versus language use. We also discuss various drawbacks with corpus-based typology, such as comparability of datasets and uneven coverage across the globe.

Keywords: morphology; dependency treebank; corpora; information theory; surprisal

Second language abstract, in Swedish: Framväxten av stora annoterade flerspråkiga korpussamlingar öppnar upp nya möjligheter för språktyologiska studier, t.ex. inom forskning på språkkomplexitet. Morfologisk information är en av flera aspekter av språkkomplexitet. I den här artikeln använder vi oss av morfologiska annoteringar från Universal Dependencies (UD)-korpusarna för att kvantifiera information i morfologi baserat på språkanvändning. Studien täcker 154 korpusar spänner över 72 språkvarianter. Vi föreslår en informationsteoretisk metod som mäter hur överraskande morfologiska annoteringar är givet tokenens ordklass eller lemma. Vi aggregerar dessa mätvärden på tokennivå till medelvärden för korpusar, vilket ger en användningsviktad uppskattning av morfologisk informationsbelastning överlag. Vi finner betydande tvärspråklig variation i morfologisk information och observerar måttliga till starka korrelationer med en relaterad informationsteoretisk mätteknik som föreslagits av Çöltekin och Rama. Däremot finner vi liten överensstämmelse med frågeformulärsbaserade typologiska mätvärden härledda från Grambank, vilket representerar ett alternativt tillvägagångssätt för tvärspråklig jämförelse baserad på grammatiska system snarare än användning. Detta illustrerar skillnaden mellan att studera den möjliga omfattningen av det grammatiska systemet kontra hur språket används. Vi diskuterar också olika nackdelar med korpusbaserad typologi, såsom jämförbarhet av datamängder och dålig global täckning.

Nyckelord: morfologi; dependency treebank; korpus; informationsteori; överraskning

***Corresponding authors: Hedvig Skirgård and Stephen Francis Mann**, Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, E-mail: hedvig_skirgard@eva.mpg.de (H. Skirgård), stephen_mann@eva.mpg.de (S.F. Mann). https://orcid.org/0000-0002-7748-2381 (H. Skirgård). https://orcid.org/0000-0002-4136-8595 (S.F. Mann)

1 Introduction

Language complexity is notoriously multifaceted. Even within the subdomain of morphological complexity there are multiple dimensions along which a language's morphology can vary, making comparison between languages difficult (Miestamo 2006b). Historically, much of the work on morphological complexity has focused on typological questionnaire surveys of morphological features and paradigms. Approaches based on corpora, however, offer complementary advantages, in that they can capture variation in usage. Over the last decades, a growing number of studies have quantified aspects of language complexity using corpus data (Bentz et al. 2016; Koplenig and Wolfer 2023; Moscoso del Prado Martín et al. 2004; Sagot 2013), building on earlier forerunners such as Greenberg (1960) and Juola (1998). Information-theoretic measurements are of particular interest in this context. Information, in the quantitative sense, is defined in terms of probabilities, and corpora provide empirical estimates of the frequencies with which different forms occur in language use.

In this study we introduce a measure of morphological complexity based on morphological feature annotations in corpora. The goal is to capture the informational load that languages typically carry in their morphology. Although a measure of this kind has been introduced previously (Çöltekin and Rama 2023), we argue that ours offers several improvements. Our measure provides a fine-grained quantification of the amount of morphological information carried by particular tokens. This goes beyond previous work that aggregates informational quantities at the level of the entire dataset by measuring the even-spreadedness of morphological features.

The structure of the paper is as follows. Section 2 briefly surveys previous work on language complexity, especially morphological complexity, and justifies the kind of corpora-based analysis we undertake. Section 3 introduces our study by defining our metric and describing the Universal Dependencies (UD) corpora to which it will be applied. This section also describes similarities and differences between our metric and Çöltekin and Rama's mean feature entropy (henceforth Ç&R's MFH). Section 4 provides the main results of our study. We determine correlations between our metric and other quantities measurable from UD corpora. We also determine correlations between these measures and others previously associated with morphological complexity but derived from Grambank rather than UD corpora. The goals are (i) to demonstrate differences between our metric and Ç&R's MFH, and (ii) to demonstrate differences between corpus-based measures of morphological complexity and those derived from grammars. Section 5 discusses limitations of our study, while Section 6 concludes.

2 Background: defining language complexity

Language complexity[1] is a multifaceted concept that can be approached in different ways (Raviv et al. 2022). One influential distinction is between absolute complexity, referring to properties of the linguistic system itself (e.g., the number of morphological distinctions), and relative complexity, concerning learning and processing difficulty for particular speakers (Miestamo 2006a, 2008). Even when focusing on absolute complexity, however, many measures are motivated by cognitive considerations. The present study is primarily concerned with absolute complexity.

Different theoretical approaches target different dimensions of complexity, yielding measurements that may correlate (Bentz et al. 2016, 2023; Çöltekin and Rama 2023), but also diverge (Lupyan and Raviv 2024). For example, complexity can be operationalized in terms of paradigmatic predictability (Ackerman et al. 2009; Cotterell et al. 2019; Round et al. 2022) or syntactic dependency length (Gibson 1998; Liu 2008).

Within morphological complexity, two commonly studied dimensions are enumerative complexity (E-complexity), capturing the number of overt morphological distinctions (Ackerman and Malouf 2013; Cotterell

1 Unfortunately, the terms "simple" and "complex" in relation to language have in some circles acquired value judgement, with so called "complex" languages being interpreted as more admirable or better in some other way (Sampson 2009: 4–5). This is nonsense: all languages are equally valid and valuable forms of human communication.

Table 1: Non-exhaustive table of different language complexity dimensions found in the research literature.

Metric	Description	Selection of relevant publications
Boundedness/fusion	How much grammatical material is morphologically bound	Sapir (1921); Greenberg (1960); Bybee (1985); Skirgård et al. (2023a); Shcherbakova et al. (2023); Lupyan and Dale (2010)
Type–token ratio	How many unique words in relation to all words	Kettunen (2014); Çöltekin and Rama (2023)
Synthesis	Morpheme to word ratio	Sapir (1921); Greenberg (1960); Bybee (1985); Easterday et al. (2021)
Enumerative complexity	How many grammatical distinctions there are	Greenberg (1960); Ackerman and Malouf (2013); Shcherbakova et al. (2023); Skirgård et al. (2023a); Lupyan and Dale (2010); Çöltekin and Rama (2023)
Informativity/prediction accuracy/contextual predictability	How accurately the next item in a string of language production can be predicted	Cohen-Priva and Jurafsky (2008); Cotterell et al. (2018); Frank and Bod (2011); Levshina (2022)
Redundancy	How much information is repeated	Leufkens (2023)
Systematicity/transparency	The degree of consistency of form to meaning patterns	Raviv et al. (2019); Hengeveld and Leufkens (2018); Bybee (1985)
Integrative complexity/paradigm complexity	Predictability of forms within paradigms	Bybee (1985); Round et al. (2022); Ackerman et al. (2009); Ackerman and Malouf (2013); Cotterell et al. (2019); Sagot (2013); Moscoso del Prado Martín et al. (2004); Beniamine (2021); Bonami and Beniamine (2016)
Paradigm syncretism	Degree to which distinct morphosyntactic feature combinations share surface forms	Quinn (2025); Bonami and Beniamine (2016)
Compositionality	The extent to which the meaning of larger units is built from the meanings of smaller parts	Wray and Grace (2007); Lindsay-Smith et al. (2024)
Kolmogorov complexity/description length	Compactness of descriptions; amount of information required to describe linguistic structure, typically approximated via compression or coding-based measures	Kolmogorov (1965); Juola (1998); Dahl (2004); Aceves and Evans (2024); Sagot (2013); Ehret (2016)
Ease of LLM learning	How easy is it for a large language model (LLM) to learn the language	Koplenig and Wolfer (2023)
Length of syntactic dependencies	How nested the syntactical structure is (tax on short-term memory)	Gibson (1998); Liu (2008)
Shared types	The total number of population-wide shared types	Spike (2017)

et al. 2019), and integrative complexity (I-complexity), capturing implicative relations within paradigms (Ackerman and Malouf 2013; Cotterell et al. 2019; Kusters 2003). Several dimensions of morphological structure commonly discussed in the complexity literature, such as fusion and synthesis, build on early work in the history of linguistics by Sapir (1921) and Greenberg (1960).[2] Table 1 summarizes a range of approaches to language complexity. Our study focuses on the surprisal of morphological features in corpora and can therefore be understood as a usage-based application of E-complexity, sensitive to distributional asymmetries in use.

