SIGMORPHON 2021 Shared Task on Morphological Reinflection: Generalization Across Languages

Tiago Pimentel[ʕ*] **Maria Ryskina**[ɨ*] **Sabrina Mielke**[ʒ] **Shijie Wu**[ʒ] **Eleanor Chodroff**[ʏ] **Brian Leonard**[β] **Garrett Nicolai**[ɸ] **Yustinus Ghanggo Ate**[ɜ] **Salam Khalifa**[ħ] **Nizar Habash**[ħ] **Charbel El-Khaissi**[ʃ] **Omer Goldman**[ʎ] **Michael Gasser**[ɪ] **William Lane**[ɾ] **Matt Coler**[ɢ] **Arturo Oncevay**[ɛ] **Jaime Rafael Montoya Samame**[ɦi] **Gema Celeste Silva Villegas**[ɦi] **Adam Ek**[ʛ] **Jean-Philippe Bernardy**[ʛ] **Andrey Shcherbakov**[ɔ] **Aziyana Bayyr-ool**[ʑ] **Karina Sheifer**[ɛ,ɮ,œ] **Sofya Ganieva**[ɱ,ɮ] **Matvey Plugaryov**[ɱ,ɮ] **Elena Klyachko**[ɛ,ɮ] **Ali Salehi**[ɷ] **Andrew Krizhanovsky**[ʁ] **Natalia Krizhanovsky**[ʁ] **Clara Vania**[ɐ] **Sardana Ivanova**[ɨ] **Aelita Salchak**[ʂ] **Christopher Straughn**[ɲ] **Zoey Liu**[ɬ] **Jonathan North Washington**[ɸ] **Duygu Ataman**[æ] **Witold Kieraś**[θ] **Marcin Woliński**[θ] **Totok Suhardijanto**[ɒ] **Niklas Stoehr**[ð] **Zahroh Nuriah**[ɒ] **Shyam Ratan**[ʊ] **Francis M. Tyers**[ɪ,ɛ] **Edoardo M. Ponti**[ø] **Grant Aiton**[ʃ] **Richard J. Hatcher**[ɷ] **Emily Prud'hommeaux**[ɬ] **Ritesh Kumar**[ʊ] **Mans Hulden**[χ] **Botond Barta**[ɑ] **Dorina Lakatos**[ɑ] **Gábor Szolnok**[ɑ] **Judit Ács**[ɑ] **Mohit Raj**[ʊ] **David Yarowsky**[ʒ] **Ryan Cotterell**[ð] **Ben Ambridge**[ʎ,ɕ] **Ekaterina Vylomova**[ɔ]

[ʕ]University of Cambridge [ɨ]Carnegie Mellon University [ʒ]Johns Hopkins University
[ʏ]University of York [β]Brian Leonard Consulting [ɸ]University of British Columbia
[ɜ]STKIP Weetebula [ħ]New York University Abu Dhabi [ʃ]Australian National University
[ʎ]Bar-Ilan University [ɾ]Charles Darwin University [ɢ]University of Groningen
[ɪ]Indiana University [ɛ]University of Edinburgh [ɦi]Pontificia Universidad Católica del Perú
[ʛ]University of Gothenburg [ɔ]University of Melbourne [ɛ]Higher School of Economics
[ʑ]Institute of Philology of the Siberian Branch of the Russian Academy of Sciences
[ɮ]Institute of Linguistics, Russian Academy of Sciences [ɱ]Moscow State University
[œ]Institute for System Programming, Russian Academy of Sciences
[ɷ]University at Buffalo [ʁ]Karelian Research Centre of the Russian Academy of Sciences
[ɕ]ESRC International Centre for Language and Communicative Development (LuCiD)
[ɲ]Northeastern Illinois University [ɨ]University of Helsinki [ʂ]Tuvan State University
[ɐ]New York University [ɬ]Boston College [ɸ]Swarthmore College [æ]University of Zürich
[θ]Institute of Computer Science, Polish Academy of Sciences [ɒ]Universitas Indonesia
[ʊ]Dr. Bhimrao Ambedkar University [ø]Mila/McGill University Montreal
[ð]ETH Zürich [χ]University of Colorado Boulder [ʎ]University of Liverpool
[ɑ]Budapest University of Technology and Economics
tp472@cam.ac.uk mryskina@cs.cmu.edu vylomovae@unimelb.edu.au

Abstract

This year's iteration of the SIGMORPHON Shared Task on morphological reinflection focuses on typological diversity and cross-lingual variation of morphosyntactic features. In terms of the task, we enrich UniMorph with new data for 32 languages from 13 language families, with most of them being under-resourced: Kunwinjku, Classical Syriac, Arabic (Modern Standard, Egyptian, Gulf), Hebrew, Amharic, Aymara, Magahi, Braj, Kurdish (Central, Northern, Southern), Polish, Karelian, Livvi, Ludic, Veps, Võro, Evenki, Xibe, Tuvan, Sakha, Turkish, Indonesian, Kodi, Seneca, Asháninka, Yanesha, Chukchi, Itelmen, Eibela. We evaluate six systems on the new data and conduct an extensive error analysis of the systems' predictions. Transformer-based models generally demonstrate superior performance on the majority of languages, achieving >90% accuracy on 65% of them. The languages on which systems yielded low accuracy are mainly under-resourced, with a limited amount of data. Most errors made by the systems are due to allomorphy, honorificity, and form variation. In addition, we observe that systems especially struggle to inflect multiword lemmas. The systems also produce misspelled forms or end up in repetitive loops (e.g., RNN-based models). Finally, we report a large drop in systems' performance on previously unseen lemmas.[1]

[*]The authors contributed equally

[1]The data, systems, and their predictions are available: https://github.com/sigmorphon/2021Task0

1 Introduction

Chomsky (1995) noted that if a Martian anthropologist were to visit our planet, all of our world's languages would appear as a dialect of a single language, more specifically instances of what he calls a "universal grammar". This idea—that all languages have a large inventory of shared sounds, vocabulary, syntactic structures with minor variations—was especially common among cognitive scientists. It was based on highly biased ethnocentric empirical observations, resulting from the fact that a vast majority of cognitive scientists, including linguists, focused only on the familiar European languages. Moreover, as Daniel (2011) notes, many linguistic descriptive traditions of individual languages, even isolated ones such as Russian or German, heavily rely on cross-linguistic assumptions about the structure of human language that are often projected from Latin grammars. Similarly, despite making universalistic claims, generative linguists, for a very long time, have focused on a small number of the world's major languages, typically using English as their departure point. This could be partly attributed to the fact that generative grammar follows a deductive approach where the observed data is conditioned on a general model.

However, as linguists explored more languages, descriptions and comparisons of more diverse kinds of languages began to come up, both within the framework of generative syntax as well as that of linguistic typology. Greenberg (1963) presents one of the earliest typologically informed description of "language universals" based on an analysis of a relatively larger set of 30 languages, which included a substantial proportion of data from non-European languages. Subsequently, typologists have claimed that it is essential to describe the limits of cross-linguistic variation (Croft, 2002; Comrie, 1989) rather than focus only on cross-linguistic similarities. This is especially evident from Evans and Levinson (2009), where the authors question the notion of "language universals", i.e. the existence of a common pattern, or basis, shared across human languages. By looking at cross-linguistic work done by typologists and descriptive linguists, they demonstrate that "diversity can be found at almost every level of linguistic organization": languages vary greatly on phonological, morphological, semantic, and syntactic levels. This leads us to p-linguistics (Haspelmath, 2020), a study of particular languages, including the whole variety of idiosyncratic properties present in them, which makes cross-linguistic comparison challenging.

Haspelmath (2010) suggested a distinction between descriptive categories (specific to languages) and comparative concepts. The idea was then refined and further developed with respect to morphology and realized in the UniMorph schema (Sylak-Glassman et al., 2015b). Morphosyntactic features (such as "the dative case" or "the past tense") in the UniMorph occupy an intermediate position between the descriptive categories and comparative concepts. The set of features was initially established on the basis of analysis of typological literature, and refined with the addition of new languages to the UniMorph database (Kirov et al., 2018; McCarthy et al., 2020). Since 2016, SIGMORPHON organized shared tasks on morphological reinflection (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020) that aimed at evaluating contemporary systems. Parallel to that, they also served as a platform for enriching the UniMorph database with new languages. For instance, the 2020 shared task (Vylomova et al., 2020) featured 90 typologically diverse languages derived from various linguistic resources.

This year, we are bringing many under-resourced languages (languages of Peru, Russia, India, Australia, Papua New Guinea) and dialects (e.g., for Arabic and Kurdish). The sample is highly diverse: it contains languages with templatic, concatenative (fusional and agglutinative) morphology. In addition, we bring more polysynthetic languages such as Kunwinjku, Chukchi, Asháninka. Unlike previous years, we pay more attention to the conversion of the morphosynactic features of these languages into the UniMorph schema. In addition, for most languages we conduct an extensive error analysis.

2 Task Description

In this shared task, the participants were told to design a model that learns to generate morphological inflections from both a lemma and a set of morphosyntactic features of the target form. Specifically, each language in the task had its own training, development, and test splits. The training and development splits contained triples, with a lemma, a set of morphological features, and the target inflected form, while test splits only provided lemmas and morphological tags: the participants' models needed to predict the missing target form—making this a standard supervised learning task.

The target of the task, however, was to analyse how well the current state-of-the-art reinflection models could generalise across a typologically diverse set of languages. These models should, in theory, be general enough to work for natural languages of any typological patterning.[2] As such, we designed the task in three phases: a Development Phase, a Generalization Phase, and an Evaluation Phase. As the phases advanced, more data and more languages were released.

In the **Development Phase**, we provided training and development splits that should be used by participants to develop their systems. Model development, evaluation, and hyper-parameter tuning were, thus, mainly performed on these sets of languages. We will refer to these as the development languages.

In the **Generalization Phase**, we provided training and development splits for new languages where approximately half were genetically related (belonged to the same family) and half were genetically unrelated (either isolates or belonging to different families) to the development languages. These languages (and their families) were kept as a surprise throughout the first (development) phase and were only announced later on. As the participants were only given a few days with access to these languages before the submission deadline, we expected that the systems couldn't be radically improved to work on them—as such, these languages allowed us to evaluate the generalization capacity of the re-inflection models, and how well they performed on new typologically unrelated languages.

Finally, in the **Evaluation Phase**, the participants' models were evaluated on held-out test forms from all of the languages of the previous phases. The languages from the Development Phase and the Generalization Phase were evaluated simultaneously. The only difference between the development and generalization languages was that participants had more time to construct their models for the languages released in the Development Phase. It follows that a model could easily favor or overfit to the phenomena that are more frequent in the languages presented in the Development Phase, especially if the parameters were shared across languages. For instance, a model based on the morphological patterning of Indo-European languages may end up with a bias towards suffixing and would struggle to learn prefixing or circumfixation, and the degree of the bias only becomes apparent during experimentation on other languages whose inflectional morphology patterns differ. Further, the model architecture itself could also explicitly or implicitly favor certain word formation types (suffixing, prefixing, etc.).

[2]For example, Tagalog verbs exhibit circumfixation; thus, a model with a strong inductive bias towards suffixing would likely not work well for Tagalog.

3 Description of the Languages

3.1 Gunwinyguan

The Gunwinyguan language family consists of Australian Aboriginal languages spoken in the Arnhem Land region of Australia's Northern Territory.

3.1.1 Gunwinggic: Kunwinjku

This data set contains one member of this family: a dialect of Bininj Kunwok called **Kunwinjku**. Kunwinjku is a polysynthetic language with mostly agglutinating verbal morphology. A typical verb there might look like *Aban-yawoith-warrgah-marne-ganj-ginje-ng* '1/3PL-again-wrong-BEN-meat-cook-PP' ("I cooked the wrong meat for them again"). As shown, the form has several prefixes and suffixes attached to the stem. As in other Australian languages, long vowels are typically represented by double characters, and trills with "*rr*".[3] According to Evans' (2003) analysis, the verb template contains 12 affix slots which include two incorporated noun classes, and derivational affixes such as the benefactive and comitative. The data included in this set are verbs extracted from the Kunwinjku translation of the Bible using the morphological analyzer from Lane and Bird (2019) and manually verified by human annotators.

3.2 Afro-Asiatic

The Afro-Asiatic language family is represented by the Semitic subgroup.

3.2.1 Semitic: Classical 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

[3]More details: https://en.wikipedia.org/wiki/Transcription_of_Australian_Aboriginal_languages.

Family	Genus	ISO 639-3	Language	Source of Data	Annotators
			Development		
Afro-Asiatic	Semitic	afb	Gulf Arabic	Khalifa et al. (2018)	Salam Khalifa, Nizar Habash
	Semitic	amh	Amharic	Gasser (2011)	Michael Gasser
	Semitic	ara	Modern Standard Arabic	Taji et al. (2018)	Salam Khalifa, Nizar Habash
	Semitic	arz	Egyptian Arabic	Habash et al. (2012)	Salam Khalifa, Nizar Habash
	Semitic	heb	Hebrew (Vocalized)	Wiktionary	Omer Goldman
	Semitic	syc	Classic Syriac	SEDRA	Charbel El-Khaissi
Arawakan	Southern Arawakan	ame	Yanesha	Duff-Trip (1998)	Arturo Oncevay, Gema Celeste Silva Villegas
	Southern Arawakan	cni	Asháninka	Zumaeta Rojas and Zerdin (2018); Kindberg (1980)	Arturo Oncevay, Jaime Rafael Montoya Samame
Austronesian	Malayo-Polynesian	ind	Indonesian	KBBI, Wikipedia	Clara Vania, Totok Suhardijanto, Zahroh Nuriah
	Malayo-Polynesian	kod	Kodi	Ghanggo Ate (2021)	Yustinus Ghanggo Ate, Garrett Nicolai
Aymaran	Aymaran	aym	Aymara	Coler (2014)	Matt Coler, Eleanor Chodroff
Chukotko-Kamchatkan	Northern Chukotko-Kamchatkan	ckt	Chukchi	Chuklang; Tyers and Mishchenkova (2020)	Karina Sheifer, Maria Ryskina
	Southern Chukotko-Kamchatkan	itl	Itelmen		Karina Sheifer, Sofya Ganieva, Matvey Plugaryov
Gunwinyguan	Gunwinggic	gup	Kunwinjku	Lane and Bird (2019)	William Lane
Indo-European	Indic	bra	Braj	Raw data from Kumar et al. (2018)	Shyam Ratan, Ritesh Kumar
	Slavic	bul	Bulgarian	UniMorph (Kirov et al., 2018, Wiktionary)	Christo Kirov
	Slavic	ces	Czech	UniMorph (Kirov et al., 2018, Wiktionary)	
	Iranian	ckb	Central Kurdish (Sorani)	Alexina project	Ali Salehi
	Germanic	deu	German	UniMorph (Kirov et al., 2018, Wiktionary)	
	Iranian	kmr	Northern Kurdish (Kurmanji)	Alexina project	
	Indic	mag	Magahi	Raw data from (Kumar et al., 2014, 2018)	Mohit Raj, Ritesh Kumar
	Germanic	nld	Dutch	UniMorph (Kirov et al., 2018, Wiktionary)	
	Slavic	pol	Polish	Woliński et al. (2020); Woliński and Kieraś (2016)	Witold Kieraś, Marcin Woliński
	Romance	por	Portuguese	UniMorph (Kirov et al., 2018, Wiktionary)	
	Slavic	rus	Russian	UniMorph (Kirov et al., 2018, Wiktionary)	Ekaterina Vylomova
	Romance	spa	Spanish	UniMorph (Kirov et al., 2018, Wiktionary)	
	Iranian	sdh	Southern Kurdish	Fattah (2000, native speakers)	Ali Salehi
Iroquoian	Northern Iroquoian	see	Seneca	Bardeau (2007)	Richard J. Hatcher, Emily Prud'hommeaux, Zoey Liu
Trans–New Guinea	Bosavi	ail	Eibela	Aiton (2016b)	Grant Aiton, Edoardo Maria Ponti, Ekaterina Vylomova
Tungusic	Tungusic	evn	Evenki	Kazakevich and Klyachko (2013)	Elena Klyachko
Turkic	Turkic	sah	Sakha	Forcada et al. (2011, Apertium: apertium-sah)	Francis M. Tyers, Jonathan North Washington, Sardana Ivanova, Christopher Straughn, Maria Ryskina
	Turkic	tyv	Tuvan	Forcada et al. (2011, Apertium: apertium-tyv)	Francis M. Tyers, Jonathan North Washington, Aziyana Bayyr-ool, Aelita Salchak, Maria Ryskina
Uralic	Finnic	krl	Karelian	Zaytseva et al. (2017, VepKar)	Andrew and Natalia Krizhanovsky
	Finnic	lud	Ludic	Zaytseva et al. (2017, VepKar)	Andrew and Natalia Krizhanovsky
	Finnic	olo	Livvi	Zaytseva et al. (2017, VepKar)	Andrew and Natalia Krizhanovsky
	Finnic	vep	Veps	Zaytseva et al. (2017, VepKar)	Andrew and Natalia Krizhanovsky
			Generalization (Surprise)		
Tungusic	Tungusic	sjo	Xibe	Zhou et al. (2020)	Elena Klyachko
Turkic	Turkic	tur	Turkish	UniMorph (Kirov et al., 2018, Wiktionary)	Omer Goldman and Duygu Ataman
Uralic	Finnic	vro	Võro	Wiktionary	Ekaterina Vylomova

