Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3922–3931
Marseille, 11–16 May 2020

UniMorph 3.0: Universal Morphology

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskij, Natalya Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and **David Yarowsky**

Johns Hopkins University, Google, Turin (Italy), City University of New York, University of Melbourne, University of Helsinki, Higher School of Economics, Lomonosov Moscow State University, Karelian Research Centre of RAS, Georgia Institute of Technology, University of Wisconsin, ETH Zurich, University of Colorado Boulder

Abstract

The Universal Morphology (UniMorph) project is a collaborative effort providing broad-coverage instantiated normalized morphological paradigms for hundreds of diverse world languages. The project comprises two major thrusts: a language-independent feature schema for rich morphological annotation and a type-level resource of annotated data in diverse languages realizing that schema. We have implemented several improvements to the extraction pipeline which creates most of our data, so that it is both more complete and more correct. We have added 66 new languages, as well as new parts of speech for 12 languages. We have also amended the schema in several ways. Finally, we present three new community tools: two to validate data for resource creators, and one to make morphological data available from the command line. UniMorph is based at the Center for Language and Speech Processing (CLSP) at Johns Hopkins University in Baltimore, Maryland. This paper details advances made to the schema, tooling, and dissemination of project resources since the UniMorph 2.0 release described at LREC 2018.

Keywords: morphology, lexical database, multilinguality

1. Introduction

A notable majority of the world's languages exhibit inflectional morphology. Inflected variants of a word express such distinctions as number, grammatical gender, tense, and aspect. Explicit modeling of morphology has demonstrable benefits for language modeling (Blevins and Zettlemoyer, 2019), speech recognition (Creutz et al., 2007), parsing (Hohensee and Bender, 2012; Seeker and Çetinoğlu, 2015), word embedding (Cotterell et al., 2016b), and keyword search in audio (Narasimhan et al., 2014; Adams et al., 2019). A goal of research, then, is to leverage morphological information to arrive at cross-linguistic or transferable representations for word forms. This has value both to linguists wishing to annotate and understand linguistic phenomena and also to practitioners who seek to use this information for cross-lingual transfer (Yarowsky and Ngai, 2001) or translation (Dyer et al., 2008).

The Universal Morphology (UniMorph) project, centered at the Center for Language and Speech Processing (CLSP) at Johns Hopkins University, is the largest multilingual effort to improve how NLP systems handle complex morphology across the world's languages. The collaborative project releases data sets in dozens of languages, annotated in a common tag set. A lemma (citation form) and bundle of morphological features (represented by the universal tags) are related to a particular inflected word form, as in this Italian example for *inventare*, "to create":

(*inventare*, `V;IND;PRS;3;PL`, *inventano*).

(Note that several sources of ambiguity prevent this relation from being a function.)

UniMorph has been the underlying data resource of four major shared tasks in morphology, beginning with Cotterell et al. (2016a). The 2.0 release was used for the CoNLL–SIGMORPHON 2017 shared task (Cotterell et al., 2017). A pilot version of the 3.0 data was used in the CoNLL–SIGMORPHON 2018 shared task and SIGMORPHON 2019 shared task (Cotterell et al., 2018; McCarthy et al., 2019).

The principal release of UniMorph (Kirov et al., 2016) extracted and normalized inflectional paradigms from Wiktionary, an open online dictionary. UniMorph 2.0 (Kirov et

Lang.	Lemma	Features	Inflection
`deu`	Tisch	N;DAT;MASC;PL	Tischen
`hin`	अग्निपरीक्षा	N;*non*{ACC/NOM};FEM;PL	अग्निपरीक्षाओं
`ita`	intentare	V.PTCP;MASC;PL;PST	intentati
`lat`	mālus	N;GEN;NEUT;PL	mālōrum
`mwf`	nu	V;1+INCL;SG;*non*{FUT}	thunungam
`ron`	măr	N;{DAT/GEN};DEF;FEM;PL	merelor
`tgk`	бахшидан	V;SUBJ;PL;2;PRS	бахшед
`tgl`	gupit	V;IND;BFOC;PFV	iginupit
`zul`	fika	V;BANTU6;PST+RCT	afikē

Table 1: Commonality and universality of UniMorph features. Features exist to represent case, gender, number, clusivity, Austronesian voice, tense, and others. Every triplet above is either new or updated in UniMorph 3.0: newly scraped annotations of inherent gender (`deu`, `hin`, `ron`, `lat`), new data from LINDAT/CLARIN (`ita`), correctly extracted lemmas (`lat`), schema-compliant features (`mwf`, `zul`), previously missing common words (`deu`), and new languages (`tgl`, `tgk`). At present, UniMorph data sets cover 16 language families (Afro-Asiatic, Araucanian, Austronesian, Dené–Yeniseian, Dravidian, Eskimo–Aleut, Indo-European, Kartvelian, Niger–Congo, Northwest Caucasian, Quechua, Sino-Tibetan, Southern Daly, Tungusic, Turkic, Uralic) and 2 isolates.

al., 2018) represented a major restructuring of the extraction process, leading to a higher-quality resource of adjectives, nouns, verbs, or a combination in each of 52 languages.

The work described here, representing the UniMorph 3.0 milestone, makes several contributions to further improve UniMorph's data and tools ecosystem. We've updated the the Wiktionary extraction process, accounting for issues discovered by the community, including those discussed by Gorman et al. (2019) and Kementchedjhieva et al. (2019). Leveraging straightforward heuristics extracts morphological information from HTML tables more accurately and ensures that common word forms are included, improving coverage. Further, we extract non-tabular data. Consequently, we provide annotations of inherent lexical features such as noun gender, which were previously unavailable. We include new word classes for 85 languages, and the total number of languages is now 118, including 14.8 million (lemma, features, inflection) triplets. Additional data comes from automatic conversion of existing resources, redistributed with permission. The annotation schema shared by all UniMorph data has been updated to improve representation of phenomena like syntactic blends and argument politeness. We also present automatic and semi-automatic tools for resource creators and a command-line tool that can provide paradigms and analyses for practitioners.

