Annotating and Inferring Compositional Structures in Numeral Systems Across Languages

Arne Rubehn[1], Christoph Rzymski[2], Luca Ciucci[1], Katja Bocklage[1], Alžběta Kučerová[1], David Snee[1], Abishek Stephen[3], Kellen Parker van Dam[1], Johann-Mattis List[1]

[1]Chair for Multilingual Computational Linguistics, University of Passau, Passau, Germany
[2]Department of Linguistic and Cultural Evolution, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
[3]Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic

Abstract

Numeral systems across the world's languages vary in fascinating ways, both regarding their synchronic structure and the diachronic processes that determined how they evolved in their current shape. For a proper comparison of numeral systems across different languages, however, it is important to code them in a standardized form that allows for the comparison of basic properties. Here, we present a simple but effective coding scheme for numeral annotation, along with a workflow that helps to code numeral systems in a computer-assisted manner, providing sample data for numerals from 1 to 40 in 25 typologically diverse languages. We perform a thorough analysis of the sample, focusing on the systematic comparison between the underlying and the surface morphological structure. We further experiment with automated models for morpheme segmentation, where we find allomorphy as the major reason for segmentation errors. Finally, we show that subword tokenization algorithms are not viable for discovering morphemes in low-resource scenarios.

1 Introduction

Numeral systems represented by the words for cardinal numbers used in counting are an interesting kind of linguistic data: they code a part of the lexicon of human languages that is potentially large and often exhibits a regularity that increases with higher numbers. Regularity is reflected in the *recycling* of linguistic material used to create higher numbers, where morphemes for smaller number words are often reused to motivate the formation of larger numerals. In addition, numeral systems are also maximally *distinctive*. Being used to distinguish ordinal numbers, we rarely find cases in which two distinct numbers are expressed by the same word form, even if numeral words themselves can have multiple meanings outside of the number domain (as can be easily seen when browsing number words in the *Database of Cross-Linguistic Colexifications*, Rzymski et al., 2020).

Another important aspect of numeral systems is that they are not created in an ad-hoc fashion but have instead often evolved over hundreds of years. The evolution can leave traces in numeral systems that counter-act former regularity, leading to allophonic variation in the morphemes that compose numeral words. Language contact can also feature as an important aspect of evolution, resulting in extreme cases where languages use two or more numeral systems in combination, reflecting different stages of their history.

The fact that most numeral systems are *compositional*, while at the same time being distinctive and discrete in their denotation, makes them an interesting test object for linguistic analyses that deal with lexical compositionality in the context of language change. While one would otherwise have to cope with problems resulting from various kinds of morphological and semantic variation, numeral systems can be seen as an ideal test ground for the annotation and inference of compositional structures in the lexicon of human languages. In the following, we will try to illustrate this point in more detail. After a short overview on numeral systems in the context of descriptive and computational linguistics (§ 2), we present a small collection of numeral systems along with methods that can be used to annotate numeral systems manually or to segment numeral words automatically into morphemes (§ 3). After testing these methods and reporting the results on our small cross-linguistic sample of numeral systems (§ 4) we discuss our findings and point to ideas for future work (§ 5).

2 Background

The cross-linguistic diversity of numeral systems has attracted the interest of scholars since Hervás y Panduro's comparative work (1786), which pre-

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 29–42

sented data from missionaries on many then little-known languages. Today, the most comprehensive database on numerals is Chan (2024), who collected data on more than 5,000 varieties, often provided by linguists with first-hand experience of the respective languages. The constant increase in data has allowed for the study of numeral systems from a formal (see e.g. Brandt Corstius, 1968; Hurford, 1975) and a typological perspective. The latter approach reached a turning point with Greenberg's (1978) 54 generalizations, most of which stood the test of time (Comrie, 2020).

Even though their synchronic structure may be opaque, numeral systems are diachronically motivated and are built through a limited number of cross-linguistic strategies (Heine, 1997, 18-34). They typically combine a small set of morphemes (mainly numbers, but also linking elements) according to three parameters, including (1) the choice of the base(s), (2) the operations applied to the base(s), following the implicational hierarchy: addition < multiplication < subtraction / division; and (3) the order of the morphemes (Greenberg, 1978; Moravcsik, 2017, 459-461). Despite the presumed regularity and compositionality of numeral systems, they may occasionally display gaps and ambiguities (Comrie, 1997, 2005, 79-80).

The most common bases are 'five', 'ten', and 'twenty', whose conceptual sources are, respectively, the fingers of the hand, of both hands, and of all hands and toes (Heine, 1997, 19-24; on finger counting and its cultural variability, see Bender and Beller, 2012). Decimal systems are the most frequent worldwide, followed by vigesimal and quinary systems (Skirgård et al., 2023). Languages can employ more than one base, resulting in hybrid numeral systems.

While languages with no numerals or only the number 'one' are rare (Hammarström, 2010), the numeral systems of many languages, particularly in South America, New Guinea and Australia, are restricted to a few numerals (Moravcsik, 2017, 459). According to Dixon (2012, 71-72), this indicates that the speakers did not count and enumeration was not the primary use of these number words. Hammarström (2008) observed that pidgins and creoles tend to have more complex numeral systems than the global average, with their frequent origin as trade languages being a potential contributing factor. Numeral systems often developed out of contact, which usually comes with societal change, and borrowing may also involve the lowest numbers (Dixon, 2012, 75-77).

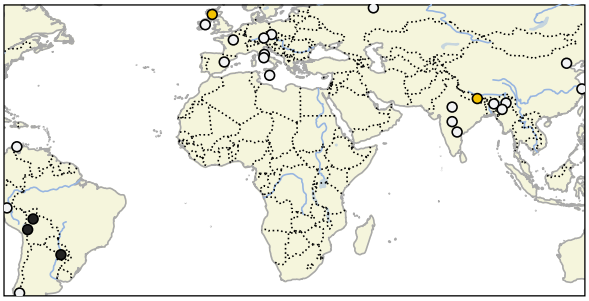

Figure 1: Geographical distribution of the languages in our sample, indicating the numeral bases they employ (white: 10, black: 5 and 10, orange: 10 and 20).

While numeral systems all over the world have been quite intensively investigated in the past, very few computational studies (Calude and Verkerk, 2016; Cathcart, 2025) formalize approaches to model compositionality and reveal motivation patterns underlying individual number words. Recent advances in the annotation of lexical motivation patterns (Hill and List, 2017) and the automated segmentation of words into morphemes (Goldsmith et al., 2017) open new possibilities for a computational investigation of numeral systems that we will discuss in more detail in the following.

