SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection

Ekaterina Vylomova[ə] **Jennifer White**[ʕ] **Elizabeth Salesky**[ʒ] **Sabrina J. Mielke**[ʒ] **Shijie Wu**[ʒ]
Edoardo Ponti[ʕ] **Rowan Hall Maudslay**[ʕ] **Ran Zmigrod**[ʕ] **Josef Valvoda**[ʕ] **Svetlana Toldova**[ɛ]
Francis Tyers[ɪ,ɛ] **Elena Klyachko**[ɛ] **Ilya Yegorov**[ɱ] **Natalia Krizhanovsky**[ʁ] **Paula Czarnowska**[ʕ]
Irene Nikkarinen[ʕ] **Andrew Krizhanovsky**[ʁ] **Tiago Pimentel**[ʕ] **Lucas Torroba Hennigen**[ʕ]
Christo Kirov[ɐ] **Garrett Nicolai**[ɓ] **Adina Williams**[ɸ] **Antonios Anastasopoulos**[ɬ]
Hilaria Cruz[ʎ] **Eleanor Chodroff**[ʏ] **Ryan Cotterell**[ʕ,ð] **Miikka Silfverberg**[ɓ] **Mans Hulden**[χ]
[ə]University of Melbourne [ʕ]University of Cambridge [ʒ]Johns Hopkins University
[ɛ]Higher School of Economics [ɱ]Moscow State University [ʁ]Karelian Research Centre
[ɐ]Google AI [ɓ]University of British Columbia [ɸ]Facebook AI Research
[ɬ]Carnegie Mellon University [ɪ]Indiana University [ʎ]University of Louisville
[ʏ]University of York [ð]ETH Zürich [χ]University of Colorado Boulder
ekaterina.vylomova@unimelb.edu.au ryan.cotterell@ethz.inf.ch

Abstract

A broad goal in natural language processing (NLP) is to develop a system that has the capacity to process any natural language. Most systems, however, are developed using data from just one language such as English. The SIGMORPHON 2020 shared task on morphological reinflection aims to investigate systems' ability to generalize across typologically distinct languages, many of which are low resource. Systems were developed using data from 45 languages and just 5 language families, fine-tuned with data from an additional 45 languages and 10 language families (13 in total), and evaluated on all 90 languages. A total of 22 systems (19 neural) from 10 teams were submitted to the task. All four winning systems were neural (two monolingual transformers and two massively multilingual RNN-based models with gated attention). Most teams demonstrate utility of data hallucination and augmentation, ensembles, and multilingual training for low-resource languages. Non-neural learners and manually designed grammars showed competitive and even superior performance on some languages (such as Ingrian, Tajik, Tagalog, Zarma, Lingala), especially with very limited data. Some language families (Afro-Asiatic, Niger-Congo, Turkic) were relatively easy for most systems and achieved over 90% mean accuracy while others were more challenging.

1 Introduction

Human language is marked by considerable diversity around the world. Though the world's languages share many basic attributes (e.g., Swadesh, 1950 and more recently, List et al., 2016), grammatical features, and even abstract implications (proposed in Greenberg, 1963), each language nevertheless has a unique evolutionary trajectory that is affected by geographic, social, cultural, and other factors. As a result, the surface form of languages varies substantially. The morphology of languages can differ in many ways: Some exhibit rich grammatical case systems (e.g., 12 in Erzya and 24 in Veps) and mark possessiveness, others might have complex verbal morphology (e.g., Oto-Manguean languages; Palancar and Léonard, 2016) or even "decline" nouns for tense (e.g., Tupi–Guarani languages). Linguistic typology is the discipline that studies these variations by means of a systematic comparison of languages (Croft, 2002; Comrie, 1989). Typologists have defined several dimensions of morphological variation to classify and quantify the degree of cross-linguistic variation. This comparison can be challenging as the categories are based on studies of known languages and are progressively refined with documentation of new languages (Haspelmath, 2007). Nevertheless, to understand the potential range of morphological variation, we take a closer look at three dimensions here: fusion, inflectional synthesis, and position of case affixes (Dryer and Haspelmath, 2013).

Fusion, our first dimension of variation, refers to the degree to which morphemes bind to one another in a phonological word (Bickel and Nichols, 2013b). Languages range from strictly isolating (i.e., each morpheme is its own phonological word) to concatenative (i.e., morphemes

Proceedings of the Seventeenth SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 1–39

https://doi.org/10.18653/v1/P17

bind together within a phonological word); nonlinearities such as ablaut or tonal morphology can also be present. From a geographic perspective, isolating languages are found in the Sahel Belt in West Africa, Southeast Asia and the Pacific. Ablaut–concatenative morphology and tonal morphology can be found in African languages. Tonal–concatenative morphology can be found in Mesoamerican languages (e.g., Oto-Manguean). Concatenative morphology is the most common system and can be found around the world. Inflectional synthesis, the second dimension considered, refers to whether grammatical categories like tense, voice or agreement are expressed as affixes (synthetic) or individual words (analytic) (Bickel and Nichols, 2013c). Analytic expressions are common in Eurasia (except the Pacific Rim, and the Himalaya and Caucasus mountain ranges), whereas synthetic expressions are used to a high degree in the Americas. Finally, affixes can variably surface as prefixes, suffixes, infixes, or circumfixes (Dryer, 2013). Most Eurasian and Australian languages strongly favor suffixation, and the same holds true, but to a lesser extent, for South American and New Guinean languages (Dryer, 2013). In Mesoamerican languages and African languages spoken below the Sahara, prefixation is dominant instead.

These are just three dimensions of variation in morphology, and the cross-linguistic variation is already considerable. Such cross-lingual variation makes the development of natural language processing (NLP) applications challenging. As Bender (2009, 2016) notes, many current architectures and training and tuning algorithms still present language-specific biases. The most commonly used language for developing NLP applications is English. Along the above dimensions, English is productively concatenative, a mixture of analytic and synthetic, and largely suffixing in its inflectional morphology. With respect to languages that exhibit inflectional morphology, English is relatively impoverished.[1] Importantly, English is just one morphological system among many. A larger goal of natural language processing is that the system work for *any* presented language. If an NLP system is trained on just one language, it could be missing important flexibility in its ability to account for cross-linguistic morphological variation.

[1]Note that many languages exhibit no inflectional morphology e.g., Mandarin Chinese, Yoruba, etc.: Bickel and Nichols (2013a).

In this year's iteration of the SIGMORPHON shared task on morphological reinflection, we specifically focus on typological diversity and aim to investigate systems' ability to generalize across typologically distinct languages many of which are low-resource. For example, if a neural network architecture works well for a sample of Indo-European languages, should the same architecture also work well for Tupi–Guarani languages (where nouns are "declined" for tense) or Austronesian languages (where verbal morphology is frequently prefixing)?

2 Task Description

The 2020 iteration of our task is similar to CoNLL-SIGMORPHON 2017 (Cotterell et al., 2017) and 2018 (Cotterell et al., 2018) in that participants are required to design a model that learns to generate inflected forms from a lemma and a set of morphosyntactic features that derive the desired target form. For each language we provide a separate training, development, and test set. More historically, all of these tasks resemble the classic "wug"-test that Berko (1958) developed to test child and human knowledge of English nominal morphology.

Unlike the task from earlier years, this year's task proceeds in three phases: a Development Phase, a Generalization Phase, and an Evaluation Phase, in which each phase introduces previously unseen data. The task starts with the **Development Phase**, which was an elongated period of time (about two months), during which participants *develop* a model of morphological inflection. In this phase, we provide training and development splits for 45 languages representing the Austronesian, Niger-Congo, Oto-Manguean, Uralic and Indo-European language families. Table 1 provides details on the languages. The **Generalization Phase** is a short period of time (it started about a week before the Evaluation Phase) during which participants fine-tune their models on new data. At the start of the phase, we provide training and development splits for 45 new languages where approximately half are genetically related (belong to the same family) and half are genetically unrelated (are isolates or belong to a different family) to the languages presented in the Development Phase. More specifically, we introduce (surprise) languages from Afro-Asiatic, Algic, Dravidian, Indo-European, Niger-Congo, Sino-Tibetan,

Siouan, Songhay, Southern Daly, Tungusic, Turkic, Uralic, and Uto-Aztecan families. See Table 2 for more details.