2. Wiktionary Extraction

The general framework for constructing UniMorph data has been a three-step process:

1. Isolate tables of morphological information from Wiktionary, a free online dictionary covering hundreds of languages.
2. Pair each cell in each table to a bundle of inflectional features.
3. Convert the cell–bundle pairs to UniMorph format.

While Kirov et al. (2016) used statistical and visual heuristics to identify relevant morphological information, Kirov et al. (2018) represented a major restructuring, purposefully extracting tables while relying on shape-based heuristics to map positions to features.[1] To construct UniMorph data for a language, annotators manually mapped table cells to inflectional features in one sample table for each of the small number of table layouts that existed, and the annotations we automatically applied to all lemmas that shared the sample table's layout.

More recently, a critique of morphological inflection generation systems (Gorman et al., 2019) suggested several ways to improve the UniMorph data. It found that 25% of the sample they considered from Cotterell et al. (2017) was either incorrectly scraped or incorrectly recorded in Wiktionary. Several lemmas or word classes were systemically incorrect—Romanian adjectives, Hungarian nouns, Finnish nouns, and Latin lemmas. Gorman et al. (2019) suggested that the inflection generation task was under-specified without inherent features like animacy or noun gender, at least in those languages where such properties determine patterns of inflection—they "rediscovered what linguists have long known: certain allomorphic patterns cannot be predicted from the form of lemmata alone". In a related vein, Kementchedjhieva et al. (2019) note that the limited coverage of UniMorph prevented it from aiding in a bilingual lexicon induction task in Bulgarian.

In this work, we extend Kirov et al. (2018) to improve upon the issues noted above, overhauling the extraction process while continuing to use human annotators. Foremost, we improved the way in which relevant morphological tables are isolated. Additionally, we improved the identification of lemmas, to account for cases where the page title is not the lemma. Finally, we extract morphological information from *non-tabular* regions of the Wiktionary page. As a consequence, (a) the extracted data are cleaner, addressing criticism from Gorman et al. (2019); and (b) the data are more complete, addressing the critique by Kementchedjhieva et al. (2019).

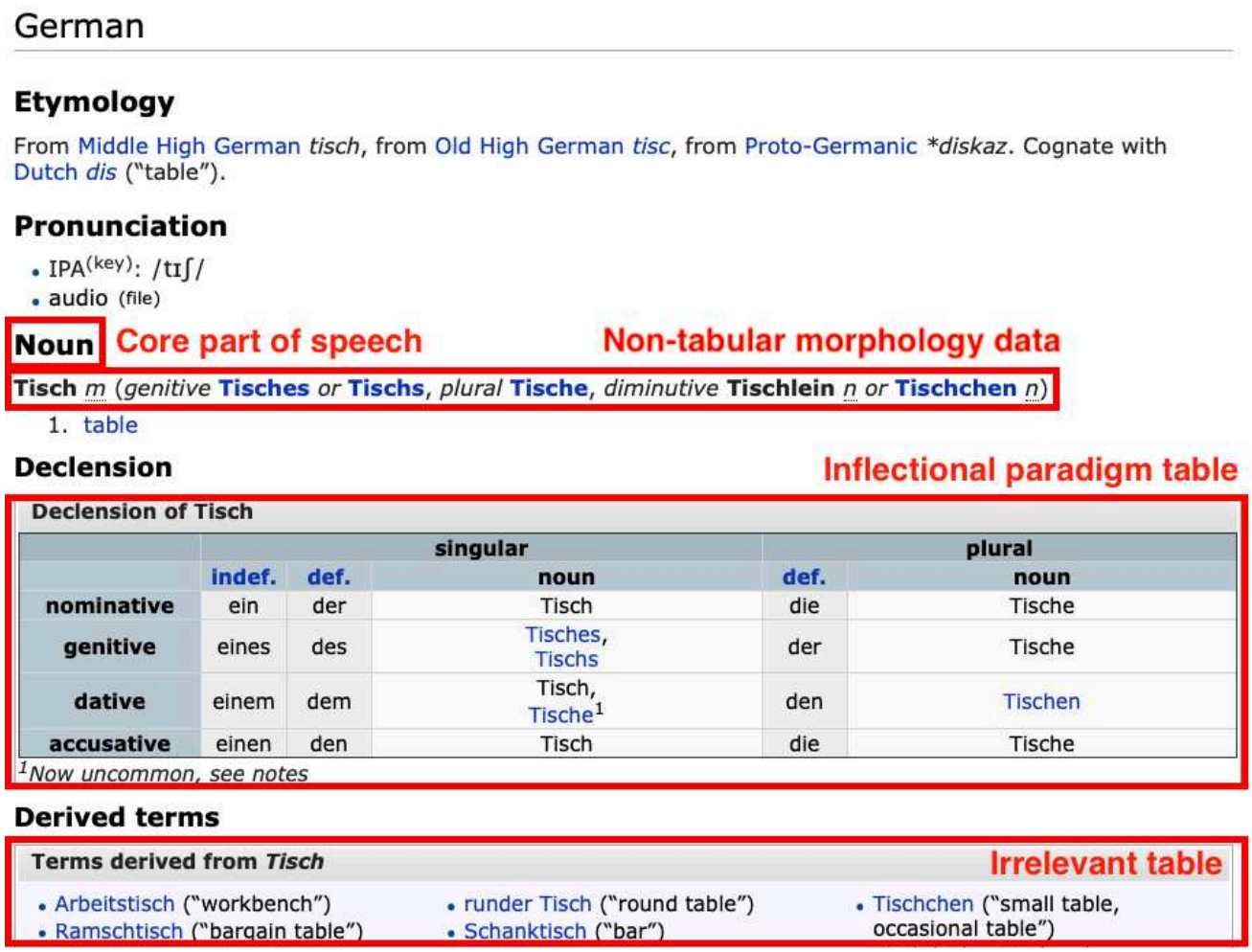
2.1. Upgraded Table Parsing

The process of Kirov et al. (2018) for parsing morphological data *concatenated* all HTML tables within the section of a Wiktionary page corresponding to a particular lemma in a particular language. Figure 1 illustrates a problem with this approach: tables which do not contain inflectional paradigms are also caught in this dragnet. Depending on how many of these additional table types exist, combining them with the desired table can result in more possible overall table dimensions, with fewer lemmas assigned to each. This means annotators have to work harder to annotate more sample tables.