In this study, we are interested in the information of morphological features, as defined by usage frequencies in corpora. Our information-theoretic corpus-based metric can be seen as a probabilistic extension of absolute complexity, generalizing Miestamo's (2006b, 2008) notion to account for actual language use. As with previous studies, we take morphological complexity to be related to the difficulty of language processing and learning (Bentz and Berdicevskis 2016; Cohen-Mimran et al. 2013; Hyönä et al. 1995; Mehravari et al. 2015).

Languages with richer morphology tend to exhibit larger inventories of distinct word forms and more complex paradigmatic structure, including non-compositional patterns such as suppletion or stem alternations (Baerman

2 Estimating language complexity (e.g., Fusion) using texts as opposed to surveying grammatical descriptions appears already in Greenberg (1960).

et al. 2015; Bentz et al. 2016; Dahl 2004). Traditional approaches to measuring morphological complexity are often based on surveying grammatical rules and distinctions (Lupyan and Dale 2010; Shcherbakova et al. 2023) – the possible extent of the system. However, this approach does not capture frequency patterns which can be important for testing certain language complexity hypotheses. For example, both Spanish and Turkish grammars distinguish three demonstrative forms (Spanish: *este/ese/aquel*, Butt et al. 2019: 87; Turkish: *bu/şu/o*, Schaaik 2020: 67). In actual usage, however, these distinctions are unevenly used. In the UD dataset Spanish-AnCora, *aquel* accounts for only 5 % of occurrences, compared to 26 % for *ese* and 69 % for *este* (Taulé et al. 2008). In Turkish-Penn, *bu* dominates with 85 % of occurrences, while *o* and *şu* together account for less than 15 %. Although both languages exhibit the same number of demonstratives (n = 3), their usage distributions differ, with potential consequences for learning and processing. Such distributional asymmetries suggest that usage-based measures, such as the one presented in this study, provide a complementary perspective on morphological complexity by weighting distinctions according to their empirical prominence in language use (cf. Bentz et al. 2016; Levshina et al. 2023).

3 Study outline

We quantify one aspect of language complexity by studying how much information is typically encoded in morphology in language use. An increase in morphological information can be understood as an increase in the informational commitments made in the grammatical form. Our metric, described in Section 3.3 and defined in more detail in Appendix A.2, captures the amount of information carried by morphological features across 72 language varieties spanning 154 datasets from the UD corpora collection (Zeman et al. 2024).

3.1 Data

We use Universal Dependencies v2.14 (Zeman et al. 2024; see further citations in Appendix D). We process the UD datasets before calculating morpho-surprisal (removing empty nodes, resolving multiwords, inserting dummy values for missing features, etc.); see Appendix A.2 for details. We use Glottolog v5.0 (Hammarström et al. 2024) for information about languages such as their location. For comparison with Grambank metrics, we use Grambank v1.0 (Skirgård et al. 2023a, 2023b). All code and data are freely available; see Appendix D.

3.2 Similar studies

This study follows similar theoretical principles to studies involving the Grambank-based metrics Fusion (out of all Grambank features that target bound morphology, what proportion are answered positively?) and Informativity[3] (out of all Grambank features which target if a grammatical distinction is made, how many are answered yes?).[4] Both approaches seek to cross-linguistically compare amount of morphology and information in morphology, but instead of using a survey of grammars, we use metrics calculated on annotated corpora. As discussed earlier, corpora-based studies have the advantage of capturing greater nuance in usage, which can be meaningful when studying diversity and change (Bentz et al. 2016; Greenberg 1960; Levshina 2022). We note that one of the current drawbacks with corpora-based typology is a considerably smaller sample size. For example, Grambank contains more than 2,000 languages and UD fewer than 200. Sample bias is also a concern, with smaller and non-Eurasian languages being under-represented in corpora collections like UD (see Figure 3). For comparison with our metric, we compute the two Grambank metrics according to established procedures (Shcherbakova et al. 2023; Skirgård 2025; Skirgård et al. 2023a) using Grambank v1.0 (see Appendix A.3).

3 The term "informativity" is found in the literature with at least two senses: how many grammatical distinctions a grammar makes possible (Shcherbakova et al. 2023; Skirgård et al. 2023a) and the average contextual unpredictability (Cohen-Priva and Jurafsky 2008). These two senses are distinct. When we refer to the Grambank Informativity metric, we refer to the first one.

4 For more details on these metrics, please see Skirgård et al. (2023a) and Skirgård (2025).

A similar approach to ours was pursued by Çöltekin and Rama (2023). They calculated several complexity metrics on UD datasets, including one designed to capture morphological information. Ç&R were interested in how different measures of complexity may latch onto the same underlying properties of language. They analysed eight measures, from a simple count of the type–token ratio to the far more involved assessment of how accurately a machine learning model can predict an inflected word from its lemma and morphological features. The measure most relevant for our purposes is MORPHOLOGICAL FEATURE ENTROPY (MFH), which captures how evenly spread morphological features are throughout a UD dataset. Entropy connects to complexity in a similar way to our metrics: if a few features dominate the corpus then they are more predictable. Since predictability is intuitively connected to complexity, the even-spreadedness of morphological features in a corpus can indicate the morphological complexity of a language. Ç&R's MFH has also been used in further studies, such as Bentz et al. (2023). The similarity between our approach and Ç&R's MFH was accidental; we only discovered their paper after formulating our study. Given the similarity of Ç&R's MFH with the current study, we describe this metric in more detail in Appendix A.1 and we directly compare results of both metrics (more details in Appendix B). We see our study as an improvement on Ç&R's MFH as we believe that our handling of missing features, aggregation levels, and other aspects are preferable. We acknowledge that they are similar in spirit and results, and we appreciate Çağrı Çöltekin and Taraka Rama making their code available and responding to questions.[5]

Lastly, a measure similar to ours was previously defined by Sproat et al. (2014). Their focus is efficiency, which they relate to differences in informational content across strings when normalized by length. They discuss different methods of quantifying both "information" and "space" in order to obtain ratios $\frac{\text{information}}{\text{space}}$ for different languages. One of their measures of information weights tokens by the SURPRISAL of their morphological features (given their bespoke annotation scheme). The surprisal of an event is a function of its probability: it is the logarithm of the reciprocal, written $\log \frac{1}{p}$, so that events with lower probability have higher surprisal. Using a customized dataset, they report results for eight languages (based on fewer than 1,000 sentences per language). Their work is conceptually similar to Ç&R's MFH, as well as our own. Unfortunately we cannot include a direct comparison because the original results are based on a small sample and we lack enough details to reproduce their procedure.

3.3 Overview of our metrics

Our approach consists of several metrics that measure information in morphological features in corpora. UD datasets (Zeman et al. 2024) contain token annotations for morphological features, universal parts of speech (UPOS), and lemmas which we can use to this end. Table 2 illustrates a Turkish sentence from the UD Turkish-Penn dataset (Kuzgun et al. 2020), where tokens are annotated for categories such as case, number, tense, aspect, and person.

In this study, we are interested not only in how many morphological features a token bears on average, but also in how informative those features are given their usage frequencies. For example, in Turkish-Penn the feature Number is often marked as Singular (87 %), whereas Tense is more evenly distributed (Past 50 %, Present 45 %, Future 5 %). From an information-theoretic perspective, Tense is more informative than Number, as its values are less predictable: the entropy of Tense is $0.5 \log \frac{1}{0.5} + 0.45 \log \frac{1}{0.45} + 0.05 \log \frac{1}{0.05} = 1.234$ bits, whereas the entropy of Number is only $0.87 \log \frac{1}{0.87} + 0.13 \log \frac{1}{0.13} = 0.557$ bits. For a given token, if the value of Tense is Past, that is less surprising than if it was Future $\left(\log \frac{1}{0.5} = 1 \text{ bit vs. } \log \frac{1}{0.05} = 4.322 \text{ bits} \right)$.

We measure token-level surprisal of morphological feature values and average across datasets to obtain a point estimate of morphological information (henceforth morpho-surprisal). The exact procedure for calculating our metrics can be found in Appendix A.2.

5 We would particularly like to thank Çöltekin and Rama for corresponding with us so that we could confirm our understanding of their procedure. Any remaining misunderstandings are our own.

Table 2: Sentence 15-0000 from Universal Dependencies dataset Turkish-Penn (Kuzgun et al. 2020). UPOS = universal part of speech; lemma = base form; token = specific word.