Finally, test splits for all 90 languages are released in the **Evaluation Phase**. During this phase, the models are evaluated on held-out forms. Importantly, the languages from both previous phases are evaluated simultaneously. This way, we evaluate the extent to which models (especially those with shared parameters) overfit to the development data: a model based on the morphological patterning of the Indo-European languages may end up with a bias towards suffixing and will struggle to learn prefixing or infixation.

3 Meet our Languages

In the 2020 shared task we cover 15 language families: Afro-Asiatic, Algic, Austronesian, Dravidian, Indo-European, Niger-Congo, Oto-Manguean, Sino-Tibetan, Siouan, Songhay, Southern Daly, Tungusic, Turkic, Uralic, and Uto-Aztecan.[2] Five language families were used for the Development phase while ten were held out for the Generalization phase. Tab. 1 and Tab. 2 provide information on the languages, their families, and sources of data. In the following section, we provide an overview of each language family's morphological system.

3.1 Afro-Asiatic

The Afro-Asiatic language family, consisting of six branches and over 300 languages, is among the largest language families in the world. It is mainly spoken in Northern, Western and Central Africa as well as West Asia and spans large modern languages such as Arabic, in addition to ancient languages like Biblical Hebrew. Similarly, some of its languages have a long tradition of written form, while others have yet to incorporate a writing system. The six branches differ most notably in typology and syntax, with the Chadic language being the main source of differences, which has sparked discussion of the division of the family (Frajzyngier, 2018). For example, in the Egyptian and Semitic branches, the root of a verb may not contain vowels, while this is allowed in Chadic. Although only four of the six branches, excluding Chadic and Omotic, use a prefix and suffix in conjugation when adding a subject to a verb, it is considered an important characteristic of the family. In addition, some of the families in the phylum use tone to encode tense, modality and number among others. However, all branches use objective and passive suffixes. Markers of tense are generally simple, whereas aspect is typically distinguished with more elaborate systems.

3.2 Algic

The Algic family embraces languages native to North America—more specifically the United States and Canada—and contain three branches. Of these, our sample contains Cree, the language from the largest genus, Algonquian, from which most languages are now extinct. The Algonquian genus is characterized by its concatenative morphology. Cree morphology is also concatenative and suffixing. It distinguishes between impersonal and non-impersonal verbs and presents four apparent declension classes among non-impersonal verbs.

3.3 Austronesian

The Austronesian family of languages is largely comprised of languages from the Greater Central Philippine and Oceanic regions. They are characterized by limited morphology, mostly prefixing in nature. Additionally, tense–aspect affixes are predominantly seen as prefixes, though some suffixes are used. In the general case, verbs do not mark number, person, or gender. In Māori, verbs may be suffixed with a marker indicating the passive voice. This marker takes the form of one of twelve endings. These endings are difficult to predict as the language has undergone a loss of word-final consonants and there is no clear link between a stem and the passive suffix that it employs (Harlow, 2007).

3.4 Dravidian

The family of Dravidian languages comprises several languages which are primarily spoken across Southern India and Northern Sri Lanka, with over 200 million speakers. The shared task includes Kannada and Telugu. Dravidian languages primarily use the SOV word order. They are agglutinative, and primarily use suffixes. A Dravidian verb indicates voice, number, tense, aspect, mood and person, through the affixation of multiple suffixes. Nouns indicate number, gender and case.

[2] The data splits are available at https://github.com/sigmorphon2020/task0-data/

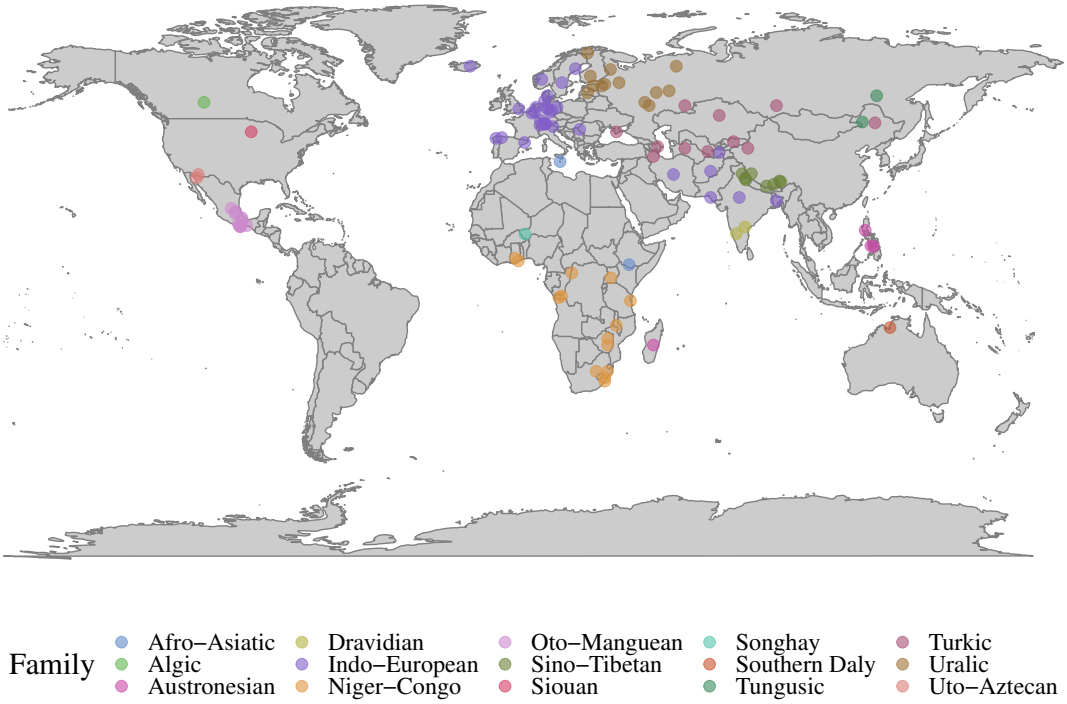

Figure 1: Languages in our sample colored by family.

3.5 Indo-European

Languages in the Indo-European family are native to most of Europe and a large part of Asia—with our sample including languages from the genera: Germanic, Indic, Iranian, and Romance. This is (arguably) the most well studied language family, containing a few of the highest-resource languages in the world.

Romance The Romance genus comprises of a set of fusional languages evolved from Latin. They traditionally originated in Southern and Southeastern Europe, though they are presently spoken in other continents such Africa and the Americas. Romance languages mark tense, person, number and mood in verbs, and gender and number in nouns. Inflection is primarily achieved through suffixes, with some verbal person syncretism and suppletion for high-frequency verbs. There is some morphological variation within the genus, such as French, which exhibits comparatively less inflection, and Romanian has comparatively more—it still marks case.

Germanic The Germanic genus comprises several languages which originated in Northern and Northwestern Europe, and today are spoken in many parts of the world. Verbs in Germanic languages mark tense and mood, in many languages person and number are also marked, predominantly through suffixation. Some Germanic languages exhibit widespread Indo-European ablaut. The gendering of nouns differs between Germanic languages: German nouns can be masculine, feminine or neuter, while English nouns are not marked for gender. In Danish and Swedish, historically masculine and feminine nouns have merged to form one common gender, so nouns are either common or neuter. Marking of case also differs between the languages: German nouns have one of four cases and this case is marked in articles and adjectives as well as nouns and pronouns, while English does not mark noun case (although Old English, which also appears in our language sample, does).

Indo-Iranian The Indo-Iranian genus contains languages spoken in Iran and across the Indian subcontinent. Over 1.5 billion people worldwide speak an Indo-Iranian language. Within the Indo-European family, Indo-Iranian languages belong to the Satem group of languages. Verbs in Indo-Iranian languages indicate tense, aspect, mood, number and person. In languages such as Hindi verbs can also express levels of formality. Noun gender is present in some Indo-Iranian languages, such as Hindi, but absent in languages such as Persian. Nouns generally are marked for case.