Table 1: Development and surprise languages used in the shared task.

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.

3.2.2 Semitic: Arabic

Modern Standard Arabic (MSA, `ara`) is the primarily written form of Arabic which 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 state of diglossia (Ferguson, 1959) whether in spoken or written form. Arabic dialects vary amongst 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 refers to the dialects spoken by the indigenous populations of the members of the Gulf Cooperation Council, especially regions on the Arabian Gulf.

Similar to other Semitic languages, Arabic is a templatic language. A word consists of a templatic stem (root and pattern) and a number of affixes and clitics. Verb lemmas in Arabic inflect for person, gender, number, voice, mood, and aspect. Nominal lemmas inflect for gender, number, case, and state. Those features are realized through both the templatic patterns and the concatenative affixations. Arabic words also take on a number of clitics: attachable prepositions, conjunctions, determiners, and pronominal objects and possessives. In this work, we do not include clitics as a part of the paradigms, as they heavily increase the size of the paradigms. We made the exception to add the *Al* determiner particle in order to be consistent with commonly used tokenizations for Arabic treebanks—Penn Arabic Treebank (Maamouri et al., 2004) and Arabic Universal Dependencies (Taji et al., 2017). For MSA, the paradigms inflect for all the above-mentioned features, while for EGY and GLF they inflect for the above-mentioned features except for voice, mood, case, and state. We use the functional (grammatical) gender and number for MSA and GLF, but the form-based gender and number for EGY, since the resources we used did not have EGY functional gender and number (Alkuhlani and Habash, 2011).

We generated all the inflection tables from the morphological analysis databases using the generation component provided by CamelTools (Obeid et al., 2020). We extracted all the verb, noun, and adjective lemmas from a number of annotated corpora and selected those that are already in the morphological analysis databases. For MSA, we used the CALIMA-STAR database (Taji et al., 2018), based on the SAMA database (Maamouri et al., 2010), and the PATB (Maamouri et al., 2004) as the sources of lemmas. For EGY, we used the CALIMA-EGY database (Habash et al., 2012) and the ARZTB (Maamouri et al., 2012) as the sources of lemmas. For GLF, we used the Gulf verb analyzer (Khalifa et al., 2017) for verbs, and for both nouns and adjectives we extracted all the annotations from the Annotated Gumar Corpus (Khalifa et al., 2018).

3.2.3 Semitic: Hebrew

As Syriac, 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 work we present data in vocalized Hebrew, in order to examine the models' ability to handle Hebrew's full-fledged Semitic morphological system.

Hebrew verbs belong to 7 major classes (*Binyanim*) with many subclasses depending on the phonological features of the root's consonants. Verbs inflect for number, gender, and tense-mood, while the nominal inflection tables include definiteness and possessor.

The provided inflection tables are largely identical to those of the past years' shared tasks, scraped from Wiktionary, with the addition of the verbal nouns and all forms being automatically vocalized.

3.2.4 Semitic: Amharic

Amharic is the most spoken and best-resourced among the roughly 15 languages in the Ethio-Semitic branch of South Semitic. Unlike most other Semitic languages, but like other Ethio-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, both by templates and affixes, but these are treated as separate lemmas in the data. Other verb affixes (or clitics, depending on the analysis) indicate object person, gender, and number; negation; relativization; conjunctions; and, on relativized forms, prepositions and definiteness. None of these are included in the data.

Nouns and adjectives share most of their morphology and are often not clearly distinguished. Nouns and adjectives inflect for definiteness, number, and possession. Gender is only explicit when the masculine or feminine singular definite suffixes are present; most nouns have no inherent gender. Nouns and adjectives also have prepositional prefixes (or clitics) and accusative suffixes, which are not included in the data.

The data for the shared task were generated by the HornMorpho generator (Gasser, 2011), an FST weighted with feature structures. Common orthographic variants of the lemmas and common variant plural forms of nouns are included. In these cases, the variants are distinguished with the LGSPEC1 and LGSPEC2 features. Predictable orthographic variants are not included.

3.3 Aymaran

The Aymaran family has two branches: Southern Aymaran (which is the branch described in this contribution, as represented by **Mulyaq' Aymara**) and Central Aymaran (Jaqaru).[4] Aymaran has no external relatives. The neighboring and overlapping Quechuan family is often erroneously believed to be related.

[4]Sometimes Cauqui (also spelled "Kawki"), a language spoken by less than ten elders in Cachuy, Canchán, Caipán, and Chavín, is considered to be a third Aymaran language but it may be more accurate to consider it a Jaqaru dialect.

3.3.1 Aymaran: Aymara

Aymara is spoken mainly in Andean communities in the region encompassing Bolivia and Peru from the north of Lake Titicaca to the south of Lake Poopó, extending westward to the valleys of the Pacific coast and eastward to the Yunga valleys. It has roughly two million speakers, over half of whom are Bolivian. The rest reside mainly in Peru, with small communities in Chile and Argentina. Aymara is a highly agglutinative, suffix-only language. Nouns are inflected for grammatical number, case, and possessiveness. As Coler (2010) notes, Aymara has 11–12 grammatical cases, depending on the variety (as in some varieties the locative and genitive suffixes have merged and in others they have not). The case suffix is attached to the last element of a noun phrase. Verbs present relatively complex paradigms, with dimensions such as grammatical person (marking both subject and direct object), number, tense (simple, future, recent past, distal past), mood (evidentials, two counterfactual paradigms, and an imperative paradigm). Moreover, Aymara has a variety of suffixes which change the grammatical category of the word. Words can change grammatical category multiple times.[5]

3.4 Indo-European

The Indo-European language family is the parent family of most of the European and Asian languages. In this iteration of the shared task, we enrich the data with languages from Indo-Aryan, Iranian, and Slavic groups. Iranian and Indo-Aryan are recognised as distinct subgroups of Indo-European. Characteristic retentions and innovations make Iranian and Indo-Aryan language families diverged and distinct from each other (Jain and Cardona, 2007).

3.4.1 Indo-Aryan, or Indic: Magahi, Braj

The Indian subcontinent is the heartland of where the Indo-Aryan languages are spoken. This area is also referred to as South Asia and encompasses India, Pakistan, Bangladesh, Nepal, Bhutan and the islands of Sri Lanka and Maldives (Jain and Cardona, 2007). Magahi and Braj, which belong to the Indo-Aryan language family, are under our observation.

Magahi comes under the Magadhi group of the middle Indo-Aryan which includes Bhojpuri and

[5]Tags' conversion into UniMorph: https://github.com/unimorph/aym/blob/main/Muylaq'AymaraUnimorphConversion.tsv

Maithili. While the exact classification within this subgroup is still debatable, most accepted analyses put it under one branch of the Eastern group of languages which includes Bangla, Asamiya, and Oriya (Grierson and Konow, 1903). Magahi speech area is mainly concentrated in the Eastern Indian states of Bihar and Jharkhand, but it also extends to the adjoining regions of Bengal and Odisha (Grierson, 1903).

There is no grammatical gender and number agreement in Magahi, though sex-related gender derivation commonly occurs for animate nouns like */laika/* (boy) and */laiki/* (girl). Number is also marked on nouns, and it affects the form of case markers and postpositions in certain instances (Lahiri, 2021). Moreover, it has a rich system of verbal morphology to show the tense, aspect, person, and honorific agreement with the subject as well as the addressee.

In the present dataset, the inflectional paradigms for verbs show the honorificity level of both the subjects and the addressees, and also the person of the subject, the tense and aspect markers. The inflectional paradigms for nouns and adjectives are generated on the basis of the inflectional marker used for expressing case, familiarity, plurality, and (sometimes) gender within animate nouns. Pronouns are marked for different cases and honorificity levels. These paradigms are generated on the basis of a manually annotated corpus of Magahi folktales.

We used a raw dataset from the literary domain. First, we annotated the dataset with the Universal Dependency morphological feature tags at token level using the CoNLL-U editor (Heinecke, 2019). We then converted the annotated dataset into the UniMorph schema using the script available for converting UD data into the UniMorph tagset (McCarthy et al., 2018). To finalize the data, we manually validated the dataset against the UniMorph schema (Sylak-Glassman et al., 2015a).

Brajbhasha, or Braj is one of the Indo-Aryan languages spoken in the Western Indian states of Uttar Pradesh, Madhya Pradesh, and Rajasthan. Grierson (1908) groups Brajbhasha under Western Hindi of the Central Group in the Indo-Aryan family, along with other languages like Hindustani, Bangaru, Kannauji, and Bundeli. Braj is not generally used in education or for any official purposes in any Braj spoken state, but it has a very rich literary tradition. Also in order to preserve, promote, publish and popularise the literary tradition of Braj, the local state government of Rajasthan has set up the Braj Bhasha Akademi (Braj Bhasha Academy) in Jaipur. Along with this, some individuals, local literary and cultural groups, and language enthusiasts at the local level also bring out publications in Braj (Kumar et al., 2018). In all of the above sources, bhakti poetry[6] constitutes a large proportion of the traditional literature of Braj (Pankaj, 2020).

As in the case of other Indo-Aryan languages, Braj is also rich in morphological inflections. The dataset released for the present task contains two sets of inflectional paradigms with morphological features for nouns and verbs. Nominal lemmas in Braj are inflected for gender (masculine and feminine) and number (singular and plural); verb lemmas take gender (masculine and feminine), number (singular and plural), person (first, second and third), politeness/honorificity (formal and informal), tense (present, past and future), and aspect (perfective, progressive, habitual and prospective) markings. Among these, the politeness feature is marked for showing honorificity and formality. More generally, a formal/polite marker is used for strangers and the elite class, while informal/neutral markers are used for family and friends.

In order to generate the morphological paradigms, we have used the data from the literary domain, annotated at the token level in the CoNLL-U editor (Heinecke, 2019). The dataset was initially annotated using the Universal Dependencies morphological feature set and then automatically converted to the UniMorph schema using the script provided by McCarthy et al. (2018). Finally, the converted dataset was manually validated and edited to conform to the constraints and conventions of UniMorph to arrive at the final labels.

3.4.2 Iranian: Kurdish

The Iranian branch is represented by **Kurdish**. Among Western Iranian languages, Kurdish is the term covering the largest group of related dialects. Kurdish comprises three main subgroup dialects, namely Northern Kurdish (including Kurmanji), Central Kurdish (including Sorani), and Southern Kurdish. Sorani Kurdish, spoken in Iran and Iraq, is known for its morphological split ergative system. There are two sets of morphemes traditionally described as agreement markers: clitic markers and

[6]This is dedicated to Indian spiritual and mythological imagination as being associated with Lord Krishna.

verbal affixes, which are verbal agreement markers, or the copula. The distribution of these formatives can be described as ergative alignment, although mandatory agent indexing has led some scholars to refer to the Sorani system as post- or remnant-ergative (Jügel, 2009). Note that Sorani nominals do not feature case marking. The single argument of an intransitive verb is an affix while the transitive verbs have a tense-sensitive alignment. With transitive verbs, agents are indexed by affixes in the present tense and with clitics in the past tense. On the other hand, the object is indexed with a clitic in the present tense and an affix in the past tense. In addition, Sorani also has the so-called experiencer-subject verbs, with which both the agent and the object are marked with clitic markers. Like other Iranian languages, Sorani also features a series of light-verb constructions which are composed using the verbs *kirdin* 'to do' or *bun* 'to be'. In the light verb constructions, the agent is marked with an affix in the present tense, while a clitic marks the subject in the past tense. Southern Kurdish features all the same verbs types, clitics and affixes, while the alignment pattern can be completely different due to a nominative-accusative alignment system. The usage of agreement markers with affixes is widely predominant in Southern Kurdish and clitics can be used to mark the possessives.

Both dialects of Kurdish allow for clitic and affix stacking marking the agent and the object of a verb. In Sorani, for instance, *dit=yan-im* 'They saw me' uses a clitic and an affix to mark the agent and the object, and *wist=yan=im* 'I wanted them' marks both the agent and the object with clitics. Ditransitive verbs can be formed by a transitive verb and an applicative marker. For instance, a ditransitive three-participant verb *da-m-în=î-yê* 'He gave them to me' marks the recipient and the object with affixes, and the agent is marked with a clitic in the presence of an applicative (*yê*). A separate set of morphological features is needed to account for such structures, in which the verb dictates the person marker index as subject, agent, object or recipient.

3.4.3 Slavic: Polish

The Slavic genus comprises a group of fusional languages evolved from Proto-Slavic and spoken in Central and Eastern Europe, the Balkans and the Asian parts of Russia from Siberia to the Far East. Slavic languages are most commonly divided into three major subgroups: East, West, and South. All three are represented in this dataset, with Polish and Czech being the typical West Slavic languages, Russian being the most prominent East Slavic language, and Bulgarian representing the Eastern part of the South Slavic group. Slavic languages are characterized by a rich verbal and nominal inflection system. Typically, verbs mark tense, person, gender, aspect, and mood. Nouns mark gender, number, and case, although in Bulgarian and Macedonian cases are reduced to only nominative and vocative. Masculine nouns additionally mark animacy.

