University of Groningen

UniMorph 4.0

Batsuren, Khuyagbaatar; Goldman, Omer; Khalifa, Salam; Habash, Nizar; Kieraś, Witold; Bella, Gábor; Leonard, Brian; Nicolai, Garrett; Gorman, Kyle; Ate, Yustinus Ghanggo

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Early version, also known as pre-print

Publication date:
2022

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Batsuren, K., Goldman, O., Khalifa, S., Habash, N., Kieraś, W., Bella, G., Leonard, B., Nicolai, G., Gorman, K., Ate, Y. G., Ryskina, M., Mielke, S. J., Budianskaya, E., El-Khaissi, C., Pimentel, T., Gasser, M., Lane, W., Raj, M., Coler, M., ... Vylomova, E. (2022). *UniMorph 4.0: Universal Morphology*. arXiv.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 20-11-2022

arXiv:2205.03608v2 [cs.CL] 10 May 2022

UniMorph 4.0: Universal Morphology

Khuyagbaatar Batsuren[ɯ*] **Omer Goldman**[λ*] **Salam Khalifa**[ʏ] **Nizar Habash**[ħ]
Witold Kieraś[θ] **Gábor Bella**[ɣ] **Brian Leonard**[β] **Garrett Nicolai**[ɓ] **Kyle Gorman**[ŋ]
Yustinus Ghanggo Ate[ɜ] **Maria Ryskina**[ɬ] **Sabrina Mielke**[ʒ] **Elena Budianskaya**[ь]
Charbel El-Khaissi[ʃ] **Tiago Pimentel**[ʕ] **Michael Gasser**[ʔ] **William Lane**[ɾ] **Mohit Raj**[ʊ]
Matt Coler[ɢ] **Jaime Rafael Montoya Samame**[ﬁ] **Delio Siticonatzi Camaiteri**[ɟ]
Esaú Zumaeta Rojas[ɟ] **Didier López Francis**[ɟ] **Arturo Oncevay**[ᴇ] **Juan López Bautista**[ɟ]
Gema Celeste Silva Villegas[ﬁ] **Lucas Torroba Hennigen**[ʕ] **Adam Ek**[ɠ] **David Guriel**[λ]
Peter Dirix[ʋ] **Jean-Philippe Bernardy**[ɠ] **Andrey Scherbakov**[ɘ] **Aziyana Bayyr-ool**[ʑ]
Antonios Anastasopoulos[ʄ] **Roberto Zariquiey**[ﬁ] **Karina Sheifer**[ɛ,ь,ɶ] **Sofya Ganieva**[ɱ,ь]
Hilaria Cruz[з] **Ritván Karahóǧa**[ʛ] **Stella Markantonatou**[ʛ] **George Pavlidis**[ʛ]
Matvey Plugaryov[ɱ,ь] **Elena Klyachko**[ɛ,ь] **Ali Salehi**[ɷ] **Candy Angulo**[ﬁ] **Jatayu Baxi**[ʌ]
Andrew Krizhanovsky[ʁ] **Natalia Krizhanovsky**[ʁ] **Elizabeth Salesky**[ʒ] **Clara Vania**[ɐ]
Sardana Ivanova[ɨ] **Jennifer White**[ʕ] **Rowan Hall Maudslay**[ʕ] **Josef Valvoda**[ʕ]
Ran Zmigrod[ʕ] **Paula Czarnowska**[ʕ] **Irene Nikkarinen**[ʕ] **Aelita Salchak**[ʂ] **Brijesh Bhatt**[ʌ]
Christopher Straughn[ɲ] **Zoey Liu**[ŧ] **Jonathan North Washington**[ɸ] **Yuval Pinter**[ᵞ]
Duygu Ataman[ɐ] **Marcin Woliński**[θ] **Totok Suhardijanto**[þ] **Anna Yablonskaya**[ɛ]
Niklas Stoehr[ð] **Hossep Dolatian**[ʏ] **Zahroh Nuriah**[þ] **Shyam Ratan**[ʊ] **Francis M. Tyers**[ʔ,ɛ]
Edoardo M. Ponti[ø] **Grant Aiton**[ʃ] **Aryaman Arora**[ɖ] **Richard J. Hatcher**[ɷ]
Ritesh Kumar[ʊ] **Jeremiah Young**[ɔ] **Daria Rodionova**[ɛ] **Anastasia Yemelina**[ɛ]
Taras Andrushko[ɛ] **Igor Marchenko**[ɛ] **Polina Mashkovtseva**[ɛ] **Alexandra Serova**[ɛ]
Emily Prud'hommeaux[ŧ] **Maria Nepomniashchaya**[ɛ] **Fausto Giunchiglia**[ɣ]
Eleanor Chodroff[ɤ] **Mans Hulden**[χ] **Miikka Silfverberg**[ɓ] **Arya D. McCarthy**[ʒ]
David Yarowsky[ʒ] **Ryan Cotterell**[ð] **Reut Tsarfaty**[λ] **Ekaterina Vylomova**[ɘ]

[ɯ]National University of Mongolia [λ]Bar-Ilan University [ʒ]Johns Hopkins University [ɣ]University of Trento
[ɤ]University of York [ɬ]Carnegie Mellon University [β]Brian Leonard Consulting [ʔ]Indiana University
[ɓ]University of British Columbia [ʌ]Dharmsinh Desai University [ħ]New York University Abu Dhabi
[ʕ]University of Cambridge [ɠ]University of Gothenburg [ɔ]University of Oregon [ʃ]Australian National University
[ʛ]ILSP/Athena RC [ɢ]University of Groningen [ʋ]KU Leuven [з]University of Louisville [ᴇ]University of Edinburgh
[ﬁ]Pontificia Universidad Católica del Perú [ɟ]Universidad Católica Sedes Sapientiae, Filial Atalaya
[ʑ]Institute of Philology of the Siberian Branch of the Russian Academy of Sciences [ɱ]Moscow State University
[ŧ]Boston College [ɛ]Higher School of Economics [ь]Institute of Linguistics, Russian Academy of Sciences
[æ]University of Zürich [ɜ]STKIP Weetebula [ɶ]Institute for System Programming, Russian Academy of Sciences
[ɷ]University at Buffalo [ʁ]Karelian Research Centre of the Russian Academy of Sciences [ɸ]Swarthmore College
[ɕ]ESRC International Centre for Language and Communicative Development(LuCiD) [ɐ]New York University
[ɲ]Northeastern Illinois University [ɨ]University of Helsinki [ʂ]Tuvan State University [ɖ]Georgetown University
[ɾ]Charles Darwin University [θ]Institute of Computer Science, Polish Academy of Sciences [þ]Universitas Indonesia
[ʏ]Stony Brook University [ʊ]Dr. Bhimrao Ambedkar University [ø]Mila/McGill University Montreal
[χ]University of Colorado Boulder [ʎ]University of Liverpool [ŋ]Graduate Center, City University of New York
[ʄ]George Mason University [ᵞ]Ben-Gurion University of the Negev [ð]ETH Zürich [ɘ]University of Melbourne
khuyagbaatar@num.edu.mn omer.goldman@gmail.com vylomovae@unimelb.edu.au

Abstract

The Universal Morphology (UniMorph) project is a collaborative effort providing broad-coverage instantiated normalized morphological inflection tables 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. This paper presents the expansions and improvements made on several fronts over the last couple of years (since McCarthy et al. (2020)). Collaborative efforts by numerous linguists have added 67 new languages, including 30 endangered languages. We have implemented several improvements to the extraction pipeline to tackle some issues, e.g. missing gender and macron information. We have also amended the schema to use a hierarchical structure that is needed for morphological phenomena like multiple-argument agreement and case stacking, while adding some missing morphological features to make the schema more inclusive. In light of the last UniMorph release, we also augmented the database with morpheme segmentation for 16 languages. Lastly, this new release makes a push towards inclusion of derivational morphology in UniMorph by enriching the data and annotation schema with instances representing derivational processes from MorphyNet.

1. Introduction

Developing categories that allow for cross-linguistic comparison is one of the most challenging tasks in linguistic typology. Typologists have proposed dimensions of cross-linguistic variation such as fusion (Bickel and Nichols, 2013a), inflectional synthesis (Bickel and Nichols, 2013b), position of case affixes (Dryer, 2013), number of cases (Iggesen, 2013), and others, and these dimensions and descriptions are being progressively refined (Haspelmath, 2007).

Evans and Levinson (2009) critically discuss the idea of "linguistic universals", demonstrating extensive diversity across all levels of linguistic organization. The distinction between *g-linguistics*, a study of Human Language in general, and *p-linguistics*, a study of particular languages, including their idiosyncratic properties, is discussed in Haspelmath (2021). The UniMorph annotation schema (Sylak-Glassman et al., 2015b), and this work in particular, is an attempt to balance the trade-off between descriptive categories and comparative concepts through a more fine-grained analysis of languages (Haspelmath, 2010). The initial schema (Sylak-Glassman et al., 2015a) was based on the analysis of typological literature and included 23 dimensions of meaning (such as tense, aspect, grammatical person, number) and over 212 features (such as past/present for tense or singular/plural for number). The first release of the UniMorph database included 8 languages extracted from the English edition of Wiktionary (Kirov et al., 2016; Cotterell et al., 2016). The database has been augmented with 52 and 66 new languages in versions 2.0 and 3.0, respectively (Kirov et al., 2018; McCarthy et al., 2020). UniMorph 3.0 introduced many under-resourced languages derived from various linguistic sources. Prior to each release, all language datasets were included in part in the SIGMORPHON shared tasks on morphological reinflection (Cotterell et al., 2016; Cotterell et al., 2017; Cotterell et al., 2018; McCarthy et al., 2019). The current release includes languages of the 2020–2021 shared tasks (Vylomova et al., 2020; Pimentel et al., 2021). Unlike previous versions, linguistic data comes from grammar descriptions and finite-state models.