Development

Family	Genus	ISO 639-3	Language	Source of Data
Austronesian	Barito	mlg (plt)	Malagasy	Kasahorow (2015a)
	Greater Central Philippine	ceb	Cebuano	Reyes (2015)
	Greater Central Philippine	hil	Hiligaynon	Santos (2018)
	Greater Central Philippine	tgl	Tagalog	NIU (2017)
	Oceanic	mao (mri)	Māori	Moorfield (2019)
Indo-European	Germanic	ang	Old English	UniMorph
	Germanic	dan	Danish	UniMorph
	Germanic	deu	German	UniMorph
	Germanic	eng	English	UniMorph
	Germanic	frr	North Frisian	UniMorph
	Germanic	gmh	Middle High German	UniMorph
	Germanic	isl	Icelandic	UniMorph
	Germanic	nld	Dutch	UniMorph
	Germanic	nob	Norwegian Bokmål	UniMorph
	Germanic	swe	Swedish	UniMorph
Niger-Congo	Bantoid	kon (kng)	Kongo	Kasahorow (2016)
	Bantoid	lin	Lingala	Kasahorow (2014a)
	Bantoid	lug	Luganda	Namono (2018)
	Bantoid	nya	Chewa	Kasahorow (2019a)
	Bantoid	sot	Sotho	Kasahorow (2020)
	Bantoid	swa (swh)	Swahili	Kasahorow (2012b)
	Bantoid	zul	Zulu	Kasahorow (2015b)
	Kwa	aka	Akan	Imbeah (2012)
	Kwa	gaa	Gã	Kasahorow (2012a)
Oto-Manguean	Amuzgoan	azg	San Pedro Amuzgos Amuzgo	Feist and Palancar (2015)
	Chichimec	pei	Chichimeca-Jonaz	Feist and Palancar (2015)
	Chinantecan	cpa	Tlatepuzco Chinantec	Feist and Palancar (2015)
	Mixtecan	xty	Yoloxóchitl Mixtec	Feist and Palancar (2015)
	Otomian	ote	Mezquital Otomi	Feist and Palancar (2015)
	Otomian	otm	Sierra Otomi	Feist and Palancar (2015)
	Zapotecan	cly	Eastern Chatino of San Juan Quiahije	Cruz et al. (2020)
	Zapotecan	ctp	Eastern Chatino of Yaitepec	Feist and Palancar (2015)
	Zapotecan	czn	Zenzontepec Chatino	Feist and Palancar (2015)
	Zapotecan	zpv	Chichicapan Zapotec	Feist and Palancar (2015)
Uralic	Finnic	est	Estonian	UniMorph
	Finnic	fin	Finnish	UniMorph
	Finnic	izh	Ingrian	UniMorph
	Finnic	krl	Karelian	Zaytseva et al. (2017)
	Finnic	liv	Livonian	UniMorph
	Finnic	vep	Veps	Zaytseva et al. (2017)
	Finnic	vot	Votic	UniMorph
	Mari	mhr	Meadow Mari	Arkhangelskiy et al. (2012)
	Mordvin	mdf	Moksha	Arkhangelskiy et al. (2012)
	Mordvin	myv	Erzya	Arkhangelskiy et al. (2012)
	Saami	sme	Northern Sami	UniMorph

Table 1: Development languages used in the shared task.

3.6 Niger–Congo

Our language sample includes two genera from the Niger–Congo family, namely Bantoid and Kwa languages. These have mostly exclusively concatenative fusion, and single exponence in verbal tense–aspect–mood. The inflectional synthesis of verbs is moderately high, e.g. with 4-5 classes per word in Swahili and Zulu. The locus of marking is inconsistent (it falls on both heads and dependents), and most languages are are predominantly prefixing. Full and partial reduplication is attested in most languages. Verbal person–number markers tend to be syncretic.

As for nominal classes, Bantoid languages are characterized by a large amount of grammatical genders (often more than 5) assigned based on both semantic and formal rules, whereas some Akan languages (like Ewe) lack a gender system. Plural tends to be always expressed by affixes or other morphological means. Case marking is generally absent or minimal. As for verbal classes, aspect is grammaticalized in Akhan (Kwa) and Zulu (Bantoid), but not in Luganda and Swahili (Bantoid). Both past and future tenses are inflectional in Bantoid languages. 2-3 degrees of remoteness can be distinguished in Zulu and Luganda, but not in Swahili. On the other hand, Akan (Kwa) has no opposition between past and non-past. There are

Generalization (Surprise)

Family	Genus	ISO 639-3	Language	Source of Data
Afro-Asiatic	Semitic	mlt	Maltese	UniMorph
	Lowland East Cushitic	orm	Oromo	Kasahorow (2017)
	Semitic	syc	Syriac	UniMorph
Algic	Algonquian	cre	Plains Cree	Hunter (1923)
Tungusic	Tungusic	evn	Evenki	Klyachko et al. (2020)
Turkic	Turkic	aze (azb)	Azerbaijani	UniMorph
	Turkic	bak	Bashkir	UniMorph
	Turkic	crh	Crimean Tatar	UniMorph
	Turkic	kaz	Kazakh	Nabiyev (2015); Turkicum (2019a)
	Turkic	kir	Kyrgyz	Aytnatova (2016)
	Turkic	kjh	Khakas	UniMorph
	Turkic	tuk	Turkmen	Abdulin (2016); US Embassy (2018)
	Turkic	uig	Uyghur	Kadeer (2016)
	Turkic	uzb	Uzbek	Abdullaev (2016); Turkicum (2019b)
Dravidian	Southern Dravidian	kan	Kannada	UniMorph
	South-Central Dravidian	tel	Telugu	UniMorph
Indo-European	Indic	ben	Bengali	UniMorph
	Indic	hin	Hindi	UniMorph
	Indic	san	Sanskrit	UniMorph
	Indic	urd	Urdu	UniMorph
	Iranian	fas (pes)	Persian	UniMorph
	Iranian	pus (pst)	Pashto	UniMorph
	Iranian	tgk	Tajik	UniMorph
	Romance	ast	Asturian	UniMorph
	Romance	cat	Catalan	UniMorph
	Romance	frm	Middle French	UniMorph
	Romance	fur	Friulian	UniMorph
	Romance	glg	Galician	UniMorph
	Romance	lld	Ladin	UniMorph
	Romance	vec	Venetian	UniMorph
	Romance	xno	Anglo-Norman	UniMorph
	West Germanic	gml	Middle Low German	UniMorph
	West Germanic	gsw	Swiss German	Egli-Wildi (2007)
	North Germanic	nno	Norwegian Nynorsk	UniMorph
Niger-Congo	Bantoid	sna	Shona	Kasahorow (2014b); Nandoro (2018)
Sino-Tibetan	Bodic	bod	Tibetan	Di et al. (2019)
Siouan	Core Siouan	dak	Dakota	LaFontaine and McKay (2005)
Songhay	Songhay	dje	Zarma	Kasahorow (2019b)
Southern Daly	Murrinh-Patha	mwf	Murrinh-Patha	Mansfield (2019)
Uralic	Permic	kpv	Komi-Zyrian	Arkhangelskiy et al. (2012)
	Finnic	lud	Ludic	Zaytseva et al. (2017)
	Finnic	olo	Livvi	Zaytseva et al. (2017)
	Permic	udm	Udmurt	Arkhangelskiy et al. (2012)
	Finnic	vro	Võro	Iva (2007)
Uto-Aztecan	Tepiman	ood	O'odham	Zepeda (2003)

Table 2: Surprise languages used in the shared task.

no grammatical evidentials.

3.7 Oto-Manguean

The Oto-Manguean languages are a diverse family of tonal languages spoken in central and southern Mexico. Even though all of these languages are tonal, the tonal system within each language varies widely. Some have an inventory of two tones (e.g., Chichimec and Pame) others have ten tones (e.g., the Eastern Chatino languages of the Zapotecan branch, Palancar and Léonard (2016)).

Oto-Manguean languages are also rich in tonal morphology. The inflectional system marks person–number and aspect in verbs and person–number in adjectives and noun possessions, relying heavily on tonal contrasts. Other interesting as-

pects of Oto-Manguean languages include the fact that pronominal inflections use a system of enclitics, and first and second person plural has a distinction between exclusive and inclusive (Campbell, 2016). Tone marking schemes in the writing systems also vary greatly. Some writing systems do not represent tone, others use diacritics, and others represent tones with numbers. In languages that use numbers, single digits represent level tones and double digits represent contour tones. For example, in San Juan Quiahije of Eastern Chatino number 1 represents high tone, number 4 represents low tone, and numbers 14 represent a descending tone contour and numbers 42 represent an ascending tone contour Cruz (2014).