Furthermore, tables for different parts of speech are sometimes concatenated. These larger, concatenated tables appear multiple times to annotators, labeled independently with the POS they should fill in. For example, they might see the same concatenated table in both an N (noun) and V (verb) context. It is entirely up to the annotators to *only* annotate the appropriate cells in the current context, and ignore the rest. If done correctly, only the N tags are annotated and picked up in an N context, and only the V tags are picked up in a V context. However, the N and V tables may sometimes appear in one order in the sample annotators work with, but in an opposite order on some Wiktionary pages. This breaks the positional heuristic and leads to a pairing of the wrong cells with the annotator's inflectional features.

We leverage information already encoded in the webpage to distinguish inflectional from irrelevant tables, as well as to separate tables for different parts of speech. The inflectional tables are automatically generated, and every corresponding HTML `<table>` element carries the attribute `class="inflection-table"`. Thus, for a given language (defined by the section following an `h2` tag), we isolate all inflection tables with the BeautifulSoup Python package. This leads to much more regular patterns of paradigms. For instance, the 2.0 process identified 171 table shapes for German among nouns, verbs, and adjectives. The most common noun shape occurs 9,323 times. The 3.0 process, by contrast, finds only 14 shapes among these parts of speech. The most common noun shape occurs 14,181 times. This

[1] The underlying insight was that only a handful of table shapes are used for each language.



German

Etymology

From Middle High German *tisch*, from Old High German *tisc*, from Proto-Germanic **diskaz*. Cognate with Dutch *dis* ("table").

Pronunciation

- IPA(key): /tɪʃ/
- audio (file)

Noun — Core part of speech — Non-tabular morphology data

Tisch *m* (*genitive* **Tisches** *or* **Tischs**, *plural* **Tische**, *diminutive* **Tischlein** *n* *or* **Tischchen** *n*)

1. table

Declension — Inflectional paradigm table

Declension of Tisch

	singular			plural	
	indef.	def.	noun	def.	noun
nominative	ein	der	Tisch	die	Tische
genitive	eines	des	Tisches, Tischs	der	Tische
dative	einem	dem	Tisch, Tische[1]	den	Tischen
accusative	einen	den	Tisch	die	Tische

[1] *Now uncommon, see notes*

Derived terms

Terms derived from *Tisch* — Irrelevant table

- Arbeitstisch ("workbench")
- Ramschtisch ("bargain table")
- runder Tisch ("round table")
- Schanktisch ("bar")
- Tischchen ("small table, occasional table")

Figure 1: Wiktionary, our primary data source, includes several pieces of data that are relevant to inflectional morphology and agreement, as well as those which are not. The relevant information is distributed across tables, subheadings, and additional content rows. (Above: Subset of the page for *Tisch*, which appears only as a German noun. Character sequences appearing in several languages, as well as those with ambiguous parts of speech, have more complicated pages.)
Previously, UniMorph extraction relied on only the inflectional paradigm table, though additional tables on the page may interfere. Now, UniMorph extraction correctly isolates inflection tables, and we also extracts non-tabular morphology data—both inherent features and also simple paradigms.

dramatically increases the reward per unit of annotation time. Importantly, the remaining shapes are also extremely regular, so annotation errors are less likely.

We also pair these tables with their part of speech and lemma in a different way. We search for the correspondingly annotated preceding tags in the HTML, again looking for particular `class` attributes. Automatically identifying the parts of speech allows us to extract more of these. In addition to adjectives, nouns, and verbs, we were able to isolate determiners, numerals, pronouns, and proper nouns. Automatically extracting the lemma helps to better pair lemmas and forms. For instance, the Latin adjective *malus* 'unpleasant' and noun *mālus* 'apple tree' were both present on the page titled "malus". Separately identifying the lemmas prevents an underspecification problem noted by Gorman et al. (2019): that accents on word-forms cannot be inferred without accents on the lemma.[2]

[2]While classical Latin writers did not make use of macrons, they are used by modern scholars.

Finally, a bug in the previous extraction process caused languages' data to be read into other languages' files whose names are their suffixes. For instance, "Greek" contained data from "Ancient Greek", and "French" contained data from "Middle French". Filtering and rerunning our extraction process eliminated these erroneously grouped paradigms.

2.2. Parsing Non-Tabular Data

While most of the morphological information in Wiktionary can be scraped from HTML tables, Figure 1 shows that inherent features like noun gender or animacy are annotated outside of these tables in structured parenthetical lists. For several languages, some parts of speech are only annotated in non-tabular form. Spanish noun paradigms typically contain only two cells: the singular and plural forms. These are included in parenthetical lists, similar to Figure 1. While

non-tabular, this format *is* structured and thus can be parsed.[3] In the same style as for tabular data, we use a language-specific pattern-matching approach to extract inherent features, as well as other inflected forms of words. In this release, we present nouns and adjectives extracted in this style for 19 languages from the Indo-European family, which were not available through Wiktionary's tables. Seven of these augment our new languages; the remainder add new parts of speech to existing languages. This fills a major outstanding gap in UniMorph's coverage.

The additional inherent features are included as separate files for each language. This helps distinguish between the two classes of features, as inherent features may not be pertinent for all tasks. One can combine the inherent with the inflectional features via a standard table join on the lemma and part of speech.

3. Schema Updates

A key contribution of the UniMorph project is a universal feature schema for morphological annotation, amenable to cross-linguistic comparison and projection. In the 3.0 release of UniMorph, we introduce several (backwards-incompatible) changes to expand the schema.