Polish data was obtained via a conversion from the largest Polish morphological dictionary (Woliński et al., 2020) which is also used as the main data source in the morphological analysis. Table 10 presents a simplified mapping from the original flexemic tagset of Polish (Przepiórkowski and Woliński, 2003) to the UniMorph schema. The data for the remaining three Slavic languages were obtained from Wiktionary.

3.5 Uralic: Karelian, Livvi, Ludic, Veps, Võro

The Uralic languages are spoken from the north of Siberia in Russia to Scandinavia and Hungary. They are agglutinating with some subgroups displaying fusional characteristics (e.g., the Sámi languages). Many of the languages have vowel harmony. Many of the larger case paradigms are made up of spatial cases, sometimes with distinctions for direction and position. Further, most of the languages have possessive suffixes, which can express possession or agreement in non-finite clauses.

We use **Karelian**, **Ludic**, **Livvi**, **Veps**, and **Võro** in the shared task. All the data except Võro were exported from the Open corpus of Veps and Karelian languages (VepKar). Veps and Karelian are agglutinative languages with rich suffixal morphology. All inflectional categories in these languages are formed by attaching one or more affixes corresponding to different grammatical categories to the stem.

The presence of one or two stems in the nominal parts of speech and verbs is essential when constructing word forms in the Veps and Karelian languages (Novak, 2019, 57). In these languages, to build the inflected forms of nouns and verbs, one needs to identify one or two word stems. There are formalized (algorithmic) ways to determine the stem, although not for all words (Novak et al., 2020,

684).

Note that in the Ludic and Livvi dialects of the Karelian language and in the Veps language, reflexive forms of verbs have their own paradigm. Thus, one set of morphological rules is needed for reflexive verbs and another set for non-reflexive verbs.

Võro represents the South Estonian dialect group. Similar to other Uralic languages, it has agglutinative, primarily suffixal, morphology. Nouns inflect for grammatical case and number. The current shared task sample contains noun paradigm tables derived from Wiktionary.[7]

3.6 Tungusic

The Tungusic genus comprises a group of agglutinative languages spoken from Central and Eastern Siberia to the Far East over the territories of Russia and China. The genus is considered to be a member of the Altaic (or Transeurasian) language family by some researchers, although this is disputed. Tungusic languages are commonly divided into two or three branches (see Oskolskaya et al. (2021) for discussion).

3.7 Tungusic: Evenki and Xibe

The dataset presents two Tungusic languages, namely **Evenki** and **Xibe**, belonging to different branches in any approach, with Xibe being quite aberrant from other Tungusic languages. Tungusic languages are characterized by rich verbal and nominal inflection and demonstrate vowel harmony. Typically verbs mark tense, person, aspect, voice and mood. Nouns mark number, case and possession.

Inflection is achieved through suffixes. Evenki is a typical agglutinative language with almost no fusion whereas Xibe is more fusional.

The Evenki data was obtained by conversion from a corpus of oral Evenki texts (Kazakevich and Klyachko, 2013), which uses IPA. The Xibe data was obtained by conversion from a Universal Dependency treebank compiled by Zhou et al. (2020), which contains textbook and newspaper texts. Xibe texts use the traditional script.

[7]The tag conversion schema for Uralic languages is provided here: https://docs.google.com/spreadsheets/d/1RjO_J22yDB5FH5C24ej7sGGbeFAjcIadJA6ML55tsOI/edit.

3.8 Turkic

3.8.1 Siberian Turkic: Sakha and Tuvan

The Turkic languages of Siberia, spoken mostly within the Russian Federation, range from vulnerable to severely endangered (Eberhard et al., 2021) and represent several branches of Turkic with varying degrees of relatedness (Баскаков, 1969; Tekin, 1990; Schönig, 1999). They have rich agglutinating morphology, like other Turkic languages, and share many grammatical properties (Washington and Tyers, 2019).

In this shared task, the Turkic languages of this area are represented by **Tuvan** (Sayan Turkic) and **Sakha** (Lena Turkic). For both languages, we make use of the lexicons of the morphological transducers built as part of the Apertium open-source project (Khanna et al., to appear in 2021; Washington et al., to appear in 2021). We use the transducers for Tuvan[8] (Tyers et al., 2016; Washington et al., 2016) and Sakha[9] (Ivanova et al., 2019, to appear in 2022) as morphological generators, extracting the paradigms for all the verbs and nouns in the lexicon. We manually design a mapping between the Apertium tagset and the UniMorph schema (Table 8), based on the system descriptions and additional grammar resources (Убрятова et al. (1982) for Sakha and Исхаков and Пальмбах (1961); Anderson and Harrison (1999); Harrison (2000) for Tuvan). Besides the tag mapping, we also include a few conditional rules, such as marking definiteness for nouns in the accusative and genitive cases.

Since the UniMorph schema in its current version is not well-suited to capture the richness of Turkic morphology, we exclude many forms with morphological attributes that do not have a close equivalent in UniMorph. We also omit forms with affixes that are considered quasi-derivational rather than inflectional, such as the desiderative /-ksA/ in Tuvan (Washington et al., 2016), with the exception of the negative marker. These constraints greatly reduce the sizes of the verbal paradigms: the median number of forms per lemma is 234 and 87 for Tuvan and Sakha respectively, compared to roughly 5,700 forms per lemma produced by either generator. Our tag conversion and paradigm filtering code is publicly released.[10]

[8]https://github.com/apertium/apertium-tyv/

[9]https://github.com/apertium/apertium-sah/

[10]https://github.com/ryskina/apertium2unimorph

3.8.2 Turkic: Turkish

One of the further west Turkic languages, Turkish is part of the Oghuz branch, and, like the other languages of this family, it is highly agglutinative.

In this work, we vastly expanded the 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.

3.9 Austronesian

3.9.1 Malayo-Polynesian: Indonesian

Indonesian or *Bahasa Indonesia* is the official language of Indonesia. It belongs to the Austronesian language family and it is written with the Latin script.

Indonesian does not mark grammatical case, gender, or tense. Words are composed from their roots through affixation, compounding, or reduplication. The four types of Indonesian affixes are prefixes, suffixes, circumfixes (combination of prefixes and suffixes), and infixes (inside the base form). Indonesian uses both full and partial reduplication processes to form words. Full reduplication is often used to express the plural forms of nouns, while partial reduplication is typically used to derive forms that might have a different category than their base forms. Unlike English, the distinction between inflectional and derivational morphological processes in Indonesian is not always clear (Pisceldo et al., 2008).

In this shared task, the Indonesian data is created by bootstrapping the data from an Indonesian Wikipedia dump. Using a list of possible Indonesian affixes, we collect unique word forms from Wikipedia and analyze them using MorphInd (Larasati et al., 2011), a morphological analyzer tool for Indonesian based on an FST. We manually create a mapping between the MorphInd tagset and the UniMorph schema. We then use this mapping and apply some additional rule-based formulas created by Indonesian linguists to build the final dataset (Table 9).

3.9.2 Malayo-Polynesian: Kodi/Kodhi

Kodi or Kodhi [kodi] is spoken in Sumba Island, eastern Indonesia (Ghanggo Ate, 2020). Regarding its linguistic classification, Kodi belongs to the Central-Eastern subgroup of Austronesian, related to Sumba-Hawu languages. Based on the linguistic fieldwork observations done by Ghanggo Ate (2020), it may be tentatively concluded that there are only two Kodi dialects: Kodi Bhokolo and Mbangedho-Mbalaghar. Even though some work has been done on Kodi (Ghanggo Ate, to appear in 2021), it remains a largely under-documented language. Further, Kodi is vulnerable or threatened because Indonesian, the prestigious national language, is used in most sociolinguistic domains outside the domestic sphere.

A prominent linguistic feature of Kodi is its clitic system, which is pervasive in various syntactic categories—verbs, nouns, and adjectives—and marks person (1, 2, 3) and number (SG vs. PL). In addition, Kodi contains four sets of pronominal clitics that agree with their antecedent: NOM(inative) proclitics, ACC(usative) enclitics, DAT(ive) enclitics and GEN(initive) enclitics. Interestingly, these clitic sets are not markers of NOM, ACC, DAT, or GEN grammatical case—as in Malayalam or Latin—but rather identify the head for TERM relations (subject and object). Thus, by default, pronominal clitics are core grammatical arguments reflecting subject and object.

For the analyses of the features of Kodi clitics, the data freshly collected in the fieldwork funded by the Endangered Language Fund is annotated. Then, the collected data is converted to the UniMorph task format, which has the lemmas, the word forms, and the morphosyntactic features of Kodi.

3.10 Iroquoian

3.10.1 Northern Iroquoian: Seneca

The Seneca language is an indigenous Native American language from the Iroquoian (Hodinöhšöni) language family. Seneca is considered critically endangered and is currently estimated to have fewer than 50 first-language speakers left, most of whom are elders. The language is spoken mainly in three reservations located in Western New York: Allegany, Cattaraugus, and Tonawanda.

Seneca possesses highly complex morphological features, with a combination of both agglutinative and fusional properties. The data presented here consists of inflectional paradigms for Seneca verbs,

the basic structure of which is composed of a verb base that describes an event or state of action. In virtually all cases, the verb base would be preceded by a pronominal prefix which indicates the agent, the patient, or both for the event or state, and followed by an aspect suffix which usually marks a habitual or a stative state.

(1) ha skatkwë s
it he laugh HAB
He laughs.

In some other scenarios, for instance, when the verb is describing a factual, future or hypothetical event, a modal prefix is attached before the pronominal prefix and the aspect suffix marks a punctual state instead. The structures and orders of the prefixes can be more complicated depending on, e.g., whether the action denoted by the verb is repetitive or negative; these details are realized by adding a prepronominal prefix before the modal prefix.

3.11 Arawakan

3.11.1 Southern Arawakan: Asháninka

Asháninka is an Arawak language with more than 70,000 speakers in Central and Eastern Peru and in the state of Acre in Eastern Brazil, in a geographical region located between the eastern foothills of the Andes and the western fringe of the Amazon basin (Mihas, 2017; Mayor Aparicio and Bodmer, 2009). Although it is the most widely spoken Amazonian language in Peru, certain varieties, such as Alto Perené, are highly endangered.

It is an agglutinating, polysynthetic, verb-initial 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. Other notable linguistic features of the language include a distinction between alienably and inalienably possessed nouns, obligatory marking of reality status (realis/irrealis) on the verb, a rich system of applicative suffixes, serial verb constructions, and pragmatically conditioned split intransitivity.

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.

3.11.2 Southern Arawakan: Yanesha

Yanesha is an Amazonian language from the Pre-Andine subgroup of Arawakan family (Adelaar and Muysken, 2004), spoken in Central Peru by between 3 and 5 thousand people. It has two linguistic variants that correspond to the upriver and downriver areas, both mutually intelligible.

Yanesha is an agglutinating, polysynthetic language with a VSO word order. Nouns and verbs are the two major word classes while the adjective word class is questionable due to the absence of non-derived forms. The verb is the most morphologically complex word class and the only obligatory constituent of a clause (Dixon and Aikhenvald, 1999).

Among other typologically remarkable features, we find that the language lacks the distinction in grammatical gender, the subject cross-referencing morphemes and one of the causatives are prefixes; all other verbal affixes are suffixes, and nouns and classifiers may be incorporated in the verb (Wise, 2002).

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.

3.12 Chukotko-Kamchatkan

The Chukotko-Kamchatkan languages, spoken in the far east of the Russian Federation, are represented in this dataset by two endangered languages, **Chukchi** and **Itelmen** (Eberhard et al., 2021).

3.12.1 Chukotko-Kamchatkan: Chukchi

Chukchi is a polysynthetic language that exhibits polypersonal agreement, ergative–absolutive alignment, and a subject–object–verb basic word order in transitive clauses (Tyers and Mishchenkova, 2020). We use the data of the Amguema corpus, available through the Chuklang website,[11] comprised of transcriptions of spoken Chukchi in the Amguema variant. The Amguema data had been annotated in the CoNLL-U format by Tyers and Mishchenkova (2020), and we convert it to the UniMorph format using the conversion system of McCarthy et al. (2018).

3.12.2 Chukotko-Kamchatkan: Itelmen

Itelmen is a language spoken on the western coast of the Kamchatka Peninsula. The language is con-

[11] https://chuklang.ru/

sidered to be highly endangered since it stopped been transferred from elders to youth ∼50 years ago (most are Russian-speaking monolinguals). The language is agglutinative and primarily uses suffixes. For instance, the plural form of a noun is expressed by the suffix *-ʔn*. We note that the plural form only exists in four grammatical cases (NOM, DAT, LOC, VOC).[12] The same plural suffix transforms a noun into an adjective. Verbs mark both subjects (with prefixes and suffixes) and objects (with suffixes). For instance, the first person subject is marked by attaching the prefix *t-* and the suffix *-čen* (Volodin, 1976).[13] The Itelmen data presented in the task was collected through fieldwork and manually annotated according to the UniMorph schema.

3.13 Trans-New Guinea

3.13.1 Bosavi: Eibela

Eibela, or Aimele, is an endangered language spoken by a small (∼300 speakers) community in Lake Campbell, Western Province, Papua New Guinea. Eibela morphology is exclusively suffixing. Verbs conjugate for tense, aspect, mood, evidentiality and exhibit complex paradigms with a high degree of irregularity. Generally, verbs cab be grouped into three classes based on their stems. Verbal inflectional classes present various kinds of stem alternations and suppletion. As Aiton (2016b) notes, the present and past forms are produced either through stem changes or by a concatenative suffix. In some cases, the forms can be quite similar (such as *na:gla:* 'be sick.PST' and *na:glɛ* 'be sick.PRS'). The future tense forms are typically inflected using suffixes. The current sample has been derived from interlinear texts from Aiton (2016a) and contains mostly partial paradigms.

4 Data Preparation

As in the previous editions, each instance in the provided training and development is in a form of a triple (lemma, tag, inflected form). The test set, on the other hand, was released with only lemmas and tags (i.e. without the target inflections). Producing these data sets required a few extra steps, which we discuss in this section.

Conversion into the UniMorph schema. After the data collection was finalised for the above languages, we converted them to the UniMorph schema—canonicalising them in the process.[14] This process consisted mainly of typo corrections (e.g. removing an incorrectly placed space in a tag, "`PRIV ` " → "`PRIV`"), removing redundant tags (e.g. duplicated verb annotation, "`V;V.PTCP`" → "`V.PTCP`"), and fixing tags to conform to the UniMorph schema (e.g. "`2;INCL`" → "`2+INCL`"). These changes were implemented via language-specific Bash scripts. Given this freshly converted data, we canonicalised its tag annotations, making use of `https://github.com/unimorph/um-canonicalize`. This process sorts the inflection tags into their canonical order and verifies that all the used tags are present in the ground truth UniMorph schema, flagging potential data issues in the process.