3 Materials and Methods

3.1 Sample of Numeral Systems

We collected the cardinal numbers from 1 to 40 in 25 typologically diverse languages from Eurasia and Southern America, spanning ten different language families. Most language families, with the exception of Indo-European (12 languages) and Sino-Tibetan (5 languages), are represented by a single language. Table 1 provides a comprehensive overview of the languages covered in the sample, accompanied by a geographical visualization in Figure 1.

Most languages employ a decimal system, reflecting that the number 10 is by far the most common base. Three languages in our sample – Aymara, Cavineña, and Paraguayan Guaraní – make use of the number 5 as a base. They represent a hybrid between quinal and decimal systems, since the word for 10 is monomorphemic and used to express multiples of 10. Furthermore, two languages of our sample (Lamjung Yolmo and Scottish Gaelic) have retained a vigesimal system used in parallel to a

Family	Branch	Language	Base
Afro-Asiatic	Semitic	Maltese	10
Araucanian	—	Mapudungun	10
Arawak	Ta-Arawak	Wayuu	10
Aymaran	—	Aymara	5 / 10
Dravidian	Southern	Telugu	10
Indo-European	Balto-Slavic	Czech	10
		Russian	10
	Celtic	Irish	10
		Scottish Gaelic	10 / 20
	Germanic	German	10
	Indo-Iranian	Assamese	10
		Hindi	10
		Sanskrit	10
	Romance	French	10
		Italian	10
		Latin	10
		Spanish	10
Pano-Takanan	Takanan	Cavineña	5 / 10
Quechuan	Quechua I	Huallaga Quechua	10
Sino-Tibetan	Bodic	Lamjung Yolmo	10 / 20
	Brahmaputran	Uipo (Maringic)	10
	Patkaian	Makyam	10
	Sinitic	Mandarin Chinese	10
		Shanghainese	10
Tupian	Tupí-Guaraní	Paraguayan Guaraní	5 / 10

Table 1: Overview of languages covered in the sample, with their genetic classification and primary bases for counting.

decimal system, which results in alternating forms for numbers higher than 20.

All data were collected, annotated, and curated in a collaborative manner, such that the data for each language were thoroughly reviewed by at least two scholars: the responsible annotator for the given language, and at least one reviewer. The data were then aggregated and deployed as a unified dataset conforming to the Cross-Linguistic Data Formats (CLDF, Forkel et al., 2018; Forkel and List, 2020). Automated tests accounted for the structural integrity of the data (e.g. ensuring that one cognate ID does not map to more than one underlying form; the annotation format is described in detail in § 3.3).

3.2 Representing Numeral Systems in Tables

The CLDF specification builds on CSVW, a standard for tabular data on the web (https://csvw.org; Gower, 2021) that extends simple tabular data, typically represented in the form of CSV files, by metadata that can be used to specify the content of tabular data in various ways, including the combination of multiple tables in a relational database. Given that numeral data can be easily treated as *lexical data*, typically provided in the form of wordlists, we represent number systems as extended CLDF wordlists that build on the extended wordlist formats introduced by the Lexibank repository (List et al., 2022; Blum et al., 2025). Lexibank wordlists represent individual word forms as triples consisting of a *language*, a *concept*, and a *form*. In order to compare data from different sources, Lexibank makes use of reference catalogs that link language varieties to Glottolog (https://glottolog.org; Hammarström et al., 2025), map concepts to Concepticon (https://concepticon.clld.org; List et al., 2025a), and represent phonetic transcriptions compatible with the subset of the IPA proposed by the Cross-Linguistic Transcription Systems (CLTS) reference catalog (https://clts.clld.org; List et al., 2021).

While following Lexibank in assembling our exploratory database of numeral systems, we extend the format by adding new layers of annotation that help us to make individual analyses of the numeral systems explicit through annotation. As a first step, we rigorously split words into morphemes by adding morpheme boundary markers to all multi-morphemic words (using the plus symbol – + – as a boundary marker). As a second step, we identify language-internal partial cognates in all numeral systems in order to mark the degree by which morphemes are reused to build new numeral expressions (see List et al., 2016 on partial cognates). In other words, we annotate which morphemes recur across several forms by assigning a unique numerical ID to each morpheme. As a third step of analysis, we add *morpheme glosses* to the data to add human-readable semantic descriptors to all morphemes (Hill and List, 2017; Schweikhard and List, 2020). As a fourth step, we make use of *inline-alignments* in order to handle allomorphs by distinguishing underlying forms from surface forms (List, 2025). As a fifth step, we conduct *phonetic alignment analyses* (List, 2014) of all language-internal cognate morphemes, in order to facilitate the comparison of allomorphic variants that differ in length.

Table 2 shows how our annotations are rendered in tabular form, with examples for annotated numerals from German and French. The column *Segments* provides phonetic transcriptions, segmented into sounds, using a space as boundary marker, and secondarily segmented into morphemes, using the plus symbol as a boundary marker. The transcriptions use *inline alignments* (List, 2025) to align the surface forms with their underlying forms. Inline alignments (first introduced by List 2021 and later tested on Old Chinese etymologies by Pulini and

Language	Concept	Form	Segments	Cognates	Morphemes
German	one	*eins*	aɪ n s	1	ONE
German	two	*zwei*	ts v aɪ	2	TWO
German	three	*drei*	d r aɪ	3	THREE
German	twenty one	*einundzwanzig*	ai n -/s + ʊ n -/d + ts v a n + ts ɪ ç	1 4 5 6	ONE and TWEN TY
German	thirty two	*zweiunddreißig*	ts v aɪ + ʊ n -/d + d r aɪ + s/ts ɪ ç	2 4 3 6	THREE and THREE TY
French	one	*un*	œ̃	1	ONE
French	two	*deux*	d ø	2	TWO
French	three	*trois*	t ʁ w a	3	THREE
French	twenty one	*vingt-et-un*	v ɛ̃ t + e + œ̃	4 5 1	TWENTY and ONE
French	thirty two	*trente-deux*	t ʁ -/w -/ɑ + ɑ̃ t + d ø	3 6 2	THREE TY TWO

Table 2: Illustration of the format used to annotate morpheme boundaries along with allomorphic variation, language internal cognates, and morpheme glosses.

List 2024) use the slash symbol (/) in order to distinguish a surface sound (shown to the left of the slash) from its corresponding underlying sound. As an example, consider the transcription of German [aɪ n -/s] 'one' in the word for 'twenty one' in the table, where [s] is treated as the underlying form, while the surface form does not show this sound (which is marked by using the gap-symbol - before the slash). The notion of surface form and underlying form is strictly *technical*. We assume that one morpheme with multiple allomorphs has only one underlying form, which must consistently be aligned with all surface forms. We do not claim that this handling shows any cognitive or historical truth, but we aim for an annotation that would ideally be meaningful from a diachronic and cognitive perspective.