The UniMorph schema was designed by a broad typological survey in a top-down fashion (Sylak-Glassman et al., 2015). It breaks morphosyntax into over 200 features distributed across 23 dimensions of meaning. The tags are based on the Leipzig glossing standard, and they retain high compatibility with the Universal Dependencies morphological annotations. (As constructive evidence, McCarthy et al. (2018) present a deterministic conversion of Universal Dependencies morphological annotations into the UniMorph schema, with high token-averaged recall.) The features are globally unique, so that a given feature pertains to exactly one dimension of meaning.

3.1. Deficiencies and absences

In annotating the additional languages and parts of speech available in the UniMorph 3.0 release, we identified several necessary modifications to the schema. Foremost, we need to handle syntactic blends—situations when the notion of a word breaks down. Beyond this, it's common for users to treat a morphological feature bundle as a single string instead of a set, so we standardized feature order within string-format bundles to preserve their uniqueness.

3.2. Syntactic blends

Previous versions of UniMorph were limited to nouns, verbs, adjectives, and adverbs. In UniMorph 3.0 we introduce annotated forms of function words—determiners, adpositions, and conjunctions. With this expansion, a discussion of what consists a word (the basic unit of annotation) becomes necessary. In written-English language terminology, a word typically refers to a string of characters separated by whitespace. This usually corresponds to a spoken notion of word understood as the fundamental level of syntax. However,

[3]Some of the data present in UniMorph 2.0, namely English Verbs, needed to be extracted in this way. For UniMorph 3.0, we have applied the process en masse.

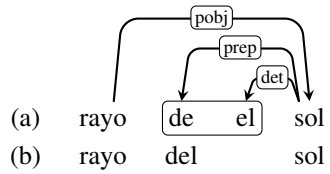


Figure 2: A Spanish word-form token (b) corresponding to two words in a non-phrasal correspondence (a).

words may diverge from the whitespace-delimited heuristic via compounding, where compositional phrases connect at the surface level, such as "whitespace-delimited". At this point our unit of reference becomes a token, which is still subject to inflection, and compounds retain their syntactic atomicity (in this case, an adjectival phrase). A second phenomenon, not present in English and not common in open-class lexical categories, is **syntactic fusion**. This is where a single token performs the task of two or more syntactic words which do not form a constituent together, such as Spanish *del*, which adds together the adposition *de* with the determiner *el*. From the perspective of a word-level resource like UniMorph, this creates a problem in that neither of the blended words can be considered a head of the token, necessary for providing the main lemma and part-of-speech for a compound.

One available solution is to follow conventions from syntactic annotations at the sentence level. In Universal Dependencies (UD) (Nivre et al., 2016, and later versions), syntactic fusion is disassembled: the token is annotated as a range of syntactic words which are then connected via dependencies to the rest of the sentence. Crucially, the words comprising a token might not form a subtree in the dependency structure (see Figure 2). The French Treebank (FTB) (Abeillé et al., 2003) separates the annotations into one part for each syntactic unit, joined by a plus sign. Another possibility is to annotate syntactically-fused word types as only the part-of-speech amenable to inflection. In the Spanish example *del*, this would mean the type is dominated by the determiner, and annotated as a `DET`. In a reinflection scenario, this will be translated into requiring *del* $-$ `MASC` $+$ `FEM` $=$ *de la*, being conscious of the fact that this item does not exhibit a full single-token inflectional paradigm. In this sense, it would be similar to English *walked* $-$ `PST` $+$ `FUT` $=$ *will walk*.

We opt for an approach based on the FTB syntax: we refresh the lemma tagging schema to allow for multi-POS items. Since there is no structural preference towards one of the fused forms in the syntactically-fused token, we will in this case default to displaying them at the order in which the respective forms manifest themselves in the fused form. Introducing a new pipe separator "`|`", the lemma for *del* will be annotated as `ADP|DET;MASC;SG`. A reinflection of this lemma from `MASC` to `FEM` would remain "de la".

That UniMorph now recognizes that lemmas and word forms can be segmented opens up possibilities for our morphological dictionaries. In the case of clitics or agglutinative formations, it is both reasonable and natural to map morphological features onto segments of a word form. Parallel structure in the segmentation of the word form and the feature bundle

Lemma	Features	Inflection	Segmentation
abrir	V|PTCP;PRS|PRO;ACC;MASC;3;SG	abr | iéndolo	abr | iéndo | lo
abrir	V|PTCP;PRS|PRO;3;{SG/PL};REFL	abriéndose	abr | iéndo | se
abrir	V|IMP;1;PL;POS|PRO;1;PL;REFL	abramonos	abr | amo | nos

Table 2: Including syntax into the UniMorph schema that segments features opens the possibility of segmented and aligned featural representation, as with these several forms of the Spanish verb *abrir* "to open".

(perhaps both demarcated by the pipe character "|") can clarify the morphological role of each affix. For instance, in the Spanish example *abriéndolo* "opening it", it may be natural to separate the clitic and align features accordingly, and even to separate the stem from the ending, as in Table 2. In other cases of fusion, affix boundaries may not be as clear; selecting a segmentation requires subscribing to a particular theory. We leave the future role of segmentation in UniMorph open to discussion by the community.

3.3. Canonical Ordering of Dimensions

While the UniMorph feature bundle represents a set of features, it is common for practitioners to treat the bundle as a single, atomic string. Similarly, for disjunction of features, the order carries no meaning: {GEN/DAT} and {DAT/GEN} carry the same semantics—but practitioners do not parse these. They treat feature bundles as raw strings. In order to improve matching for these users, we now impose a canonical ordering of the 23 UniMorph dimensions of meaning. While several possible orderings exist, with varying amounts of linguistic motivation, we opt for a simple one. The part of speech is always the first feature in the bundle.[4] Beyond this, features are ordered alphabetically according to their dimension's name. Number always precedes person, and mood always precedes tense. Language-specific tags (LGSPEC01, LGSPEC02, ...) follow all other dimensions, and they are lexicographically ordered.

4. New Languages and Data

For the UniMorph 3.0 milestone, we have added new languages scraped from Wiktionary, as well as language data gathered from additional external sources or synthesized automatically. Ultimately, we've more than doubled the amount of languages for which we have data. The current total is 118 languages from 16 major families, as well as two isolates.