Data splitting. Given the canonicalised data as described above, we removed all instances with duplicated `<lemma; tags>` pair—as these instances were ambiguous with respect to their target inflected form—and removed all forms other than verbs, nouns, or adjectives. We then capped the dataset sizes to a maximum of 100,000 instances per language, subsampling when necessary. Finally, we create a 70–10–20 train–dev–test split per language, splitting the data across these sets at the instance level (as opposed to, e.g., the lemma one). As such, the information about a lemma's declension or inflection class is spread out across these train, dev and test sets, making this task much simpler than if one had to predict the entire class from the lemma's form alone, as done by, e.g., Williams et al. (2020) and Liu and Hulden (2021).

5 Baseline Systems

The organizers provide four neural systems as baselines, a product of two models and optional data augmentation. The first model is a transformer (Vaswani et al., 2017, `TRM`), and the second model is an adaption of the transformer to character-level transduction tasks (Wu et al., 2021, `CHR-TRM`), which holds the state-of-the-art on the 2017 SIGMORPHON shared task data. Both models follow the hyperparameters of Wu et al. (2021). The optional data augmentation follows the technique proposed by Anastasopoulos and Neubig (2019). Rely-

[12] `https://postnauka.ru/longreads/156195`

[13] `http://148.202.18.157/sitios/publicacionesite/pperiod/funcion/pdf/11-12/289.pdf`

[14] The new languages are included into the UniMorph data: `https://unimorph.github.io/`

	$\mathcal{L}$	V		N		ADJ		V.CVB		V.PTCP		V.MSDR	
Development	afb	19,861	2,184	7,595	2,996	4,208	1,510	–	–	–	–	–	–
	amh	20,254	670	20,280	1,599	829	195	4,096	668	–	–	668	668
	ara	31,002	635	53,365	1,703	58,187	742	–	–	–	–	–	–
	arz	8,178	1,320	10,533	3,205	6,551	1,771	–	–	–	–	–	–
	heb	28,635	1,041	3,666	142	–	–	–	–	–	–	847	847
	syc	596	187	724	329	158	86	–	–	261	77	–	–
	ame	1,246	184	2,359	143	–	–	–	–	–	–	–	–
	cni	5,478	150	14,448	258	–	–	–	–	–	–	–	–
	ind	8,805	2,570	5,699	2,759	1,313	731	–	–	–	–	–	–
	kod	315	44	91	14	56	8	–	–	–	–	–	–
	aym	50,050	910	91,840	656	–	–	–	–	–	–	910	910
	ckt	67	62	113	95	8	8	–	–	–	–	–	–
	itl	718	424	567	419	63	59	412	352	–	–	20	19
	gup	305	73	–	–	–	–	–	–	–	–	–	–
	bra	564	286	808	757	174	157	–	–	–	–	–	–
	bul	13,978	699	8,725	1,334	13,050	435	423	423	17,862	699	1,692	423
	ces	33,989	500	44,275	3,167	48,370	1,458	2,518	360	5,375	360	–	–
	ckb	14,368	112	1,882	142	–	–	–	–	289	112	–	–
	deu	64,438	2,390	73,620	9,543	–	–	–	–	4,777	2,390	–	–
	kmr	6,092	301	135,604	14,193	–	–	–	–	397	150	783	301
	mag	442	145	692	664	77	76	–	–	6	6	3	3
	nld	32,235	2,149	–	–	21,084	2,844	–	–	2,148	2,148	–	–
	pol	40,396	636	12,313	894	23,042	424	625	614	50,772	446	15,456	633
	por	133,499	1,884	–	–	–	–	–	–	9,420	1,884	–	–
	rus	33,961	2,115	54,153	4,747	46,268	1,650	3,188	2,107	5,486	2,138	–	–
	spa	132,702	2,042	–	–	–	–	2,042	2,042	8,184	2,046	–	–
	see	5,430	140	–	–	–	–	–	–	–	–	–	–
	ail	940	365	339	249	32	24	–	–	–	–	–	
	evn	2,106	961	3,661	2,249	446	393	612	390	716	517	–	–
	sah	20,466	237	122,510	1,189	–	–	2,832	236	–	–	–	–
	tyv	61,208	314	81,448	970	–	–	9,336	314	–	–	–	–
	krl	108,016	1,042	1,118	107	213	24	–	–	3,043	1,021	–	–
	lud	57	31	125	77	1	1	–	–	–	–	–	–
	olo	72,860	649	55,281	2,331	12,852	538	–	–	1,762	575	–	–
	vep	55,066	712	69,041	2,804	16,317	560	–	–	2,543	705	–	–
Surprise	sjo	135	99	49	41	16	16	86	69	78	65	51	44
	tur	97,090	190	44,892	992	1,440	20	–	–	–	–	–	–
	vro	–	–	1,148	41	–	–	–	–	–	–	–	–

Table 2: Number of samples and unique lemmata (the second number in each column) in each word class in the shared task data, aggregated over all splits. Here: "V" – verbs, "N" – nouns, "ADJ" – adjectives, "V.CVB" – converbs, "V.PTCP" – participles, "V.MSDR" – masdars.

ing on a simple character-level alignment between the lemma and the form, this technique replaces shared substrings of length > 3 with random characters from the language's alphabet, producing hallucinated lemma–tag–form triples. Data augmentation (+AUG) is applied to languages with fewer than 10K training instances, and 10K examples are generated for each language.

6 Submitted Systems

GUClasp The system submitted by Team GUClasp is based on the architecture and data augmentation technique presented by Anastasopoulos and Neubig (2019). More specifically, the team implemented an encoder–decoder model with an attention mechanism. The encoder processes a character sequence using an LSTM-based RNN with attention. Tags are encoded with a self-attention (Vaswani et al., 2017) position-invariant module. The decoder is an LSTM with separate attention mechanisms for the lemma and the tags. GUClasp focus their efforts on exploring strategies for training a multilingual model, in particular, they implement the following strategies: curriculum learning with competence (Platanios et al., 2019) based on character frequency and

$\mathcal{L}$	BME	GUClasp	TRM	TRM+AUG	CHR-TRM	CHR-TRM+AUG
			Development			
afb	92.39	81.71	94.88	94.88	**94.89**	**94.89**
amh	98.16	93.81	99.37	99.37	**99.45**	**99.45**
ara	99.76	94.86	99.74	99.74	**99.79**	**99.79**
arz	95.27	87.12	**96.71**	**96.71**	96.46	96.46
heb	97.46	89.93	99.10	99.10	**99.23**	**99.23**
syc	21.71	10.57	35.14	34.29	**36.29**	34.57
ame	82.46	55.94	87.43	**87.85**	87.15	86.19
cni	99.5	93.36	**99.90**	**99.90**	99.88	99.88
ind	81.31	55.68	**83.61**	**83.61**	83.30	83.30
kod	94.62	87.1	**96.77**	95.70	95.70	**96.77**
aym	**99.98**	99.97	**99.98**	**99.98**	**99.98**	**99.98**
ckt	44.74	52.63	26.32	55.26	28.95	**57.89**
itl	32.4	31.28	38.83	**39.66**	38.55	39.11
gup	14.75	21.31	59.02	**63.93**	55.74	60.66
bra	58.52	56.91	53.38	**59.81**	59.49	58.20
bul	98.9	96.46	**99.63**	**99.63**	99.56	99.56
ces	98.03	94.00	**98.24**	**98.24**	98.21	98.21
ckb	99.46	96.60	99.94	99.94	**99.97**	**99.97**
deu	**97.98**	91.94	97.43	97.43	97.46	97.46
kmr	**98.21**	98.09	98.02	98.02	98.01	98.01
mag	70.2	72.24	66.94	**73.47**	70.61	72.65
nld	98.28	94.91	98.89	98.89	**98.92**	**98.92**
pol	99.54	98.52	99.67	99.67	**99.70**	**99.70**
por	99.85	99.11	**99.90**	**99.90**	99.86	99.86
rus	**98.07**	94.32	97.55	97.55	97.58	97.58
spa	99.82	97.65	99.86	99.86	**99.90**	**99.90**
see	78.28	40.97	**90.65**	89.64	90.01	88.63
ail	6.84	6.46	12.17	11.79	10.65	**12.93**
evn	51.9	51.5	57.65	58.05	57.85	**59.12**
sah	99.95	99.69	99.93	99.93	**99.97**	**99.97**
tyv	99.97	99.78	99.95	99.95	**99.97**	**99.97**
krl	99.88	98.50	**99.90**	**99.90**	**99.90**	**99.90**
lud	**59.46**	**59.46**	16.22	45.95	27.03	45.95
olo	**99.72**	98.2	99.67	99.67	99.66	99.66
vep	**99.72**	97.05	99.65	99.65	99.70	99.70
			Surprise			
sjo	35.71	15.48	35.71	**47.62**	45.24	42.86
tur	**99.90**	99.49	99.36	99.36	99.35	99.35
vro	94.78	87.39	97.83	**98.26**	97.83	97.39

Table 3: Accuracy for each language on the test data.

model loss, predicting Levenshtein operations (`copy`, `delete`, `replace` and `add`) as a multi-task objective going from lemma to inflected form, label smoothing based on other characters in the same language (language-wise label smoothing), and scheduled sampling (Bengio et al., 2015).

BME Team `BME`'s system is an LSTM encoder-decoder model based on the work of Faruqui et al. (2016), with three-step training where the model is first trained on all languages, then fine-tuned on each language family, and finally fine-tuned on individual languages. A different type of data augmentation technique inspired by Neuvel and Fulop (2002) is also used in the first two steps. Team `BME` also perform ablation studies and show that the augmentation techniques and the three training steps often help but sometimes have a negative effect.

7 Evaluation

Following the evaluation procedure established in the previous shared task iterations, we compare all systems in terms of their test set accuracy. In addition, we perform an extensive error analysis for most languages.

8 Results

As Table 3 demonstrates, most systems achieve over 90% accuracy on most languages, with transformer-based baseline models demonstrating superior performance on all language families except Uralic. Two Turkic languages, Sakha and Tuvan, achieve particularly high accuracy of 99.97%.

This is likely due to the data being derived from morphological transducers where certain parts of verbal paradigms were excluded (see Section 3.8.1). On the other hand, the accuracy on Classical Syriac, Chukchi, Itelmen, Kunwinjku, Braj, Ludic, Eibela, Evenki, Xibe is low overall. Most of them are under-resourced and have very limited amounts of data—indeed, the Spearman correlation between the transformer model's performance and a language's training set size is of roughly 77%.

Analysis for each POS

Tables 13 to 18 in the Appendix provide the accuracy numbers for each word class. Verbs and nouns are the most represented classes in the dataset. For under-resourced languages such as Classical Syriac, Itelmen, Chukchi, Braj, Magahi, Evenki, Ludic, nouns are predicted more accurately, most likely due to the larger number of samples and smaller paradigms. Still, the models' performance is relatively stable across POS tags, the Pearson correlation between all models' performance on the verb and the noun data is 86%, while the noun–adjective performance correlation is 89% and the verb–adjective one is 84%. The most stable model across POS tags, at least according to these correlations, is BME, with an 87%, 96% and 91% Pearson correlation for verb–noun, noun–adjective, and verb–adjective performance accuracies respectively.

Analysis for out-of-vocabulary lemmas

Table 4 shows the differences in performance between the lemmas present both in training and test sets and the "unknown" lemmas. A closer inspection of this table shows that while BME and GUClasp models have an accuracy gap of 5.5% and 3% respectively between previously known and unknown lemmas, the transformer-based architectures show an accuracy gap from 9% to 16%. This larger gap, however, is partly explained by the better performance of the transformer-based models on previously seen lemmas (around 75%, while BME's performance is 71% and GUClasp's is 66%). The performance on the previously unseen lemmas, on the other hand, is mostly driven by data augmentation. The models without data augmentation have an accuracy around 60% on these lemmas, while all other models achieve around 65% on previously unseen lemmas. This is in line with the findings of Liu and Hulden (2021), who show that the transducer's performance on previously seen

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG	
afb	94.24	82.35	96.31	96.31	96.47	96.47	
	75.04	75.69	81.40	81.40	80.09	80.09	
amh	98.15	93.77	99.38	99.38	99.44	99.44	
	100.00	100.00	98.07	98.07	100.00	100.00	
ara	99.78	94.90	99.78	99.78	99.82	99.82	*
	50.00	27.77	33.33	33.33	33.33	33.33	
arz	95.66	86.70	97.08	97.08	96.80	96.80	
	89.91	92.79	91.64	91.64	91.64	91.64	
heb	97.46	89.92	99.09	99.09	99.23	99.23	
	-	-	-	-	-	-	
syc	28.89	14.06	46.38	43.72	46.76	44.10	
	0	0	1.14	5.74	4.59	5.74	
ame	82.45	55.93	87.43	87.84	87.15	86.18	
	-	-	-	-	-	-	
cni	99.50	93.35	99.90	99.90	99.87	99.87	
	100.00	100.00	100.00	100.00	100.00	100.00	
ind	82.29	55.18	84.90	84.90	84.49	84.49	
	68.83	61.90	67.09	67.09	67.96	67.96	
kod	94.62	90.32	100.00	98.92	98.92	100.00	
	-	-	-	-	-	-	
aym	99.97	99.96	99.97	99.97	99.97	99.97	
	-	-	-	-	-	-	
ckt	50.00	75.00	50.00	75.00	50.00	75.00	*
	44.11	50.00	23.52	52.94	26.47	55.88	
itl	29.03	23.22	34.83	38.06	36.12	33.54	
	34.97	37.43	41.87	40.88	40.39	43.34	
gup	14.28	21.42	62.50	64.28	58.92	64.28	
	20.00	20.00	20.00	60.00	20.00	20.00	*
bra	29.29	30.30	28.28	32.32	32.32	28.28	
	72.16	69.33	65.09	72.64	72.16	72.16	
bul	98.94	96.50	99.67	99.67	99.60	99.60	
	37.50	25.00	37.50	37.50	37.50	37.50	*
ces	98.02	94.00	98.23	98.23	98.21	98.21	
	-	-	-	-	-	-	
ckb	99.45	96.60	99.93	99.93	99.96	99.96	
	-	-	-	-	-	-	
deu	97.98	91.93	97.43	97.43	97.45	97.45	
	94.73	89.47	94.73	94.73	94.73	94.73	*
kmr	98.21	98.09	98.01	98.01	98.00	98.00	
	-	-	-	-	-	-	
mag	37.50	43.05	37.50	38.88	43.05	43.05	
	83.81	84.39	79.19	87.86	82.08	84.97	
nld	98.32	94.91	98.92	98.92	98.96	98.96	
	88.00	92.00	90.00	90.00	90.00	90.00	
pol	99.55	98.54	99.68	99.68	99.70	99.70	
	81.25	62.50	81.25	81.25	81.25	81.25	*
por	99.84	99.11	99.89	99.89	99.85	99.85	
	-	-	-	-	-	-	
rus	98.06	94.31	97.54	97.54	97.57	97.57	
	100.00	100.00	100.00	100.00	100.00	100.00	
spa	99.81	97.64	99.86	99.86	99.89	99.89	
	-	-	-	-	-	-	
see	78.27	40.97	90.65	89.64	90.00	88.63	
	-	-	-	-	-	-	
ail	5.88	5.88	11.76	11.17	12.35	12.35	
	8.60	7.52	12.90	12.90	7.52	13.97	
evn	46.88	44.88	52.66	53.00	53.66	53.33	
	59.46	61.47	65.15	65.66	64.15	67.83	
sah	99.95	99.68	99.93	99.93	99.97	99.97	
	-	-	-	-	-	-	
tyv	99.97	99.78	99.95	99.95	99.96	99.96	
	-	-	-	-	-	-	
krl	99.88	98.50	99.92	99.92	99.92	99.92	
	100.00	100.00	58.33	58.33	50.00	50.00	*
lud	62.50	56.25	37.50	37.50	50.00	43.75	*
	57.14	61.90	0	52.38	9.52	47.61	
olo	99.72	98.20	99.71	99.71	99.69	99.69	
	100.00	94.73	71.05	71.05	73.68	73.68	
vep	99.75	97.06	99.67	99.67	99.72	99.72	
	88.29	92.55	92.55	92.55	91.48	91.48	
sjo	45.65	10.86	54.34	47.82	60.86	41.30	
	23.68	21.05	13.15	47.36	26.31	44.73	
tur	99.90	99.49	99.35	99.35	9 *9.35	99.35	
	-	-	-	-	-	-	
vro	94.78	87.39	97.82	98.26	97.82	97.39	
	-	-	-	-	-	-	

Table 4: Accuracy comparison for the lemmas known from the training set (black numbers) vs. unknown lemmas (red numbers). Groups having <20 unique lemmas are marked with asterisks.

words can be greatly improved by simply training the models to perform the trivial task of copying random lemmas during training—a method somewhat related to data augmentation.