The work described here, representing the UniMorph 4.0 milestone, makes several contributions to further improve the UniMorph data and tools. First, we include inflection tables for 67 new languages and extend the datasets for 31 languages, increasing the total number of languages to 182. We note that the upcoming decade 2022–2032 has been announced as the Decade on Indigenous Languages,[1] and in this release we are enriching the UniMorph database with 30 endangered languages, as listed by UNESCO.[2] Second, we update the annotation schema to improve representation of phenomena such as polypersonal agreement and case stacking. Third, we provide morpheme segmentation data for 16 languages. Fourth, we introduce morpheme-annotated dataset of derivational morphology in 30 languages. Finally, we release new automatic validation tool to evaluate UniMorph against Universal Dependencies treebanks (Nivre et al., 2016). On the whole, UniMorph 4.0 covers 182 languages (as shown in Figure 1), 122M inflections, and 769K derivations.

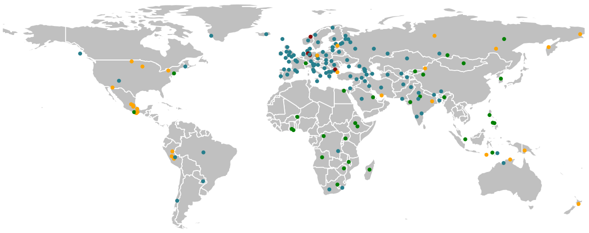


Figure 1: The UniMorph 4.0 languages (Oranges are endangered, dark reds are historic, greens are new languages, and blues are old languages).

2. Schema Updates

2.1. Hierarchical Annotation

The major structural change to the annotation schema in this release is the introduction of a hierarchical feature structure, following Guriel et al. (2022), instead of the flat structure that characterized the schema thus far. The shift is done to allow smoother incorporation of data for some non-western languages while keeping it easy to process. Specifically, the hierarchy is needed to annotate case stacking, polypersonal agreement, and more—treatment of some of which is impossible under the current system.

Verb forms with polypersonal agreement agree with more than one argument of the verb. In contrast to most western languages, where the verb agrees only with the subject (in the nominative case), verbs in many languages may agree with up to four different arguments. The existing schema attributes nominative features directly to the verbs in languages where only nominative agreement exists. Thus, for example, the English form *drinks* is annotated as V;PRS;3;SG, where the nominative-related features 3;SG are on the same level as PRS. However, for languages with poly-personal agreement a case specification is needed, and the solution is to mark that in a composite feature like ARGAC1S for a case where a form agrees with the verb's accusative argument which is 1st person singular.

The updated schema places the treatment of both cases on equal ground, while unpacking the composite feature string to a decomposable feature structure. Following Anderson (1992), features are *layered* such that some features may be composed of another set of features from the same feature inventory. We employ this structure to annotate every argument as a complex feature that includes all features pertaining to that argument.

*The authors contributed equally

[1] https:///en.unesco.org/news/upcoming-decade-indigenous-languages-2022-2032-focus-indigenous-language-users-human-rights

[2] http://www.unesco.org/languages-atlas/index.php

Language	Form	Hierarchical Schema	Flat Schema
English	drinks	V;PRS;NOM(3,SG)	V;PRS;3;SG
Georgian	გაგიშვებთ	V;FUT;NOM(1,PL);ACC(2,SG)	V;FUT;ARGNO1P;ARGAC2S
Hebrew	**עמדתה**	N;SG;PSSD;PSS(3,SG,FEM)	N;SG;PSSD;PSS3SF
Russian	собакам	N;DAT(PL)	N;DAT;PL
Evenki	ӈинакиннундуле	N;ALL(COM(SG))	—
Turkish	kedisini	N;ACC(SG;PSSD;PSS(1,SG))	N;SG;ACC;PSSD;PSS1S

Table 1: Example hierarchically annotated forms, including treatment of arguments, cases or both.

The aforementioned feature ARGAC1S is thus replaced with the composite feature ACC(1,SG), and a form that was formerly annotated as V;PRS;ARGNO3P;ARGAC2S is annotated as V;PRS;NOM(3,PL);ACC(2,SG). This solution applies not only to poly-personal agreement, but to any case in which annotation of a single form requires more than one person-number-gender feature bundle, like in the case of possessed nominals. See Table 1 for detailed examples.

Another case that requires hierarchical annotation is case stacking. In this phenomenon a noun takes the case suffix of its nominal head in addition to its own case suffix. For example in Evenki:

(1) асаткандула ӈинакиннундуле
asatkan-dula nginakin-nun-dule
girl.ALL dog.COM.ALL
'to the girl with the dog'

In these cases, the order of the cases is essential, but it cannot be captured by a flat unordered set of features. Therefore, in the updated schema cases are applied on top of the other nominal features and a form that was formerly tagged as N;SG;NOM would now be tagged as N;NOM(SG). This allows application of multiple cases in an order-preserving manner such that N;ALL(COM(SG)) is different from N;COM(ALL(SG)). For backward compatibility, the previous flat schema will continue to be maintained, although it cannot treat all forms in some extreme cases.

2.2. Derivational Morphology

UniMorph 4.0 releases a dataset of derivational morphology in 30 languages, annotated with morphemes and morphological features. The lemma (source word form) and derivation (target word form) are related to particular morphological annotation features represented by common part-of-speech tags and morpheme, as in the Italian example of *morfologia* 'morphology' and *morfologico* 'morphological':

(*morfologia*, *morfologico*, N:ADJ, '*-ico*'),

and in the French example of *décrit* 'to describe' and *susdécrit* 'above described':

(*décrit*, *susdécrit*, V:ADJ, '*sus-*').

Compared to state-of-the-art derivational resources (Vidra et al., 2019; Kyjánek et al., 2019), this dataset provides explicit morphemes between source and target word forms. With these morphemes, subword tokenization (Sennrich et al., 2016; Mielke et al., 2021) can be advanced to dictionary-based morpheme segmentation for derivationally rich languages like English and French. The extraction process and results of the derivational dataset are presented in Section 3.2.

2.3. New Morphosyntactic Features

Mood. The UniMorph schema (Sylak-Glassman et al., 2015a) combines imperative and jussive moods under one tag (IMP). This creates inconsistencies for languages such as Arabic. In Modern Standard Arabic (MSA), a verb can be perfective, imperfective or imperative (often marked as their aspect). Perfective verbs are always indicative, imperative verbs don't usually express mood, and imperfective verbs can be indicative, subjunctive, or jussive. To be able to transparently describe verbs in MSA, we split the imperative–jussive tag into two tags: imperative (IMP) and jussive (JUS), to accommodate imperative verbs and imperfective–jussive verbs.

Argument Marking. While working on indigenous languages of the Americas, Australia and Russia, we augmented the schema with the following features for argument marking: NO1, NO2, NO3, NO3F, NO3M, AC1, AC2, AC3 (no number specified), NO1PI, NO1PE (adding inclusivity), AC1D, AC2D, AC3D (adding dual number).[3]

Possession. We added the following tags: PSS1I (1st person inclusive), PSS3F, PSS3M (gender-specific tags), PSSRS and PSSRP (reflexive singular and plural).

2.4. Paradigm Classes in Russian

Aiming to establish a more granular performance analysis of (re)inflection models, we developed an application that infers possible inflection classes for each lemma present in UniMorph. By using this application, one may annotate each lemma with a set of known inflection paradigms that match all inflection samples present for a given lemma. To use this a technique, one needs a list of possible paradigms to be considered.

As a case study, we extracted a list of known inflection paradigms for Russian from the Russian edition of Wik-

[3] Although the annotation guidelines dictate that all argument marking features have an ARG prefix, in practice it is omitted for all argument features.