UPOS	Lemma	Token	Morphological features
ADJ	devasa	*Devasa*	
ADJ	ölçek	*ölçekli*	
ADJ	yeni	*yeni*	
NOUN	kanun	*kanunda*	Case=Loc|Number=Sing|Person=3
ADJ	kullan	*kullanılan*	
ADJ	karmaşık	*karmaşık*	
CCONJ	ve	*ve*	
ADJ	çetrefil	*çetrefilli*	
NOUN	dil	*dil*	Case=Nom|Number=Sing|Person=3
NOUN	kavga	*kavgayı*	Case=Acc|Number=Sing|Person=3
VERB	bulan	*bulandırdı*	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past|VerbForm=Fin|Voice=Cau

Translation: 'The complex and complicated language in the massive new law has muddied the fight.'

We calculate feature and value frequencies conditional on either UPOS or lemma (henceforth aggregation level). This serves several methodological purposes. First, features need not have comparable functions across parts of speech (e.g., Number on nouns vs. verbs). Second, feature distributions often pattern into groups such as UPOS. Besides UPOS, there can be other classes of words which pattern regularly (e.g., animates versus inanimates). Because other linguistically relevant classes (e.g., animacy, verb type) are not consistently annotated across languages,[6] lemma serves as a complementary more fine-grained aggregation level. Third, aggregating by UPOS or lemma better reflects language use, since users are not facing situations where any feature category can appear for any word.

We acknowledge that "lemma" is as an elusive category psycholinguistically (Meyer 2026), but in this study we believe it contributes a meaningful way of structuring the lexicon at a finer level compared to UPOS and is preferable to treating all tokens as one group. In principle any way of grouping tokens can be implemented in our approach (e.g., Appendix C.1.2.1), and our metric can be calculated on any dataset with the necessary formatting (Appendix A.2.1).

All tokens of a given UPOS or lemma are assigned the same set of feature categories. Missing feature values are filled with a dummy value ("unassigned"), reflecting the fact that absence of annotation itself may carry information. For instance, in Czech-PDT, 96 % of adjectives are marked for Polarity; the remaining tokens are assigned "Polarity = unassigned" when aggregation level is UPOS. If all tokens of an UPOS or lemma lack features, a single dummy feature is assigned ("unassigned = unassigned").

We define eight variants of morpho-surprisal based on three binary parameters. The first is aggregation level (UPOS or lemma). The second concerns whether we use all features in a dataset, or only those centrally defined ("core" features as defined by the UD project; see Appendix C.1.1). The third concerns whether we treat the features separately (e.g., Case = Ins, Degree = Pos) or as a concatenated string ("featstring"; e.g., "Case = Ins|Degree = Pos"). This parameter is included to partially address conditional frequency (see Appendix C.2). All variants of morpho-surprisal are listed in Table 3.

3.4 Additional metrics

We also calculate related metrics which are commonly reported in other corpora-based studies of morphological complexity. For example, we expect that a dataset that contains a higher absolute number of morphological

6 In Appendix C.1.2.1 we explore "subclasses" within UPOS. This approach is experimental and is included in the appendix as sensitivity analysis.

Table 3: Variants of our metric.

Aggregation level	All features or only core features	Features separated or featstring
Lemma	All	Features separated
Lemma	Core only	Features separated
UPOS	All	Features separated
UPOS	Core only	Features separated
Lemma	All	Featstring
Lemma	Core only	Featstring
UPOS	All	Featstring
UPOS	Core only	Featstring

feature categories is more likely to exhibit higher levels of surprisal of morphological features per token as there are more features that can be utilized. The additional metrics are:

- type–token ratio (TTR): the number of unique tokens divided by the total number of tokens
- lemma–token ratio (LTR): the number of unique lemmas divided by the total number of tokens
- mean number of morphological features per token (calculated without "dummy" features)
  a) all features
  b) core features only
- mean token surprisal: mean surprisal per token given all tokens, regardless of UPOS/lemma or morphological features

We also report some simple statistics for each dataset:

- number of types (# Types)
- number of tokens (# Tokens)
- number of unique feature categories (# Feat cat)
  a) all features
  b) core features only

Finally, we compare with metrics from other studies:

- Ç&R's MFH, with small modifications (see Appendices A.1, A.1.1, and B)
- Grambank metrics (see Appendix A.3):
  - Fusion (degree of morphological boundness as measured by Grambank features)
  - Informativity (number of grammatical distinctions as measured by Grambank features)

3.5 Comparison: Ç&R's MFH and ours

There are several differences between Ç&R's MFH and our metrics. We list a few key differences here; for more, see Appendix B.

Firstly, Çöltekin and Rama remove tokens with missing features (for example *ölçekli*, *kullanılan*, *karmaşık*, and *ve* in Table 2), whereas we keep them. For datasets of languages with little morphology, this makes a large difference in the number of tokens analysed and therefore the averages.

Secondly, instead of calculating the average surprisal of features (e.g., Case = Acc) across the entire sample of tokens, we split features into their component category/value pairs (e.g., "Case" and "Acc"), and calculate the sum of the surprisals of feature values (e.g., "Acc") given a feature category (e.g., "Case") given an aggregation level (UPOS or lemma) for a given token. This produces the token's morpho-surprisal: the amount of information it carries by virtue of its morphological feature markings or lack thereof compared to relevant other tokens. Once we have a measure of morpho-surprisal for each token, we can average this across the entire dataset to obtain mean morpho-surprisal per token.

4 Results

We report the results in scatterplot matrices (SPLOMs) and maps.[7] The SPLOM visualizations compare different measurements of the same data in a pairwise fashion. The lower-left portion of each SPLOM shows scatterplots for the pairwise comparisons of measurements. The upper-right portion shows the Spearman's rank correlation coefficient of the corresponding pair (see Appendix E for details). The diagonal of the plot shows histograms of each measure. The upper and right edges of the diagram list the names of each measure, while the left and lower edges display axis values for the scatterplots.

Significant correlations ($p < 0.05$) are marked with an asterisk. Because these SPLOM plots comprise multiple comparisons, we used the Holm–Bonferroni method (Holm 1979) to adjust p values in order to control the family-wise error rate where relevant (see Appendix F for more details). An absolute correlation coefficient larger than 0.7 is commonly interpreted as a strong relationship, and above 0.9 very strong (Schober et al. 2018). We have marked significant correlations above 0.7 in red bold text. For each pair of measurements, we compare as many overlapping data points without missing data as possible. The number of data points compared for each pair is reported in the upper triangle of the SPLOM as n.[8]

UD datasets can vary considerably in terms of genre, which can have an impact on the measurements we are calculating. The UD datasets that belong to the Parallel Universal Dependencies (PUD) subset are more comparable in genre as they are translational equivalents. Therefore, we report results both for all UD datasets and for the PUD datasets only. See Appendix C.1.3 for more information.

4.1 Our morpho-surprisal metrics

First, we consider the relationship between the eight different new metrics defined in this paper (Figures 1 and 2). All correlate strongly with each other ($\geq 0.8^*$), regardless of whether we consider all datasets or only PUD. This suggests that while they differ in specific implementation, they are picking up on the same characteristics of the datasets (and thereby languages) in question.[9]

The pair of morpho-surprisal metrics that differ the most from each other are:[10]

- mean sum surprisal of features, aggregation level = UPOS, all features (Figure 3(a))
- mean sum surprisal of featstring, aggregation level = lemma, core features only (Figure 3(b))

These two variants of our morpho-surprisal metric are also maximally distinct (features vs. featstring, UPOS vs. lemma, all features vs. core only), so it makes sense that they would correlate less than other pairs. To save space, we will include only these two in further plots as they represent the variation of our custom metrics well. Figure 3 shows the distribution of these two metrics across the datasets linked to language locations. These distributions

7 We matched UD datasets to glottocodes and used geographical locations of languages found in Glottolog 5.0 (Hammarström et al. 2024). The points have been jittered somewhat to avoid overlaps for datasets of the same or nearby languages which would obscure information. Unfortunately there are no UD datasets included in our study representing languages of the Americas or the central or eastern Pacific regions, which is why these are not displayed on the maps.

8 When comparing between measurements based on the UD datasets, each point represents a single UD dataset. When comparing against the Grambank metrics Fusion and Informativity, the UD dataset metrics (including Ç&R's MFH) are matched to entries in Grambank via glottocodes and averaged over language.