Analysis for the most challenging inflections

Table 5 shows the accuracy of the submitted systems on the "most challenging" test instances, where all four baselines failed to predict the target form correctly.

Frequently observed types of such cases include:

- Unusual alternations of some letters in particular lexemes which are hard to generalize;
- Ambiguity of the target forms. Certain lemmas allow some variation in forms, while the test set only lists a single exclusive golden form for each (lemma, tags) combination. In most cases, multiple acceptable forms may be hardly distinguishable in spoken language. For instance, they may only differ by an unstressed vowel or be orthographic variants of the same form.
- Multiword expressions are challenging when agreement is required. UniMorph does not provide dependency information, however, the information can be inferred from similar samples or other parts of the same lemma paradigm. The system's ability to make generalizations from a sequence down to its subsequences essentially depends on its architecture.
- Errors in the test sets. Still, a small percentage of errors come from the data itself.

$\mathcal{L}$	BME	GUClasp
afb	19.61	14.23
amh	6.81	15.90
ara	49.05	5.66
arz	20.28	18.11
heb	15.90	13.63
syc	0	.50
ame	6.45	4.83
cni	33.33	0
ind	19.94	15.60
aym	33.33	0
ckt	0	0
itl	3.50	2.33
gup	0	0
bra	9.27	9.27
bul	18.42	34.21
ces	35.59	16.57
ckb	100.00	0
deu	55.59	13.14
kmr	39.36	10.10
mag	4.00	6.00
nld	11.30	14.78
pol	47.61	30.15
por	27.27	0
rus	46.18	20.93
spa	64.00	12.00
see	7.89	3.94
ail	.99	.99
evn	5.55	4.78
sah	25.00	0
tyv	80.00	20.00
krl	78.94	47.36
lud	17.64	23.52
olo	64.61	23.07
vep	58.46	9.23
sjo	9.37	6.25
tur	93.37	88.39
vro	33.33	0

Table 5: Test accuracy for each language for the samples where none of baseline systems succeeds to produce correct prediction.

9 Error Analysis

As Elsner et al. (2019) note, accuracy-level evaluation might be sufficient to compare model variants but does not provide much insight into the understanding of morphological systems and their learnability. Therefore, we now turn to a more detailed analysis of mispredictions made by the systems. For the purpose of this study, we will rely on the error type taxonomy proposed by Gorman et al. (2019) and Muradoglu et al. (2020).

9.1 Evenki and Xibe

For the Evenki language, GUClasp tends to shorten the resulting words, sometimes generating theoretically impossible forms. However, in several cases, the result is practically correct but only in case of a different dialect, such as *abullən* instead of *abuldən*. The performance is better for nominal wordforms (74.27 accuracy for nouns only vs. 30.55 for verbs only). This is perhaps due to the higher regularity of nominal forms. BME is performing slightly better for the Evenki language, with errors in vowel harmony (such as *ahatkanmo* instead of *ahatkanmə*). In contrast with GUClasp, it tends to generate longer forms, adding unnecessary suffixes. The problems with dialectal forms can be found as well. The performance for Xibe is worse for both systems, though BME is better, despite the simpler morphology—perhaps it is due to the complexity of the Xibe script. At least in one instance, one of the systems generated a form with a Latin letter *n* instead of Xibe letters.

9.2 Syriac

Both GUClasp and BME generated 350 nominal, verbal and adjectival forms with less than 50% ac-

curacy. This includes forms that are hypothetically correct despite being historically unattested (e.g., *abydwtkwn* 'your (M.PL) loss'). Both systems performed better on nominal and adjectival forms than verbal forms. This may be explained by the higher morphological regularity of nominal forms relative to verbal forms; nominal/adjectival inflections typically follow linear affixation rules (e.g., suffixation) while verbal forms follow the same rules in addition to non-concatenative processes. Further, both systems handle lexemes with two or three letters (e.g., *dn* 'to judge') poorly compared to longer lexemes (e.g., *bṭnwt'* 'conception'). Where both systems generate multiple verbal forms for the same lexeme, the consonantal root is inconsistent. Finally, as expected, lexicalised phrases (e.g., *klnš* 'everyone', derived from the high-frequency contraction *kl* 'every' and *nš* 'person') and homomorphs (e.g., *ql'* 'an expression (n.)' or 'to fry (v.)') are handled poorly. Comparatively, the BME system performed worse than GUClasp, especially in terms of vowel diacritic placement and consonant doubling, which are consistently hypercorrected in both cases (e.g., *ḥbyb'* > *ḥabbbbay*; *ḥyltn'* > *ḥaallto'*).

9.3 Amharic

Both submitted systems performed well on the Amharic data, BME (98.16% accuracy) somewhat better than GUClasp (93.81% accuracy), though neither outperformed the baseline models.

For both systems, target errors represented a significant proportion of the errors, 32.35% for BME, 24.08% for GUClasp. Many of these involved alternative plural forms of nouns. The data included only the most frequent plural forms when there were alternatives, sometimes excluding uncommon but still possible forms. In some cases only an irregular form appeared in the data, and the system "erroneously" predicted the regular form with the suffix *-(w)oč*, which also correct. For example, BME produced *hawaryawoč*, the regular plural of *hawarya* 'apostle', instead of the expected irregular plural *hawaryat*. Another source of target errors was the confusion resulting from multiple representations for the phonemes /h,ʔ,s,s'/ in the Amharic orthography. Again, the data included only the common spelling for a given lemma or inflected form, but alternative spellings are usually also attested. Many of the "errors" consisted of predicting correct forms with one of these phonemes spelled differently than in the expected form.

The largest category of errors for both systems (unlike the baseline systems) were allomorphy errors, 51.76% for BME, 62.65% for GUClasp. Most of these resulted from the confusion between vowels in verbal templates. Particularly common were vowel errors in jussive-imperative (IMP) forms. Most Amharic verb lemmas belong to one of two inflection classes, each based on roots consisting of three consonants. The vowels in the templates for these categories are identical in the perfective (PRF), imperfective (IPFV), and converb (V.CVB) forms, but differ in the infinitive (V.MSDR) and jussive-imperative, where class A has the vowels *.i.ə* and class B has the vowels *.ə.i.*

Both systems also produced a significant number of silly errors—incorrect forms that could not be explained otherwise. Most of these consisted of consonant deletion, replacing one consonant with another, or repeating a consonant–vowel sequence.

9.4 Polish

Polish is among languages for which both systems and all the baselines achieved the highest accuracy results. BME, with 99.54% accuracy, is doing slightly better than GUClasp (98.52%). However, neither system exceeds the baseline results (99.67–99.70%).

Most of the errors made by both systems were already noted and classified by Gorman et al. (2019) and follow from typical irregularities in Polish. For example, masculine nouns have two GEN.SG suffixes: *-a* and *-u*. The latter is typical for inanimate nouns but the former could be used both with animate and inanimate nouns, which makes it highly unpredictable and causes production of incorrect forms such as *negatywa*, *rabunka* instead of *negatywu* 'negative', *rabunku* 'robbery'. Both systems are vulnerable to such mistakes. Another example would be the GEN.PL forms of plurale tantum nouns, which could have *-ów* or zero suffix, leading to errors such as: *tekstyli*, *wiktuał* instead of *tekstyliów* 'textiles', *wiktuałów* 'victuals'. Some loan words in Polish have fully (*mango*, *marines*, *monsieur*) or partially (*millenium*, in singular only) syncretic inflectional paradigms. This phenomenon is hard to predict, as the vast majority of Polish nouns inflect regularly. Both systems tend to produce inflected forms of those nouns according to their regular endings, which would be otherwise correct if not for their syncretic paradigms.

One area in which BME returns significantly bet-

ter results than GUClasp are imperative forms. Polish imperative forms follow a few different patterns involving some vowel alternations but in general are fairly regular. For the 364 imperative forms in the test data set, BME produced only 12 errors, mostly excusable and concerning existing phonetic alternations which could cause some problems even for native or L2 speakers. GUClasp, however, produced 61 erroneous imperative forms, some of them being examples of overgeneralization of the zero suffix pattern for first person singular imperatives (*wyjaśn* instead of *wyjaśnij* for the verb WYJAŚNIĆ 'explain').

Interestingly, both systems sometimes produce forms that are considered incorrect in standard Polish, but are quite often used colloquially by native speakers. Both BME and GUClasp generated the form *podniesą się* (instead of *podniosą się* 'they will rise'). Moreover, GUClasp generated the form *podeszłeś* (instead of *podszedłeś* 'you came up').

9.5 Russian

Similar to Polish and many other high-resource languages, the accuracy of all systems on Russian is high, with BME being the best-performing model (98.07%). Majority of errors consistently made by all systems (including the baseline ones) are related to the different inflections for animate and inanimate nouns in the accusative case. In particular, UniMorph does not provide the corresponding animacy feature for nouns, an issue that has also been reported previously by Gorman et al. (2019).

The formation of the short forms of adjectives and participles with -ен- and -енн- is another source of misprediction. The systems either generate an incorrect number of н, as in *умерена (should be умеренна 'moderate'), or fail to attach the suffix in cases that require some repetition in the production, as in *жертвен (should be жертвенен 'sacrificial'), i.e. generation stops after the first ен is produced. In addition to that, the systems often mispredict alternations of е and ё, as in *ошеломлённы instead of ошеломлены 'overwhelmed'. The same error also occurs in the formation of past participle forms such as *покормлённый (should be покормленный 'fed'). Further, we also observe it in noun declension, more specifically, in the prediction of the singular instrumental form: *слесарём (should be слесарем 'locksmith'), *гостьёй (should be гостьей, 'female guest'). Additionally, we observe more errors in the prediction of the instrumental case forms, mainly due to allomorphy. In many cases, the systems would have benefited from observing stress patterns or grammatical gender. For instance, consider the feminine акварель 'aquarelle' and the masculine пароль 'password'. In order to make a correct prediction, a model should either be explicitly provided with the grammatical gender, or a partial paradigm (e.g., the dative and genitive singular slots) for the corresponding lemma should be observed in the training set. Indeed, the latter is often the case, but the systems still fail to make a correct inference.

Finally, multiword expressions present themselves as an extra challenge to the models. In most cases, the test lemmas also appeared in the training set, therefore the systems could infer the dependency information from other parts of the same lexeme. Still, Russian multiword expessions appeared to be harder to inflect, probably as they show richer combinatory diversity. For instance, электромагнитное взаимодействие 'electromagnetic interaction' for the plural instrumental case slot is mispredicted as *электромагнитными взаимодействия, i.e the adjective part is correct while the noun form is not. As Table 7 illustrates, the accuracy gap in predicting multiword expressions with lemmas in- or out-of-vocabulary is quite large.

9.6 Ludic

The Ludic language, in comparison with the Karelian, Livvi and Veps languages, has the smallest number of lemmas (31 verbs, 77 nouns and 1 adjective) and has the lowest accuracy (16–60%). Therefore, the incomplete data is the main cause of errors in the morphological analyzers working with the Ludic dialect ('target errors' in the error taxonomy proposed by Gorman et al.).

9.7 Kurdish

Testing on the Sorani data yields high accuracy values, although there are errors in some cases. More than 200 lemmas and 16K samples were generated. Both BME and GUClasp generate regular nominal and verbal forms with a high accuracy of 99.46% and 96.6% respectively, although neither system exceeds the baseline results of 99.94% and 99.97%. Kurdish has a complex morphological system with defective paradigms and second-position person markers. Double clitic and affix-clitic construc-

tions can mark subjects or objects in a verbal construction and ditranstives are made with applicative markers. Such morphological structures can be the reason for the few issues that still occur.

9.8 Tuvan and Sakha

Both BME and GUClasp predict the majority of the inflected forms correctly, achieving test accuracies of over 99.6% on both Tuvan and Sakha, with BME performing slightly better on both languages. The remaining errors are generally caused by misapplications of morphophonology, either by the prediction system or by the data generator itself.

Since the forms treated as ground truth were automatically generated by morphological transducers (§3.8.1), the mismatch between the prediction and the reference might be due to 'target errors' where the reference itself is wrong (Gorman et al., 2019). For the BME system, target errors account for 1/8 disagreement cases for Tuvan and 3/13 for Sakha, although for all of them the system's prediction is indeed incorrect as well. For GUClasp, the reference is wrong in 19/62 cases for Tuvan (four of them also have an incorrect lemma, which makes it impossible to judge the correctness of any inflection) and 43/90 for Sakha. Interestingly, GUClasp actually predicts the correct inflected form for 27/43 and 3/15[15] target error cases for Sakha and Tuvan, respectively.

The actual failure cases for both BME and GUClasp are largely allomorphy errors, per Gorman et al.'s classification. Common problems include consonant alternation (Sakha *охсусуҥ instead of охсуһуҥ), vowel harmony (Tuvan *ижиарлер instead of ижигерлер) and vowel/null alternation (Tuvan *шымынар силер instead of шымныр силер). Unadapted loanwords that entered the languages through Russian (e.g. Sakha педагог 'pedagogue', принц 'prince', наследие 'heritage') are also frequent among the errors for both systems.

9.9 Ashaninka and Yanesha

For Ashaninka, the high baseline scores (over 99.8%) could be attributed to the relatively high regularity of the (morpho)phonological rules in the language. In this language, the BME system achieved comparable performance with 99.5%, whereas GUClasp still achieved a robust accuracy of 93.36%.

[15] 19 target errors excluding the 4 unjudgeable cases.