Family	Genus	ISO	Language	Source of Data	Annotators	Lemmas/Forms
Afro-Asiatic	Semitic	afb	Gulf Arabic	Khalifa et al. (2018)	Salam Khalifa, Nizar Habash	6,345/24,077
	Semitic	amh	Amharic	Gasser (2011)	Michael Gasser	2,461/46,224
	Semitic	arz	Egyptian Arabic	Habash et al. (2012)	Salam Khalifa, Nizar Habash	6,004/17,009
	Cushitic	orm	Oromo	Kasahorow (2017)	Irene Nikkarinen	92/2,046
Algic	Algonquian	cre*	Plains Cree	Hunter (1923)	Eleanor Chodroff	32/9,577
Arawakan	Southern Arawakan	ame*	Yanesha'	Duff-Trip (1998)	Gema Celeste Silva Villegas, Juan López Bautista, Didier López Francis, Roberto Zariquiey, Arturo Oncevay	327/3,767
	Southern Arawakan	cni*	Asháninka	Zumaeta Rojas and Zerdin (2018; Kindberg (1980)	Jaime Rafael Montoya Samame, Esaú Zumaeta Rojas, Delio Siticonatzi C., Roberto Zariquiey, Arturo Oncevay	407/20,070
Austronesian	Malayo-Polynesian	ind	Indonesian	KBBI, Wikipedia	Clara Vania, Totok Suhardijanto, Zahroh Nuriah	3,877/27,714
		kod*	Kodi	Ghanggo Ate (2021a)	Yustinus Ghanggo Ate, Garrett Nicolai	64/463
	Greater Central Philippine	ceb	Cebuano	Reyes (2015)	Ran Zmigrod	97/618
		hil	Hiligaynon	Santos (2018)	Ran Zmigrod	97/1,256
		tgl	Tagalog	NIU (2017)	Jennifer White	344/2,912
	Oceanic	mri*	Māori	Moorfield (2019)	Jennifer White	104/214
	Barito	mlg	Malagasy	Kasahorow (2015a)	Jennifer White	159/644
Aymaran	Aymaran	aym	Aymara	Coler (2014)	Matt Coler, Eleanor Chodroff	3,410/336,341
Chukotko-Kamchatkan	Northern Chukotko-Kamchatkan	ckt*	Chukchi	Chuklang; Tyers and Mishchenkova (2020)	Karina Sheifer, Maria Ryskina	197/243
	Southern Chukotko-Kamchatkan	itl*	Itelmen		Karina Sheifer, Sofya Ganieva, Matvey Plugaryov	1,636/2,701
Gunwinyguan	Gunwinggic	gup*	Kunwinjku	Lane and Bird (2019)	William Lane	73/307
Indo-European	Indic	asm	Assamese	Wiktionary	Khuyagbaatar Batsuren, Aryaman Arora	1,877/94,147
		bra	Braj	Kumar et al. (2018)	Shyam Ratan, Ritesh Kumar	1,246/1,821
		mag*	Magahi	Kumar et al. (2014)	Mohit Raj, Ritesh Kumar	1,612/2,194
		guj	Gujarati	Baxi et al. (2021);Wiktionary	Jatayu Baxi, Brijesh S. Bhatt, Khuyagbaatar Batsuren, Aryaman Arora	6,995/19,404
		hsi*	Kholosi	Arora and Etebari (2021)	Aryaman Arora	49/174
	Germanic	afr	Afrikaans	Dirix (2022)	Peter Dirix	179,941/309,558
		gsw	Swiss German	Egli-Wildi (2007)	Ryan Cotterell	145/2067
		got	Gothic	Wiktionary	Khuyagbaatar Batsuren (KB)	4,126/102,083
		goh	Old High German	Wiktionary	Jeremiah Young; KB	482/7,248
		non	Old Norse	Wiktionary	Jeremiah Young; KB	2,520/98,185
	Slavic	slk	Slovak	Hajič and Hric (2017)	Witold Kieraś	366,183/28,428,612
		hsb*	Upper Sorbian	Fraser (2020)	Taras Andrushko, Igor Marchenko	310/400
		poma	Pomak	under review	Ritván Karahóǧa, Stella Markantonatou, Georgios Pavlidis, Antonios Anastasopouos	233,533/6,557,759
Iroquoian	Northern Iroquoian	see*	Seneca	Bardeau (2007)	Richard J. Hatcher, Emily Prud'hommeaux, Zoey Liu	5,430/140
Koreanic	Koreanic	kor	Korean	Wiktionary	Maria Nepomniashchaya, Daria Rodionova, Anastasia Yemelina	2,686/241,323
Mongolic	Mongolic	khk	Khalkha Mongolian	Munkhjargal et al. (2016; Batsuren et al. (2019)	Khuyagbaatar Batsuren	2,085/14,592
Niger–Congo	Bantoid	kon	Kongo	Kasahorow (2016)	Jennifer White	200/828
		lin	Lingala	Kasahorow (2014a)	—	57/228
		lug	Luganda	Namono (2018)	Edoardo M. Ponti	89/4,895
		nya	Chewa	Kasahorow (2019a)	Ryan Cotterell	227/4,370
		sot	Sotho	Kasahorow (2020)	—	26/494
		sna	Shona	Kasahorow (2014b; Nandoro (2018)	Rowan Hall Maudslay	86/3,030
	Kwa	aka	Akan	Imbeah (2012)	Tiago Pimentel	96/4,182
		gaa	Gã	Kasahorow (2012a)	Tiago Pimentel	95/909

Table 2: Inflectional paradigms: new languages (Endangered languages are marked with *)

tionary.[4] The resource provides tables of patterns which represent declension and conjugation classes as they were defined by Zaliznyak (2003). We merged imported patterns into a list of records each represented as a triple

[4] https://ru.wiktionary.org/

Family	Genus	ISO	Language	Source of Data	Annotators	Lemmas/Forms
Oto-Manguean	Amuzgoan	azg*	San Pedro Amuzgos Amuzgo	Feist et al. (2015c)	Antonis Anastasopoulos	332/12,204
	Chichimec	pei*	Chichimeca-Jonaz	Feist and Palancar (2015b)	Antonis Anastasopoulos	123/15,120
	Chinantecan	cpa*	Tlatepuzco Chinantec	Feist and Palancar (2015e)	Antonis Anastasopoulos	697/7,893
	Mixtecan	xty	Yoloxóchitl Mixtec	Feist et al. (2015a)	Antonis Anastasopoulos	594/3,057
	Otomian	ote*	Mezquital Otomi	Feist and Palancar (2015d)	Antonis Anastasopoulos	2,028/33,162
	Otomian	otm*	Sierra Otomi	Feist and Palancar (2015c)	Antonis Anastasopoulos	1,909/31,380
	Zapotecan	cly*	Eastern Chatino of San Juan Quiahije	Cruz et al. (2020)	Hilaria Cruz, Antonis Anastasopoulos	185/4,716
	Zapotecan	ctp*	Eastern Chatino of Yaitepec	Feist et al. (2015d)	Antonis Anastasopoulos	223/3,796
	Zapotecan	czn*	Zenzontepec Chatino	Feist et al. (2015b)	Antonis Anastasopoulos	386/1,567
	Zapotecan	zpv*	Chichicapan Zapotec	Feist and Palancar (2015a)	Antonis Anastasopoulos	379/1,164
Pano-Tacana	Pano	shp*	Shipibo-Konibo	James et al. (1993);Valenzuela (2003)	Candy Angulo, Roberto Zariquiey, Arturo Oncevay	2,111/14,588
Siouan	Core Siouan	dak*	Dakota	LaFontaine and McKay (2005)	Eleanor Chodroff	537/3,766
Songhay	Songhay	dje	Zarma	Kasahorow (2019b)	Ran Zmigrod	27/84
Trans-New Guinea	Bosavi	ail*	Eibela	Aiton (2016)	Grant Aiton, Edoardo Maria Ponti, Ekaterina Vylomova	642/2,718
Tungusic	Tungusic	evn*	Evenki	Kazakevich and Klyachko (2013)	Elena Klyachko	4,495/11,371
	Tungusic	sjo*	Xibe	Zhou et al. (2020)	Elena Klyachko	1,892/3,054
Turkic	Turkic	sah	Sakha	Forcada et al. (2011, Apertium: apertium-sah)	Francis M. Tyers, Jonathan North Washington, Sardana Ivanova, Christopher Straughn, Maria Ryskina	5,622/590,765
	Turkic	tyv	Tuvan	Forcada et al. (2011, Apertium: apertium-tyv)	Francis M. Tyers, Jonathan North Washington, Aziyana Bayyr-ool, Aelita Salchak, Maria Ryskina	5,032/586,180
	Turkic	kir	Kyrgyz	(Aytnatova, 2016)	Eleanor Chodroff	98/5,544
	Turkic	uig	Uyghur	(Kadeer, 2016)	Eleanor Chodroff	90/8,178
	Turkic	uzb	Uzbek	(Abdullaev, 2016; Turkicum, 2019b)	Eleanor Chodroff	428/36,031
Uralic	Finnic	vro*	Võro	Iva (2007)	Ekaterina Vylomova	63/512
Uto-Aztecan	Tepiman	ood*	O'odham	Zepeda (2003)	Eleanor Chodroff	370/1,628
Yeniseian	Northern Yeniseian	ket*	Ket	Ket corpus	Elena Budianskaya, Polina Mashkovtseva, Alexandra Serova	349/1,184
Constructed	—	epo	Esperanto	Wiktionary	Arya D. McCarthy	1,945/58,350

Table 3: Inflectional paradigms: new languages (continuation; Endangered languages are marked with *)

consisting of the following:

- paradigm identifier (formed from a respective paradigm name given in Wiktionary);
- relevant UniMorph grammatical tags in their canonical order;
- word form pattern which usually contains one or more variable parts shared to other grammatical forms within the same paradigm.

We also developed an application that finds matching paradigms for every lemma in the UniMorph database by finding the intersection of matching paradigms over all {lemma, form, features} triplets observed for each given lemma in a UniMorph data file. Normally, multiple inflected forms occur for each lemma, which enables finding precise paradigms for most lemmas. Nevertheless, some ambiguity remains in many cases in Russian, due to the existence of numerous subtle variants in similar paradigms.