3.8 Sino-Tibetan

The Sino-Tibetan family is represented by the Tibetan language. Tibetan uses an abugida script and contains complex syllabic components in which vowel marks can be added above and below the base consonant. Tibetan verbs are inflected for tense and mood. Previous studies on Tibetan morphology (Di et al., 2019) indicate that the majority of mispredictions produced by neural models are due to allomorphy. This is followed by generation of nonce words (impossible combinations of vowel and consonant components).

3.9 Siouan

The Siouan languages are located in North America, predominantly along the Mississippi and Missouri Rivers and in the Ohio Valley. The family is represented in our task by Dakota, a critically endangered language spoken in North and South Dakota, Minnesota, and Saskatchewan. The Dakota language is largely agglutinating in its derivational morphology and fusional in its inflectional morphology with a mixed affixation system (Rankin et al., 2003). The present task includes verbs, which are marked for first and second person, number, and duality. All three affixation types are found: person was generally marked by an infix, but could also appear as a prefix, and plurality was marked by a suffix. Morphophonological processes of fortition and vowel lowering are also present.

3.10 Songhay

The Songhay family consists of around eleven or twelve languages spoken in Mali, Niger, Benin, Burkina Faso and Nigeria. In the shared task we use Zarma, the most widely spoken Songhay language. Most of the Songhay languages are predominantly SOV with medium-sized consonant inventories (with implosives), five phonemic vowels, vowel length distinctions, and word level tones, which also are used to distinguish nouns, verbs, and adjectives (Heath, 2014).

3.11 Southern Daly

The Southern Daly is a small language family of the Northern Territory in Australia that consists of two distantly related languages. In the current task we only have one of the languages, Murrinh-patha (which was initially thought to be a language isolate). Murrinh-patha is classified as polysynthetic with highly complex verbal morphology. Verbal roots are surrounded by prefixes and suffixes that indicate tense, mood, object, subject. As Mansfield (2019) notes, Murrinh-patha verbs have 39 conjugation classes.

3.12 Tungusic

Tungusic languages are spoken principally in Russia, China and Mongolia. In Russia they are concentrated in north and eastern Siberia and in China in the east, in Manchuria. The largest languages in the family are Xibe, Evenki and Even; we use Evenki in the shared task. The languages are of the agglutinating morphological type with a moderate number of cases, 7 for Xibe and 13 for Evenki. In addition to case markers, Evenki marks possession in nominals (including reflexive possession) and distinguishes between alienable and inalienable possession. In terms of morphophonological processes, the languages exhibit vowel harmony, consonant alternations and phonological vowel length.

3.13 Turkic

Languages of the Turkic family are primarily spoken in Central Asia. The family is morphologically concatenative, fusional, and suffixing. Turkic languages generally exhibit back vowel harmony, with the notable exception of Uzbek. In addition to harmony in backness, several languages also have labial vowel harmony (e.g., Kyrgyz, Turkmen, among others). In addition, most of the languages have dorsal consonant allophony that accompanies back vowel harmony. Additional morphophonological processes include vowel epenthesis and voicing assimilation. Selection of the inflectional allomorph can frequently be determined

from the infinitive morpheme (which frequently reveals vowel backness and roundedness) and also the final segment of the stem.

3.14 Uralic

The Uralic languages are spoken in Russia from the north of Siberia to Scandinavia and Hungary in Europe. They are agglutinating with some subgroups displaying fusional characteristics (e.g., the Sámi languages). Many of the languages have vowel harmony. The languages have almost complete suffixal morphology and a medium-sized case inventory, ranging from 5–6 cases to numbers in the high teens. Many of the larger case paradigms are made up of spatial cases, sometimes with distinctions for direction and position. Most of the languages have possessive suffixes, which can express possession, or agreement in non-finite clauses. The paradigms are largely regular, with few, if any, irregular forms. Many exhibit complex patterns of consonant gradation—consonant mutations that occur in specific morphological forms in some stems. Which gradation category a stem belongs to in often unpredictable. The languages spoken in Russia are typically SOV, while those in Europe have SVO order.

3.15 Uto-Aztecan

The Uto-Aztecan family is represented by the Tohono O'odham (Papago–Pima) language spoken along the US–Mexico border in southern Arizona and northern Sonora. O'odham is agglutinative with a mixed prefixing and suffixing system. Nominal and verbal pluralization is frequently realized by partial reduplication of the initial consonant and/or vowel, and occasionally by final consonant deletion or null affixation. Processes targeting vowel length (shortening or lengthening) are also present. A small number of verbs exhibit suppletion in the past tense.

4 Data Preparation

4.1 Data Format

Similar to previous years, training and development sets contain triples consisting of a lemma, a target form, and morphosyntactic descriptions (MSDs, or morphological tags).[3] Test sets only contain two fields, i.e., target forms are omitted. All data follows UTF-8 encoding.

[3]Each MSD is a set of features separated by semicolons.

4.2 Conversion and Canonicalization

A significant amount of data for this task was extracted from corresponding (language-specific) grammars. In order to allow cross-lingual comparison, we manually converted their features (tags) into the UniMorph format (Sylak-Glassman, 2016). We then canonicalized the converted language data[4] to make sure all tags are consistently ordered and no category (e.g., "Number") is assigned two tags (e.g., singular and plural).[5]

4.3 Splitting

We use only noun, verb, and adjective forms to construct training, development, and evaluation sets. We de-duplicate annotations such that there are no multiple examples of exact lemma-form-tag matches. To create splits, we randomly sample 70%, 10%, and 20% for train, development, and test, respectively. We cap the training set size to 100k examples for each language; where languages exceed this (e.g., Finnish), we subsample to this point, balancing lemmas such that all forms for a given lemma are either included or discarded. Some languages such as Zarma (dje), Tajik (tgk), Lingala (lin), Ludian* (lud), Māori (mao), Sotho (sot), Võro (vro), Anglo-Norman (xno), and Zulu (zul) contain less than 400 training samples and are extremely low-resource.[6] Tab. 6 and Tab. 7 in the Appendix provide the number of samples for every language in each split, the number of samples per lemma, and statistics on inconsistencies in the data.

5 Baseline Systems

The organizers provided two types of pre-trained baselines. Their use was optional.

5.1 Non-neural

The first baseline was a non-neural system that had been used as a baseline in earlier shared tasks on morphological reinflection (Cotterell et al., 2017, 2018). The system first heuristically extracts lemma-to-form transformations; it assumes that these transformations are suffix- or prefix-based.

[4]Using the UniMorph schema canonicalization script https://github.com/unimorph/um-canonicalize

[5]Conversion schemes and canonicalization scripts are available at https://github.com/sigmorphon2020/task0-data

[6]We also note that Ludian contained inconsistencies in data due to merge of various dialects.

A simple majority classifier is used to apply the most frequent suitable transformation to an input lemma, given the morphological tag, yielding the output form. See Cotterell et al. (2017) for further details.

5.2 Neural

Neural baselines were based on a neural transducer (Wu and Cotterell, 2019), which is essentially a hard monotonic attention model (`mono-*`). The second baseline is a transformer (Vaswani et al., 2017) adopted for character-level tasks that currently holds the state-of-the-art on the 2017 SIGMORPHON shared task data (Wu et al., 2020, `trm-*`). Both models take the lemma and morphological tags as input and output the target inflection. The baseline is further expanded to include the data augmentation technique used by Anastasopoulos and Neubig (2019, `-aug-`) (conceptually similar to the one proposed by Silfverberg et al. (2017)). Relying on a simple character-level alignment between lemma and form, this technique replaces shared substrings of length $>$ 3 with random characters from the language's alphabet, producing hallucinated lemma–tag–form triples. Both neural baselines were trained in mono- (`*-single`) and multilingual (shared parameters among the same family, `*-shared`) settings.

6 Competing Systems

As Tab. 3 shows, 10 teams submitted 22 systems in total, out of which 19 were neural. Some teams such as **ETH Zurich** and **UIUC** built their models on top of the proposed baselines. In particular, **ETH Zurich** enriched each of the (multilingual) neural baseline models with exact decoding strategy that uses Dijkstra's search algorithm. **UIUC** enriched the transformer model with synchronous bidirectional decoding technique (Zhou et al., 2019) in order to condition the prediction of an affix character on its environment from both sides. (The authors demonstrate positive effects in Oto-Manguean, Turkic, and some Austronesian languages.)