The columns *Cognates* and *Morphemes* provide information on language-internal cognates in the form of morphemes that are reused. Here, the *Cognates* column employs numerical identifiers, following the format proposed by List et al. (2016), while the functionally identical *Morphemes* column provides semantic glosses that help in making the lexical motivation underlying the formation of numerals transparent. This annotation, which provides explicit glosses for all morphemes constituting a word, was originally developed to make language-internal cognate relations more explicit (Hill and List, 2017). By now, however, it has been shown to be also very useful to provide rudimentary annotations of lexical motivation patterns (Brid et al., 2022).

3.3 Computer-Assisted Annotation

While the annotations shown in Table 2 can be easily carried out with the help of a spreadsheet editor or directly in text files, we use the web-based EDICTOR tool for the annotation of numeral data (List et al., 2025b). Originally, EDICTOR was designed to facilitate the process of creating multilingual comparative wordlists (List, 2017). Since Version 3.0 (List and van Dam, 2024), however, EDICTOR has been substantially extended to help with the annotation of lexical motivation patterns. Improvements include – among others – a visual rendering of inline alignments, sound sequences, cognate sets, and morpheme glosses, combined with annotation helpers for manual morpheme segmentation, as well as several sanity checks that increase the consistency of human annotation.

3.4 Automated Morpheme Segmentation

The task of unsupervised morpheme segmentation – automatically inferring a language's morphological structure from unannotated corpus data – has received notable attention in the field of Natural Language Processing, especially in the late 1990's and early 2000's (Hammarström and Borin, 2011). While those models were developed with a different background in mind, assuming the presence of relatively large training corpora, numeral systems naturally lend themselves as an interesting use case for morpheme segmentation models due to their high degree of compositionality. Therefore, we experiment with simple morpheme segmentation techniques to observe their performance in a transfer setting with much less data, but an extraordinarily strong morphological signal.

The first formalization of an algorithm for morpheme segmentation reaches back to Harris (1955) who proposed the so-called *Letter Successor Variety* (LSV) as a predictability measure at each position within a word. The underlying assumption is that the continuation of a word should be fairly predictable within a morpheme, but much harder to

predict at a morpheme boundary. Several proposals have been made to improve upon LSV. Hafer and Weiss (1974) suggest measuring predictability in terms of entropy rather than type variety (Letter Successor Entropy, LSE). They also propose Letter Predecessor Variety (or Entropy) as a logical inversion of LSV, processing each word backwards. Hammarström (2009) proposes *Letter Successor Max-Drop*, measuring how likely the most frequent continuation of a word is in comparison to all other potential continuations. Çöltekin (2010) suggests normalizing LSV by word position to account for the fact that LSV usually becomes smaller towards the end of a word. We experiment with all these different flavors of LSV, but report only LSE, since it performs best on average and all LSV variations show similar patterns in general. Following Hafer and Weiss (1974), we also experiment with a simple model that considers every possible prefix and suffix (in a computational sense) of a word form as a morpheme if and only if it appears as a complete word in the data. Using this simple measure, List (2023) reports promising results in inferring partial colexifications from multilingual wordlists which seem to advance concept embeddings substantially (Rubehn and List, 2025b).

A line of research that can be seen as complementary to LSV-based approaches formalizes the task of morpheme segmentation as a *minimum description length* (MDL) problem (Goldsmith, 2001). The basic idea behind MDL is to define a description length as a combination of basic tokens and rules to derive complex forms from the basic vocabulary. This notion is especially interesting on theoretical grounds, since the complexity of numeral systems can also be measured in terms of MDL (Hammarström, 2008). In an ideal setting, an MDL-based segmentation model is therefore expected to accurately infer and model the compositional structure of numeral systems. Representing this family of morpheme segmentation algorithms, we run our experiments with the Morfessor Baseline model (Creutz and Lagus, 2002, 2005; Virpioja et al., 2013).

3.5 Subword Tokenization

Algorithms for *subword tokenization* form an integral preprocessing step of state-of-the-art language models, since they effectively reduce the vocabulary size and avoid the occurrence of out-of-vocabulary items. While these tokenization methods in principle make downstream applications more flexible, it can at least be doubted whether the inferred subwords concord with the language's morphological structure (Batsuren et al., 2024). We apply three popular algorithms for subword tokenization on our multilingual numeral data: Byte-Pair-Encoding (BPE; Gage, 1994; Sennrich et al., 2016), WordPiece (Schuster and Nakajima, 2012), and Unigram tokenization (Kudo, 2018).

3.6 Evaluation

All models described in § 3.4 and § 3.5 are trained on unannotated and unsegmented representations of the numeral lists. The predicted segmentations are then evaluated against our manual annotations which serve as a gold standard. Since all models are inherently monolingual, each language is processed and evaluated independently.

Predicted segmentations can directly be evaluated against the gold standard using *precision* and *recall* (Virpioja et al., 2011). While we are aware of more sophisticated evaluation metrics for morphological analyses (Spiegler and Monson, 2010), we argue that simply calculating boundary precision and recall (BPR) is sufficient in our use case, since we investigate small corpora with hardly ambiguous morphological patterns. Due to its simplicity, BPR is readily interpretable, rendering it the ideal evaluation metric for our use case.

We run all experiments on two different representations of the numeral lists, relying on the *surface* and *underlying* forms respectively (see § 3.3 for details on the two representations). The former is a faithful representation of the actually observable word forms and therefore reflects a "real-world" use case for segmentation models. The latter is an artificially construed "ideal" setting that removes allophonic and allomorphic variation, that is, variation that needs to be explained on a different level than morphology. Comparing these two settings allows for a fine-grained evaluation of morpheme segmentation models, enabling us to assess the share of segmentation errors caused by allomorphy.

3.7 Implementation

The data were annotated using EDICTOR 3.1 (List et al., 2025b), and validated and compiled using CLDFBench (Forkel and List, 2020). The visualization in Figure 1 was created using CLDFViz (Forkel, 2024). All experiments regarding automated morpheme segmentation were run in Python, using LinSe (Forkel and List, 2024) to conveniently

	Average		Highest		Lowest	
	S	**U**	**S**	**U**	**S**	**U**
Morphemes	21.8	13.5	48	20	10	7
Expressivity	5.6	7.9	10.6	15	1.4	3.4
Opacity	1.60		3.18		1	
Code Length	2.53		3.83		1.68	

Table 3: Overview of statistics about the different numeral systems. **S** and **U** refer – where applicable – to surface vs. underlying forms.

represent the internal structure of word forms in different granularities. Morfessor was run from its Python package (Virpioja et al., 2013), all other models were implemented from scratch and are available through MorSeg, a package for morpheme segmentation in multi- and monolingual wordlists (Rubehn and List, 2025a). All data and code accompanying this study are made available in the supplementary material.