3. New Languages and Data

3.1. Inflectional Paradigms

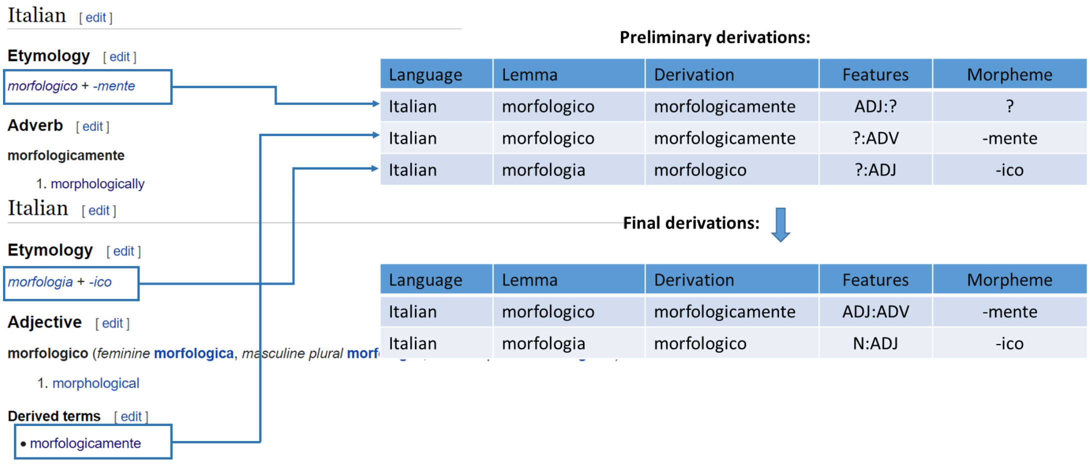
For the UniMorph 4.0 milestone, we have added new languages scraped from linguistic resources such as Surrey Morphology Group databases (Feist et al., 2015c), Apertium morphological analysers (Tyers et al., 2010), and other language grammars. The current release of inflectional paradigms cover about 122 million inflections in 182 languages in total.

3.1.1. New Languages

In the UniMorph 4.0 release, we introduce 67 new languages from 22 families: Afro-Asiatic, Algic, Arawakan, Austronesian, Aymaran, Chukotko-Kamchatkan, Gunwinyguan, Indo-European, Iroquoian, Koreanic, Mongolic, Niger–Congo, Oto-Manguean, Pano-Tacana, Siouan, Songhay, Trans-New Guinea, Tungusic, Turkic, Uralic, Uto-Aztecan, and Yeniseian,

Family	Genus	ISO	Language	Source of Data	Annotators	Lemmas/Forms
Afro-Asiatic	Semitic	ara	Standard Arabic	Taji et al. (2018)	Salam Khalifa, Nizar Habash	11,676/418,010
	Semitic	heb	Hebrew (Vocalized)	Wiktionary	Omer Goldman	1,183/33,178
	Semitic	heb	Hebrew (Unvocalized)	Sade et al. (2018)	Anna Yablonskaya	6,499/14,454
	Semitic	syc	Classic Syriac	SEDRA	Charbel El-Khaissi	3,299/31,972
Indo-European	Iranian	ckb	Central Kurdish (Sorani)	Alexina project	Ali Salehi	274/22,990
	Iranian	sdh	Southern Kurdish	Fattah (2000, native speakers)	Ali Salehi	1/189
	Slavic	pol	Polish	Woliński et al. (2020; Woliński and Kieraś (2016)	Witold Kieraś, Marcin Woliński	274,550/13,882,543
	Slavic	ces	Czech	Hajič et al. (2020)	Witold Kieraś	824,074/50,284,287
Niger-Congo	Bantoid	swc	Swahili	Kasahorow (2012b)	Jennifer White	97/4,949
	Bantoid	zul	Zulu	Kasahorow (2015b)	—	87/500
Turkic	Turkic	tur	Turkish	UniMorph (Kirov et al., 2018, Wiktionary)	Omer Goldman and Duygu Ataman	3,579/570,420
	Turkic	kaz	Kazakh	(Nabiyev, 2015; Turkicum, 2019a), Polish Wiktionary	Eleanor Chodroff, Khuyagbaatar Batsuren	1,755/40,283
Uralic	Finnic	krl	Karelian	Boyko et al. (2021, VepKar)	Andrew Krizhanovsky Natalia Krizhanovsky Elizabeth Salesky	10,842/411,271
	Finnic	lud	Ludic	Boyko et al. (2021, VepKar)		6,751/11,313
	Finnic	olo	Livvi	Boyko et al. (2021, VepKar)		27,676/1,199,149
	Finnic	vep	Veps	Boyko et al. (2021, VepKar)		18,618/815,676
Kartvelian	Karto-Zan	kat	Georgian	Guriel et al. (2022)	David Guriel	118/21,055

Table 4: Inflectional paradigms: augmented languages.



Preliminary derivations:

Language	Lemma	Derivation	Features	Morpheme
Italian	morfologico	morfologicamente	ADJ:?	?
Italian	morfologico	morfologicamente	?:ADV	-mente
Italian	morfologia	morfologico	?:ADJ	-ico

Final derivations:

Language	Lemma	Derivation	Features	Morpheme
Italian	morfologico	morfologicamente	ADJ:ADV	-mente
Italian	morfologia	morfologico	N:ADJ	-ico

Figure 2: The Wiktionary extraction process of derivational paradigms

and the Esperanto constructed language, as shown in Table 2 and 3. Of these new languages, 30 are endangered.[5] Extended details on some of the languages can be found in Appendix A.

3.1.2. Augmented Languages

The data for a handful of existing languages was expanded in several dimensions. In most cases the expansion included additions of new inflection tables from various sources, but for some languages data was added by expanding existing inflection tables (e.g. Turkish), by adding for more dialects (e.g. Arabic), or by accounting for orthographic variations (e.g. Hebrew). See Table 4 for details.

For some languages the additional data is much larger. For example, the new Czech data consists of about 50M analyzed word forms from Hajič et al. (2020), compared to the 135k existing forms, and some Uralic languages' data grew from a few hundred forms to about a million using the VepKar corpus (Boyko et al., 2021).

3.2. Derivational Paradigms

Language-specific editions of Wiktionary contain large amounts of derivational data, typically in two forms: *etymology templates* and *derived terms* (see Figure 2). Building on prior results from the *MorphyNet* project (Batsuren et al., 2021), we have implemented an extraction mechanism from both kinds of sections, covering 12 Wiktionary editions and 30 languages.

We managed to extract 4.3 million preliminary derivations, as reported in Table 5. We considered such derivations as 'preliminary' because they are both redundant

[5] http://www.unesco.org/languages-atlas/index.php

Wiktionary edition	Etymology	Derived terms
English	683,351	1,116,122
French	17,784	475,843
Finnish	16,727	23,516
Hungarian	9,358	n.a
Polish	n.a	1,200,228
Russian	n.a	303,052
German	n.a	244,032
Czech	n.a	178,383
Italian	n.a	40,020
Portuguese	n.a	12,667
Catalan	n.a	7,069
Serbo-Croatian	n.a	4,271
Total	727,220	3,605,203

Table 5: Preliminary incomplete derivations extracted from 12 editions of Wiktionary

and incomplete: some derivations are provided multiple times, but may lack indications for certain derivational features, such as parts of speech or affixes, as shown in Figure 2. For example, the etymology section of the Italian *'morfologia → morfologico'* does not provide the part of speech of the source lemma, while *'morfologico → morfologicamente'* is provided in two different ways.

In order to obtain final and complete derivations, we automatically fused the preliminary instances and eliminated duplicates. As a final result, shown in Table 6, we inferred 769,102 derivations and 12,420 affixes for 30 languages of 10 genera.

3.3. Morpheme Segmentation

The schema update of UniMorph 3.0 (McCarthy et al., 2020) introduced segmentation structure of inflected forms along with segmented morphological features, as in Figure 3(c). UniMorph 4.0 extends this data structure by complete morphological analysis for 16 languages. Segmentations were computed using language-specific inflectional morpheme datasets representing the inflection network between word forms, as shown in Figure 3(b). Each node of this network represents a unique set of morphological features, and each directed edge represents the fact that the target form is an inflection of the source. Each row of Figure 3(b) corresponds to an edge of the network, with each item in the *Morphemes* column implementing the inflection. For example, in Hungarian all plural dative noun N;DAT;PL word forms are inflected from the plural nominal N;NOM;PL forms by one of the suffixes *-ak,-ek,-ok,-ök,-k*. Such morpheme tables were created by language expert contributors for 16 languages. Using the morpheme tables, we algorithmically (recursively) segment each inflected word form in UniMorph. This method is very effective with regular inflection cases for the 16 languages considered. In order to cover irregular inflections (Gorman et al., 2019), we implemented custom segmentation rules for these languages. In total, 15 million segmentations were computed for 16 languages, as shown in Table 7.

Languages	Lemmas	Derivations	Morphemes
English	67,412	225,131	2,445
Russian	11,922	93,039	575
French	12,473	72,952	636
Italian	18,650	58,848	749
Polish	6,518	58,711	405
Finnish	18,142	36,843	446
Czech	4,875	32,336	318
German	8,070	29,381	465
Hungarian	14,566	28,177	832
Spanish	9,159	25,080	490
Dutch	7,810	13,506	366
Portuguese	6,076	11,774	387
Romanian	6,929	11,039	382
Swedish	2,190	9,244	217
Serbo-Croatian	4,916	8,553	429
Catalan	5,492	8,284	241
Ukraine	5,212	6,650	105
Irish	3,719	6,417	270
Latin	3,429	5,889	689
Latvian	1,869	4,235	91
Bokmal	2,310	3,238	227
Danish	2,137	3,021	184
Galician	1,995	2,832	230
Greek	1,842	2,575	372
Nynorsk	1,542	2,131	217
Armenian	1,527	2,009	130
Kazakh	1,348	1,965	91
Scottish-Gaelic	1,346	1,837	80
Turkish	1,248	1,776	122
Mongolian	1,410	1,629	229
Total	236,134	769,102	12,420

Table 6: Final derivations of 30 languages, released in UniMorph 4.0

Related work on segmentation or extracting lexical information from Wiktionary include the Wikinflection project (Metheniti and Neumann, 2020), the DBnary project (Sérasset, 2015), MorphoChallenge data (Kurimo et al., 2010), JWKTL (Zesch et al., 2008), EtymDB-2.0 (Fourrier and Sagot, 2020), and Yawipa (Wu and Yarowsky, 2020a; Wu and Yarowsky, 2020b).