A few teams further improved models that were among top performers in previous shared tasks. **IMS** and **Flexica** re-used the hard monotonic attention model from (Aharoni and Goldberg, 2017). **IMS** developed an ensemble of two models (with left-to-right and right-to-left generation order) with a genetic algorithm for ensemble search (Haque et al., 2016) and iteratively provided hallucinated data. **Flexica** submitted two neural systems. The first model (`flexica-02-1`) was multilingual (family-wise) hard monotonic attention model with improved alignment strategy. This model is further improved (`flexica-03-1`) by introducing a data hallucination technique which is based on phonotactic modelling of extremely low-resource languages (Shcherbakov et al., 2016). **LTI** focused on their earlier model (Anastasopoulos and Neubig, 2019), a neural multi-source encoder–decoder with two-step attention architecture, training it with hallucinated data, cross-lingual transfer, and romanization of scripts to improve performance on low-resource languages. **DeepSpin** reimplemented gated sparse two-headed attention model from Peters and Martins (2019) and trained it on all languages at once (massively multilingual). The team experimented with two modifications of the softmax function: sparsemax (Martins and Astudillo, 2016, `deepspin-02-1`) and 1.5-entmax (Peters et al., 2019, `deepspin-01-1`).

Many teams based their models on the transformer architecture. **NYU-CUBoulder** experimented with a vanilla transformer model (`NYU-CUBoulder-04-0`), a pointer-generator transformer that allows for a copy mechanism (`NYU-CUBoulder-02-0`), and ensembles of three (`NYU-CUBoulder-01-0`) and five (`NYU-CUBoulder-03-0`) pointer-generator transformers. For languages with less than 1,000 training samples, they also generate hallucinated data. **CULing** developed an ensemble of three (monolingual) transformers with identical architecture but different input data format. The first model was trained on the initial data format (lemma, target tags, target form). For the other two models the team used the idea of lexeme's principal parts (Finkel and Stump, 2007) and augmented the initial input (that only used the lemma as a source form) with entries corresponding to other (non-lemma) slots available for the lexeme. The **CMU Tartan** team compared performance of models with transformer-based and LSTM-based encoders and decoders. The team also compared monolingual to multilingual training in which they used several (related and unrelated) high-resource languages for low-resource language training.

Although the majority of submitted systems

Team	Description	System	Model Features			
			Neural	Ensemble	Multilingual	Hallucination
Baseline	Wu and Cotterell (2019)	mono-single	✓			
		mono-aug-single	✓			✓
		mono-shared	✓		✓	
		mono-aug-shared	✓		✓	✓
	Wu et al. (2020)	trm-single	✓			
		trm-aug-single	✓			✓
		trm-shared	✓		✓	
		trm-aug-shared	✓		✓	✓
CMU Tartan	Jayarao et al. (2020)	cmu_tartan_00-0	✓			✓
		cmu_tartan_00-1	✓		✓	✓
		cmu_tartan_01-0	✓			✓
		cmu_tartan_01-1	✓		✓	✓
		cmu_tartan_02-1	✓		✓	✓
CU7565	Beemer et al. (2020)	CU7565-01-0				
		CU7565-02-0				
CULing	Liu and Hulden (2020)	CULing-01-0	✓	✓		
DeepSpin	Peters and Martins (2020)	deepspin-01-1	✓		✓	
		deepspin-02-1	✓		✓	
ETH Zurich	Forster and Meister (2020)	ETHZ00-1	✓		✓	
		ETHZ02-1	✓		✓	
Flexica	Scherbakov (2020)	flexica-01-0				
		flexica-02-1	✓		✓	
		flexica-03-1	✓		✓	✓
IMS	Yu et al. (2020)	IMS-00-0	✓	✓		✓
LTI	Murikinati and Anastasopoulos (2020)	LTI-00-1	✓		✓	✓
NYU-CUBoulder	Singer and Kann (2020)	NYU-CUBoulder-01-0	✓	✓		✓
		NYU-CUBoulder-02-0	✓			✓
		NYU-CUBoulder-03-0	✓	✓		✓
		NYU-CUBoulder-04-0	✓			✓
UIUC	Canby et al. (2020)	uiuc-01-0	✓			

Table 3: The list of systems submitted to the shared task.

were neural, some teams experimented with non-neural approaches showing that in certain scenarios they might surpass neural systems. A large group of researchers from **CU7565** manually developed finite-state grammars for 25 languages (CU7565-01-0). They additionally developed a non-neural learner for all languages (CU7565-02-0) that uses hierarchical paradigm clustering (based on similarity of string transformation rules between inflectional slots). Another team, **Flexica**, proposed a model (flexica-01-0) conceptually similar to Hulden et al. (2014), although they did not attempt to reconstruct the paradigm itself and treated transformation rules independently assigning each of them a score based on its frequency and specificity as well as diversity of the characters surrounding the pattern.[7]

7 Evaluation

This year, we instituted a slightly different evaluation regimen than in previous years, which takes into account the statistical significance of differences between systems and allows for an informed comparison across languages and families better than a simple macro-average.

The process works as follows:

1. For each language, we rank the systems according to their accuracy (or Levenshtein distance). To do so, we use paired bootstrap resampling (Koehn, 2004)[8] to only take statistically significant differences into account. That way, any system which is the same (as assessed via statistical significance) as the best performing one is also ranked 1st for that language.

2. For the set of languages where we want collective results (e.g. languages within a linguistic genus), we aggregate the systems' ranks and

[7] English plural noun formation rule "* → *s" has high diversity whereas past tense rule such as "*a* → *oo*" as in *(understand, understood)* has low diversity.

[8] We use 10,000 samples with 50% ratio, and $p < 0.005$.

Individual Language Rankings								Final Ranking					
cly		ctp		czn		zpv			avg	#1	#3	#4	#6
uiuc	(1)	CULing	(1)	deepspin	(1)	NYU-CUB	(1)	uiuc	1	4			
trm-single	(1)	uiuc	(1)	uiuc	(1)	CULing	(1)	trm-single	1	4			
CULing	(3)	trm-single	(1)	IMS	(1)	deepspin	(1)	CULing	1.5	3	1		
deepspin	(3)	IMS	(4)	NYU-CUB	(1)	uiuc	(1)	deepspin	2.25	2	1	1	
NYU-CUB	(3)	deepspin	(4)	CULing	(1)	trm-single	(1)	NYU-CUB	2.25	2	1	1	
IMS	(6)	NYU-CUB	(4)	trm-single	(1)	IMS	(1)	IMS	3	2	0	1	1

Table 4: Illustration of our ranking method, over the four Zapotecan languages. Note: The final ranking is based on the actual counts (#1,#2, etc), not on the system's average rank.

re-rank them based on the amount of times they ranked 1st, 2nd, 3rd, etc.

Table 4 illustrates an example of this process using four Zapotecan languages and six systems.

8 Results

This year we had four winning systems (i.e., ones that outperform the best baseline): `CULing-01-0`, `deepspin-02-1`, `uiuc-01-0`, and `deepspin-01-1`, all neural. As Tab. 5 shows, they achieve over 90% accuracy. Although `CULing-01-0` and `uiuc-01-0` are both monolingual transformers that do not use any hallucinated data, they follow different strategies to improve performance. The strategy proposed by `CULing-01-0` of enriching the input data with extra entries that included non-lemma forms and their tags as a source form, enabled their system to be among top performers on all language families; `uiuc-01-0`, on the other hand, did not modify the data but rather changed the decoder to be bidirectional and made family-wise fine-tuning of each (monolingual) model. The system is also among the top performers on all language families except Iranian. The third team, **DeepSpin**, trained and fine-tuned their models on all language data. Both models are ranked high (although the sparsemax model, `deepspin-02-1`, performs better overall) on most language groups with exception of Algic. Sparsemax was also found useful by **CMU-Tartan**. The neural ensemble model with data augmentation from **IMS** team shows superior performance on languages with smaller data sizes (under 10,000 samples). **LTI** and **Flexica** teams also observed positive effects of multilingual training and data hallucination on low-resource languages. The latter was also found useful in the ablation study made by **NYU-CUBoulder** team. Several teams aimed to address particular research questions; we will further summarize their results.