9 A reviewer rightly pointed out that one cannot infer from a correlation between two variables that both are tracking the same underlying characteristic. For example, age and wage are correlated but do not track any single property in a population. Here we are relying on the fact that each of our metrics is defined in terms of the same formula, applied to the same datasets, but with variations in precise settings (UPOS/lemma, all features/core features, features separated/featstring). The strong correlations between the outcome of these implementations demonstrate, we think, that there is a general property of each dataset – what we have described as the "morpho-surprisal" – that each specific implementation is capturing an aspect of.

10 We note that there is another pair in the larger dataset (not PUD) that are as uncorrelated (0.84*): mean sum surprisal of features, aggregation = UPOS, core features versus mean sum surprisal of featstring, aggregation level = lemma, all features. As this pairing is very similar to our chosen one, and our chosen pairing is the least correlated in the PUD dataset as well, we will focus on it alone.

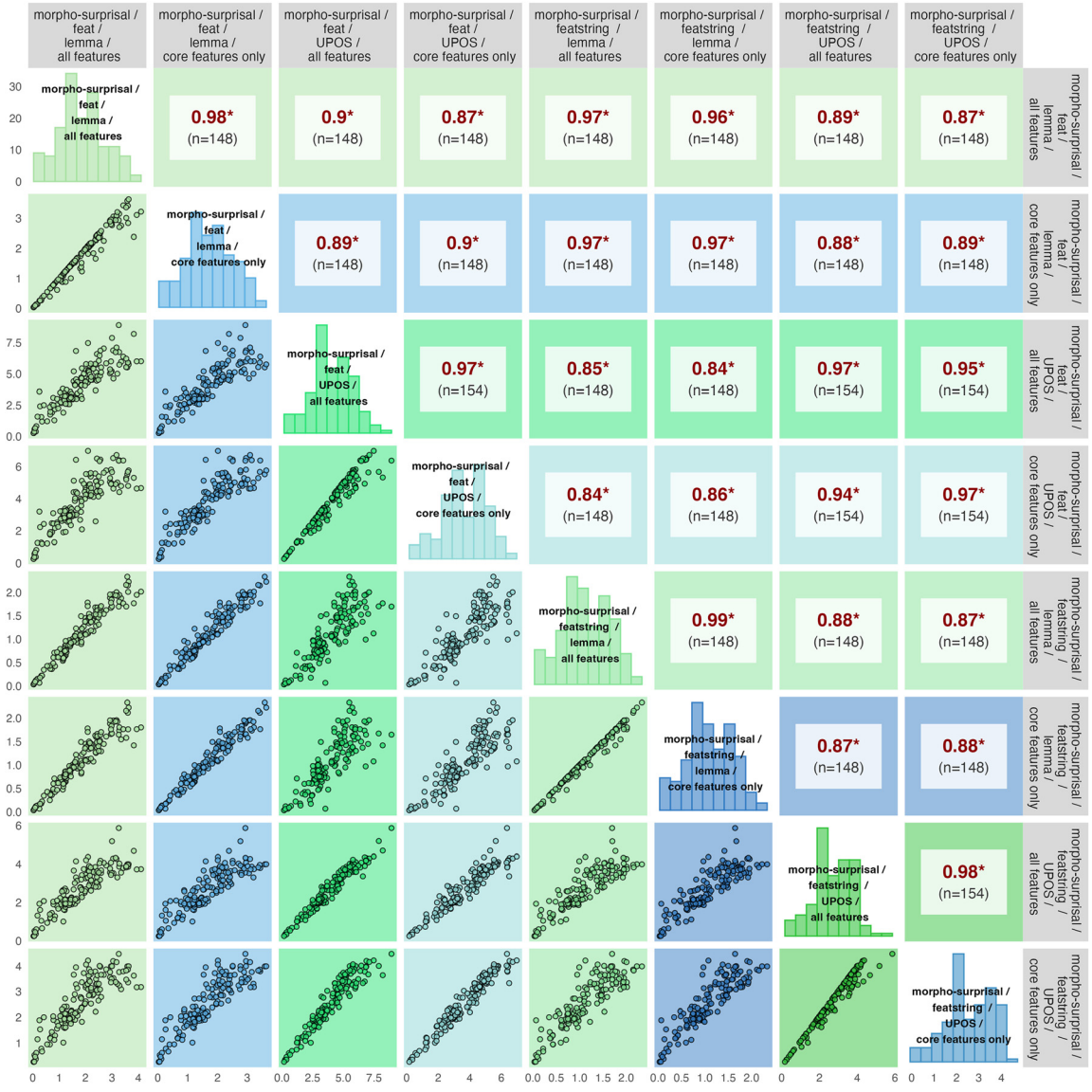


Figure 1: Scatterplot matrix of our eight custom metrics of morpho-surprisal over all UD datasets in the study.

are similar to the dimension "degree of synthesis" (Comrie 1989; Croft 1990; Greenberg 1960; Sapir 1921) often referenced in linguistic typology – "hot spots" in eastern Europe and Western Asia and lower values in Southeast Asia. However, as several key areas are missing or severely under-represented (primarily the Americas, Africa, and Australia), the seeming alignment of our values with patterns discussed in linguistics literature is indicative rather than conclusive.

We also compare our metrics to other statistics of the dataset (TTR, number of feature categories, and more). Due to space constraints, we cannot include these SPLOMs and discussion in full here; please see Appendix G. Both representative custom metrics correlate moderately to strongly ($\geq 0.69^*$) with mean number of features per token (for all datasets and for PUD only). This is expected: the more features tokens have, the more chances there are to be surprised.

For the PUD datasets, there were also moderate to strong correlations between one of the custom metrics (mean sum surprisal of features, aggregation level = UPOS, all features) and two other statistics: TTR (0.68*) and number of types (0.73*). This can be explained by the fact that the more morphology a language has, the

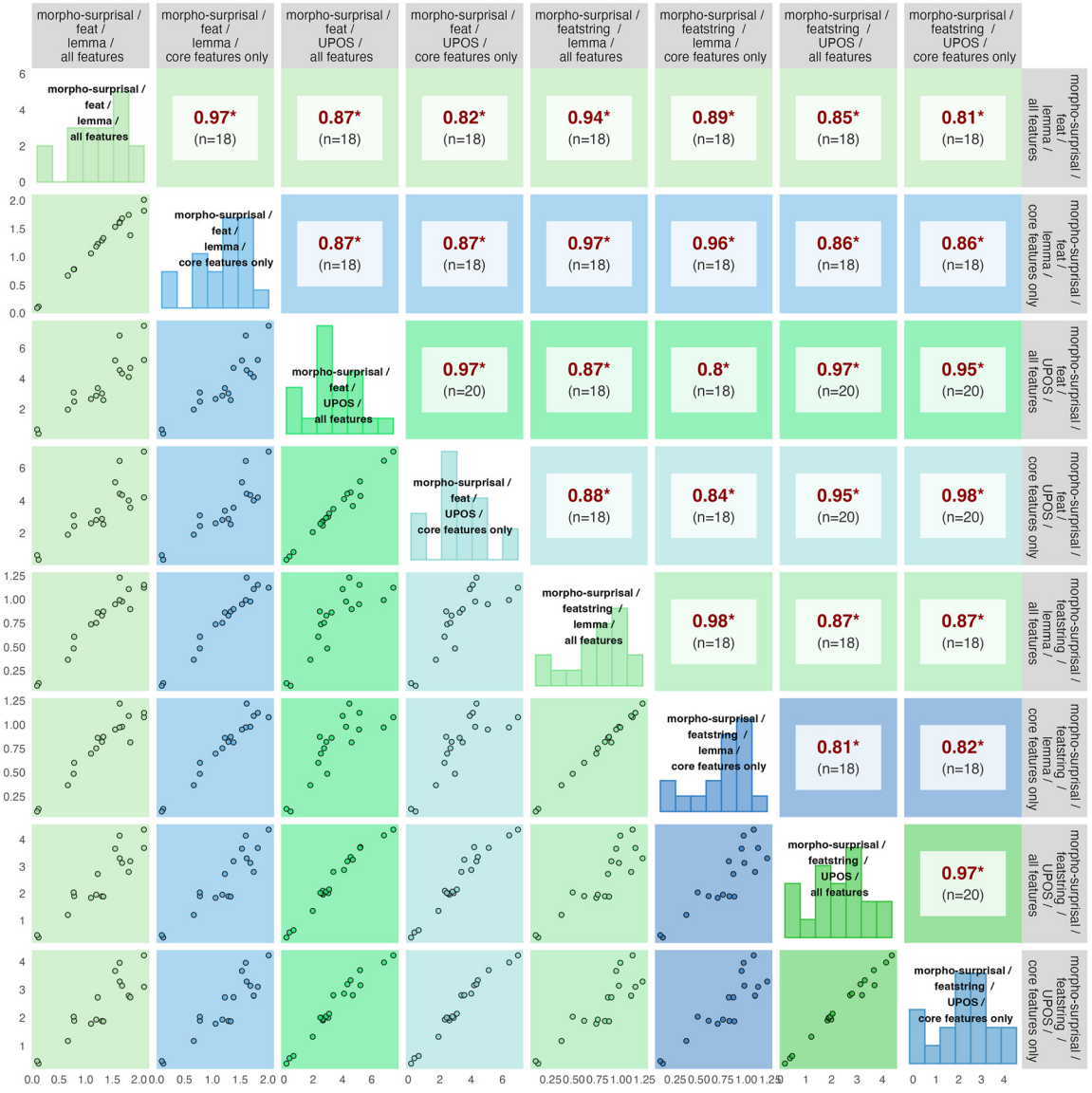
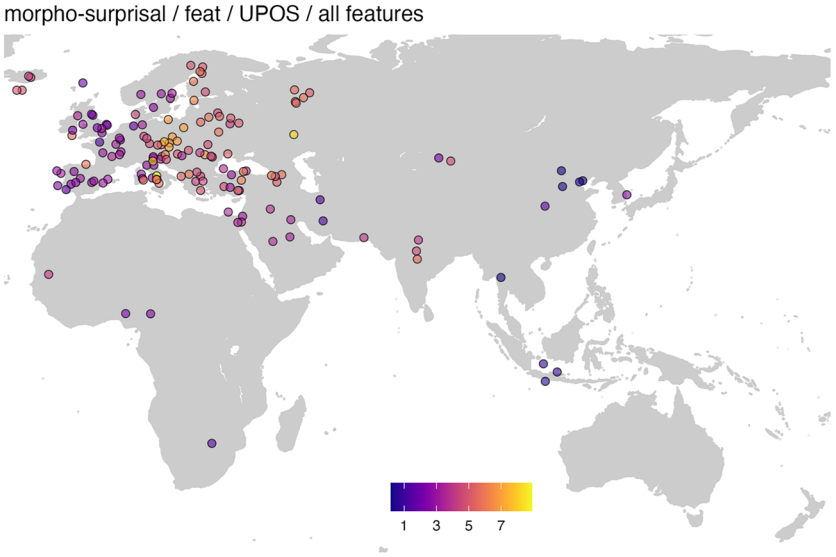