4. Validation tool

Evaluation of morphological databases' quality is a challenging task due to the weird and irregular morphological aspects of languages (Gorman et al., 2019). Given millions of inflections in languages such as Finnish and Russian, manual evaluation is often time-consuming and cost-inefficient. In this release, we extend an existing UniMorph validation tool[6], developed by McCarthy et al. (2018). With this extension, we can compute the precision, recall, and F-measure for all part-of-speech categories of UniMorph resources. It complements the tools released in McCarthy et al. (2020) for canonicalization and flagging common annotation errors.

[6] https://github.com/unimorph/ud-compatibility

(a) UniMorph 3.0

Lemma	Form	Features
légy	légy	`N;NOM;SG`
légy	legyek	`N;NOM;PL`
légy	legyeknek	`N;DAT;PL`

(b) Morpheme Table

Source Form	Morphemes	Target Form
`N;NOM;SG`	-ök;-ok;-ek;-ak;-k	`N;NOM;PL`
`N;NOM;PL`	-nak;-nek	`N;DAT;PL`

(c) UniMorph 4.0 with Segmentation

Lemma	Form	Features	Segmentation
légy	légy	`N;NOM;SG`	—
légy	legyek	`N|NOM;PL`	légy | ek
légy	legyeknek	`N|PL|DAT`	légy | ek | nek

Figure 3: Segmentation process

Language	Lemmas	Forms/Segmentations
Finnish	81,729	3,708,296
Serbo-Croatian	68,757	1,760,095
Latin	50,949	1,440,506
Russian	36,387	1,321,024
Spanish	65,565	1,289,324
Hungarian	38,067	1,016,819
Czech	33,348	816,956
Italian	89,763	712,021
Polish	36,940	663,545
English	396,772	649,594
German	39,275	490,331
French	52,711	453,229
Portuguese	39,029	376,341
Catalan	14,979	158,922
Swedish	12,508	131,599
Mongolian	2,085	14,592
Total	1,058,864	15,003,194

Table 7: UniMorph 4.0 languages with segmentations

With this validation tool, we evaluated five high-resource languages—English, Latin, French, Russian, and Spanish—against the UD treebanks (Silveira et al., 2014; Haug and Jøhndal, 2008; Guillaume et al., 2019; Lyashevskaya et al., 2019; Taulé and Recasens, 2008) (Table 8). UniMorph 3.0 data results in high precision between 97.2% and 99.8% but at low recall rates from 10.8% to 43.3%. An important reason for these low recall rates was that UniMorph 3.0 was based on the data extracted 4–5 years ago. Since then, Wiktionary has been constantly improved by the Wiktionarians. Another crucial reason was the fact that UniMorph 3.0 had no inflections for adjectives and nouns for English, French, and Spanish. In addition, Latin inflections lack the entire class of deponent verbs and Russian inflections miss lexical features, e.g., gender for nouns and perfective/imperfective aspects for verbs. In both Latin and Russian, participles have no morphological features on case, gender, and number. By incorporating these into the extraction pipeline, we extracted new data from Wiktionary on these five languages and conducted the evaluation again. As shown in Table 8, recall rates were significantly improved to 61.5–89.7% while maintaining high quality at 95.2–99.3%. With this approach, we have so far extended and improved 17 existing languages of UniMorph.

Language	UniMorph	Recall	Precision	F_1
English	v3.0	24.6	98.6	39.4
	v4.0	71.6	99.7	83.4
Latin	v3.0	43.3	97.2	59.9
	v4.0	76.3	98.1	85.3
French	v3.0	20.6	98.5	34.1
	v4.0	79.7	97.9	87.9
Russian	v3.0	10.8	97.4	19.4
	v4.0	61.5	95.2	74.7
Spanish	v3.0	32.1	99.8	48.6
	v4.0	89.7	99.3	94.3

Table 8: Automatic validation of UniMorph v3.0 and v4.0 on UD Treebanks for five languages

5. Conclusion

The UniMorph project represents a massively multilingual effort at cataloguing the world's inflectional and derivational morphology. Here, we present UniMorph 4.0 which has several improvements and expansions both in terms of contents and scopes over the previous release. First, a large community of linguists from all over the world contributed to the UniMorph project over the last few years, resulting in 67 new languages (including 30 endangered languages) and an extension of inflectional data on existing 31 languages. Second, we amended the schema with a hierarchical structure necessary for morphological phenomena like multiple-argument agreement and case stacking, while adding missing morphological features to make the schema more inclusive. Third, we introduced morpheme-annotated derivational paradigms, covering 769K derivations in 30 languages from 10 genera. Fourth, we added morpheme segmentation for 16 languages. Finally, we implemented an automatic validation tool to evaluate the UniMorph data against the Universal Dependencies treebanks. With all these efforts, the new release becomes more accurate and complete. The data and tools are published under an open source license at `unimorph.github.io`. The project welcomes continued contributions from the community.

Acknowledgments

OG and RT wish to thank ERC grant no. 677352.

References

Abdullaev, D. (2016). *Uzbek language: 100 Uzbek verbs conjugated in common tenses*. CreateSpace Independent Publishing Platform, Online.

Aiton, G. W. (2016). *A grammar of Eibela: a language of the Western Province, Papua New Guinea*. Ph.D. thesis, James Cook University.

Anderson, S. R. (1992). *A-morphous morphology*. Number 62. Cambridge University Press.

Arka, I. W. (2002). Voice systems in the Austronesian languages of Nusantara: Typology, symmetricality and undergoer orientation. *Linguistik Indonesia*, 21(1):113–139.

Arora, A. and Etebari, A. (2021). *Kholosi Dictionary*.

Aytnatova, A. (2016). *Kyrgyz Language: 100 Kyrgyz Verbs Fully Conjugated in All Tenses*. CreateSpace Independent Publishing Platform, Online.

Bardeau, P. E. W. (2007). *The Seneca Verb: Labeling the Ancient Voice*. Seneca Nation Education Department, Cattaraugus Territory.

Batsuren, K., Ganbold, A., Chagnaa, A., and Giunchiglia, F. (2019). Building the Mongolian WordNet. In *Proceedings of the 10th Global Wordnet Conference*, pages 238–244, Wroclaw, Poland, July. Global Wordnet Association.

Batsuren, K., Bella, G., and Giunchiglia, F. (2021). MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online, August. Association for Computational Linguistics.

Baxi, J., Bhatt, D., et al. (2021). Morpheme boundary detection & grammatical feature prediction for gujarati: Dataset & model. *arXiv preprint arXiv:2112.09860*.

Bickel, B. and Nichols, J. (2013a). Fusion of selected inflectional formatives. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Bickel, B. and Nichols, J. (2013b). Inflectional synthesis of the verb. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Boyko, T., Zaitseva, N., Krizhanovskaya, N., Krizhanovsky, A., Novak, I., Pellinen, N., Rodionova, A., and Trubina, E. (2021). The linguistic corpus VepKar is a language refuge for the Baltic-Finnish languages of Karelia. *Transactions of the Karelian Research Centre of the Russian Academy of Sciences*, (7):100–115.

Coler, M. (2014). *A grammar of Muylaq'Aymara: Aymara as spoken in Southern Peru*. Brill.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). 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., 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. J., 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.

Cruz, H., Anastasopoulos, A., and Stump, G. (2020). A resource for studying chatino verbal morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2820–2824, Marseille, France, May. European Language Resources Association.

Dirix, P. (2022). The need for a large(r) Afrikaans treebank. In Ian Bekker et al., editors, *"Ex Africa semper aliquid novi": Linguistic shorts in honour of Andries Coetzee on his 50th birthday*. Stellenbosch Papers in Linguistics Plus, Stellenbosch.

R. M. W. Dixon et al., editors. (1999). *The Amazonian languages (Cambridge Language Surveys)*. Cambridge University Press.

Dryer, M. S. (2013). Position of case affixes. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Duff-Trip, M. (1998). *Diccionario Yanesha' (Amuesha)-Castellano*. Lima: Instituto Lingüístico de Verano.

Egli-Wildi, R. (2007). *Züritüütsch verstaa - Züritüütsch rede*. Küsnacht.

Evans, N. and Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448.

Fattah, I. (2000). *Les dialectes kurdes méridionaux: étude linguistique et dialectologique*. Acta Iranica : Encyclopédie permanente des études iraniennes. Peeters.

Feist, T. and Palancar, E. L. (2015a). Oto-manguean inflectional class database: Chichicapan Zapotec. University of Surrey.

Feist, T. and Palancar, E. L. (2015b). Oto-manguean inflectional class database: Chichimec. University of Surrey.

Feist, T. and Palancar, E. L. (2015c). Oto-manguean inflectional class database: Eastern Highland Otomi. University of Surrey.

Feist, T. and Palancar, E. L. (2015d). Oto-manguean inflectional class database: Mezquital Otomi. University of Surrey.