System	Rank	Acc
uiuc-01-0	**2.4**	**90.5**
deepspin-02-1	2.9	**90.9**
BASE: trm-single	2.8	90.1
CULing-01-0	3.2	**91.2**
deepspin-01-1	3.8	**90.5**
BASE: trm-aug-single	3.7	90.3
NYU-CUBoulder-04-0	7.1	88.8
NYU-CUBoulder-03-0	8.9	88.8
NYU-CUBoulder-02-0	8.9	88.7
IMS-00-0	10.6	89.2
NYU-CUBoulder-01-0	9.6	88.6
BASE: trm-shared	10.3	85.9
BASE: mono-aug-single	7.5	88.8
cmu_tartan_00-0	8.7	87.1
BASE: mono-single	7.9	85.8
cmu_tartan_01-1	9.0	87.1
BASE: trm-aug-shared	12.5	86.5
BASE: mono-shared	10.8	86.0
cmu_tartan_00-1	9.4	86.5
LTI-00-1	12.0	86.6
BASE: mono-aug-shared	12.8	86.8
cmu_tartan_02-1	10.6	86.1
cmu_tartan_01-0	10.9	86.6
flexica-03-1	16.7	79.6
ETHZ-00-1	20.1	75.6
**CU7565-01-0*	*24.1*	*90.7*
flexica-02-1	17.1	78.5
**CU7565-02-0*	*19.2*	*83.6*
ETHZ-02-1	17.0	80.9
flexica-01-0	24.4	70.8
Oracle (Baselines)		96.1
Oracle (Submissions)		97.7
Oracle (All)		97.9

Table 5: Aggregate results on all languages. **Bolded** results are the ones which beat the best baseline. $*$ and *italics* denote systems that did not submit outputs in all languages (their accuracy is a partial average).

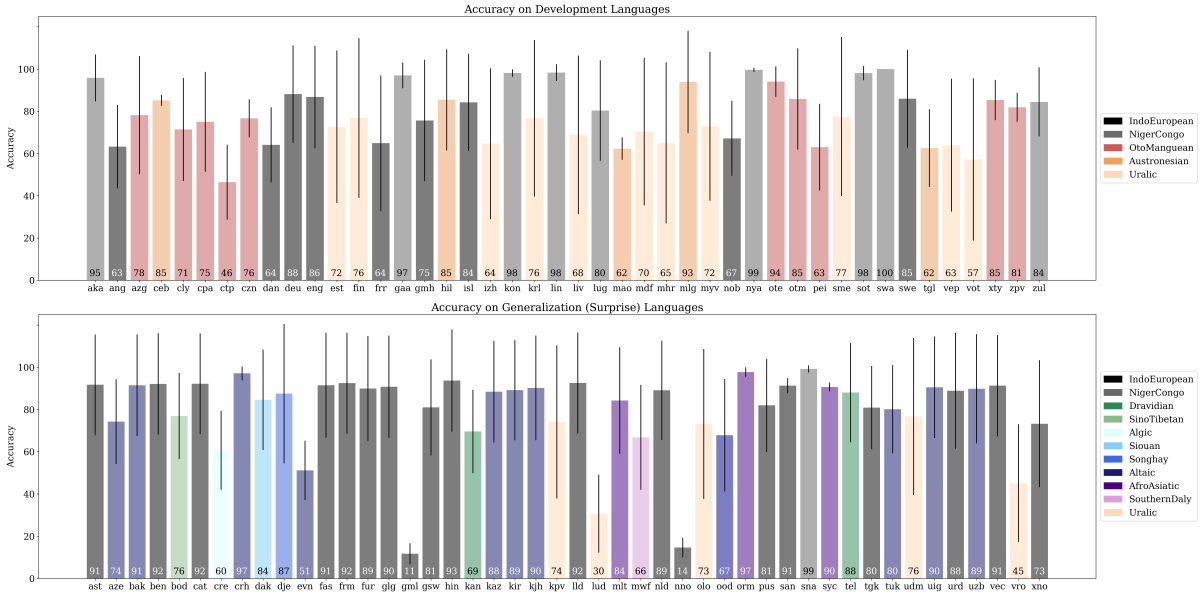

Figure 2: Accuracy by language averaged across all the final submitted systems with their standard deviations. Language families are demarcated by color, with accuracy on development languages (top), and generalization languages (bottom).

Is developing morphological grammars manually worthwhile? This was the main question asked by **CU7565** who manually designed finite-state grammars for 25 languages. Paradigms of some languages were relatively easy to describe but neural networks also performed quite well on them even with a limited amount of data. For low-resource languages such as Ingrian and Tagalog the grammars demonstrate superior performance but this comes at the expense of a significant amount of person-hours.

What is the best training strategy for low-resource languages? Teams that generated hallucinated data highlighted its utility for low-resource languages. Augmenting the data with tuples where lemmas are replaced with non-lemma forms and their tags is another technique that was found useful. In addition, multilingual training and ensembles yield extra gain in terms of accuracy.

Are the systems complementary? To address this question, we evaluate oracle scores for baseline systems, submitted systems, and all of them together. Typically, as Tables 8–21 in the Appendix demonstrate, the baselines and the submissions are complementary - adding them together increases the oracle score. Furthermore, while the full systems tend to dominate the partial systems (that were designed for a subset of languages, such as `CU7565-01-0`), there are a number of cases where the partial systems find the solution when the full systems don't - and these languages often then get even bigger gains when combined with the baselines. This even happens when the accuracy of the baseline is very high - Finnish has baseline oracle of 99.89; full systems oracle of 99.91; submission oracle of 99.94 and complete oracle of 99.96, so an ensemble might be able to improve on the results. The largest gaps in oracle systems are observed in Algic, Oto-Manguean, Sino-Tibetan, Southern Daly, Tungusic, and Uto-Aztecan families.[9]

Has morphological inflection become a solved problem in certain scenarios? The results shown in Fig. 2 suggest that for some of the development language families, such as Austronesian and Niger-Congo, the task was relatively easy, with most systems achieving high accuracy, whereas the task was more difficult for Uralic and Oto-Manguean languages, which showed greater variability in level of performance across submitted systems. Languages such as Ludic (lud), Norwegian Nynorsk (nno), Middle Low German

[9]Please see the results per language here: https://docs.google.com/spreadsheets/d/1ODFRnHuwN-mvGtzXA1sNdCi-jNqZjiE-i9jRxZCK0kg/edit?usp=sharing

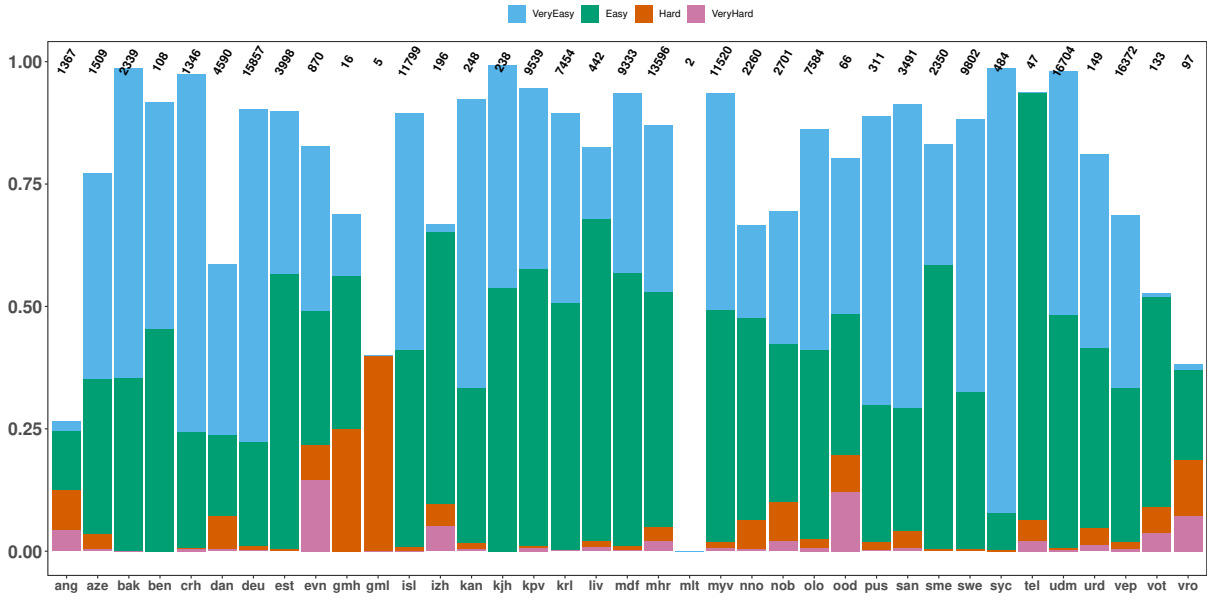

Figure 3: Difficulty of Nouns: Percentage of test samples falling into each category. The total number of test samples for each language is outlined on the top of the plot.