4.1. New Wiktionary Languages

We extend UniMorph by covering more languages already present in Wiktionary. Nine annotators worked to annotate morphological paradigms extracted from 50 new languages. As with the other languages, we used the table-based extraction strategy to map HTML table cells to slots in morphological paradigms. Data for most of these languages was used in the CoNLL–SIGMORPHON 2018 shared task on morphological inflection (Cotterell et al., 2018) and the SIGMORPHON 2019 shared task on cross-lingual transfer of inflection (McCarthy et al., 2019). In aggregate, the new languages extracted from Wiktionary span 13 language families. We present the distribution of these languages, along with the number of paradigms and word forms, in Table 3.

4.2. The LINDAT/CLARIN Morphological Dictionaries

Morphological dictionaries exist for several European languages and are publicly hosted on the LINDAT/CLARIN repository of Charles University under a CC BY-NC-SA 4.0 license. Particularly, these exist for English (Grella, 2014), French (Grella, 2019a), German (Grella, 2019b), Italian (Grella, 2018a; Grella, 2018b), and Spanish (Grella, 2019c). These resources were constructed by hand or by algorithmic generation using rules of a language's morphology, e.g. (Lepschy and Lepschy, 1981). We have automatically converted these dictionaries to our morphological schema, and we distribute the UniMorph-annotated dictionaries (with permission) under the same license. The sizes of these resources are given in Table 4. At the extreme, the LINDAT/CLARIN Italian data set adds 378,882 new lemmas (for a total of 384,848) and 2,112,952 new word forms (for a total of 2,622,526) to UniMorph's Italian data set.

A common grammatical dichotomy for word types is into the categories of content words (which carry necessary information and meaning) versus function words (which are a grammatical necessity). In addition to including a large number of content words in these high-resource languages which were not available through Wiktionary, the data sets from LINDAT/CLARIN have extensive annotations for function words in the languages. For instance, the Italian data contains 999 single-word forms and 2,501 multi-words forms which are function words. Function words remain an area of active exploration (Kim et al., 2019). Their inclusion builds toward the possibility of *complete* morphological representations of text (as opposed to only content words), based on morphological analysis. Previous iterations of UniMorph data did not include function words. Together, these benefits reduce the frequency of out-of-vocabulary words when UniMorph data is used in downstream tasks.

4.3. Low-Resource Languages of Russia, Latvia, and Australia

Some languages and dialects are extremely low-resource, and the data for such languages mainly comes from linguists who study them. For instance, VepKar (Zaytseva et al., 2017) shared their corpora for languages and dialects commonly spoken in the Republic of Karelia, Russia and Finland (Karelian, Livvi, Veps, Ludian). Many other minority languages of Russia such as Erzya, Komi, Moksha, Meadow Mari,

[4] This convention has already been obeyed historically in the UniMorph data; we now formalize it.

Language	Family	Lemmas	Forms
Adyghe$^{\mathcal{W}}$	Caucasian	1,666	20,475
Ancient Greek$^{\mathcal{W}}$	Hellenic	2,409	41,593
Asturian$^{\mathcal{W}}$	Romance	436	29,797
Azeri$^{\mathcal{W}}$	Iranian	340	8,004
Bashkir$^{\mathcal{W}}$	Turkic	1,084	12,168
Belarusian$^{\mathcal{W}}$	Slavic	1027	16,113
Breton$^{\mathcal{W}}$	Celtic	44	2,294
Classical Syriac$^{\mathcal{W}}$	Semitic	160	3,652
Cornish$^{\mathcal{W}}$	Celtic	9	469
Crimean Tatar$^{\mathcal{W}}$	Turkic	1,230	7,514
Erzya$^{\mathcal{L}}$	Uralic	13,035	107,164
Evenki$^{\mathcal{L}}$	Tungusic	1,264	29,745
Friulian$^{\mathcal{W}}$	Romance	168	8,071
Galician$^{\mathcal{W}}$	Romance	486	36,801
Greenlandic$^{\mathcal{W}}$	Inuit	23	368
Ingrian$^{\mathcal{W}}$	Uralic	50	1,099
Kabardian$^{\mathcal{W}}$	Caucasian	250	3,092
Kannada$^{\mathcal{W}}$	Dravidian	159	6,402
Karelian$^{\mathcal{W}+\mathcal{L}}$	Uralic	2,303	113,731
Kashubian$^{\mathcal{W}}$	Slavic	37	509
Kazakh$^{\mathcal{W}}$	Turkic	26	357
Khakas$^{\mathcal{W}}$	Turkic	75	1,200
Komi$^{\mathcal{L}}$	Uralic	18,541	90,940
Ladin$^{\mathcal{W}}$	Romance	180	7,656
Livonian$^{\mathcal{L}}$	Uralic	9,783	221,652
Livvi$^{\mathcal{L}}$	Uralic	15,297	63,815
Ludian$^{\mathcal{L}}$	Uralic	119	401
Maltese$^{\mathcal{W}}$	Semitic	112	3,584
Manx$^{\mathcal{W}}$	Celtic	1	14
Mapudungun$^{\mathcal{W}}$	Araucanian	26	783
Meadow Mari$^{\mathcal{L}}$	Uralic	16,196	101,457
Middle French$^{\mathcal{W}}$	Romance	603	36,970
Middle High German$^{\mathcal{W}}$	Germanic	29	708
Middle Low German$^{\mathcal{W}}$	Germanic	52	1,513
Moksha$^{\mathcal{L}}$	Uralic	10,152	66,250
Murrinhpatha$^{\mathcal{L}}$	Australian	29	1,110
Neapolitan$^{\mathcal{W}}$	Romance	40	1,808
Norman$^{\mathcal{W}}$	Romance	5	280
North Frisian$^{\mathcal{W}}$	Germanic	51	3,204
Occitan$^{\mathcal{W}}$	Romance	174	8,316
Old Armenian$^{\mathcal{W}}$	Indo-European	4,300	97,181
Old Church Slavonic$^{\mathcal{W}}$	Slavic	152	4,148
Old English$^{\mathcal{W}}$	Germanic	1867	41,589
Old French$^{\mathcal{W}}$	Romance	1,700	12,3374
Old Irish$^{\mathcal{W}}$	Celtic	49	1,089
Old Saxon$^{\mathcal{W}}$	Germanic	863	22,287
Pashto$^{\mathcal{W}}$	Iranian	395	6,945
Sanskrit$^{\mathcal{W}}$	Indo-Aryan	917	33,847
Selkup$^{\mathcal{L}}$	Uralic	4,779	15,862
Swahili$^{\mathcal{W}}$	Bantu	100	10,092
Tajik$^{\mathcal{W}}$	Iranian	75	77
Tatar$^{\mathcal{W}}$	Turkic	1,283	7,832
Telugu$^{\mathcal{W}}$	Dravidian	127	1,548
Tibetan$^{\mathcal{L}}$	Sino-Tibetan	1,410	5,696
Turkmen$^{\mathcal{W}}$	Turkic	68	810
Udmurt$^{\mathcal{L}}$	Uralic	25,042	201,708
Uzbek$^{\mathcal{W}}$	Turkic	15	1,260
Venetian$^{\mathcal{W}}$	Romance	368	18,227
Veps$^{\mathcal{L}}$	Uralic	3,455	134,137
Votic$^{\mathcal{W}}$	Uralic	55	1,430
West Frisian$^{\mathcal{W}}$	Germanic	85	1,429
Yiddish$^{\mathcal{W}}$	Germanic	803	7,986
Zulu$^{\mathcal{W}}$	Bantu	566	49,119