Feist, T. and Palancar, E. L. (2015e). Oto-manguean inflectional class database: Tlatepuzco Chinantec. University of Surrey.

Feist, T., Palancar, E. L., Amith, J., and Castillo García, R. (2015a). Oto-manguean inflectional class database: Yoloxóchitl Mixtec. University of Surrey.

Feist, T., Palancar, E. L., and Campbell, E. (2015b). Oto-manguean inflectional class database: Zenzontepec Chatino. University of Surrey.

Feist, T., Palancar, E. L., and Fermin, T. (2015c). Oto-manguean inflectional class database: San Pedro Amuzgos Amuzgo. University of Surrey.

Feist, T., Palancar, E. L., and Rasch, J. (2015d). Oto-manguean inflectional class database: Yaitepec Chatino. University of Surrey.

Ferguson, C. F. (1959). Diglossia. *Word*, 15(2):325–340.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Fourrier, C. and Sagot, B. (2020). Methodological aspects of developing and managing an etymological lexical resource: Introducing EtymDB-2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3207–3216, Marseille, France, May. European Language Resources Association.

Fraser, A. (2020). Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online, November. Association for Computational Linguistics.

Gasser, M. (2011). HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. In *Proceedings of the Conference on Human Language Technology for Development*, Alexandria, Egypt.

Ghanggo Ate, Y. (2021a). *Documentation of Kodi*. New Haven: Endangered Language Fund.

Ghanggo Ate, Y. (2021b). Reduplication in Kodi: A paradigm function account. *Word Structure 14(3)*, 14(3):312–353.

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.

Grierson, G. A. and Konow, S. (1903). *Linguistic Survey of India*. Calcutta Supt., Govt. Printing.

Guillaume, B., de Marneffe, M.-C., and Perrier, G. (2019). Conversion et améliorations de corpus du français annotés en universal dependencies. *Traitement Automatique des Langues*, 60(2):71–95.

Guriel, D., Goldman, O., and Tsarfaty, R. (2022). Morphological reinflection with multiple arguments: An extended annotation schema and a georgian case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, May. Association for Computational Linguistics.

Habash, N., Eskander, R., and Hawwari, A. (2012). A morphological analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada, June. Association for Computational Linguistics.

Hajič, J. and Hric, J. (2017). MorfFlex SK 170914. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., and Štěpánková, B. (2020). MorfFlex CZ 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Haspelmath, M. (2007). Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1):119–132.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Haspelmath, M. (2021). General linguistics must be based on universals (or non-conventional aspects of language). *Theoretical Linguistics*, 47(1-2):1–31.

Haug, D. T. and Jøhndal, M. (2008). Creating a parallel treebank of the old indo-european bible translations. In *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 27–34.

Hunter, J. (1923). *A Lecture on the Grammatical Construction of the Cree Language. Also Paradigms of the Cree Verb (Original work published 1875*. The Society for Promoting Christian Knowledge, London.

Iggesen, O. A. (2013). Number of cases. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Imbeah, P. K. (2012). *102 Akan Verbs*. CreateSpace Independent Publishing Platform, Online.

Iva, S. (2007). *Võru kirjakeele sõnamuutmissüsteem*. Ph.D. thesis.

Jain, D. and Cardona, G. (2007). *The Indo-Aryan Languages*. Routledge.

James, L., Lauriault, E., and Day, D. (1993). *Diccionario Shipibo-Castellano*. Instituto Lingüístico de Verano.

Kadeer, A. (2016). *Uyghur language: 94 Uyghur verbs in common tenses*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2012a). *102 Ga Verbs*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2012b). *102 Swahili Verbs*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2014a). *102 Lingala Verbs: Master the Simple Tenses of the Lingala*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2014b). *102 Shona Verbs: Master the simple tenses of the Shona language*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2015a). *Modern Malagasy Verbs: Master the Simple Tenses of the Malagasy Language*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2015b). *Modern Zulu Verbs: Master the simple tenses of the Zulu language*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2016). *Modern Kongo Verbs: Master the Simple Tenses of the Kongo Language*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2017). *Modern Oromo Dictionary: Oromo-English, English-Oromo*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2019a). *Modern Chewa Verbs: Master the basic tenses of Chewa*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2019b). *Modern Zarma Verbs: Master the basic tenses of Zarma*. CreateSpace Independent Publishing Platform, Online.

Kasahorow. (2020). *Modern Sotho Verbs: Master the basic tenses of Sotho (Sotho dictionary*. CreateSpace Independent Publishing Platform, Online.

Kazakevich, O. and Klyachko, E. (2013). Creating a multimedia annotated text corpus: a research task (Sozdaniye multimediynogo annotirovannogo korpusa tekstov kak issledovatelskaya protsedura). In *Proceedings of International Conference Computational linguistics 2013*, pages 292–300.

Khalifa, S., Habash, N., Eryani, F., Obeid, O., Abdulrahim, D., and Al Kaabi, M. (2018). A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Kindberg, L. (1980). *Diccionario asháninca*. Lima: Instituto Lingüístico de Verano.

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.

Kirov, C., Cotterell, R., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., Faruqui, M., Mielke, S. J., 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).

Kumar, R., Lahiri, B., and Alok, D. (2014). Developing LRs for Non-scheduled Indian Languages: A Case of Magahi. In *Human Language Technology Challenges for Computer Science and Linguistics*, Lecture Notes in Computer Science, pages 491–501. Springer International Publishing, Switzerland. original-date: 2014.

Kumar, R., Lahiri, B., Ojha, D. A. A. K., Jain, M., Basit, A., and Dawar, Y. (2018). Automatic identification of closely-related Indian languages: Resources and experiments. In *Proceedings of the 4th Workshop on Indian Language Data Resource and Evaluation (WILDRE-4)*, Paris, France, may. European Language Resources Association (ELRA).

Kurimo, M., Virpioja, S., and Turunen, V. T. (2010). Proceedings of the morpho challenge 2010 workshop.

Kyjánek, L., Žabokrtský, Z., Ševčíková, M., and Vidra, J. (2019). Universal derivations kickoff: A collection of harmonized derivational resources for eleven languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 101–110, Prague, Czechia, September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

LaFontaine, H. and McKay, N. (2005). *550 Dakota Verbs*. Minnesota Historical Society Press, Online.

Lahiri, B. (2021). *The Case System of Eastern Indo-Aryan Languages: A Typological Overview*. Routledge.

Lane, W. and Bird, S. (2019). Towards a robust morphological analyzer for Kunwinjku. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 1–9, Sydney, Australia, 4–6 December. Australasian Language Technology Association.

Levin, T. and Polinsky, M. (2019). Morphology in Austronesian. In *Oxford Research Encyclopedia of Linguistics*.

Lyashevskaya, O., Droganova, K., Zeman, D., Alexeeva, M., Gavrilova, T., Mustafina, N., Shakurova, E., et al. (2019). Universal dependencies for russian: A new syntactic dependencies tagset.

McCarthy, A. D., Silfverberg, M., Cotterell, R., Hulden, M., and Yarowsky, D. (2018). Marrying universal dependencies and universal morphology. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*.

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.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., Arkhangelskiy, T., Krizhanovsky, N., Krizhanovsky, A., Klyachko, E., Sorokin, A., Mansfield, J., Ernštreits, V., Pinter, Y., Jacobs, C. L., Cotterell, R., Hulden, M., and Yarowsky, D. (2020). UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France, May. European Language Resources Association.

Metheniti, E. and Neumann, G. (2020). Wikinflection corpus: A (better) multilingual, morpheme-annotated inflectional corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3905–3912, Marseille, France, May. European Language Resources Association.

Mielke, S. J., Alyafeai, Z., Salesky, E., Raffel, C., Dey, M., Gallé, M., Raja, A., Si, C., Lee, W. Y., Sagot, B., et al. (2021). Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*.

Moorfield, J. C. (2019). *Te Aka Online Māori Dictionary*. Online.

Munkhjargal, Z., Chagnaa, A., and Jaimai, P. (2016). Morphological transducer for mongolian. In *International Conference on Computational Collective Intelligence*, pages 546–554. Springer.

Nabiyev, T. (2015). *Kazakh Language: 101 Kazakh Verbs*. Preceptor Language Guides, Online.

Namono, M. (2018). *Luganda Language: 101 Luganda Verbs*. CreateSpace Independent Publishing Platform, Online.

Nandoro, I. (2018). *Shona Language: 101 Shona Verbs*. CreateSpace Independent Publishing Platform, Online.

NIU, C. f. S. A. S. (2017). *Table of Tagalog Verbs*. CreateSpace Independent Publishing Platform, Online.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., et al. (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.

Pedrós, T. (2018). Ashéninka y asháninka: ¿de cuántas lenguas hablamos? *Cadernos de Etnolingüística*, 6(1):1–30.

Pimentel, T., Ryskina, M., Mielke, S. J., Wu, S., Chodroff, E., Leonard, B., Nicolai, G., Ghanggo Ate, Y., Khalifa, S., Habash, N., Goldman, O., Gasser, M., Lane, W., Coler, M., Oncevay, A., Montoya Samame, J. R., Silva Villegas, G. C., Ek, A., Bernardy, J.-P., Shcherbakov, A., Bayyr-ool, A., Sheifer, K., Ganieva, S., Plugaryov, M., Klyachko, E., Salehi, A., Krizhanovsky, A., Krizhanovsky, N., Vania, C., Ivanova, S., Salchak, A., Straughn, C., Liu, Z., North Washington, J., Ataman, D., Kieraś, W., Woliński, M., Suhardijanto, T., Stoehr, N., Nuriah, Z., Ratan, S., Tyers, F. M., Ponti, E. M., Aiton, G., Hatcher, R. J., Prud'hommeaux, E., Kumar, R., Hulden, M., Barta, B., Lakatos, D., Szolnok, G., Ács, J., Raj, M., Yarowsky, D., Cotterell, R., Ambridge, B., and Vylomova, E. (2021). Sigmorphon 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259.