4 Analysis and Results

4.1 Sample Data of Coded Numeral Systems

Table 3 summarizes the results of computing different types of metrics based on surface and underlying forms across all languages in our sample (Table 6 in the appendix provides metrics for individual languages). In the table, we introduce three simple metrics – *expressivity*, *opacity*, and *length* – to get a better understanding of the data and the strategies to form higher numbers from basic morphemes. First, we measure the average *morpheme expressivity* of a language by counting how many different numbers are formed using this morpheme. For instance, Mandarin *wǔ* 'five' is used in the formation of the numbers 5, 15, 25, and 35 and therefore has an expressivity of 4. Expressivity is averaged over all morphemes found in a language's numeral system. For the rare cases where a language has multiple forms for the same number, expressivity is weighted accordingly. *Opacity* describes the ratio between allomorphic variants and morphemes, measuring the degree of allomorphy in a system. The lowest score is 1, with each morpheme in a language surfacing with the same form. Finally, the *average coding length* measures how many morphemes are used to form a word.

On theoretical grounds, the minimum amount of morphemes required in a numeral system is the base of that system. That means, a decimal system needs at least 10 different morphemes to be fully expressive. Indeed, our sample covers three languages – Mandarin, Mapudungun, and Huallaga Quechua – that use such a minimal decimal system to express the numbers up to 40. This observation holds true on both the surface and the underlying level, indicating that exactly these languages lack any kind of allomorphy. Mandarin is often taken as a prime example for a perfectly transparent and symmetric numeral system: Complex numerals are simply formed by concatenating the simple numbers from 1 to 10. For example, *twenty three* in Mandarin is *èr shí sān*, literally *two ten three* ($2 * 10 + 3$).

On the other side of the spectrum, we find Assamese with 20 different morphemes and Hindi with 48 distinct morphs, the highest value for the respective category. This aligns with the general impression that Indo-Aryan languages feature some of the most complex and opaque numeral systems of the world (Hammarström, 2008; Cathcart, 2025).

We observe a wide range of morpheme opacity. With Uipo, Huallaga Quechua, Mandarin, and Mapudungun, four languages in our sample have the lowest possible opacity of 1.0, thus not featuring any allomorphy in their numeral systems. On the other hand, the language with the highest opacity is still Hindi with a value of 2.82, followed by Lamjung Yolmo, Telugu, and Sanskrit. From these extreme cases, the impression might arise that the opacity correlates with the size of the underlying morpheme inventory. However, across the entire dataset, no significant correlation between these two metrics could be found.

The expressivity of morphemes and their allomorphic variants, on the other hand, shows a significant negative correlation with the number of morphemes. The interpretation is straightforward: The fewer morphemes are available in a system, the more expressive they need to be, and the more they will be used. It is therefore not surprising that exactly those three languages that employ a base of 5 (Aymara, Cavineña, and Paraguayan Guaraní) rank the highest in terms of expressivity on the surface and the underlying level. On the low end of expressivity, we again find Hindi and Assamese, as well as the modern Romance languages French, Italian, and Spanish.

Based on these correlations, one might expect that the average coding length is also directly dependent on the size of the morpheme inventory, since less available morphemes should – in theory – require longer word forms. However, no significant correlation between these two metrics could

be found. There is only a significant correlation between the coding length and the morpheme expressivity. Considering that our sample is heavily biased towards decimal systems, and that even the systems that employ other bases show traces of decimal coding, we cannot interpret these effects as a result of different numeral bases. Instead, this seems to result from oblique marking (connecting morphemes with particles like 'and' or 'with') which can happen independently of the numeral base.

Finally, we experiment with *type-token ratio* (TTE) and *entropy*, which have been proposed as measures of morphological complexity in the past (Bentz et al., 2017; Çöltekin and Rama, 2023). These metrics are not able to capture any aspect of complexity in our sample, since they correlate almost perfectly with the number of morphemes. We therefore conclude that in this special setting, TTE and entropy are dependent on the vocabulary size alone, which is probably due to the fact that morphemes in numeral systems by and large do not follow a Zipfian distribution, as is the case for words in natural language corpora.

4.2 Automated Morpheme Segmentation

Table 4 reports the overall performance of three models for automated morpheme segmentation on the individual languages, both for those cases where surface forms were passed to the algorithms, and where underlying forms were taken as the basis of analysis. From the model family based on Letter Successor Variety, we only report Letter Successor/Predecessor Entropy, which generally performed best.

The most obvious (and unsurprising) observation is that all models perform better on the underlying form than on the surface forms. Since it is a well-known issue in the literature that automated methods are challenged by allomorphy (Hammarström and Borin, 2011; Virpioja et al., 2011), this does not seem too surprising to us. Comparing the average scores of the models, however, shows that allomorphy is the biggest source of error for the analysis on surface forms, which naturally is the common use case for those models. By extension, it does not come as a surprise that opacity significantly correlates with how well the models perform on the surface forms, as shown in Figure 2.

Model	Surface Forms	Underlying Forms
Morfessor	0.74	0.88
LSPE	0.72	0.83
Affix	0.72	0.88

Table 4: Average F_1 scores of morpheme segmentation algorithms.

But even on the underlying forms – an "ideal" scenario in which allomorphy does not exist – there are notable differences in how well the morphological structure is detected by the models. Particularly interesting is the case of Uipo. This numeral system poses a big challenge for Morfessor and the Affix model, which both only achieve an F_1-score of 0.4 (while achieving a perfect precision of 1.0!). A closer look at the language data reveals that Uipo has a complex numeral system, in which even the numbers between 2 and 9 consist of two morphemes, a prefix and a stem. The number 6 for example is [t^h ɘ + r u k], but both morphemes are only used to form the number six (and by extension, numbers that are formed using 'six'). Without any further knowledge of the language, it is very hard if not impossible to recognize the underlying compositionality, leading to a massive undersegmentation by the models at hand. On the other hand, the high score of LSPE on Uipo – which may come as a surprise – can be described as a coincidental byproduct of the present morphophonology. As generally typical for South-East Asian languages, Uipo only allows the simple syllable structure CV(C), and each syllable in Uipo is a morpheme at the same time. Since there are more consonants than vowels, the continuation of a word is much less predictable at the start of a new syllable. LSPE can therefore accurately predict *syllable* boundaries, which happen to be morpheme boundaries as well.