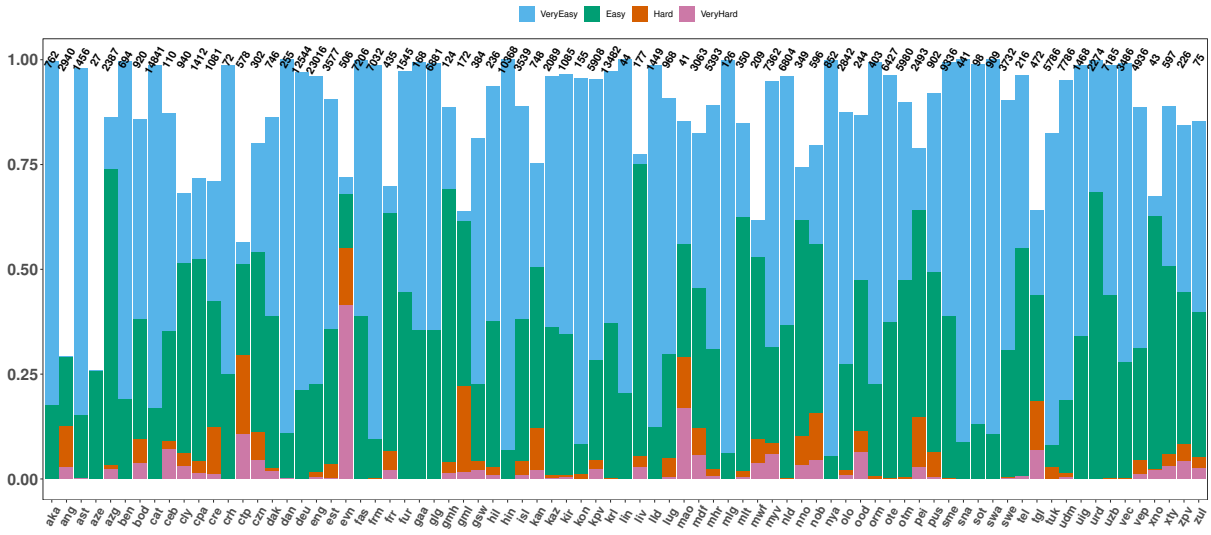

Figure 4: Difficulty of Verbs: Percentage of test samples falling into each category. The total number of test samples for each language is outlined on the top of the plot.

(gml), Evenki (evn), and O'odham (ood) seem to be the most challenging languages based on simple accuracy. For a more fine-grained study, we have classified test examples into four categories: "very easy", "easy", "hard", and "very hard". "Very easy" examples are ones that all submitted systems got correct, while "very hard" examples are ones that no submitted system got correct. "Easy" examples were predicted correctly for 80% of systems, and "hard" were only correct in 20% of systems. Fig. 3, Fig. 4, and Fig. 5 represent percentage of noun, verb, and adjective samples that fall into each category and illustrate that most language samples are correctly predicted by majority of the systems. For noun declension, Old English (ang), Middle Low German (gml), Evenki (evn), O'odham (ood), Võro (vro) are the most difficult (some of this difficulty comes from language data inconsistency, as described in the following section). For adjective declension, Classic Syriac presents the highest difficulty (likely due to its limited data).

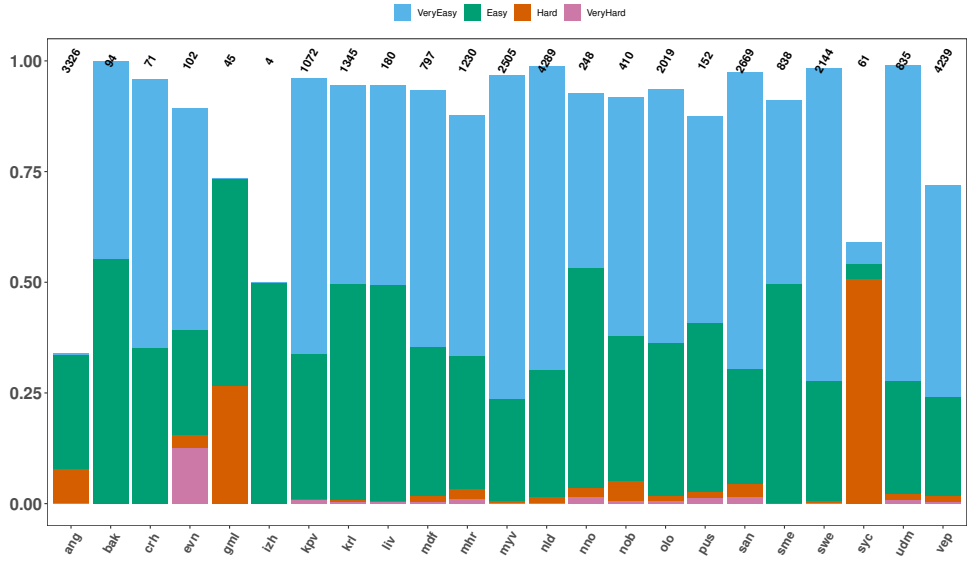

Figure 5: Difficulty of Adjectives: Percentage of test samples falling into each category. The total number of test samples for each language is outlined on the top of the plot.

9 Error Analysis

In our error analysis we follow the error type taxonomy proposed in Gorman et al. (2019). First, we evaluate systematic errors due to inconsistencies in the data, followed by an analysis of whether having seen the language or its family improved accuracy. We then proceed with an overview of accuracy for each of the language families. For a select number of families, we provide a more detailed analysis of the error patterns.

Tab. 6 and Tab. 7 provide the number of samples in the training, development, and test sets, percentage of inconsistent entries (the same lemma–tag pair has multiple infected forms) in them, percentage of contradicting entries (same lemma–tag pair occurring in train and development or test sets but assigned to different inflected forms), and percentage of entries in the development or test sets containing a lemma observed in the training set. The train, development and test sets contain 2%, 0.3%, and 0.6% inconsistent entries, respectively. Azerbaijani (aze), Old English (ang), Cree (cre), Danish (dan), Middle Low German (gml), Kannada (kan), Norwegian Bokmål (nob), Chichimec (pei), and Veps (vep) had the highest rates of inconsistency. These languages also exhibit the highest percentage of contradicting entries. The inconsistencies in some Finno-Ugric languages (such as Veps and Ludic) are due to dialectal variations.

The overall accuracy of system and language pairings appeared to improve with an increase in the size of the dataset (Fig. 6; see also Fig. 7 for accuracy trends by language family and Fig. 8 for accuracy trends by system). Overall, the variance was considerable regardless of whether the language family or even the language itself had been observed during the Development Phase. A linear mixed-effects regression was used to assess variation in accuracy using fixed effects of language category, the size of the training dataset (log count), and their interactions, as well as random intercepts for system and language family accuracy.[10] Language category was sum-coded with three levels: development language–development family, surprise language–development family, or surprise language–surprise family.

A significant effect of dataset size was observed, such that a one unit increase in log count corresponded to a 2% increase in accuracy (β = 0.019, $p < 0.001$). Language category type also significantly influenced accuracy: both development languages and surprise languages from development families were less accurate on average ($\beta_{dev-dev}$ = -0.145, $\beta_{sur-dev}$ = -0.167, each $p < 0.001$). These main effects were, however, significantly modulated by interactions with dataset size: on top of the main effect of dataset size, accuracy for development languages increased an additional $\approx$ 1.7% ($\beta_{dev-dev \times size}$ = 0.017, $p < 0.001$) and accuracy for surprise languages from development families

[10] Accuracy should ideally be assessed at the trial level using a logistic regression as opposed to a linear regression. By-trial accuracy was however not available at analysis time.

increased an additional $\approx$ 2.9% ($\beta_{sur-dev \times size}$ = 0.029, $p < 0.001$).