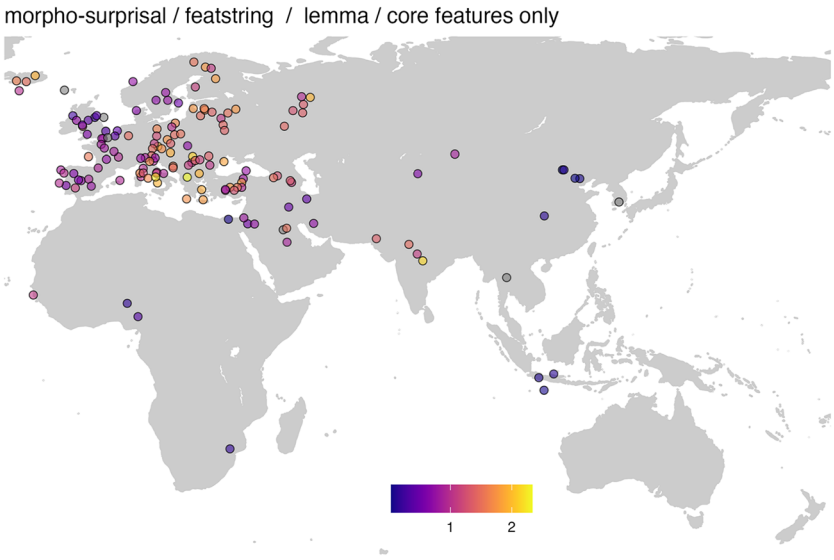
Figure 2: Scatterplot matrix of our eight custom metrics of morpho-surprisal over only the PUD-datasets.

more unique types can be generated in a corpora and the more morphological features there are to be surprised by. However, this relationship is not evident when considering all datasets (<0.36). This is most likely due to PUD datasets being translational equivalents. Because of this the number of types in PUD datasets is more likely to track relative morphological diversity compared to all datasets which may be of varying lengths and genres.

Our second representative custom metric (mean sum surprisal of featstring, aggregation level = lemma, core features only) does not show a strong correlation with TTR or number of types, even in the PUD datasets. This is consistent with the fact that lemma-level aggregation combined with a restriction to core features removes much of the variance. By contrast, UPOS-level aggregation pools across lemmas and preserves cross-lemma variation, which makes it more possible for number of types to co-vary with our morpho-surprisal metric.



(a)



(b)

Figure 3: Maps showing two of the custom metrics. (a) Mean sum surprisal of features, aggregation level = UPOS, all features. (b) Mean sum surprisal of featstring, aggregation level = lemma, core features only.

4.2 Comparison with Ç&R's MFH

An empirical comparison of Ç&R's MFH metric with our two representative morpho-surprisal metrics, absolute number of feature categories, and TTR can be seen in Figure 4.[11] The correlations between our metrics and Ç&R's MFH are moderate to strong (0.58–0.83*). For further details of the conceptual differences between the metrics, see Section 3.5 and Appendices A.1 and B.

11 Note that we implement Ç&R's MFH slightly differently from the original version. See Appendix A.1.1 for details.

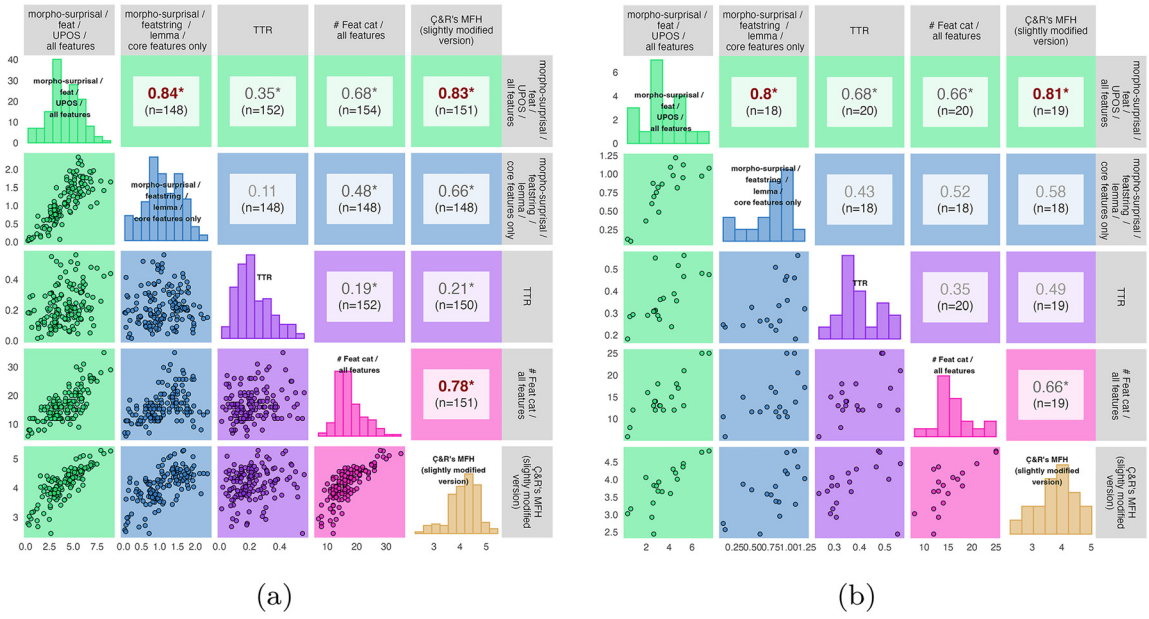


Figure 4: Scatterplot matrices comparing two of our custom morpho-surprisal metrics and Çöltekin and Rama's (2023) mean feature entropy (MFH), type–token ratio (TTR), and the total number of feature categories. (a) All datasets. (b) PUD datasets only.