Reyes, D. (2015). *Cebuano Language: 101 Cebuano Verbs*. CreateSpace Independent Publishing Platform, Online.

Sade, S., Seker, A., and Tsarfaty, R. (2018). The hebrew universal dependency treebank: Past present and future. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 133–143.

Santos, A. (2018). *Hiligaynon Language. 101 Hiligaynon Verbs*. CreateSpace Independent Publishing Platform, Online.

Sennrich, R., Haddow, B., and Birch, A. (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.

Sérasset, G. (2015). Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Silveira, N., Dozat, T., De Marneffe, M.-C., Bowman, S. R., Connor, M., Bauer, J., and Manning, C. D. (2014). A gold standard dependency corpus for english. In *LREC*, pages 2897–2904. Citeseer.

Sylak-Glassman, J., Kirov, C., Post, M., Que, R., and Yarowsky, D. (2015a). A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In Cerstin Mahlow et al., editors, *Systems and Frameworks for Computational Morphology*, pages 72–93, Cham. Springer International Publishing.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (2015b). 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.

Taji, D., Khalifa, S., Obeid, O., Eryani, F., and Habash, N. (2018). An Arabic morphological analyzer and generator with copious features. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 140–150, Brussels, Belgium, October. Association for Computational Linguistics.

Taulé, M. and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish.

Turkicum. (2019a). *The Kazakh Verbs: Review Guide*. Preceptor Language Guides, Online.

Turkicum. (2019b). *The Uzbek Verbs: Review Guide*. CreateSpace Independent Publishing Platform, Online.

Tyers, F. and Mishchenkova, K. (2020). Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Tyers, F. M., Sánchez-Martínez, F., Ortiz Rojas, S., Forcada, M. L., et al. (2010). Free/open-source resources in the apertium platform for machine translation research and development.

Valenzuela, P. (2003). *Transitivity in Shipibo-Konibo Grammar*. Ph.D. thesis, University of Oregon, July.

Vidra, J., Žabokrtský, Z., Ševčíková, M., and Kyjánek, L. (2019). DeriNet 2.0: Towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 81–89, Prague, Czechia, September. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.

Vylomova, E., White, J., Salesky, E., Mielke, S. J., Wu, S., Ponti, E. M., Hall Maudslay, R., Zmigrod, R., Valvoda, J., Toldova, S., Tyers, F., Klyachko, E., Yegorov, I., Krizhanovsky, N., Czarnowska, P., Nikkarinen, I., Krizhanovsky, A., Pimentel, T., Torroba Hennigen, L., Kirov, C., Nicolai, G., Williams, A., Anastasopoulos, A., Cruz, H., Chodroff, E., Cotterell, R., Silfverberg, M., and Hulden, M. (2020). SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online, July. Association for Computational Linguistics.

Woliński, M. and Kieraś, W. (2016). The on-line version of Grammatical Dictionary of Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, pages 2589–2594, Portorož, Slovenia. European Language Resources Association (ELRA), European Language Resources Association (ELRA).

Woliński, M., Saloni, Z., Wołosz, R., Gruszczyński, W., Skowrońska, D., and Bronk, Z. (2020). *Słownik gramatyczny języka polskiego*. Warsaw, 4th edition. http://sgjp.pl.

Wu, W. and Yarowsky, D. (2020a). Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France, May. European Language Resources Association.

Wu, W. and Yarowsky, D. (2020b). Wiktionary normalization of translations and morphological information. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4683–4692, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Zaliznyak, A. A. (2003). *Grammaticheskij slovar' russkogo jazyka [The grammar dictionary of Russian]*. Русские словари.

Zepeda, O. (2003). *A Tohono O'odham grammar (Original work published 1983)*. University of Arizona Press, Online.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Zhou, H., Chung, J., Kübler, S., and Tyers, F. (2020). Universal Dependency treebank for Xibe. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 205–215, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Zumaeta Rojas, E. and Zerdin, G. A. (2018). *Guía teórica del idioma asháninka*. Lima: Universidad Católica Sedes Sapientiae.

A. Languages Details

Semitic

Arabic Modern Standard Arabic (MSA, ara) is the primarily written form of Arabic and is used in all official communication means. In contrast, Arabic dialects are the primarily spoken varieties of Arabic, and the increasingly written varieties on unofficial social media platforms. Dialects have no official status despite being widely used. Both MSA and the dialects coexist in a sate of diglossia (Ferguson, 1959) whether in spoken or written form. Arabic dialects vary among themselves and are different from MSA in most linguistic aspects (phonology, morphology, syntax, and lexical choice). In this work we provide inflection tables for (MSA, ara), Egyptian Arabic (EGY, arz), and Gulf Arabic (GLF, afb). Egyptian Arabic is the variety of Arabic spoken in Egypt. Gulf Arabic is referred to the dialects spoken by the indigenous populations of the member states of the Gulf Cooperation Council, especially those in regions on the Arabian Gulf.

Syriac Classical Syriac is a dialect of the Aramaic language and is attested as early as the 1st century

CE. As with most Semitic languages, it displays non-concatenative morphology involving primarily tri-consonantal roots. Syriac nouns and adjectives are conventionally classified into three ‘states’—Emphatic, Absolute, Construct—which loosely correlate with the syntactic features of definiteness, indeterminacy and the genitive. There are over 10 verbal paradigms that combine affixation slots with inflectional templates to reflect tense (past, present, future), person (first, second, third), number (singular, plural), gender (masculine, feminine, common), mood (imperative, infinitive), voice (active, passive), and derivational form (i.e., participles). Paradigmatic rules are determined by a range of linguistic factors, such as root type or phonological properties. The data included in this set was relatively small and consisted of 1,217 attested lexemes in the New Testament, which were extracted from *Beth Mardutho: The Syriac Institute*’s lexical database, SEDRA.

Hebrew is a member of the Northwest Semitic branch, and, like Syriac and Arabic, it is written using an abjad where the vowels are sparsely marked in unvocalized text. This fact entails that in unvocalized data the complex ablaut-extensive non-concatenative Semitic morphology is somewhat watered down as the consonants of the root frequently appear consecutively with the alternating vowel unwritten. In this release we added data in vocalized Hebrew, in order to examine the models’ ability to handle Hebrew’s full-fledged Semitic morphological system.
The inflection tables are largely identical to those included in UniMorph 3.0, scraped from Wiktionary, with the addition of the verbal nouns and all forms being automatically vocalized.

Amharic is the most spoken among the roughly 15 languages in the Ethio-Semitic branch of South Semitic. Unlike most other Semitic languages, it is written in the Ge’ez (Ethiopic) script, an abugida in which each character represents either a consonant-vowel sequence or a consonant in the syllable coda position. Like other Semitic languages, Amharic displays both affixation and non-concatenative template morphology. Verbs inflect for subject person, gender, and number and tense/aspect/mood. Voice and valence are also marked, but these are treated as separate lemmas in the data. Other verb affixes, which are not included in the data, indicate object person, gender, and number; negation; and relativization. Nouns and adjectives share most of their morphology and are often not clearly distinguished. Nouns and adjectives inflect for definiteness, number, and possession. Nouns and adjectives also have prepositional prefixes and accusative suffixes, which are not included in the data.

Turkic

Turkish is part of the Oghuz branch, and it is highly agglutinative, like the other languages of this family.
This release vastly expanded the pre-existing UniMorph inflection tables. As with the Siberian Turkic languages, it was necessary to omit many forms from the paradigm as the UniMorph schema is not well-suited for Turkic languages. For this reason, we only included the forms that may appear in main clauses. Other than this limitation, we tried to include all possible tense-aspect-mood combinations, resulting in 30 series of forms, each including 3 persons and 2 numbers. The nominal coverage is less comprehensive and includes forms with case and possessive suffixes.

Indo-European

The Indo-European language family consists most of European and Asian languages. South Asia that encompasses India, Pakistan, Bangladesh, Nepal, Bhutan, Sri Lanka and Maldives is referred to as the heartland of Indo-Aryan or Indic languages are spoken (Jain and Cardona, 2007). We enrich the data with two languages Magahi and Braj from Indo-Aryan or Indic languages which are spoken in Indian states.

Indo-Aryan: Braj bhasha, or Braj is spoken in the Western Indian states of Uttar Pradesh, Rajasthan and Madhya Pradesh, which is one of the Indo-Aryan languages. Braj is highly inflectional language in this language family. We have used the data from the literary domain (Kumar et al., 2018). The final dataset contains 1,821 wordforms and 1,246 lexemes including nouns, verbs and adjectives. our analysis of the language has shown that there are 34 possible forms for verbs, 3 forms for adjectives and 2 forms for nouns. As is clear from this, in the first phase, we have preferred breadth (i.e. represent larger number of lexemes) over depth (i.e. only a few wordforms of most of the lexemes are represented) in the current version.