The case of Yanesha is different, as the baseline only peaked at 87.43%, whereas the BME and GUClasp systems underperformed with 82.46% and 55.94%, respectively. The task for Yanesha is harder, as the writing tradition is not transparent enough to predict some rules. For instance, large and short vowels are written in a similar way, always with a single vowel, and the aspirated vowels are optionally marked with a diacritic. These distinctions are essential at the linguistic level, as they allow one to explain the morphophonological processes, such as the syncope of the weak vowels in the flexionated forms (*po'kochllet* instead of *po'kchellet*). We also observe allomorphy errors, for instance, predicting *phomchocheñ* instead of *pomchocheñ* (from *mochocheñets* and V;NO2;FIN;REAL). The singular second person prefix has *ph-*, *pe-* and *p-* as allomorphs, each with different rules to apply. Moreover, there are some spelling issues as well, as the diacritic and the apostrophe are usually optional. For instance, the spellings *wapa* or *wápa* (to come where someone is located) are both correct. It is important to note that the orthographic standards are going to be revised by the Peruvian government to reduce the ambiguous elements.

9.10 Magahi

The transformer baseline with data augmentation (TRM+AUG) achieved the highest score, with GUClasp taking the second place (with 72.24%) and the base transformer yielding the lowest score of 66.94%. For Magahi, the results do not vary too much between systems, and the clearest performance boost seems to arise from the use of data augmentation. The low score of the TRM baseline is caused by the scarcity of data and the diversity in the morphophonological structure. Prediction errors on Magahi include incorrect honorificity, mispredicting plural markers, and spelling errors.

Honorificity: the systems predict forms lacking the honorific marker */-ak/*. For example, */puchhal/* ('asked') is predicted instead of */puchhlak/* ('asked'), or */bital/* ('passed time') instead of */bitlak/* ('passed time').

Plural marker: the systems' predictions omit the plural markers */-van/* and */-yan/*, similarly to the case of the honorific markers discussed above. For example, */thag/* ('con') is produced instead of */thagwan/* ('con').

Spelling errors: the predicted words do not oc-

cur in the Magahi language. The predictions also do not show any specific error pattern.

We thus conclude that the performance of the baseline systems is greatly affected by the morphological structure of Magahi. Also, some language-specific properties of Magahi are not covered by the scope of the UniMorph tagset. For example, consider the following pair:

(*/dekh/*, */dekhlai/*, 'V;3;PRF;INFM;LSGSPEC1', 'see')

(*/dekh/*, */dekhlak/*, 'V;3;PRF;INFM;LGSPEC2', 'see')

Here both forms exhibit morphological features that are not defined in the default annotation schema. Morphologically, the first form indicates that the speaker knows the addressee but not intimately (or there is a low level of intimacy), while the second one signals a comparatively higher level of intimacy. Such aspects of the Magahi morphology are challenging for the systems.

9.11 Braj

For the low-resource language Braj, both submitted systems performed worse than the baseline systems. `BME` achieved 58.52% prediction accuracy, slightly outperforming `GUClasp` with 56.91%. As for the baseline systems, `CHR-TRM` scored highest with 59.49% accuracy and `TRM` scored lowest with 53.38%. Among Indo-European languages, the performance of the `BME`, `GUClasp`, and the baseline systems is lowest for Braj. The low accuracy and the larger number of errors are broadly due to misprediction and misrepresentation of the morphological units and the smaller data size.

`BME`, `GUClasp`, and the baseline systems generated 311 nominal, verbal, and adjectival inflected forms from existing lemmas. In these outputs, the misprediction and misrepresentation errors are morphemic errors, already included/classified by Gorman et al. (2019). The findings of our analysis of both the gold data and the predictions of all systems highlight several common problems for nominal, verbal, and adjectival inflected forms. Common errors, mispredicted by all models, include morphemes of gender (masculine and feminine: for the noun *akhabaaree* instead of *akhabaar* 'newspaper', for the verb *arraa* instead of *arraaii* 'shout', and for the adjective *mithak* instead of *mithakeey* 'ancient'); morphemes of number (singular and plural: for the noun *kahaanee* instead of *kahaaneen* 'story', for the verb *utaran* instead of *utare* 'get off', for the adjective *achchhe* instead of *achchhau* 'good'); morphemes of honorificity (formal and informal: *suni* instead of *sunikain* 'listen', *rahee* instead of *raheen* 'be' and *karibau* instead of *kari* 'do', etc.).

A portion of these errors is also caused by the inflection errors in predicting and generating multiword expressions (MWEs) (e.g. *aannd* instead of *aannd-daayak* 'comfortable'). Apart from the mentioned error types, the systems also made silly errors (e.g. *uthi* instead of *uthaay* 'get up', *kathan* instead of *kathanopakathan* 'conversation', *karaah* instead of *karaahatau* 'groan', *keeee* instead of *keenee* 'do', and *grahanave* instead of *grahan* 'accept', etc.) and spelling errors (e.g. *dhamaky* instead of *dhamakii* 'threat', *laau* or *laauy* instead of *liyau* 'take', *saii* instead of *saanchii* 'correct', and *samajhat* instead of *samajhaayau* 'explain', etc.) as classified by Gorman et al. (2019). Under all of the models, the majority of errors were silly or spelling errors.

9.12 Aymara

All systems achieved high accuracy (99.98%) on this language. The few errors are mainly due to the inconsistency in the initial data annotation. For instance, the form *uraqiw* is listed as a lemma while it can only be understood as being comprised of two morphemes: *uraqi-w(a)* 'it (is) land'. The root, or the nominative unmarked form, is *uraqi* 'land'. The *-wa* is presumably the declarative suffix. The nucleus of this suffix can be lost owing to the complex morphophonemic rules which operate at the edges of phrases. In addition, the accusative form *uraqi* is incorrect since the accusative is marked by subtractive disfixation, therefore, *uraq* is the accusative inflected form.

9.13 Eibela

Eibela seems to be one of the most challenging languages, probably due to its small data size and sparsity. Since it has been extracted from interlinear texts, a vast majority of its paradigms are partial, and this certainly makes the task more difficult. A closer look at system outputs reveals that many errors are related to misprediction of vowel length. For instance, *toːmulu* is inflected in N;ERG as *tomulɛ* instead of *toːmuːlɛː*.

9.14 Kunwinjku

The data for Kunwinjku is relatively small and contains verbal paradigms only. Test accuracies range from 14.75% (`BME`) to 63.93% (`TRM+AUG`).

In this language, many errors were due to incorrect spelling and missing parts of transcription. For instance, for the second person plural non-past of the lemma *borlbme*, TRM predicts **ngurriborlbme* instead of *ngurriborle*. Interestingly, BME mispredicts most forms due to the looping effect described by Shcherbakov et al. (2020). In particular, it starts producing sequences such as **ngarrrrrrrrrrrrrmbbbijj* (should be *karribelbmerrinj*) or **ngadjarridarrkddrrdddrrmerri* (should be *karriyawoyhdjarrkbidyikarrmerrimeninj*).

10 Discussion

Reusing transformation patterns

In most cases, morphological transformations may be properly carried out by matching a lemma against a pattern containing fixed characters and variable (wildcard) character sequences. A morphological inflection may be described in terms of inserting, deleting, and/or replacing fixed characters. When a test sample follows such a regular transformation pattern observed in the training set, it usually becomes significantly easier to track down a correct target form. Table 6 demonstrates the difference in performance w.r.t. whether the required transformation pattern was immediately witnessed in any training sample. To enumerate the possible patterns, we used the technique described by Scherbakov (2020). It is worth emphasizing that the presence of a matching pattern by itself does not guarantee that achieving high accuracy would be straightforward, because in a vast majority of cases there are multiple alternative patterns. Choosing the correct one is a challenging task.

Multi-word forms

Inflecting multi-word expressions is one of the most challenging tasks. However, in the the shared task dataset, almost all multi-word lemmas found in the test set are also present in the training set, which made the task easier to solve.

The systems were quite successful at predicting the multi-word forms if the required transformation was directly observed in a training example. Otherwise, the prediction accuracy significantly degraded. Table 7 shows the multi-word lemma transformation accuracies. From these results we further notice that while all systems' performance degrades on the previously unseen multi-word inflection patterns, this degradation is considerably smaller for the transformer-based baselines (except

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG	
afb	96.43	87.25	97.70	97.70	97.70	97.70	
	76.93	60.47	84.06	84.06	84.14	84.14	
amh	98.78	96.71	99.55	99.55	99.58	99.58	
	95.51	81.45	98.58	98.58	98.86	98.86	
ara	99.89	97.06	99.90	99.90	99.90	99.90	
	97.97	65.28	97.56	97.56	98.22	98.22	
arz	96.59	91.86	97.96	97.96	97.83	97.83	
	80.04	32.51	82.26	82.26	80.54	80.54	
heb	98.70	94.84	99.61	99.61	99.74	99.74	
	92.22	69.15	96.93	96.93	97.09	97.09	
syc	85.71	82.14	85.71	85.71	85.71	82.14	*
	16.14	4.34	30.74	29.81	31.98	30.43	
ame	87.93	75.63	92.11	92.80	92.80	90.95	
	74.40	26.96	80.54	80.54	78.83	79.18	
cni	99.86	94.69	100.00	100.00	99.94	99.94	
	93.53	71.55	98.27	98.27	98.70	98.70	
ind	81.67	56.08	83.98	83.98	83.70	83.70	
	36.00	4.00	36.00	36.00	32.00	32.00	
kod	98.82	97.64	100.00	98.82	98.82	100.00	
	50.00	12.50	100.00	100.00	100.00	100.00	*
aym	99.98	99.98	99.99	99.99	99.99	99.99	
	50.00	0	0	0	0	0	*
ckt	73.91	82.60	34.78	82.60	39.13	86.95	
	0	6.66	13.33	13.33	13.33	13.33	*
itl	54.80	51.44	62.50	65.86	62.50	64.90	
	1.33	3.33	6.00	3.33	5.33	3.33	
gup	27.77	44.44	66.66	88.88	72.22	83.33	*
	9.30	11.62	55.81	53.48	48.83	51.16	
bra	75.96	73.81	68.66	76.82	75.53	72.96	
	6.41	6.41	7.69	8.97	11.53	14.10	
bul	99.05	96.92	99.73	99.73	99.66	99.66	
	81.44	42.26	87.62	87.62	87.62	87.62	
ces	98.34	95.00	98.43	98.43	98.47	98.47	
	78.97	34.22	86.12	86.12	82.32	82.32	
ckb	99.43	97.22	99.93	99.93	99.96	99.96	
	99.66	90.42	100.00	100.00	100.00	100.00	
deu	98.11	92.50	97.58	97.58	97.59	97.59	
	78.10	10.94	76.11	76.11	77.61	77.61	
kmr	98.22	98.13	98.01	98.01	98.01	98.01	
	84.21	26.31	100.00	100.00	84.21	84.21	*
mag	86.91	90.05	80.62	89.00	85.34	89.00	
	11.11	9.25	18.51	18.51	18.51	14.81	
nld	98.39	95.31	99.00	99.00	99.06	99.06	
	78.46	24.61	78.46	78.46	75.38	75.38	
pol	99.65	98.91	99.80	99.80	99.80	99.80	
	96.12	86.86	95.90	95.90	96.66	96.66	
por	99.89	99.40	99.93	99.93	99.91	99.91	
	88.07	22.93	91.74	91.74	84.40	84.40	
rus	98.40	95.36	97.82	97.82	97.89	97.89	
	73.72	18.36	77.55	77.55	75.00	75.00	
spa	99.86	97.97	99.93	99.93	99.94	99.94	
	93.47	46.19	89.13	89.13	92.39	92.39	
see	88.14	66.75	96.39	96.13	96.13	94.32	
	72.83	26.74	87.48	86.05	86.62	85.49	
ail	32.14	39.28	39.28	42.85	39.28	60.71	
	3.82	2.55	8.93	8.08	7.23	7.23	
evn	67.75	67.84	73.62	74.79	73.98	76.06	
	6.92	5.12	12.30	10.51	12.05	11.02	
sah	99.96	99.80	99.95	99.95	99.98	99.98	
	98.57	83.88	96.68	96.68	98.10	98.10	
tyv	99.97	99.85	99.96	99.96	99.97	99.97	
	99.76	95.20	99.28	99.28	99.76	99.76	
krl	99.92	99.08	99.91	99.91	99.92	99.92	
	98.24	78.34	99.20	99.20	99.04	99.04	
lud	91.66	87.50	25.00	70.83	41.66	70.83	
	0	7.69	0	0	0	0	*
olo	99.80	98.77	99.82	99.82	99.82	99.82	
	97.77	83.56	95.91	95.91	95.45	95.45	
vep	99.76	97.67	99.72	99.72	99.75	99.75	
	97.75	66.32	96.20	96.20	97.06	97.06	
sjo	50.00	22.22	48.14	64.81	61.11	55.55	
	10.00	3.33	13.33	16.66	16.66	20.00	
tur	99.93	99.64	99.39	99.39	99.38	99.38	
	97.12	83.81	94.96	94.96	95.68	95.68	
vro	99.14	99.14	98.29	99.14	98.29	99.14	
	90.26	75.22	97.34	97.34	97.34	95.57	

Table 6: Accuracy comparison for fragment substitutions that could be observed in the training set (black numbers) vs. more complex transformations (red numbers). Groups having <20 unique lemmas are marked with asterisks.

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
aym	100.00	99.90	100.00	100.00	100.00	100.00
	-	-	-	-	-	-
bul	95.00	93.21	100.00	100.00	100.00	100.00
	81.81	63.63	100.00	100.00	100.00	100.00
ces	75.00	55.00	80.00	80.00	80.00	80.00
	71.42	57.14	100.00	100.00	100.00	100.00
ckb	100.00	100.00	100.00	100.00	100.00	100.00
	-	-	-	-	-	-
deu	88.57	74.28	97.14	97.14	100.00	100.00
	71.87	0	93.75	93.75	87.50	87.50
kmr	98.76	98.71	95.34	95.34	95.51	95.51
	-	-	-	-	-	-
nld	50.00	50.00	50.00	50.00	50.00	50.00
	-	-	-	-	-	-
pol	99.85	99.60	99.97	99.97	99.91	99.91
	98.28	92.12	98.28	98.28	98.28	98.28
rus	90.93	87.91	96.45	96.45	96.05	96.05
	56.55	27.04	72.13	72.13	72.13	72.13
tur	99.77	98.63	95.67	95.67	95.58	95.58
	90.90	59.09	77.27	77.27	86.36	86.36

Table 7: Accuracy for MWE lemmata in each language on the test data. The numbers in black correspond to fragment substitutions that could be observed in the training set, while the red numbers correspond to more complex transformations.

for Turkish), implying that these models can better generalise to previously unseen patterns.

Allomorphy

The application of wrong (albeit valid) inflectional transformations by the models (allomorphy) is present in most analysed languages. These allomorphy errors can be further divided into two groups: (1) when an inflectional tag itself allows for multiple inflection patterns which must be distinguished by the declension/inflection class to which the word belongs, and (2) when the model applies an inflectional rule that is simply invalid for that specific tag. These errors are hard to analyse, however. The first is potentially unavoidable without extra information, as declension/inflection classes are not always fully predictable from word forms alone (Williams et al., 2020). The second type of allomorphic error, on the other hand, is potentially fixable. In our analysis, however, we did not find any concrete patterns to when the models make this second (somewhat arbitrary) type of mistake.