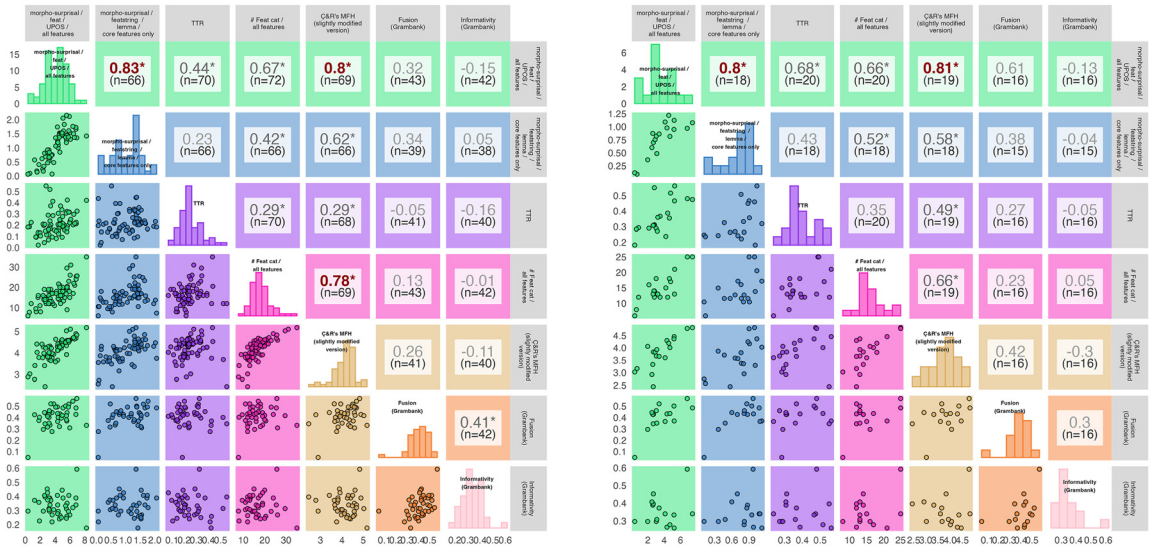
Why should our metric not correlate perfectly with Ç&R's MFH? Firstly, there are whole datasets as well as parts of datasets that cannot be included in calculating Ç&R's MFH but where our approach functions. This is due to differences in data processing decisions. For example, tokens without morphological features, without lemmas, or without a string for the token field are excluded from their analysis. We keep these in as they can still be useful for computing the information in morphology across datasets (see Appendix A.2). This make a difference for isolating languages like Mandarin Chinese, as only tokens with any morphology are taken into account for computing MFH. If a language has many tokens with no morphological features and these are all excluded, then the metric is only computed on a particular subset of the tokens instead of over all tokens. We can see the possible consequences of this if we compare Mandarin Chinese with Portuguese, which are commonly described as having little and moderate morphology respectively. If we rank UD datasets by their MFH values, then Portuguese-PUD appears lower than Chinese-PUD (rank 7 vs. 9). This is not the case for our morpho-surprisal scores, where Chinese-PUD ranks lower than Portuguese-PUD as we would expect given common descriptions of their grammatical systems.

Secondly, the first of our representative metrics (mean sum surprisal of features, aggregation level = UPOS, all features) assigns a wider range of values to the datasets, some of which are lower and some higher than mean feature entropy. This is because our UPOS-based measure covers a wider range of values than mean feature entropy does (see Figure B in Appendix B.8), giving a more fine-grained measure of morphological complexity. For more information on the differences between our two approaches methodologically, see Appendix B.

4.3 Grambank metrics

Finally, we compare our custom metrics and Ç&R's MFH to the Grambank metrics Fusion and Informativity. We also include TTR and number of feature categories as anchors to the dataset attributes. The correlations between our metrics and the Grambank metrics are not significant (Figure 5(a)). We note also that Ç&R's MFH also does not correlate significantly with either of the two Grambank metrics.

The lack of correlation between the two Grambank metrics and the corpus-based metrics points to a systematic divergence between corpus-based estimates of morphological behaviour and measures derived from the



(a) (b)

Figure 5: Scatterplot matrices comparing two of our custom morpho-surprisal metrics with two Grambank metrics, MFH and other general metrics. The metrics are aggregated over language rather than UD-dataset.. (a) All datasets. (b) PUD datasets only.

descriptions of the grammatical system, suggesting that realized patterns of marking in usage are not straightforwardly predictable from the presence of grammatical rules alone.

5 Limitations

5.1 Dummy feature assignment at the UPOS level

Because a lack of marking can carry morphological information, we assign "dummy" features to tokens of the same aggregation level (UPOS or lemma) based on certain rules. Unfortunately, when aggregating at the UPOS level this can lead to dummy features being assigned to tokens that intuitively ought not receive them. That is because UPOS are broad categories: different subclasses of token within the same part of speech may be associated with different features. In such a case our method assigns to all members of one subgroup dummy values of features that belong properly to another.

For example, tokens representing the word *the* in the corpus English-GENTLE get the following feature values when grouped by UPOS and considering core features only:

- Definite=Def
- Degree=unassigned
- Number=unassigned
- Polarity=unassigned
- PronType=Art

"Def" and "Art" mark *the* as the definite article. But it seems odd to describe *the* as carrying information about Polarity. The tokens receive this dummy feature assignment (Polarity=unassigned) because another token of the same UPOS (DET) carries this feature. In this case, it is specifically the word *no* which has a value for Polarity ("Neg"). Since some determiners are explicitly marked for polarity and others are not, in order to quantify the amount of information conveyed about polarity on encountering a determiner we capture the distinction with a dummy feature value. One can argue that UPOS DET is too broad a category, that it ought to be broken down into

smaller subclasses with distinct morphological behaviour. Unfortunately, no such categorization is available across UD datasets.

The good news is that the information quantity contributed to *the* by Polarity is very small. Vastly more determiners have "unassigned" than "Neg" for Polarity (1165 vs. 35). All these non-marking determiners therefore see a very small informational contribution: it is unsurprising when you encounter a determiner without Polarity. When *no* is encountered, the negative Polarity is much more surprising and contributes more significantly to that token's information quantity.

We have made no attempt to correct what might be seen as "mistakes" in tagging of morphological features in the treebanks themselves: we have only established a workflow designed to estimate morphological information quantities given reasonable groupings. The approach to measuring information in morphology suggested in this paper can be applied to other datasets besides UD. We provide R code with generalized functions in supplementary material and guidance in Appendix A.2.1.

Dummy feature assignment can lead to unintuitive attribution of morphological features to tokens. Our view is that information can be carried by the absence of morphological marking and that it is relevant to standardize within groups (UPOS or lemma) – this is what dummy assignment is designed to address. Cases where this can be problematic include situations where not all morphological features are strictly relevant for all tokens of the same UPOS or lemma. For our study, we saw it as necessary to differentiate in some way between groups of tokens, instead of treating them all as one big group. The systematic groupings that were most readily available to us in a consistent format across UD datasets were UPOS and lemma. As there are fewer unique UPOS tags than lemmas, the issues discussed above are less of a concern when the aggregation level is set to lemma. We note however that the morpho-surprisal scores when aggregating over UPOS and lemma are strongly correlated, suggesting that this issue is not a major concern.

In addition, we carried out an experimental sensitivity analysis defining a new aggregation level: groups of tokens within the same UPOS which are consistently marked for the same feature categories. We call these groups "subclasses" and this approach avoids the issue with dummy feature values altogether. The results strongly correlate with the other metrics ($\geq 0.74^{*}$), especially with the metrics that are aggregated over UPOS. This suggests that the effect of dummy values is not detrimental. For more details, see Appendix C.1.2.1.

We recognize that other solutions exists, such as making use of other datasets besides UD. We leave to future work the possibility of estimating morphological information aggregating in other ways.

5.2 Artificial lemmas

Some UD datasets have a sizeable proportion of missing lemmas. This becomes problematic for calculating the total number of unique lemmas, LTR, and four of our custom morpho-surprisal metrics which use lemma as the aggregation level.

We ameliorate this problem in two ways, both of which can affect comparability. First, we simply omit those six metrics for datasets with more than 40 % missing lemmas (e.g., Thai-PUD, English-GUMReddit, and Faroese-FarPaHC). Second, for the datasets with less than 40 % missing lemmas we impute missing lemmas with the token value (plus underscore UPOS, as with all other lemmas). We chose this threshold because there is a jump from 32 % to 58 % missing lemmas across the datasets in question. Manual inspection of a few datasets that remain suggests many of the tokens that lack lemmas are forms that do not inflect much or at all (e.g., Turkish adjectives). They may have been left out because the lemma would be considered identical to the token. We have made this assumption explicit by filling in the lemma with the token value, allowing us to calculate the six metrics on these datasets.