Indo-Aryan: Magahi comes under the Magadhi group of the middle Indo-Aryan language which is spoken mainly in Eastern Indian states of Bihar and Jharkhand and also to the adjoining region of Bengal and Odisha (Grierson and Konow, 1903). Magahi has no grammatical gender agreement, though animate nouns like /laika/ (boy) and /laiki/ (girl) show sex-related gender derivation, noun also carry number marker that affects the form of case markers and postposition in certain instances (Lahiri, 2021). The language has a rich and diverse system of verbal morphology to show the honorific agreement, tense, aspect, person, resulting in as many as 24 distinct forms of verbs, 19 forms of aux and 4 forms of nouns. We have used a dataset from the literary domain in order to extract the inflectional paradigm of nouns and verbs. The present dataset contains 1,612 lexemes and 2,194 wordforms which includes noun, verb, adjective, conjunction, adverb etc.

West Slavic: Upper Sorbian is a West Slavic language spoken by Sorbs in Germany in the historical province of Upper Lusatia, which is today part of Saxony. It is a minority language with about 13,000 speakers (Ethnologue). The Upper Sorbian dataset contains 310 word forms and 400 lemmas. The data source is the corpus compiled by the Sorbian Institute and The Witaj

Sprachzentrum in Germany, that was used as a training model for an unsupervised MT task (Fraser, 2020). All conjugated parts of speech existing in the language are presented in the dataset. Adjectives, when plural or dual, are marked with case only, otherwise have gender marking, according to Upper-Sorbian grammar.

West Slavic: Czech, Polish, Slovak Data for three West Slavic languages has been added or updated from sources outside Wiktionary. These are: Polish, Czech and Slovak. All three are closely related and are highly inflectional. The Polish data comes from the *Grammatical Dictionary of Polish* (Woliński et al., 2020; Woliński and Kieraś, 2016), an extensive database consisting of inflectional paradigms for Polish lexemes. It serves both as a standalone electronic dictionary as well as a source data for morphological analysers and other applications. The dictionary allows for exporting its data in various schemes so it was possible to prepare a separate exporting path directly for the UniMorph annotation scheme. In the final data all proper names were omitted. The dataset consists of 13,882,543 wordforms of 274,550 lexemes.

The Czech and Slovak data were obtained from the LINDAT/CLARIAH repository (Hajič et al., 2020), (Hajič and Hric, 2017). Both datasets were intended for the use in morphological analysers and their grammatical information is represented in the native Czech National Corpus tagset. The datasets were converted automatically to the UniMorph scheme. Proper names as well as some archaic and non-standard wordforms were omitted. Additionally to limit the size of both data collections negated forms of nouns and adjectives which are perfectly regular were also omitted. The final Czech dataset consists of 50,284,287 wordforms of 824,074 lexemes and the Slovak one contains 28,428,612 wordforms of 366,183 lexemes.

East South Slavic: Pomak Pomak (endonym: Pomácko, Pomáhcku or other dialectic variants) is a non-standardised East South Slavic (ESS) language variety mainly spoken in the region of Greek Thrace, as well as in places of Pomak diaspora. Pomak is included in the map of the European Languages Equality Network.[7] In comparison to all ESS languages, Pomak exhibits a more profound phonological, morphological, morphosyntactic and lexical influence by Medieval and Modern Greek and, due to the predominantly Muslim religion of its speakers, a more profound lexical and phonotactical influence by Ottoman and Modern Turkish. The Pomak data were collected by linguist and native Pomak speaker Ritván Karahóǧa, under the "PHILOTIS: State-of-the-art technologies for the recording, analysis and documentation of living languages" project (MIS 5047429), which is implemented under the "Action for the Support of Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund). The final dataset includes 233,533 lemmas and a total of 6,557,759 wordforms covering adjectives, nouns, and verbs.

Uralic

In 2019–2020 generation algorithms of nominal and verbal wordform were developed for the Veps language, Livvi Karelian and Karelian Proper.[8] Due to this implementation, 2.1 million word forms were generated in the VepKar corpus in the semi-automatic mode during the last two years.

Data for Uralic languages (Karelian, Ludic, Livvi and Veps) were exported from the VepKar corpus (Boyko et al., 2021). The VepKar dataset consists of more than 2,4 million wordforms of approximately 64 thousand lemmas.

Austronesian

Austronesian languages are widely spoken throughout Taiwan, Greater Central Philippines, Madagascar, Islands of Southeast Asia, and Pacific Islands. Derivational and inflectional morphology of languages in this family rely on prefixation and suffixation; some infixation and circumfixation are also attested, as found in Tagalog and Indonesian respectively (Levin and Polinsky, 2019). In this language family, reduplication is also common (Ghanggo Ate, 2021b). In Indonesian, a morphologically rich language, prefixation, suffixation, and circumfixation function in both verb-forming and noun-forming processes. In addition, in the verbal system, main morphological exponents mark voice distinctions as well as active and passive or causatives and applicatives. For some languages whose affixes are moderate in number, clitics are pervasive and morphological exponents mark voice distinction may be lost. Kodhi/Kodi, a language of the Sumba-Hawu subgroup, is the prime example. In this language, pronouns, emphatic, perfective aspect, politeness are expressed by attaching clitics to nouns, verbs, and adjectives. In terms of pronominal clitics, they co-occur with free pronouns marking TERM relations (subjects and objects) and possession, and function like a system of agreement. Kodhi/Kodi also shows loss of Austronesian voice morphology which is typically found in Indonesian-type languages (Arka, 2002).

Iroquoian

As a member of the Iroquoian (Hodinöhšöni) language family, the Seneca language is an indigenous Native American language that is considered critically endangered. Currently the language is estimated to have fewer than 50 first-language speakers left and most of them are elders. The language is spoken mainly in three reservations located in Western New York: Allegany, Cattaraugus, and Tonawanda. Seneca has high (inflectional)

[7] https://elen.ngo/languages-map/

[8] See formalized morphological inflectional rules in Veps and Karelian: https://figshare.com/projects/VepKar/100664

morphological complexity, containing agglutinative as well as fusional properties.

Arawak and Pano-Takana

We include three languages from the Amazon region:

Asháninka is an Arawak language spoken along the rivers Tambo, Ene, Apurímac, Urubamba y Bajo Perené in Central Peruvian Amazon. It belongs to the Asháninka-Ashéninka dialect complex, which comprises more than 70,000 speakers in Central and Eastern Peru and in the state of Acre in Eastern Brazil (Pedrós, 2018). Asháninka belongs to the Nihagantsi subgroup, previously known as Campa in the literature. Asháninka is an agglutinating, polysynthetic, verb-initial language. Since it is a strongly head-marking language, the verb is the most morphologically complex word class, with a rich repertoire of aspectual and modal categories. The language lacks case marking, except for one locative suffix; grammatical relations of subject and object are indexed as affixes on the verb itself. The corpus consists of inflected nouns and verbs from the variety spoken in the Tambo river of Central Peru. The annotated nouns take possessor prefixes, locative case and/or plural marking, while the annotated verbs take subject prefixes, reality status (realis/irrealis), and/or perfective aspect.

Yanesha' is an Arawak language from the Pre-Andine branch. It is spoken in Central Peru by between 3,000 - 5,000 people. Yanesha' is an agglutinating, polysynthetic language with a VSO constituent order. Nouns and verbs are the two major parts of speech. The existence of an independent class of adjectives is questionable due to the absence of clear non-derived forms. Yanesha' is strongly head-marking and therefore the verb class is the most morphologically complex lexical class and the only obligatory constituent of a clause. (Dixon and Aikhenvald, 1999). The corpus consists of inflected nouns and verbs from both dialectal varieties. The annotated nouns take possessor prefixes, plural marking, and locative case, while the annotated verbs take subject prefixes.

Shipibo-Konibo is a Panoan language spoken by around 35,000 native speakers in the Amazon region of Peru. Its morphology is mainly agglutinating, synthetic and almost exclusively suffixing (with only a closed set of prefix related to body-part concepts) Word order is pragmatically determined, but there is some tendency towards SOV constructions. Verbs lack subject and object markers, but exhibit a relatively complex set of TAME markers. As with other Panoan language, verbs is Shipibo-Konibo are strictly transitive or intransitive, with almost no cases of labile verbs in the language. Other relevant grammatical categories for Shipibo-Konibo are participant agreement, switch reference and evidentiality. Data for Shipibo-Konibo were extracted mainly from an old dictionary (James et al., 1993) and a grammar (Valenzuela, 2003).

Koreanic

Korean is an East Asian isolate language spoken by about 80 million people. The dataset was compiled using Wiktionary inflection tables. The resulting data is 2,686 lemmas and 241,323 word forms. It consists of mostly predicates, so the resulting lemmas are mainly verbs and a smaller number of adjectives. The scraped annotated paradigms turned out to be quite similar (mainly because the adjective paradigm is a reduced verb paradigm) and do not represent all forms of verbs and adjectives. It is important to note that different types of converbs were tagged consistently.

Yeniseian

Ket is the only surviving language of the Yeniseian family with about 60 speakers of all levels of linguistic competence (Minlang). The data source is a text collection compiled during the field work of the Laboratory for Computational Lexicography of the Moscow State University, that took place between 2004 and 2009. The Ket dataset contains the word forms of 12 categories, 7 of them (ADJ, NUM, ADV, INTJ, ADP, PART, CONJ) are invariable. The complexity of the Ket verb consists in polypersonal conjugation. The case and number of all arguments object and subject are reflected in the verb.