Spelling Errors

Spelling errors are pervasive in most analysed languages, even high-resource ones. These are challenging, as they require a deep understanding of the modelled language in order to be avoided. Spelling errors are especially common in languages with vowel harmony (e.g. Tungusic), as the models have some difficulty in correctly modelling it. Another source of spelling errors are the diacritics. In Portuguese, for instance, most of the errors produced by the BME system arise due to missing acute accents, which mark stress; their use is determined by specific (and somewhat idiosyncratic) orthographic rules.

11 Conclusion

In the development of this shared task we added new data for 32 languages (13 language families) to UniMorph—most of which are under-resourced. Further, we evaluated the performance of morphological reinflection systems on a typologically diverse set of languages and performed fine-grained analysis of their error patterns in most of these languages. The main challenge for the morphological reinflection systems is still (as expected) handling low-resource scenarios (where there is little training data). We further identified a large gap in these systems' performance between the test lemmas present in the training set and the previously unseen lemmas—the latter are naturally hard test cases, but the work on reinflection models could focus on improving these results going forward, following, for instance, the work of Liu and Hulden (2021). Further, allomorphy, honorificity and multiword lemmas also pose challenges for the current models. We hope that the analysis presented here, together with the new expansion of the UniMorph resources, will help drive further improvements in morphological reinflection. Following Malouf et al. (2020), we would like to emphasize that linguistic analyses using UniMorph should be performed with some degree of caution, since for many languages it might not provide an exhaustive list of paradigms and variants.

Acknowledgements

We would like to thank Dr George Kiraz and Beth Mardutho: the Syriac Institute for their help with Classical Syriac data.

References

Willem F. H. Adelaar and Pieter C. Muysken. 2004. *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press.

Grant Aiton. 2016a. The documentation of eibela: An archive of eibela language materials from the bosavi region (western province, papua new guinea).

Grant William Aiton. 2016b. *A grammar of Eibela: a language of the Western Province, Papua New Guinea*. Ph.D. thesis, James Cook University.

Sarah Alkuhlani and Nizar Habash. 2011. A corpus for modeling morpho-syntactic agreement in Arabic: Gender, number and rationality. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 357–362, Portland, Oregon, USA. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. 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 984–996, Hong Kong, China. Association for Computational Linguistics.

Gregory David Anderson and K David Harrison. 1999. *Tyvan (Languages of the World/Materials 257)*. München: LINCOM Europa.

Phyllis E. Wms. Bardeau. 2007. *The Seneca Verb: Labeling the Ancient Voice*. Seneca Nation Education Department, Cattaraugus Territory.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 1171–1179.

Noam Chomsky. 1995. Language and nature. *Mind*, 104(413):1–61.

Matt Coler. 2010. *A grammatical description of Muylaq'Aymara*. Ph.D. thesis, Vrije Universiteit Amsterdam.

Matt Coler. 2014. *A grammar of Muylaq'Aymara: Aymara as spoken in Southern Peru*. Brill.

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

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 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. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 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. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 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. Association for Computational Linguistics.

William Croft. 2002. *Typology and universals*. Cambridge University Press.

Michael Daniel. 2011. Linguistic typology and the study of language. In *The Oxford handbook of linguistic typology*. Oxford University Press.

R. M. W. Dixon and Alexandra Y. Aikhenvald, editors. 1999. *The Amazonian languages (Cambridge Language Surveys)*. Cambridge University Press.

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

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2021. Ethnologue: Languages of the world. Twenty-fourth edition. Online version: http://www.ethnologue.com.

Micha Elsner, Andrea D Sims, Alexander Erdmann, Antonio Hernandez, Evan Jaffe, Lifeng Jin, Martha Booker Johnson, Shuan Karim, David L King, Luana Lamberti Nunes, et al. 2019. Modeling morphological learning, typology, and change: What can the neural sequence-to-sequence framework contribute? *Journal of Language Modelling*, 7.

Nicholas Evans. 2003. *Bininj Gun-wok: A Pandialectal Grammar of Mayali, Kunwinjku and Kune*. Pacific Linguistics. Australian National University.

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

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

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

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

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine translation*, 25(2):127–144.

Michael Gasser. 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.

Yustinus Ghanggo Ate. 2020. Kodi (Indonesia) - Language Snapshot. *Language Documentation and Description 19*, pages 171–180.

Yustinus Ghanggo Ate. 2021. *Documentation of Kodi*. New Haven: Endangered Language Fund.

Yustinus Ghanggo Ate. to appear in 2021. Reduplication in Kodi: A paradigm function account. *Word Structure 14(3)*.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. 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. Association for Computational Linguistics.

Joseph Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph Greenberg, editor, *Universals of Language*, pages 73–113. MIT Press.

George A. Grierson. 1908. *Indo-Aryan Family: Central Group: Specimens of the Rājasthānī and Gujarātī*, volume IX(II) of *Linguistic Survey of India*. Office of the Superintendent of Government Printing, Calcutta.

George Abraham Grierson. 1903. *Linguistic Survey of India, Vol-III*. Calcutta: Office of the Superintendent, Government of PRI.

George Abraham Grierson and Sten Konow. 1903. *Linguistic Survey of India*. Calcutta Supt., Govt. Printing.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 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. Association for Computational Linguistics.

K. David Harrison. 2000. *Topics in the Phonology and Morphology of Tuvan*. Ph.D. thesis, Yale University.

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

Martin Haspelmath. 2020. Human linguisticality and the building blocks of languages. *Frontiers in psychology*, 10:3056.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for universal dependencies treebank files. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 87–93, Paris, France. Association for Computational Linguistics.

Sardana Ivanova, Anisia Katinskaia, and Roman Yangarber. 2019. Tools for supporting language learning for Sakha. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 155–163, Turku, Finland. Linköping University Electronic Press.

Sardana Ivanova, Francis M. Tyers, and Jonathan N. Washington. to appear in 2022. A free/open-source morphological analyser and generator for Sakha. *In preparation*.

Danesh Jain and George Cardona. 2007. *The Indo-Aryan Languages*. Routledge.

Thomas Jügel. 2009. Ergative Remnants in Sorani Kurdish? *Orientalia Suecana*, 58:142–158.

Olga Kazakevich and Elena Klyachko. 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.

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim, and Meera Al Kaabi. 2018. A morphologically annotated corpus of Emirati Arabic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Salam Khalifa, Sara Hassan, and Nizar Habash. 2017. A morphological analyzer for Gulf Arabic verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 35–45, Valencia, Spain. Association for Computational Linguistics.

Tanmai Khanna, Jonathan N. Washington, Francis M. Tyers, Sevilay Bayatlı, Daniel G. Swanson, Tommi A. Pirinen, Irene Tang, and Hèctor Alòs i Font. to appear in 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*.

Lee Kindberg. 1980. *Diccionario asháninca*. Lima: Instituto Lingüístico de Verano.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh*

International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Ritesh Kumar, Bornini Lahiri, and Deepak Alok. 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.

Ritesh Kumar, Bornini Lahiri, Deepak Alok Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 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. European Language Resources Association (ELRA).

Bornini Lahiri. 2021. *The Case System of Eastern Indo-Aryan Languages: A Typological Overview*. Routledge.

William Lane and Steven Bird. 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. Australasian Language Technology Association.

Septina Dian Larasati, Vladislav Kubon, and Daniel Zeman. 2011. Indonesian morphology tool (MorphInd): Towards an Indonesian corpus. In *Systems and Frameworks for Computational Morphology - Second International Workshop, SFCM 2011, Zurich, Switzerland, August 26, 2011. Proceedings*, volume 100 of *Communications in Computer and Information Science*, pages 119–129. Springer.

Ling Liu and Mans Hulden. 2021. Can a transformer pass the wug test? Tuning copying bias in neural morphological inflection models. *arXiv preprint arXiv:2104.06483*.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proceedings of the International Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

Mohamed Maamouri, Ann Bies, Seth Kulick, Dalila Tabessi, and Sondos Krouna. 2012. Egyptian Arabic Treebank DF Parts 1-8 V2.0 - LDC catalog numbers LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21.

Mohamed Maamouri, Dave Graff, Basma Bouziri, Sondos Krouna, Ann Bies, and Seth Kulick. 2010. LDC standard Arabic morphological analyzer (SAMA) version 3.1.

Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. Association for Computational Linguistics.

Pedro Mayor Aparicio and Richard E Bodmer. 2009. *Pueblos indígenas de la Amazonía peruana*. Iquitos: Centro de Estudios Teológicos de la Amazonía.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 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. Association for Computational Linguistics.

Elena Mihas. 2017. The Kampa subgroup of the Arawak language family. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *The Cambridge Handbook of Linguistic Typology*, Cambridge Handbooks in Language and Linguistics, page 782–814. Cambridge University Press.

Saliha Muradoglu, Nicholas Evans, and Ekaterina Vylomova. 2020. Modelling verbal morphology in Nen. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 43–53, Virtual Workshop. Australasian Language Technology Association.

Sylvain Neuvel and Sean A. Fulop. 2002. Unsupervised learning of morphology without morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 31–40. Association for Computational Linguistics.

I. P. Novak, N. B. Krizhanovskaya, T. P. Boiko, and N. A. Pellinen. 2020. Development of rules of generation of nominal word forms for new-written variants of the Karelian language. *Vestnik ugrovedenia = Bulletin of Ugric Studies*, 10(4):679–691.

Irina Novak. 2019. Karelian language and its dialects. In I. Vinokurova, editor, *Peoples of Karelia: Historical and Ethnographic Essays*, pages 56–65. Periodika.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.

Sofia Oskolskaya, Ezequiel Koile, and Martine Robbeets. 2021. A Bayesian approach to the classification of Tungusic languages. *Diachronica*.

Prateek Pankaj. 2020. Reconciling Surdas and Keshavdas: A study of commonalities and differences in Brajbhasha literature. *IOSR Journal of Humanities and Social Sciences*, 25.

Femphy Pisceldo, Rahmad Mahendra, Ruli Manurung, and I Wayan Arka. 2008. A two-level morphological analyser for the Indonesian language. In *Proceedings of the Australasian Language Technology Association Workshop 2008*, pages 142–150, Hobart, Australia.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom M. Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1162–1172. Association for Computational Linguistics.

Adam Przepiórkowski and Marcin Woliński. 2003. A flexemic tagset for Polish. In *Proceedings of the 2003 EACL Workshop on Morphological Processing of Slavic Languages*, pages 33–40, Budapest, Hungary. Association for Computational Linguistics.

Andreas Scherbakov. 2020. The UniMelb submission to the 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 177–183, Online. Association for Computational Linguistics.

Claus Schönig. 1999. The internal division of modern Turkic and its historical implications. *Acta Orientalia Academiae Scientiarum Hungaricae*, pages 63–95.

Andrei Shcherbakov, Saliha Muradoglu, and Ekaterina Vylomova. 2020. Exploring looping effects in RNN-based architectures. In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 115–120, Virtual Workshop. Australasian Language Technology Association.

John Sylak-Glassman, Christo Kirov, Matt Post, Roger Que, and David Yarowsky. 2015a. A universal feature schema for rich morphological annotation and fine-grained cross-lingual part-of-speech tagging. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 72–93. Springer.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 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. Association for Computational Linguistics.

Dima Taji, Nizar Habash, and Daniel Zeman. 2017. Universal Dependencies for Arabic. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 166–176, Valencia, Spain. Association for Computational Linguistics.

Dima Taji, Salam Khalifa, Ossama Obeid, Fadhl Eryani, and Nizar Habash. 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. Association for Computational Linguistics.

Talat Tekin. 1990. A new classification of the Turkic languages. *Türk dilleri araştırmaları*, 1:5–18.

Francis Tyers, Aziyana Bayyr-ool, Aelita Salchak, and Jonathan Washington. 2016. A finite-state morphological analyser for Tuvan. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2562–2567, Portorož, Slovenia. European Language Resources Association (ELRA).

Francis Tyers and Karina Mishchenkova. 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). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

A. P. Volodin. 1976. *The Itelmen language*. Prosveshchenie, Leningrad.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 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. Association for Computational Linguistics.

Jonathan North Washington, Aziyana Bayyr-ool, Aelita Salchak, and Francis M Tyers. 2016. Development of a finite-state model for morphological processing of Tuvan. *Rodnoy Yazyk*, 1:156–187.

Jonathan North Washington, Ilnar Salimzianov, Francis M. Tyers, Memduh Gökırmak, Sardana Ivanova, and Oğuzhan Kuyrukçu. to appear in 2021. Free/open-source technologies for Turkic languages developed in the Apertium project. In *Proceedings of the International Conference on Turkic Language Processing (TURKLANG 2019)*.

Jonathan North Washington and Francis Morton Tyers. 2019. Delineating Turkic non-finite verb forms by syntactic function. In *Proceedings of the Workshop on Turkic and Languages in Contact with Turkic*, volume 4, pages 115–129.

Adina Williams, Tiago Pimentel, Hagen Blix, Arya D. McCarthy, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6682–6695, Online. Association for Computational Linguistics.

Mary Ruth Wise. 2002. Applicative affixes in Peruvian Amazonian languages. *Current Studies on South American Languages [Indigenous Languages of Latin America, 3]*, pages 329–344.

Marcin Woliński and Witold Kieraś. 2016. The online version of grammatical dictionary of Polish. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2589–2594, Portorož, Slovenia. European Language Resources Association (ELRA).

Marcin Woliński, Zygmunt Saloni, Robert Wołosz, Włodzimierz Gruszczyński, Danuta Skowrońska, and Zbigniew Bronk. 2020. *Słownik gramatyczny języka polskiego*, 4th edition. Warsaw. http://sgjp.pl.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

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

He Zhou, Juyeon Chung, Sandra Kübler, and Francis Tyers. 2020. Universal Dependency treebank for Xibe. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 205–215, Barcelona, Spain (Online). Association for Computational Linguistics.

Esaú Zumaeta Rojas and Gerardo Anton Zerdin. 2018. *Guía teórica del idioma asháninka*. Lima: Universidad Católica Sedes Sapientiae.

Н. А. Баскаков. 1969. Введение в изучение тюркских языков *[N. A. Baskakov. An introduction to Turkic language studies]*. Москва: Высшая школа.

Ф. Г. Исхаков and А. А. Пальмбах. 1961. Грамматика тувинского языка*:* Фонетика и морфология *[F. G. Iskhakov and A. A. Pal'mbakh. A grammar of Tuvan: Phonetics and morphology]*. Москва: Наука.

Е. И. Убрятова, Е. И. Коркина, Л. Н. Харитонов, and Н. Е. Петров, editors. 1982. Грамматика современного якутского литературного языка*:* Фонетика и морфология *[E. I. Ubryatova et al. Grammar of the modern Yakut literary language: Phonetics and morphology]*. Москва: Наука.