Both of these choices affect comparability. In the first case, we lose the ability to compare datasets with more than 40 % missing lemmas for all metrics. The power of statistical tests involving these metrics is therefore reduced. In the second case, we might be inadvertently adding lemmas to tokens in a suboptimal manner. This would skew both the number of lemmas, LTR, and our morpho-surprisal scores which aggregate over lemmas.

Different decisions might reasonably be made here, but the aforementioned manual checks plus the relatively low level of missing lemmas in most datasets means that we would expect the impact of our procedure to be small.

5.3 Inconsistent annotation standards across corpora

The UD project provides central documentation guiding dataset annotation. At the same time, individual dataset contributors have freedom to provide additional annotation and vary implementation. We deal with the most significant issues of this kind by excluding some datasets from our analysis. For example, we exclude datasets such as Bavarian-MaiBaam that have no feature annotations at all. Our informational measure requires feature annotations to produce a meaningful result.

While we have removed certain symbols such as emojis from each dataset before the calculation step, we elected to leave in certain other symbols such as ampersands (&), hashes (#), and at-signs (@). Manual inspection of several datasets reveals that these symbols are typically used consistently within each treebank. For example, Finnish-TDT uses the hash character as a separator between the constituent parts of compound words. Such symbols could cause problems if they are used inconsistently, but we would estimate the effect on our results to be minor.

There are remaining differences that could negatively impact the comparability of our metric. For reasons of space we could not address these in the main text. They are discussed in detail in Appendix C.

6 Discussion and conclusion

We introduced a usage-based information-theoretic estimate of morphological complexity: mean token-level surprisal of morphological features. This metric can be construed eight different ways, aggregating over UPOS or lemma, all features or core features, and separating out features or concatenating them. Across the eight variants of our metric there are strong correlations, indicating that the analytical choices between variants have little impact. Our metrics correlate moderately to strongly with Ç&R's MFH. One key difference is that our metric yields non-zero values for datasets with certain missing values (e.g., little morphology) and produces a larger range of values. Weak alignment with Grambank's Fusion and Informativity most likely reflects a difference in target: usage distributions versus possibility space. Limitations include variation in UD annotation practices, artificial lemmas, and dummy feature assignment. We discuss these in Section 5 and Appendix C. We also note that the limited number of languages represented in UD makes historical or geographical interpretations limited (compared to, for example, WALS or Grambank).

In conclusion, morpho-surprisal offers a compact, usage-based estimate of information in morphology – one facet of morphological complexity.

Acknowledgements: We are indebted to Matthew Spike for his input on this article in particular and for our discussions on "complexity" and information in general. We thank Çağrı Çöltekin and Taraka Rama for their generous aid in understanding the intricacies of their measurement. We also want to thank the audiences at the Working Group for Empirical Linguistics seminar series at the Department of Linguistics and Philology at Uppsala University, department seminars at the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology, and participants at the 2024 biennial meeting of the Association for Linguistic Typology in Singapore. We have also benefited from advice from Daan van Esch, code review by Christoph Rzymski, and discussions with Angela Chira. We also gratefully acknowledge comments on the draft by Natalie Korobzow and Deepthi Gopal. The manuscript was also greatly improved thanks to the comments of two anonymous reviewers. Finally, we would like to thank the organizers of the workshop on Dependency Grammar for Typology at the ALT meeting and editors of this special issue (Andrew Dyer, Luigi Talamo, Annemarie Verkerk,

Luca Brigada Villa, Erica Biagetti, and Diego Alves), as well as the other presenters at the workshop and the audience.

Data availability statement

Appendices, data, and source code for this project are available at:
– GitHub: https://github.com/HedvigS/UD_complexity (v1.0.1)
– Zenodo: https://zenodo.org/records/19386502 (v1.0.1)

References

Aceves, Pedro & James A. Evans. 2024. Human languages with greater information density have higher communication speed but lower conversation breadth. *Nature Human Behaviour* 8(4). 644–656.

Ackerman, Farrell, James P. Blevins & Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins & Juliette Blevins (eds.), *Analogy in grammar: Form and acquisition*, 54–82. Oxford: Oxford University Press.

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

Baerman, Matthew, Dunstan Brown & Greville G. Corbett. 2015. *Understanding and measuring morphological complexity*. Oxford: Oxford University Press.

Beniamine, Sacha. 2021. One lexeme, many classes: Inflection class systems as lattices. In Berthold Crysmann & Manfred Sailer (eds.), *One-to-many relations in morphology, syntax and semantics*, 23–51. Berlin: Language Science Press.

Bentz, Christian & Aleksandrs Berdicevskis. 2016. Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François & Philippe Blache (eds.), *26th international conference on computational linguistics (COLING 2016)*, 222–232. Osaka: COLING 2016 Organizing Committee.

Bentz, Christian, Ximena Gutierrez-Vasques, Olga Sozinova & Tanja Samardžić. 2023. Complexity trade-offs and equi-complexity in natural languages: A meta-analysis. *Linguistics Vanguard* 9(s1). 9–25.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig & Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, Thomas François & Philippe Blache (eds.), *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, 142–153. Osaka: COLING 2016 Organizing Committee.

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

Butt, John, Carmen Benjamin & Antonia Moreira Rodríguez. 2019. *A new reference grammar of Modern Spanish* (Routledge Reference Grammars), 6th edn. London: Routledge.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.

Cohen-Mimran, Ravit, Jasmeen Adwan-Mansour & Shimon Sapir. 2013. The effect of morphological complexity on verbal working memory: Results from Arabic speaking children. *Journal of Psycholinguistic Research* 42(3). 239–253.

Cohen-Priva, Uriel & Dan Jurafsky. 2008. Phone information content influences phone duration. Poster presented at the Conference on Prosody and Language processing, Cornell University, Ithaca, NY, 11–13 April.

Çöltekin, Çağrı & Taraka Rama. 2023. What do complexity measures measure? Correlating and validating corpus-based measures of morphological complexity. *Linguistics Vanguard* 9(s1). 27–43.

Comrie, Bernard. 1989. *Language universals and linguistic typology: Syntax and morphology*. Chicago: University of Chicago Press.

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

Cotterell, Ryan, Sabrina J. Mielke, Jason Eisner & Brian Roark. (2018) Are all languages equally hard to language-model? In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, vol. 2, 536–541.

Croft, William. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Dahl, Östen. 2004. *The growth and maintenance of linguistic complexity*. Amsterdam: John Benjamins.

Easterday, Shelece, Matthew Stave, Marc Allassonnière-Tang & Frank Seifart. 2021. Syllable complexity and morphological synthesis: A well-motivated positive complexity correlation across subdomains. *Frontiers in Psychology* 12. https://doi.org/10.3389/fpsyg.2021.638659.

Ehret, Katharina. 2016. *An information-theoretic approach to language complexity: Variation in naturalistic corpora*. Freiburg: Albert-Ludwigs-Universität Freiburg PhD thesis.

Frank, Stefan L. & Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science* 22(6). 829–834.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68(1). 1–76.