On the other side, Morfessor is able to perfectly predict all morpheme boundaries in four languages at the surface level (Shanghainese, Mandarin, Huallaga Quechua, Mapudungun), and in seven more languages at the underlying level. Mapudungun seems to have a particularly transparent structure, since it is the only language that all three models segment perfectly at both representation levels. This makes Morfessor the model with the highest number of completely correct segmentations at the language level, showing that it clearly has the edge over the other two approaches tested, which is also indicated by the average performance. But even in this "ideal" scenario – no allomorphy and a system that shows clear compositional structures – Morfessor cannot accurately predict all morpheme boundaries for 14 out of 25 languages. For example, in

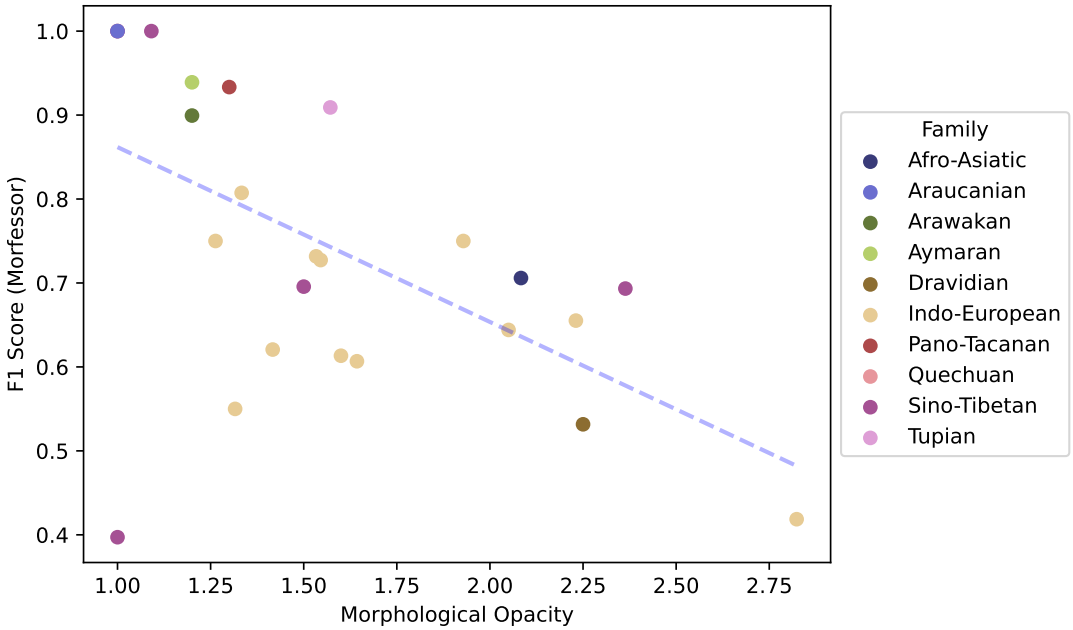

Figure 2: F1 scores of Morfessor on surface forms per language in correlation with the morphological opacity (Spearman's $\rho = -0.596$, p-value < 0.01).

the German words *zwan-zig* 'twen-ty' and *drei-ßig* 'thir-ty', the model fails to detect the morpheme boundaries, even in the underlying form where *-zig* [ts ɪ ç] and *-ßig* [s ɪ ç] are represented in the same way ({ts ɪ ç}). Generally, the model is much more prone to undersplitting than to oversplitting: On the underlying representation, it achieves a nearly perfect precision of 0.998, but a recall of only 0.80.

4.3 Subword Tokenization

Table 5 provides an overview of how accurately algorithms for subword tokenization can capture the morphological structure of the numeral systems at hand. It is evident that these models are in no way competitive with algorithms designed for the task of morphological segmentation – even the simplest segmentation algorithms outperform the subword algorithms largely. Among the subword tokenization algorithms, BPE performed the best on both levels, and the Unigram model performed worst across the board.

There are two major conceptual issues that inhibit a successful transfer of these algorithms to morpheme segmentation. First, these models only operate extremely locally – BPE and WordPiece merge bigrams based on a simple co-occurence metric, and Unigram removes unlikely *n*-grams under the assumption that the distribution of all tokens in the vocabulary is statistically independent. This prevents the models from learning relevant information about longer shared substrings, which is the foundation for all successful morpheme segmentation models. The second, and arguably strongest limiting factor is that it is unclear how to determine when a model should stop. In their intended setting, subword tokenization algorithms are designed to define an expressive vocabulary of a tractable size for downstream NLP applications. Hence, a desired vocabulary size is defined a priori, and the subword vocabulary is continually modified until the predefined size is reached. For BPE and WordPiece, the vocabulary size increases monotonically during that process, while it decreases for Unigram. In this context, vocabulary size refers to the number of unique subwords modeled by the respective tokenizer.

Model	Surface Forms	Underlying Forms
BPE	0.51	0.61
WordPiece	0.38	0.36
Unigram	0.34	0.33

Table 5: Average F_1 scores of subword tokenization algorithms for morphological segmentation.

This training set-up leads to two problems. The first is that the desired vocabulary size must be defined before running the model. For morphological segmentation, the ideal vocabulary size naturally will be the size of the morpheme inventory – but if that is already known, then no automated morphological analysis is required anymore. For the

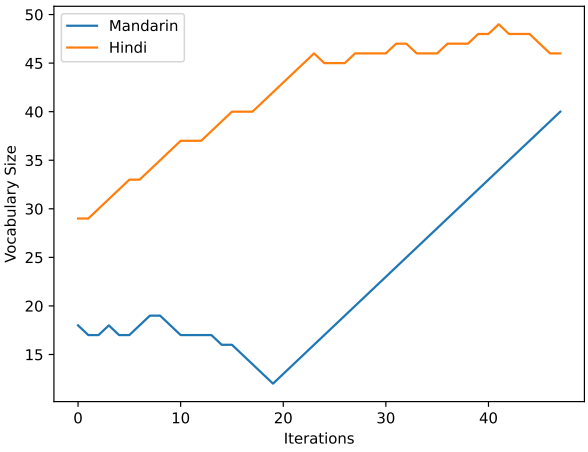

Figure 3: Vocabulary size of the BPE tokenizer for Mandarin and Hindi after each iteration.

sake of illustration, we ran the algorithms under the unrealistic assumption that the ideal vocabulary size is already known; so that each model stopped the training routine once that size was reached. The numbers shown in Table 5 therefore report the performance of an ideal setting for the models, including information that would be unknown in a practical application. BPE and WordPiece reached that ideal vocabulary size only in 11 out of 100 cases (and even then did not provide an ideal morphological segmentation by any means). An accurate reduction of the vocabulary to its minimal representation was therefore rarely achieved.