A Data conversion into UniMorph

Apertium tag	UniMorph tag	Apertium tag	UniMorph tag	Apertium tag	UniMorph tag
<p1>	1	<imp>	IMP	<px1sg>	PSS1S
<p2>	2	<ins>	INS	<px2pl>	PSS2P
<p3>	3	<iter>	ITER	<px2sg>	PSS2S
<abl>	ABL	<loc>	LOC	<px3pl>	PSS3P
<acc>	ACC	<n>	N	<px3sg>	PSS3S
<all>	ALL	<neg>	NEG	<px3sp>	PSS3S/PSS3P
<com>	COM	<nom>	NOM	<pii>	PST;IPFV
<comp>	COMPV	<aor>	NPST	<ifi>	PST;LGSPEC1
<dat>	DAT	<nec>	OBLIG	<past>	PST;LGSPEC2
<ded>	DED	<pl>	PL	<sg>	SG
<du>	DU	<perf>	PRF	<v>	V
<fut>	FUT	<resu>	PRF;LGSPEC3	<gna_cond>	V.CVB;COND
<gen>	GEN	<par>	PRT	<prc_cond>	V.CVB;COND
<hab>	HAB	<px1pl>	PSS1P		

Table 8: Apertium tag mapping to the UniMorph schema for Sakha and Tuvan. For the definitions of the Apertium tags, see Washington et al. (2016). This mapping alone is not sufficient to reconstruct the UniMorph annotation, since some conditional rules are applied on top of this conversion (see §3.8.1)

Level-1		Level-2		Level-3	
MorphInd	UniMorph	MorphInd	UniMorph	MorphInd	UniMorph
N	N	P	PL	F	FEM
		S	SG	M	MASC
				D	NEUT
P	PROPN	P	PL	1	1
		S	SG	2	2
				3	3
V	V	P	-	A	ACT
		S	-	P	PASS
C	NUM	C	-	-	-
		O	-	-	-
		D	-	-	-
A	ADJ	P	PL	P	-
		S	SG	S	-

Table 9: A simplified mapping from MorphInd tags to the UniMorph schema for Indonesian data. We follow MorphInd's three-level annotations for the mapping.

PL tag	UniMorph tag	PL tag	UniMorph tag	PL tag	UniMorph tag
pri	1	impt	IMP	perf	PFV
sec	2	imps	IMPRS	pl	PL
ter	3	inst	INS	fin	PRS/FUT
acc	ACC	imperf	IPFV	praet	PRT
adj	ADJ	m2	MASC;ANIM	sg	SG
adv	ADV	m1	MASC;HUM	sup	SPRL
com	COMPR	m3	MASC;INAN	pcon	V.CVB;PRS
dat	DAT	subst	N	pant	V.CVB;PST
pos	EQT	neg	NEG	ger	V.MSDR
loc	ESS	n	NEUT	voc	VOC
f	FEM	inf	NFIN	pact	V.PTCP;ACT
gen	GEN	nom	NOM	ppas	V.PTCP;PASS

Table 10: Simplified mapping from the original flexemic tagset of Polish used in Polish morphological analysers and corpora annotations (Przepiórkowski and Woliński, 2003) to the UniMorph schema. The mapping contains most of the POS and feature labels and does not allow to reconstruct the full conversion of the original data, as some mappings are conditional.

Xibe Universal Dependencies feature / word transliteration	UniMorph	Additional rules
ADJ	ADJ	
ADP	ADP	
ADV	ADV	
AUX	AUX	
CCONJ	CONJ	
DET	DET	
INTJ	INTJ	
NOUN	N	
NUM	NUM	
PART	PART	
PRON	PRO	
PROPN	PROPN	
PUNCT	_	excluding punctuation marks
SCONJ	CONJ	
SYM	_	excluding symbols
VERB	depends on other properties	
X		
_		
Abbr=Yes	_	
Aspect=Imp	IPFV	
Aspect=Perf	PFV	seems to be closer to PFV than to PRF
Aspect=Prog	PROG	
Case=Abl	ABL	not for adpositions
Case=Acc	ACC	not for adpositions
Case=Cmp	COMPV	not for adpositions
Case=Com	COM	not for adpositions
Case=Dat	DAT	not for adpositions
Case=Gen	GEN	not for adpositions
Case=Ins	INSTR	not for adpositions
Case=Lat	ALL	not for adpositions
Case=Loc	IN	not for adpositions
Clusivity=Ex	EXCL	
Clusivity=In	INCL	
Degree=Cmp	CMPR	
Degree=Pos	_	
Foreign=Yes	_	
Mood=Cnd	CMD=COND	for finite forms only
Mood=Imp	IMP	
Mood=Ind	IND	
Mood=Sub	SBJV	
NumType=Card	_	
NumType=Mult	POS=ADV	
NumType=Ord	POS=ADJ	
NumType=Sets	POS=ADJ	
Number=Plur	PL	
Number=Sing	SG	
Person=1	1	
Person=2	2	
Person=3	3	
Polarity=Neg	NEG	not for the negative auxiliary
Polite=Elev	_	
Poss=Yes	CMD=PSS	
PronType=Dem	CMD=DEIXIS	
PronType=Ind	_	
PronType=Int	_	
PronType=Prs	_	
PronType=Tot	_	
Reflex=Yes	_	
Tense=Fut	FUT	
Tense=Past	PST	
Tense=Pres	PRS	
Typo=Yes	_	not including typos into the resulting table

Table 11: Simplified mapping for the Xibe Universal Dependencies corpus (Pt. 1)

Xibe Universal Dependencies feature / word transliteration	UniMorph	Additional rules
VerbForm=Conv	POS=V.CVB	
VerbForm=Fin	FIN	
VerbForm=Inf	NFIN	
VerbForm=Part	POS=V.PTCP	
VerbForm=Vnoun	POS=V.MSDR	
Voice=Act	ACT	
Voice=Cau	CAUS	
Voice=Pass	PASS	
Voice=Rcp	RECP	
ateke	_	
dari	_	means ‘each, every’
eiten	_	means ‘each, every’
enteke	_	means ‘like this’
ere	PROX	
erebe	PROX	
ereci	PROX	
eremu	PROX	
geren	_	means ‘all’
harangga	_	
tenteke	_	means ‘like that’
terali	_	means ‘like that’
teralingge	_	means ‘like that’
tere	REMT	
terebe	REMT	
terei	REMT	
tesu	REMT	
tuba	_	means ‘there’
tuttu	_	means ‘like that’
uba	_	means ‘here’
ubaci	_	means ‘here’
ubai	_	means ‘here’
udu	_	means ‘some’
uttu	_	means ‘like this’

Table 12: Simplified mapping for the Xibe Universal Dependencies corpus (Pt. 2)

B Accuracy trends

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
afb	94.77	90.26	95.24	95.24	95.84	95.84
amh	89.67	87.09	94.83	94.83	94.83	94.83
ara	99.87	98.34	99.93	99.93	99.93	99.93
arz	95.65	91.39	97.31	97.31	97.07	97.07
syc	10.71	7.14	10.71	17.85	14.28	14.28
ind	80.15	69.26	85.60	85.60	84.43	84.43
kod	100.00	90.90	100.00	100.00	90.90	100.00
ckt	50.00	50.00	50.00	50.00	50.00	50.00
itl	50.00	58.33	66.66	58.33	66.66	58.33
bra	68.75	65.62	50.00	71.87	65.62	56.25
bul	99.73	96.85	100.00	100.00	99.96	99.96
ces	99.49	97.74	99.50	99.50	99.52	99.52
mag	69.23	84.61	76.92	92.30	92.30	84.61
nld	97.29	96.38	97.85	97.85	97.80	97.80
pol	99.91	99.67	99.95	99.95	100.00	100.00
rus	99.81	98.88	99.44	99.44	99.44	99.44
ail	12.50	12.50	0	12.50	12.50	12.50
evn	73.52	74.50	78.43	76.47	72.54	77.45
krl	100.00	90.69	93.02	93.02	93.02	93.02
olo	99.80	98.05	99.92	99.92	99.76	99.76
vep	99.85	97.86	99.82	99.82	99.88	99.88
sjo	66.66	66.66	66.66	100.00	100.00	100.00
tur	97.78	97.41	100.00	100.00	100.00	100.00

Table 13: Accuracy for "Adjective" on the test data.

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
syc	65.21	6.52	84.78	82.60	86.95	80.43
bul	99.60	97.90	100.00	100.00	100.00	100.00
ces	100.00	98.40	99.90	99.90	100.00	100.00
ckb	97.91	95.83	100.00	100.00	100.00	100.00
deu	94.89	91.72	95.91	95.91	96.52	96.52
kmr	100.00	100.00	100.00	100.00	100.00	100.00
nld	89.17	78.58	94.58	94.58	96.00	96.00
pol	100.00	99.92	100.00	100.00	100.00	100.00
por	99.83	98.96	99.67	99.67	99.78	99.78
rus	97.04	92.06	96.58	96.58	96.67	96.67
spa	99.93	99.03	99.35	99.35	99.29	99.29
evn	12.76	7.09	17.73	18.43	14.89	19.14
krl	100.00	98.18	100.00	100.00	100.00	100.00
olo	99.69	97.83	99.38	99.38	99.69	99.69
vep	99.02	96.67	99.21	99.21	99.21	99.21
sjo	22.22	22.22	27.77	55.55	55.55	50.00

Table 14: Accuracy for "Participle" on the test data.

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
amh	98.67	95.78	99.75	99.75	99.87	99.87
itl	19.04	13.09	25.00	20.23	21.42	21.42
bul	100.00	98.57	100.00	100.00	100.00	100.00
ces	98.97	95.47	100.00	100.00	99.38	99.38
pol	99.22	99.22	100.00	100.00	100.00	100.00
rus	99.21	97.49	97.96	97.96	99.68	99.68
spa	98.74	98.23	99.24	99.24	100.00	100.00
evn	23.38	16.12	25.00	32.25	27.41	33.06
sah	100.00	100.00	100.00	100.00	100.00	100.00
tyv	100.00	100.00	100.00	100.00	100.00	100.00
sjo	54.54	9.09	54.54	54.54	72.72	45.45

Table 15: Accuracy for "Converb" on the test data.

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
amh	88.61	79.67	95.12	95.12	97.56	97.56
heb	79.53	73.68	83.62	83.62	83.04	83.04
aym	100.00	100.00	100.00	100.00	100.00	100.00
itl	33.33	33.33	33.33	50.00	33.33	33.33
bul	98.85	98.00	100.00	100.00	100.00	100.00
kmr	99.37	100.00	98.74	98.74	100.00	100.00
pol	99.96	99.96	100.00	100.00	100.00	100.00
sjo	46.15	0	46.15	38.46	46.15	30.76

Table 16: Accuracy for "Masdar" on the test data.

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
afb	92.42	79.83	95.13	95.13	95.43	95.43
amh	98.36	91.56	99.72	99.72	99.72	99.72
ara	99.79	88.63	99.88	99.88	99.87	99.87
arz	93.31	78.08	95.16	95.16	94.98	94.98
heb	98.41	92.95	99.65	99.65	99.75	99.75
syc	11.02	4.41	25.73	21.32	27.94	25.00
ame	81.30	57.82	84.78	86.52	86.08	83.47
cni	98.73	79.51	99.72	99.72	99.63	99.63
ind	83.01	52.41	84.47	84.47	84.36	84.36
kod	93.65	92.06	100.00	98.41	100.00	100.00
aym	99.98	99.99	100.00	100.00	100.00	100.00
ckt	25.00	31.25	18.75	37.50	18.75	43.75
itl	14.96	12.92	20.40	25.17	21.08	23.12
gup	14.75	21.31	59.01	63.93	55.73	60.65
bra	31.30	29.56	24.34	27.82	27.82	30.43
bul	99.51	98.36	99.86	99.86	99.89	99.89
ces	98.88	94.97	99.54	99.54	99.40	99.40
ckb	99.72	96.44	99.96	99.96	99.96	99.96
deu	99.39	94.55	99.75	99.75	99.73	99.73
kmr	98.20	96.97	100.00	100.00	99.67	99.67
mag	36.36	38.63	42.04	42.04	44.31	39.77
nld	99.53	94.99	99.86	99.86	99.86	99.86
pol	99.57	98.22	99.76	99.76	99.74	99.74
por	99.84	99.12	99.91	99.91	99.86	99.86
rus	99.25	90.31	97.09	97.09	97.38	97.38
spa	99.82	97.55	99.90	99.90	99.92	99.92
see	78.27	40.97	90.65	89.64	90.00	88.63
ail	5.69	6.73	10.88	8.80	9.32	10.36
evn	34.70	32.03	44.90	44.66	44.90	46.35
sah	99.83	98.98	99.61	99.61	99.83	99.83
tyv	99.94	99.50	99.91	99.91	99.95	99.95
krl	99.94	98.82	99.94	99.94	99.94	99.94
lud	56.25	56.25	0	50.00	6.25	50.00
olo	99.84	99.14	99.71	99.71	99.70	99.70
vep	99.71	97.50	99.60	99.60	99.65	99.65
sjo	18.51	3.70	29.62	33.33	29.62	25.92
tur	99.96	99.98	99.17	99.17	99.17	99.17

Table 17: Accuracy for "Verb" in each language on the test data.

$\mathcal{L}$	BME	GUClasp	TRM	TRM+ AUG	CHR-TRM	CHR-TRM +AUG
afb	91.00	81.87	94.00	94.00	92.95	92.95
amh	98.46	96.30	99.24	99.24	99.31	99.31
ara	99.60	94.76	99.44	99.44	99.56	99.56
arz	96.58	91.66	97.56	97.56	97.23	97.23
heb	94.23	70.37	98.39	98.39	98.92	98.92
syc	20.00	18.57	32.85	34.28	32.14	32.85
ame	82.99	55.06	88.66	88.46	87.65	87.44
cni	99.79	98.62	99.96	99.96	99.96	99.96
ind	78.93	57.69	81.81	81.81	81.38	81.38
kod	94.73	84.21	100.00	100.00	100.00	100.00
aym	99.97	99.95	99.96	99.96	99.96	99.96
ckt	60.00	70.00	30.00	70.00	35.00	70.00
itl	64.22	66.97	71.55	71.55	72.47	72.47
bra	75.60	74.39	74.39	79.87	80.48	78.04
bul	95.27	89.54	97.92	97.92	97.47	97.47
ces	95.49	88.60	95.56	95.56	95.60	95.60
ckb	97.44	98.01	99.71	99.71	100.00	100.00
deu	96.93	89.64	95.48	95.48	95.51	95.51
kmr	98.20	98.12	97.91	97.91	97.91	97.91
mag	91.60	92.30	81.81	91.60	85.31	92.30
pol	96.30	89.71	97.07	97.07	97.35	97.35
rus	95.80	92.91	96.24	96.24	96.03	96.03
ail	9.67	4.83	17.74	20.96	14.51	20.96
evn	71.30	74.23	75.48	75.34	76.88	76.18
sah	99.97	99.78	99.97	99.97	99.99	99.99
tyv	99.98	99.93	99.96	99.96	99.97	99.97
krl	93.48	68.83	95.81	95.81	95.81	95.81
lud	61.90	61.90	28.57	42.85	42.85	42.85
olo	99.54	96.99	99.57	99.57	99.58	99.58
vep	99.72	96.50	99.66	99.66	99.70	99.70
sjo	58.33	41.66	25.00	58.33	25.00	66.66
tur	99.83	98.49	99.73	99.73	99.71	99.71
vro	94.78	87.39	97.82	98.26	97.82	97.39

Table 18: Accuracy for "Noun" in each language on the test data.