Table 3: Total number of lemmas and forms available for each new language. $\mathcal{W}$: extracted from Wiktionary; $\mathcal{L}$: data provided by linguists.

Language	Size
English	123,755
French	453,652
German	1,157,803
Italian	2,345,630
Spanish	777,438

Table 4: Number of forms in the LINDAT/CLARIN dictionaries.

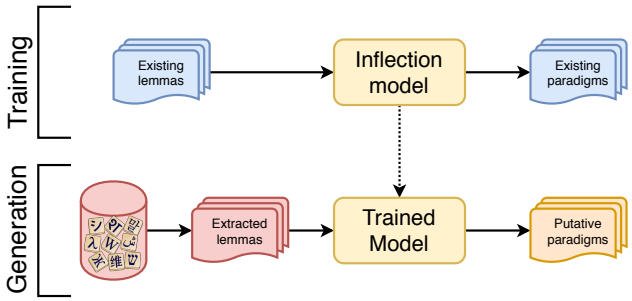


Figure 3: Generating putative UniMorph morphological paradigms for Tagalog

Udmurt as well as extra data for Eastern Armenian were derived from the UniParser project (Arkhangelskiy et al., 2012; Arkhangelskiy and Medvedeva, 2016). The data for the Selkup language, spoken by approximately 1,000 people in Yamalo-Nenets Autonomous Okrug in Russia originates from Brykina et al. (2018). Importantly, in this iteration we also enrich UniMorph with endangered members of new languages families: Tungusic and Australian. The Tungusic family is represented by Evenki derived from Klyachko et al. (2019). Another endangered language, Murrinhpatha, an Australian Aboriginal language, was extensively studied and documented in Mansfield (2019).

4.4. Augmented Data: Tagalog

We also experiment with the development of new data for morphological inflection, acquired by semi-supervision. We train a state-of-the-art morphological inflection system (Makarov and Clematide, 2018) on the Tagalog morphological data of Wicentowski (2002)—taking 189 lemmas (8,957 inflected forms) as training data. We then apply this system to the verbal lemmas in Tagalog extracted by Kirov et al. (2018). This generates a host of new putative word-forms. The process is shown in Figure 3. While this data set is machine-generated, rather than human-vetted, it may still be useful as a dictionary, or as development data for improved morphological analyzers or inflection generation systems. We chose Tagalog, an Austronesian language, for this experiment because it makes extensive use of reduplication in its verbal paradigms (Adelaar and Himmelmann, 2005). This feature is from the UniMorph languages used in the SIGMORPHON shared tasks. When tested on unseen lemmas, the exact-match accuracy of the Makarov and Clematide (2018) system is 51.48% (418 of 812) and mean reciprocal rank of correct predictions is 0.6345. However, the mean normalized edit distance between gold and predicted word-forms is often extremely low (0.136), with many of the errors

involving failure to produce unexpected stem changes. This raises the question of how to properly handle reduplicating languages in systems for morphological inflection in future years.

5. Validation Tools

UniMorph has become a large community with diverse contributors. As is natural for such large annotation endeavors, it's natural that exact compliance to the schema is a challenge, and it's easy to introduce unintentional errors in the data set construction process. As we note above, Gorman et al. (2019) found this to be true for several languages: nearly 25% of their sample was systemically incorrect, and this led a top-performing morphological inflection system to "correctly" produce the text "definite articulation" as an inflected form of a Romanian adjective. Further, several possible scraping errors have been reported by the community at large, which this data release has largely corrected.

To ameliorate this problem for future data set releases, we provide two extensible tools for validation of UniMorph morphological databases. The first is a syntactic validator and normalizer for UniMorph feature bundles. The second is a semi-automatic vetting script that flags suspicious and possibly erroneous data for manual inspection and correction. The authors used this second tool in the process of refining our new scraping process (section 2.). The tools are beneficial to the community in both creating and using UniMorph data.

5.1. A grammar of morphological tags

Our revised morphological annotation schema defines a finite grammar of tags. This allows easy validation of UniMorph feature bundles: apart from their linguistic correctness in pairing a lemma and word-form, is the tag grammatical? As one example, for the CoNLL–SIGMORPHON 2018 Shared Task (Cotterell et al., 2018), feature bundles for five languages contained a typo that led to incorrect parsing: a colon appeared instead of a semicolon for some feature bundles. While this was fixed for the SIGMORPHON 2019 shared task (McCarthy et al., 2019), we seek to prevent similar errors in the future.