The second problem results from the assumption that BPE and WordPiece lead to a monotonic increase of vocabulary size. This assumption does not hold true in the special case of numerals: Thanks to the high degree of compositionality, the smallest possible vocabulary size to construct the data is not necessarily the set of individual characters, but can be the set of employed morphemes instead. This is visualized in Figure 3: The Mandarin numerals only require 10 morphemes to construct numerals up to 40, while 19 distinct segments can be found in these forms. By subsequently merging common bigrams, the BPE algorithm is actually able to *reduce* the vocabulary size to these ten morphemes. The monotonicity assumption implied by subword algorithms therefore might be violated, and the vocabulary size might *decrease* for a while, depending on the complexity of a language's morphology and phonology. However, this is not necessarily the case, as in more opaque languages like Hindi, the vocabulary size still increases monotonically with more iterations.

5 Discussion and Conclusion

In this study, we have demonstrated an efficient, transparent, and robust workflow for the annotation and analysis of numeral systems. The workflow features a detailed annotation scheme for shared morphemes across word forms, accounts for potential allomorphy, and can be carried out in a computer-assisted manner, using a web-based annotation tool. As a result, we presented a small sample of annotated numeral systems from 25 typologically diverse languages from Eurasia and South America. We used this sample to evaluate how well unsupervised methods for automated morpheme segmentation work in extremely low-resource scenarios with an extraordinarily strong morphological signal. The results suggest that the major error source of these models is allomorphy. When this factor is accounted for, rather satisfactory morphological analyses can be inferred automatically. For future research on morpheme segmentation in low-resource scenarios, the handling of allomorphy will therefore be crucial.

Several statistical measures of numeral systems introduced here confirm intuitive correlations, such that smaller morpheme inventories necessarily entail a higher expressivity of the individual morphemes. It remains unclear, however, if a measure of morphological complexity can be inferred from our measures, since information-theoretic approaches that have been proposed to measure morphological complexity on corpus data do not convey any useful information about the morphological structure of numeral systems. Curiously, it seems that the performance of Morfessor aligns with (impressionistic) human judgement of how transparent a numeral system is. Since Morfessor is based on the Minimum Description Length (MDL, Rissanen, 1983) principle, which has been proposed as a framework for measuring complexity in numeral systems (Hammarström, 2008; Cathcart, 2025), it might serve as a useful indicator for complexity when applied on the underlying data representation.

We conclude that due to their high degree of compositionality, numerals serve as an ideal controlled sample for developing and testing the annotation and inference of morphological structures in multilingual wordlists. In the future, we hope to further expand our sample of numeral systems and test more methods for automated morpheme segmentation.

Supplementary Material

The dataset for compositional structures in numeral systems (*CoSiNuS*, Version 1.1) is curated on GitHub (https://github.com/numeralbank/cosinus) and archived with Zenodo (https://doi.org/10.5281/zenodo.15656420). The MorSeg software package is curated on GitHub (https://github.com/lingpy/morseg, Version 0.1) and archived with PyPi (https://pypi.org/project/morseg). The code that was used to run the analyses described in this study is curated on Codeberg (https://codeberg.org/calc/numeral-annotation-study, Version 1.1) and archived with Zenodo (https://doi.org/10.5281/zenodo.15672425).

Limitations

The annotation of word forms that etymologically share the same origin, but have diverged over a substantial amount of time, is not always clear and can be ambiguous. For example, consider Spanish *once* (11): There is no transparent, synchronous pattern that would combine *uno* (1) and *diez* (10) to yield this form. However, we know that this was historically the case, as proven by Latin *undecim*, which is a clear compound from *un-* (1) and *decem* (10). In Italian, this compounding strategy is still transparently visible (*un-* + *dieci* = *ùndici*). Arguably, this lexical motivation is still transparent enough in Italian to annotate it as dimorphemic form, but not in Spanish (even though the etymology and the time depth is identical). A similar case can be observed for the Gaelic languages, where the suffix for deriving tens (Irish: *déag*; Scottish: *deug*) is clearly related to the word for ten (*deich* in both languages), but the exact historical connection is unclear (Matasović, 2009, 93-94; MacBain, 1911, 130).

A further limitation to the current annotation scheme is that it linearly segments complex forms into morphemes, for example *two ten three*. The annotation does not make the underlying arithmetic process explicit: Understanding that the underlying formula would be $2 * 10 + 3$, if the word means 'twenty three', requires an additional interpretation step and is not explicitly coded in the annotation scheme.

Due to its relatively small size of 25 languages, the patterns observed in the data might not reflect universal patterns, especially considering the choice of languages. While we tried to include typologically diverse languages, we are aware that our sample is heavily biased towards Indo-European and Sino-Tibetan languages, and that the macroareas of North America, Africa, and Papunesia are not represented at all.

We furthermore observe a heavy bias towards decimal systems, and even those systems that are not primarily decimal contain some decimal structures. It is therefore impossible to systematically analyze different numeral bases beyond some impressionistic analyses. Finally, it remains an open question if (and how) the morphological complexity of a numeral system or a language in general can be measured.

Acknowledgments

This project was supported by the ERC Consolidator Grant ProduSemy (AR, LC, AK, KB, DS, JML; Grant No. 101044282, see https://doi.org/10.3030/101044282), the ERC Synergy Grant QUANTA (CR; Grant No. 951388, see https://doi.org/10.3030/951388), and the Charles University (AS; project GA UK No. 101924, see https://ufal.mff.cuni.cz/node/2690). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency (nor any other funding agencies involved). Neither the European Union nor the granting authority can be held responsible for them.

Author Contributions

CR and JML initiated the study and devised the annotation scheme. AR, CR, and JML were responsible for the data management. AR implemented the algorithms for unsupervised morpheme segmentation and subword tokenization (with contributions by AS and JML) and conducted the experiments. AR, LC, KPvD, AK, KB, and DS contributed annotated data. AR, LC, and JML wrote and revised the draft.

References

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating subword tokenization: Alien subword composition and oov generalization challenge. *arXiv preprint arXiv:2404.13292*.