Greenberg, Joseph H. 1960. A quantitative approach to the morphological typology of language. *International Journal of American Linguistics* 26(3). 178–194.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2024. Glottolog (v5.0) [dataset]. Zenodo. https://doi.org/10.5281/zenodo.10804357.
Hengeveld, Kees & Sterre Leufkens. 2018. Transparent and non-transparent languages. *Folia Linguistica* 52(1). 139–175.
Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6. 65–70.
Hyönä, Jukka, Matti Laine & Jussi Niemi. 1995. Effects of a word's morphological complexity on readers' eye fixation patterns. In John M. Findlay, Robin Walker & Robert W. Kentridge (eds.), *Eye movement research: Mechanisms, processes, and applications* (Studies in Visual Information processing 6), 445–452. Amsterdam: Elsevier.
Juola, Patrick. 1998. Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics* 5(3). 206–213.
Kettunen, Kimmo. 2014. Can type-token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics* 21(3). 223–245.
Kolmogorov, Andrei N. 1965. Three approaches to the quantitative definition of information. *Problems of Information Transmission* 1(1). 1–7.
Koplenig, Alexander & Sascha Wolfer. 2023. Languages with more speakers tend to be harder to (machine-)learn. *Scientific Reports* 13(1). https://doi.org/10.1038/s41598-023-45373-z.
Kusters, Christiaan Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection*. Utrecht: LOT PhD thesis.
Kuzgun, Aslı, Neslihan Cesur, Bilge Nas Arıcan, Merve Özçelik, Büşra Marşan, Neslihan Kara, Deniz Baran Aslan & Olcay Taner Yıldız. 2020. On building the largest and cross-linguistic Turkish dependency corpus. In *2020 innovations in intelligent systems and applications conference (ASYU)*, 1–6.
Leufkens, Sterre. 2023. Measuring redundancy: The relation between concord and complexity. *Linguistics Vanguard* 9(s1). 95–106.
Levshina, Natalia. 2022. Frequency, informativity and word length: Insights from typologically diverse corpora. *Entropy* 24(2). 280.
Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, Gabriela Garrido Rodriguez, Timothy Michael Gupton, Evan Kidd, Zoey Liu, Chiara Naccarato, Rachel Nordlinger, Anastasia Panova & Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825–883.
Lindsay-Smith, Emily, Matthew Baerman, Sacha Beniamine, Helen Sims-Williams & Erich R. Round. 2024. Analogy in inflection. *Annual Review of Linguistics* 10(1). 211–231.
Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9(2). 159–191.
Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLoS One* 5(1). https://doi.org/10.1371/journal.pone.0008559.
Lupyan, Gary & Limor Raviv. 2024. A cautionary note on sociodemographic predictors of linguistic complexity: Different measures and different analyses lead to different conclusions. In *15th international conference on the evolution of language (EVOLANG)*, 345–348. Madison, WI: Evolution of Language Conferences.
Miestamo, Matti. 2006b. On the feasibility of complexity metrics. In K. Kerge and M.-M. Sepper (eds.), *Finest linguistics: Proceedings of the Annual Finnish and Estonian Conference of Linguistics, Tallinn, May 6-7, 2004*, Tallinn: TLÜ.
Mehravari, Alison S., Darren Tanner, Emma K. Wampler, Geoffrey D. Valentine & Lee Osterhout. 2015. Effects of grammaticality and morphological complexity on the P600 event-related potential component. *PLoS One* 10(10). https://doi.org/10.1371/journal.pone.0140850.
Meyer, Antje S. 2026. The elusive lemma: On the representation of grammatical information in the mental lexicon. *Language, Cognition and Neuroscience* 41(4). 417–435.
Miestamo, Matti. 2006a. On the complexity of standard negation. *Finnish Journal of Linguistics* 19. 345–356.
Miestamo, Matti. 2008. Grammatical complexity in cross-linguistic perspective. In Matti Miestamo, Fred Karlsson & Kaius Sinnemäki (eds.), *Language complexity*, 23–41. Amsterdam: John Benjamins.
Moscoso del Prado Martín, Fermín, Aleksandar Kostić & R. Harald Baayen. 2004. Putting the bits together: An information theoretical perspective on morphological processing. *Cognition* 94(1). 1–18.
Quinn, Kyla. 2025. *An evolutionary exploration of paradigm syncretism in the world's languages*. Canberra: Australian National University PhD thesis. https://openresearch-repository.anu.edu.au/bitstreams/9715b405-9673-4482-b809-e96ee358ec1a/download (accessed 8 July 2026).
Raviv, Limor, Antje Meyer and Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B* 286(1907). https://doi.org/10.1098/rspb.2019.1262.
Raviv, Limor, Louise R. Peckre & Cedric Boeckx. 2022. What is simple is actually quite complex: A critical note on terminology in the domain of language and communication. *Journal of Comparative Psychology* 136(4). https://doi.org/10.1037/com0000328.
Round, Erich, Stephen Francis Mann, Sacha Beniamine, Emily Lindsay-Smith, Louise Esher & Matt Spike. 2022. Cognition and the stability of evolving complex morphology: An agent-based model. In *The evolution of language: Proceedings of the joint conference on language evolution (JCoLE)*, 635–642. Kanazawa: Joint Conference on Language Evolution (JCoLE).
Sagot, Benoît. 2013. Comparing complexity measures. Paper presented at the conference Computational Approaches to Morphological Complexity, Surrey Morphology Group, Paris, February 2013 https://inria.hal.science/hal-00927276v1.
Sampson, Geoffrey. 2009. A linguistic axiom challenged. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 1–18. Oxford: Oxford University Press.
Sapir, Edward. 1921. *An introduction to the study of speech*. New York: Harcourt, Brace and Company.
Schober, Patrick, Christa Boer & Lothar A. Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia* 126(5). 1763–1768.

Schaaik, Gerjan van. 2020. *The Oxford Turkish grammar*. Oxford: Oxford University Press.

Shcherbakova, Olena, Susanne Maria Michaelis, Hannah J. Haynie, Sam Passmore, Volker Gast, Russell D. Gray, Simon J. Greenhill, Damián E. Blasi & Hedvig Skirgård. 2023. Societies of strangers do not speak less complex languages. *Science Advances* 9(33). https://doi.org/10.1126/sciadv.adf770.

Skirgård, Hedvig. 2025. Aggregate metrics derived from Grambank data. *GitHub wiki article*. https://github.com/grambank/grambank/wiki/Aggregate-metrics-derived-from-Grambank-data (accessed 22 December 2025).

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O. C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023a. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). https://doi.org/10.1126/sciadv.adg6175.

Skirgård, Hedvig, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browne, Lennart Chevallier, Swintha Danielsen, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Heer Leonard, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Alexandra Marley, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya, Michael Müller, Saliha Muradoglu, Hunter Gatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Jesse Peacock, India O. C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Amalia Skilton, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Wikalier Daniel Smith, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023b. Grambank v1.0. https://doi.org/10.5281/zenodo.7740140.

Spike, Matthew. 2017. Population size, learning, and innovation determine linguistic complexity. *Proceedings of the Annual Meeting of the Cognitive Science Society* 39. https://escholarship.org/uc/item/3x3355t4 (accessed 20 July 2026).

Sproat, Richard, Bruno Cartoni, HyunJeong Choe, David Huynh, Linne Ha, Ravindran Rajakumar & Evelyn Wenzel-Grondie. 2014. A database for measuring linguistic information content. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the ninth international conference on Language resources and evaluation (LREC'14)*, 967–974. Reykjavik: European Language Resources Association. https://aclanthology.org/L14-1397/.

Taulé, Mariona, Maria Antònia Martí & Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis & Daniel Tapias (eds.), *Proceedings of the sixth international conference on language resources and evaluation (LREC'08)*, 96–101. Marrakesh: European Language Resources Association (ELRA).

Wray, Alison & George W. Grace. 2007. The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua* 117(3). 543–578.

Zeman, Daniel, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, 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, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Ansu Berg, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Esma Fatıma Bilgin Taşdemir, Kristín Bjarnadóttir, Verena Blaschke, Rogier Blokland,

Victoria Bobicev, Loïc Boizou, Johnatan Bonilla, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Yifei Chen, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Bermet Chontaeva, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Courtin Marine, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Roberto Antonio Díaz Hernández, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Hoa Do, Kaja Dobrovoljc, Caroline Döhmer, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Roald Eiselen, Marhaba Eli, Elkahky Ali, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Soudabeh Eslami, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Edith Galy, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Tanja Gaustad, Efe Eren Genç, Fabrício Ferraz Gerardi, Gerdes Kim, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, 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, Takahiro Harada, Sam Hardwick, Kim Harris, Naïma Hassert, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Vidová Hladká, Jaroslava Hlaváčová, Florinel Hociung, Diana Hoefels, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Inessa Iliadou, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Federica Iurescia, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Mayank Jobanputra, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóǧa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Barbara Kovačić, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Anne-Marie Lutgen, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Francesco Mambrini, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Maitrey Mehta, Pierre André Ménard, Gustavo Mendonça, Tatiana Merzhevich, Meurer Paul, Niko Miekka, Emilia Milano, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, 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, Victor Norrman, Alireza Nourian, Maria das Graças Volpe Nunes, Nurmi Hanna, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Annika Ott, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A. Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Alistair Plum, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Rigardt Pretorius, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Christoph Purschke, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Paolo Ruffolo, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Xulia Sánchez-Rodríguez, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Albina Sarymsakova, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Sven Sellmer, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira,

Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinþór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Tarık Emre Tıraş, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn órðarson, Vilhjálmur orsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Anishka Vissamsetty, Natalia Vlasova, Eleni Vligouridou, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, John Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Enes Yılandiloğlu, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva & Rayan Ziane. 2024. Universal dependencies 2.14. http://hdl.handle.net/11234/1-5502.