Consequently, we release a tool for both validating and normalizing these tags. This fixes incongruities in tag order (as the schema now specifies an order, making exact string-matching possible) within and across languages, imagined but invalid tags, and typos introduced in the annotation process. It uses a YAML encoding of the UniMorph features and dimensions of meaning to rearrange features into a canonical order. The canonicalization function is idempotent: if it is applied twice to the same file, the second application will not alter the file. This helps to serve as a tool for validation in addition to canonicalization.

5.2. Automated Error Flagging

We create and make public an automated Python 3 script, `autovet.py` that can be applied to any UniMorph data file. It automatically flags several types of potential errors with varying levels of logging:

1. Non-letter characters. While spaces, apostrophes, or hyphens may be common in Wiktionary data, numerals are typically indicative of scraping errors. It's often the case that a table will include a footnote; the footnote mark often needs to be deleted.

2. Rare characters. Occasionally, Wiktionary incorrectly labels the language of a page, so that Russian is marked as Romanian or Ancient Greek is marked as Latin. When these pages are scraped, they pollute the clean data of the label's language. Flagging rare characters shows where these errors may have occurred.

3. Fields with no letters. Inflectional gaps (Halle, 1973; Gorman and Yang, 2019) are often marked in Wiktionary with an em-dash and no other characters in the HTML table cell. These rows should be omitted from UniMorph.

4. Multi-word expressions. These are common in data, e.g., the light verb construction in Persian. But often these can indicate a scraping problem, where a table header or annotation is misinterpreted as an inflected form.

5. Empty fields. This error type is straightforward.

6. Ending with spaces. This error type is difficult to visually inspect but causes problems when performing exact-match string lookup.

7. Frequent forms. If a word-form appears in the paradigms of several words, it is likely a table header, not a true word-form.

8. Large edit distance. As a final heuristic, if there is a large edit distance between the lemma and its inflected form, this may suggest a scraping error.

Used in a human-in-the-loop fashion, the tool can help to refine the pipeline for creating morphological dictionaries like UniMorph. It can also be used to identify point fixes that are needed, particularly when these are idiopathic or particular to one extracted paradigm. Indeed, the authors used it in developing UniMorph 3.0.

6. UniMorph via `pip`

As a final contribution, we present the construction of a command-line interface to the UniMorph morphological databases. Feedback on the UniMorph project has suggested that even with a regular, three-column format, it may be unapproachable or confusing. This command-line tool aims to remedy these concerns by providing several standard functions—downloading data sets, searching for morphological paradigms, and performing morphological analysis of a word-form into its lemma and morphological features. With over 100 languages of UniMorph morphological dictionaries currently available, it can become tedious to download several data sets for multilingual research. Downloading via the GitHub website interface requires visiting each intended web page. Instead, the UniMorph command-line tool provides a central means to download all languages with a single `bash` command. Because each language's `git` repository is downloaded, it is straightforward to update the data as extensions or cleaning fixes become available.

Further, the two core morphological operations of inflection and analysis are facilitated by the morphological tool. When provided with a lemma, the UniMorph tool will give complete morphological paradigms from the UniMorph data. Optionally, a morphological feature bundle can be provided, which restricts the output the corresponding inflected form of the given lemma. As an inverse to this operation, a word-form provided to the tool is analyzed into candidate feature–lemma pairs.

Command-line morphological tools for out-of-lexicon words, trained on the Bible, are also provided for over 1,000 languages by (Nicolai et al., under review). These can optionally leverage the UniMorph data to "Don't guess if you know" (Tapanainen and Voutilainen, 1994). The tool is distributed through the Python Package Index and thus installable via `pip`. It can be installed with the command

```
pip install unimorph
```

on any machine with Python 3. Documentation is easily accessible in the standard way for command-line tools (i.e., the `-h` or `-help` flags).

7. Conclusion

The UniMorph project represents a massively multilingual effort at cataloguing the world's inflectional morphology. It has evolved into the largest morphological dictionary in terms of both scope and number of languages.

In the past two years, we have implemented several improvements to the extraction pipeline which creates most of our data, so that it is both more complete and more correct. We have added 66 new languages, as well as new parts of speech for 19 languages. We have amended the schema. Finally, we present three new community tools: two that vet and validate data for resource creators, and one to make our morphological data available from the command line.

UniMorph data has been the basis of several successful shared tasks on learning inflectional morphology by SIGMORPHON and CoNLL (Cotterell et al., 2016a; Cotterell et al., 2017; Cotterell et al., 2018; McCarthy et al., 2019). Further, it has demonstrable value in low-resource tagging and annotation projection (Plank et al., 2018; Shearing et al., 2018; Agić and Vulić, 2019).

The UniMorph 3.0 schema alteration and data release promise cleaner data in twice as many languages, as well as a more truly universal coverage of the world's inflectional features. The project welcomes continued contributions from the community. All data and tools are released under a permissive open-source license at `unimorph.github.io`.

Acknowledgments

We thank Djamé Seddah for important references on syntactic blend annotation. We thank Elizabeth Salesky for help with converting data. We thank Marcell Bollmann, Yova Kementchedjhieva, Matt Post, Richard Sproat, Matthew Wiesner, and Winston Wu for discussions that shaped the direction of the work. Y.P. is a Bloomberg Data Science PhD Fellow.

8. Bibliographical References

Abeillé, A., Clément, L., and Toussenel, F., (2003). *Building a Treebank for French*, pages 165–187. Springer Netherlands, Dordrecht.

Adams, O., Wiesner, M., Trmal, J., Nicolai, G., and Yarowsky, D. (2019). Induced Inflection-Set Keyword Search in Speech. *arXiv e-prints*, page arXiv:1910.12299, Oct.

Adelaar, K. A. and Himmelmann, N. (2005). *The Austronesian languages of Asia and Madagascar*, volume 7. Psychology Press.

Agić, Ž. and Vulić, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July. Association for Computational Linguistics.

Arkhangelskiy, T. and Medvedeva, M. (2016). Developing morphologically annotated corpora for minority languages of russia. In *CLiF*.