Andrea Bender and Sieghard Beller. 2012. Nature and culture of finger counting: Diversity and representational effects of an embodied cognitive tool. *Cognition*, 124(2):156–182.

Christian Bentz, Dimitrios Alikaniotis, Michael Cysouw, and Ramon Ferrer-i-Cancho. 2017. The entropy of words – learnability and expressivity across more than 1000 languages. *Entropy*, 19(6).

Frederic Blum, Carlos Barrientos, Johannes Englisch, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, and Johann-Mattis List. 2025. Lexibank 2: precomputed features for large-scale lexical data [version 1; peer review: 3 approved]. *Open Research Europe*, 5(126):1–19.

Hugo Brandt Corstius, editor. 1968. *Grammars for number words*. Reidel, Dordrecht.

Nicolás Brid, Cristina Messineo, and Johann-Mattis List. 2022. A comparative wordlist for the languages of the gran chaco, south america [version 2; peer review: 2 approved]. *Open Research Europe*, 2(90):1–17.

Andreea S. Calude and Annemarie Verkerk. 2016. The typology and diachrony of higher numerals in Indo-European: a phylogenetic comparative study. *Journal of Language Evolution*, 1(2):91–108.

Chundra Cathcart. 2025. Complexity counts: global and local perspectives on Indo-Aryan numeral systems. *arXiv preprint 2505.21510*, pages 1–30.

Eugene Chan. 2024. Numeral systems of the world's languages. `https://lingweb.eva.mpg.de/channumerals/`. Version updated on February 18, 2024.

Çağrı Çöltekin. 2010. Improving Successor Variety for Morphological Segmentation. In *Proceedings of the 20th Meeting of Computational Linguistics in the Netherlands*, volume 16, pages 13–28.

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

Bernard Comrie. 1997. Some problems in the theory and typology of numeral systems. In Bohumil Palek, editor, *Proceedings of LP'96. Typology: Prototypes, item orderings, and universals. Proceedings of the conference held in Prague August 20–22, 1996*, pages 41–56. Charles University Press, Prague.

Bernard Comrie. 2005. Endangered numeral systems. In Jan Wohlgemut and Tyro Dirksmeyer, editors, *Bedrohte Vielfalt Aspekte des Sprach(en)tods: Aspects of language death*, pages 203–230. Weissensee, Berlin.

Bernard Comrie. 2020. Revisiting Greenberg's "Generalizations about numeral systems" (1978). *Journal of Universal Language*, 21(2):43–84.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology Helsinki.

R. M. W. Dixon. 2012. *Basic linguistic theory, Vol. 3: Further grammatical topics*. Oxford University Press, Oxford.

Robert Forkel. 2024. *CLDFViz. A Python library providing tools to visualize data from CLDF datasets [Software Library, Version 1.3.0]*. Zenodo, Geneva.

Robert Forkel and Johann-Mattis List. 2020. CLDF-Bench. Give your cross-linguistic data a lift. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*, page 6997-7004, Luxembourg. European Language Resources Association (ELRA).

Robert Forkel and Johann-Mattis List. 2024. A new Python library for the manipulation and annotation of linguistic sequences. *Computer-Assisted Language Comparison in Practice*, 7(1):17–23.

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific data*, 5(1):1–10.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

John A. Goldsmith. 2001. Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

John A. Goldsmith, Jackson L. Lee, and Aris Xanthos. 2017. Computational learning of morphology. *Annual Review of Linguistics*, 3:85–106.

Robin Gower. 2021. *CSV on the Web*. Swirrl, Stirling.

Joseph H. Greenberg. 1978. Generalizations about numeral systems. In Joseph H. Greenberg, editor, *Universals of Human Language, Vol. 3: Word structure*, pages 249–295. Stanford University Press, Stanford.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information storage and retrieval*, 10(11-12):371–385.

Harald Hammarström. 2008. Complexity in numeral systems with an investigation into pidgins and creoles. In Matti Miestamo, Kaius Sinnemäki, and Fred Karlsson, editors, *Language Complexity. Typology, contact, change*, volume 94 of *Studies in Language Companion Series*. John Benjamins Publishing Company.

Harald Hammarström. 2009. *Unsupervised Learning of Morphology and the Languages of the World*. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg.

Harald Hammarström. 2010. Rarities in numeral systems. In Jan Wohlgemuth and Michael Cysouw, editors, *Rethinking universals: How rarities affect linguistic theory*, volume 45 of *Empirical Approaches to Language Typology*, pages 11–60. Mouton de Gruyter, Berlin.

Harald Hammarström and Lars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309–350.

Harald Hammarström, Martin Haspelmath, Robert Forkel, and Sebastian Bank. 2025. *Glottolog [Dataset, Version 5.2]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Zellig S Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190.

Bernd Heine. 1997. *Cognitive foundations of grammar*. Oxford University Press, New York and Oxford.

Lorenzo Hervás y Panduro. 1786. *Aritmetica delle nazioni e divisione del tempo fra l'Orientali*. Gregorio Biasini, Cesena.

Nathan W. Hill and Johann-Mattis List. 2017. Challenges of annotation and analysis in computer-assisted language comparison: A case study on Burmish languages. *Yearbook of the Poznań Linguistic Meeting*, 3(1):47–76.

James R. Hurford. 1975. *The linguistic theory of numerals*. Cambridge University Press, Cambridge.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

Johann-Mattis List. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*, pages 9–12, Valencia. Association for Computational Linguistics.

Johann-Mattis List. 2021. Using edictor 2.0 to annotate language-internal cognates in a german wordlist. *Computer-Assisted Language Comparison in Practice*, 4(4).

Johann-Mattis List. 2023. Inference of partial colexifications from multilingual wordlists. *Frontiers in Psychology*, 14(1156540):1–10.

Johann-Mattis List. 2025. Productive signs: Towards a computer-assisted analysis of evolutionary, typological, and cognitive dimensions of word families. In David Bradley, Katarzyna Dziubalska-Kołaczyk, Camiel Hamans, Ik-Hwan Lee, and Frieda Steurs, editors, *Contemporary Linguistics: Integrating Languages, Communities, and Technologies*, pages 403–412. Brill, Leiden.

Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, and Robert Forkel. 2021. *Cross-Linguistic Transcription Systems. [Dataset, Version 2.1.0]*. Max Planck Institute for the Science of Human History, Jena.

Johann-Mattis List, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray. 2022. Lexibank, a public repository of standardized wordlists with computed phonological and lexical features. *Scientific Data*, 9(316):1–31.

Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*, pages 599–605, Berlin. Association of Computational Linguistics.