Arkhangelskiy, T., Belyaev, O., and Vydrin, A. (2012). The creation of large-scale annotated corpora of minority languages using UniParser and the EANC platform. In *Proceedings of COLING 2012: Posters*, pages 83–92, Mumbai, India, December. The COLING 2012 Organizing Committee.

Blevins, T. and Zettlemoyer, L. (2019). Better character language modeling through morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1606–1613, Florence, Italy, July. Association for Computational Linguistics.

Brykina, M., Orlova, S., and Wagner-Nagy, B. (2018). INEL Selkup Corpus. Version 0.1. In *The INEL corpora of indigenous Northern Eurasian languages.*, Hamburg, December. Hamburger Zentrum für Sprachkorpora.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016a). The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany, August. Association for Computational Linguistics.

Cotterell, R., Schütze, H., and Eisner, J. (2016b). Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1651–1660, Berlin, Germany, August. Association for Computational Linguistics.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2017). CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver, August. Association for Computational Linguistics.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., McCarthy, A. D., Kann, K., Mielke, S., Nicolai, G., Silfverberg, M., Yarowsky, D., Eisner, J., and Hulden, M. (2018). The CoNLL–SIGMORPHON 2018

shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels, October. Association for Computational Linguistics.

Creutz, M., Hirsimäki, T., Kurimo, M., Puurula, A., Pylkkönen, J., Siivola, V., Varjokallio, M., Arisoy, E., Saraçlar, M., and Stolcke, A. (2007). Analysis of morph-based speech recognition and the modeling of out-of-vocabulary words across languages. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 380–387, Rochester, New York, April. Association for Computational Linguistics.

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing word lattice translation. In *Proceedings of ACL-08: HLT*, pages 1012–1020, Columbus, Ohio, June. Association for Computational Linguistics.

Gorman, K. and Yang, C. (2019). When nobody wins. In Franz Rainer, et al., editors, *Competition in Inflection and Word-Formation*, pages 169–193. Springer, Dordrecht.

Gorman, K., McCarthy, A. D., Cotterell, R., Vylomova, E., Silfverberg, M., and Markowska, M. (2019). Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 140–151, Hong Kong, China, November. Association for Computational Linguistics.

Grella, M. (2014). English dictionary. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Grella, M. (2018a). Italian content words v3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Grella, M. (2018b). Italian function words v3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Grella, M. (2019a). French morphological dictionary 0.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Grella, M. (2019b). German morphological dictionary 0.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Grella, M. (2019c). Spanish morphological dictionary 0.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4(1):3–16.

Hohensee, M. and Bender, E. M. (2012). Getting more from morphology in multilingual dependency parsing. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 315–326, Montréal, Canada, June. Association for Computational Linguistics.

Kementchedjhieva, Y., Hartmann, M., and Søgaard, A. (2019). Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3327–3332, Hong Kong, China, November. Association for Computational Linguistics.

Kim, N., Patel, R., Poliak, A., Xia, P., Wang, A., McCoy, T., Tenney, I., Ross, A., Linzen, T., Van Durme, B., Bowman, S. R., and Pavlick, E. (2019). Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kirov, C., Sylak-Glassman, J., Que, R., and Yarowsky, D. (2016). Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S., McCarthy, A., Kübler, S., Yarowsky, D., Eisner, J., and Hulden, M. (2018). UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Klyachko, E., Sorokin, A., Krizhanovskaya, N., Krizhanovsky, A., and Ryazanskaya, G. (2019). LowResourceEval-2019: A shared task on morphological analysis for low-resource languages. In *Papers from the Annual International Conference "Dialogue" (2019), Issue 18*, pages 45–62, Moscow. Computational Linguistics and Intellectual Technologies.

Lepschy, A. L. and Lepschy, G. C. (1981). *La lingua italiana*. Bompiani Milan:.

Makarov, P. and Clematide, S. (2018). UZH at CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 69–75, Brussels, October. Association for Computational Linguistics.

Mansfield, J. (2019). *Murrinhpatha morphology and phonology*, volume 653. Walter de Gruyter GmbH & Co KG.

McCarthy, A. D., Silfverberg, M., Cotterell, R., Hulden, M., and Yarowsky, D. (2018). Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium, November. Association for Computational Linguistics.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-

Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R., and Hulden, M. (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy, August. Association for Computational Linguistics.

Narasimhan, K., Karakos, D., Schwartz, R., Tsakalidis, S., and Barzilay, R. (2014). Morphological segmentation for keyword spotting. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 880–885, Doha, Qatar, October. Association for Computational Linguistics.

Nicolai, G., Lewis, D., McCarthy, A. D., Mueller, A., Wu, W., and Yarowsky, D. (under review). Fine-grained morphosyntactic analysis and generation tools for more than one thousand languages. In *Submitted to LREC 2020*.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Plank, B., Klerke, S., and Agic, Z. (2018). The best of both worlds: Lexical resources to improve low-resource part-of-speech tagging. *CoRR*, abs/1811.08757.

Seeker, W. and Çetinoğlu, Ö. (2015). A graph-based lattice dependency parser for joint morphological segmentation and syntactic analysis. *Transactions of the Association for Computational Linguistics*, 3:359–373.

Shearing, S., Kirov, C., Khayrallah, H., and Yarowsky, D. (2018). Improving low resource machine translation using morphological glosses (non-archival extended abstract). In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 132–139, Boston, MA, March. Association for Machine Translation in the Americas.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015). A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China, July. Association for Computational Linguistics.

Tapanainen, P. and Voutilainen, A. (1994). Tagging accurately - don't guess if you know. In *Fourth Conference on Applied Natural Language Processing*, pages 47–52, Stuttgart, Germany, October. Association for Computational Linguistics.

Wicentowski, R. (2002). *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*. Ph.D. thesis, Johns Hopkins University, Baltimore, Maryland.

Yarowsky, D. and Ngai, G. (2001). Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Zaytseva, N., Krizhanovsky, A., Krizhanovsky, N., Pellinen, N., and Rodionova, A. (2017). Open corpus of Veps and Karelian languages (VepKar): preliminary data collection and dictionaries. In *Corpus Linguistics-2017*, pages 172–177.