Johann-Mattis List, Annika Tjuka, Frederic Blum, Alžběta Kučerová, Carlos Barrientos Ugarte, Christoph Rzymski, Simon J. Greenhill, and Robert Forkel. 2025a. *CLLD Concepticon [Dataset, Version 3.4.0]*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Johann-Mattis List and Kellen Parker van Dam. 2024. Computer-assisted language comparison with EDICTOR 3 [invited paper]. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 1–11, Bangkok, Thailand. Association for Computational Linguistics.

Johann-Mattis List, Kellen Parker van Dam, and Frederic Blum. 2025b. *EDICTOR 3. An Interactive Tool for Computer-Assisted Language Comparison [Software Tool, Version 3.1]*. MCL Chair at the University of Passau, Passau.

Alexander MacBain. 1911. *An etymological dictionary of the Gaelic language*. Eneas Mackay, Stirling.

Ranko Matasović. 2009. *Etymological dictionary of Proto-Celtic*. Brill, Leiden, Boston.

Edith A. Moravcsik. 2017. Number. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *The Cambridge handbook of linguistic typology*, pages 440–476. Cambridge University Press, Cambridge.

Michele Pulini and Johann-Mattis List. 2024. Finding language-internal cognates in Old Chinese. *Bulletin of Chinese Linguistics*, 17(1):53–72.

Jorma Rissanen. 1983. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431.

Arne Rubehn and Johann-Mattis List. 2025a. *MorSeg: A Python package for morpheme segmentation in multi- and monolingual wordlists [Software Library, Version 0.1]*. Chair for Multilingual Computational Linguistics, University of Passau.

Arne Rubehn and Johann-Mattis List. 2025b. Partial colexifications improve concept embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria. Association of Computational Linguistics.

Christoph Rzymski, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel, and Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific Data*, 7(13):1–12.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152, Kyoto, Japan. IEEE.

Nathanael E. Schweikhard and Johann-Mattis List. 2020. Developing an annotation framework for word formation processes in comparative linguistics. *SKASE Journal of Theoretical Linguistics*, 17(1):2–26.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Hedvig Skirgård, 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, Tânia 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, and Russell D. Gray. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).

Sebastian Spiegler and Christian Monson. 2010. EMMA: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1029–1037, Beijing, China. Coling 2010 Organizing Committee.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. In *Aalto University publication series SCIENCE + TECHNOLOGY*, 25. Aalto University, Helsinki, Finland.

Sami Virpioja, Ville T Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2):45–90.

A Statistics for Individual Languages

Language	Morph.	Expressivity	Opacity	Length	Morfessor	LSPE	Affix
Maltese	25 / 12	4.24 / 8.83	2.08	2.65	0.71 / 1.00	0.64 / 0.84	0.71 / 0.84
Mapudungun	10 / 10	8.80 / 8.80	1.00	2.20	1.00 / 1.00	1.00 / 1.00	1.00 / 1.00
Wayuu	18 / 15	7.26 / 8.71	1.20	3.19	0.90 / 0.90	0.89 / 0.95	0.70 / 0.70
Aymara	12 / 10	10.58 / 12.70	1.20	3.17	0.94 / 1.00	0.67 / 0.65	0.73 / 0.74
Telugu	27 / 12	3.26 / 7.33	2.25	2.20	0.53 / 0.88	0.35 / 0.72	0.56 / 0.99
Czech	17 / 11	5.18 / 8.00	1.55	2.20	0.73 / 1.00	0.86 / 0.82	0.95 / 1.00
Russian	17 / 12	5.65 / 8.00	1.42	2.40	0.62 / 0.91	0.69 / 0.81	0.83 / 0.97
Irish	23 / 14	4.20 / 6.89	1.64	2.41	0.61 / 0.89	0.58 / 0.63	0.87 / 0.99
Scottish G.	17 / 13	6.94 / 9.08	1.31	2.95	0.64 / 0.66	0.76 / 0.90	0.61 / 0.81
German	20 / 15	5.20 / 6.93	1.33	2.60	0.81 / 0.81	0.65 / 0.95	0.80 / 0.83
Assamese	41 / 20	1.66 / 3.40	2.05	1.70	0.64 / 1.00	0.55 / 0.79	0.60 / 0.88
Hindi	48 / 17	1.40 / 3.94	2.82	1.68	0.42 / 1.00	0.46 / 0.95	0.43 / 1.00
Sanskrit	29 / 13	3.14 / 7.00	2.23	2.53	0.66 / 0.70	0.55 / 0.53	0.45 / 0.75
French	24 / 19	3.08 / 3.89	1.26	1.85	0.75 / 0.79	0.73 / 0.80	0.67 / 1.00
Italian	27 / 14	3.07 / 5.93	1.93	2.08	0.75 / 0.82	0.67 / 0.77	0.79 / 0.96
Latin	23 / 15	4.00 / 6.13	1.53	2.30	0.73 / 0.82	0.71 / 0.79	0.65 / 0.86
Spanish	25 / 19	3.92 / 5.16	1.32	2.45	0.55 / 0.87	0.82 / 0.89	0.73 / 0.97
Cavineña	13 / 10	10.46 / 13.60	1.30	3.40	0.93 / 1.00	0.46 / 0.57	0.67 / 0.68
H. Quechua	10 / 10	8.80 / 8.80	1.00	2.20	1.00 / 1.00	0.89 / 0.89	1.00 / 1.00
Lamjung Y.	26 / 11	3.58 / 8.45	2.59	2.36	0.69 / 0.89	0.89 / 1.0	0.71 / 0.89
Uipo (M.)	18 / 18	8.50 / 8.50	1.00	3.83	0.40 / 0.40	0.93 / 0.93	0.40 / 0.40
Makyam	27 / 18	4.81 / 7.22	1.50	3.25	0.70 / 0.85	0.70 / 0.71	0.45 / 0.77
Mandarin	10 / 10	8.80 / 8.80	1.00	2.20	1.00 / 1.00	1.00 / 1.00	0.96 / 0.96
Shanghainese	12 / 11	6.50 / 7.09	1.09	1.95	1.00 / 1.00	0.87 / 0.87	1.00 / 1.00
Par. Guaraní	11 / 7	9.55 / 15.00	1.57	2.62	0.91 / 1.00	0.89 / 1.00	0.99 / 1.00

Table 6: Overview of statistics about the different numeral systems for each individual language. Whenever two values are given, the left refers to the surface forms, and the right to the underlying form. Morph. indicates the number of distinct morph(eme)s in the given language. The three rightmost columns indicate the performance of automated morpheme segmentation models in terms of F_1.