Afro-Asiatic: This family was represented by three languages. Mean accuracy across systems was above average at 91.7%. Relative to other families, variance in accuracy was low, but nevertheless ranged from 41.1% to 99.0%.

Algic: This family was represented by one language, Cree. Mean accuracy across systems was below average at 65.1%. Relative to other families, variance in accuracy was low, ranging from 41.5% to 73%. All systems appeared to struggle with the choice of preverbal auxiliary. Some auxiliaries were overloaded: 'kitta' could refer to future, imperfective, or imperative. The morphological features for mood and tense were also frequently combined, such as SBJV+OPT (subjunctive plus optative mood). While the paradigms were very large, there were very few lemmas (28 impersonal verbs and 14 transitive verbs), which may have contributed to the lower accuracy. Interestingly, the inflections could largely be generated by rules.[11]

Austronesian: This family was represented by five languages. Mean accuracy across systems was around average at 80.5%. Relative to other families, variance in accuracy was high, with accuracy ranging from 39.5% to 100%. One may notice a discrepancy among the difficulty in processing different Austronesian languages. For instance, we see a difference of over 10% in the baseline performance of Cebuano (84%) and Hiligaynon (96%).[12] This could come from the fact that Cebuano only has partial reduplication while Hiligaynon has full reduplication. Furthermore, the prefix choice for Cebuano is more irregular, making it more difficult to predict the correct conjugation of the verb.

Dravidian: This family was represented by two languages: Kannada and Telugu. Mean accuracy across systems was around average at 82.2%. Relative to other families, variance in accuracy was high: system accuracy ranged from 44.6% to 96.0%. Accuracy for Telugu was systematically higher than accuracy for Kannada.

Indo-European: This family was represented by 29 languages and four main branches. Mean accuracy across systems was slightly above average at 86.9%. Relative to other families, variance in accuracy was very high: system accuracy ranged from 0.02% to 100%. For Indo-Aryan, mean accuracy was high (96.0%) with low variance; for Germanic, mean accuracy was slightly below average (79.0%) but with very high variance (ranging from 0.02% to 99.5%), for Romance, mean accuracy was high (93.4%) but also had a high variance (ranging from 23.5% to 99.8%), and for Iranian, mean accuracy was high (89.2%), but again with a high variance (ranging from 25.0% to 100%). Languages from the Germanic branch of the Indo-European family were included in the Development Phase.

Niger–Congo: This family was represented by ten languages. Mean accuracy across systems was very good at 96.4%. Relative to other families, variance in accuracy was low, with accuracy ranging from 62.8% to 100%. Most languages in this family are considered low resource, and the resources used for data gathering may have been biased towards the languages' regular forms, as such this high accuracy may not be representative of the "easiness" of the task in this family. Languages from the Niger–Congo family was included in the Development Phase.

Oto-Manguean: This family was represented by nine languages. Mean accuracy across systems was slightly below average at 78.5%. Relative to other families, variance in accuracy was high, with accuracy ranging from 18.7% to 99.1%. Languages from the Oto-Manguean family were included in the Development Phase.

Sino-Tibetan: This family was represented by one language, Bodic. Mean accuracy across systems was average at 82.1%, and variance across systems was also very low. Accuracy ranged from 67.9% to 85.1%. The results are similar to those in Di et al. (2019) where majority of errors relate to allomorphy and impossible combinations of Tibetan unit components.

Siouan: This family was represented by one language, Dakota. Mean accuracy across systems was

[11] Minor issues with the encoding of diacritics were identified, and will be corrected for release.

[12] We also note that some Hiligaynon entries contained multiple lemma forms ("bati/batian/pamatian") for a single entry. We decided to leave it since we could not find any more information on which of the lemmas should be selected as the main. A similar issue was observed in Chichicapan Zapotec.

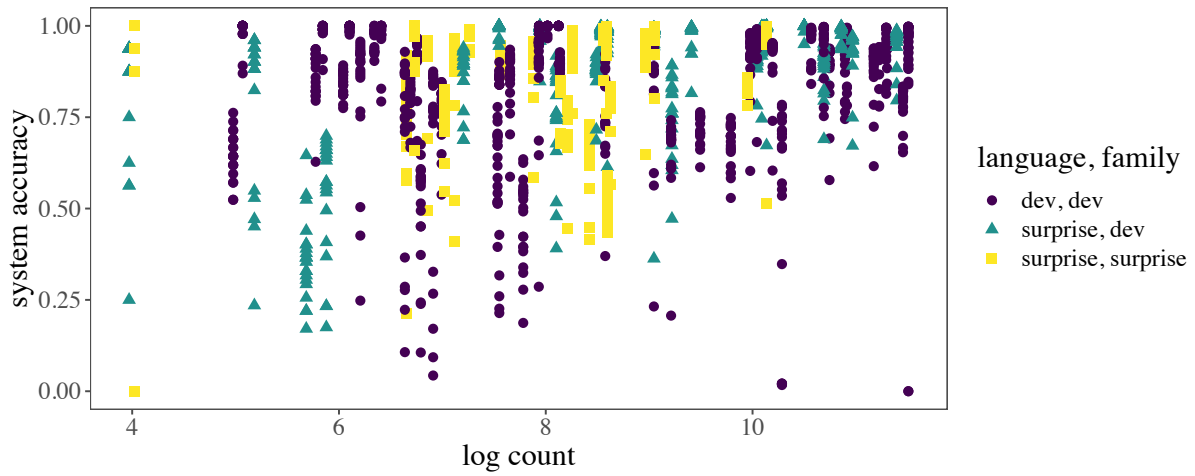

Figure 6: Accuracy for each system and language by the log size of the dataset. Points are color-coded according to language type: development language – development family, surprise language – development family, surprise language – surprise family.

above average at 89.4%, and variance across systems was also low, despite the range from 0% to 95.7%. Dakota presented variable prefixing and infixing of person morphemes, along some complexities related to fortition processes. Determining the factor(s) that governed variation in affix position was difficult from a linguist's perspective, though many systems were largely successful. Success varied in the choice of the first or second person singular allomorphs which had increasing degrees of consonant strengthening (e.g., /wa/, /ma/, /mi/ /bde/, /bdu/ for the first person singular and /ya/, /na/, /ni/, /de/, or /du/ for the second person singular). In some cases, these fortition processes were overapplied, and in some cases, entirely missed.

Songhay: This family was represented by one language, Zarma. Mean accuracy across systems was above average at 88.6%, and variance across systems was relatively high. Accuracy ranged from 0% to 100%.

Southern Daly: This family was represented by one language, Murrinh-Patha. Mean accuracy across systems was below average at 73.2%, and variance across systems was relatively high. Accuracy ranged from 21.2% to 91.9%.

Tungusic: This family was represented by one language, Evenki. The overall accuracy was the lowest across families. Mean accuracy was 53.8% with very low variance across systems. Accuracy ranged from 43.5% to 59.0%. The low accuracy is due to several factors. Firstly and primarily, the dataset was created from oral speech samples in various dialects of the language. The Evenki language is known to have rich dialectal variation. Moreover, there was little attempt at any standardization in the oral speech transcription. These peculiarities led to a high number of errors. For instance, some of the systems synthesized a wrong plural form for a noun ending in /-n/. Depending on the dialect, it can be /-r/ or /-l/, and there is a trend to have /-hVl/ for borrowed nouns. Deducing such a rule as well as the fact that the noun is a loanword is a hard task. Other suffixes may also have variable forms (such as /-kVllu/ vs /-kVldu/ depending on the dialect for the 2PL imperative. Some verbs have irregular past tense forms depending on the dialect and the meaning of the verb (e. g. /o:-/ 'to make' and 'to become'). Next, various dialects exhibit various vowel and consonant changes in suffixes. For example, some dialects (but not all of them) change /w/ to /b/ after /l/, and the systems sometimes synthesized a wrong form. The vowel harmony is complex: not all suffixes obey it, and it is also dialect-dependent. Some suffixes have variants (e. g., /-sin/ and /-s/ for SEMEL (semelfactive)), and the choice between them might be hard to understand. Finally, some of the mistakes are due to the markup scheme scarcity. For example, various past tense forms are all annotated as PST, or there are several comitative suffixes all annotated as COM. Moreover, some features are present in the word form but they receive no annotation at all. It is worth mentioning that some of the predictions could theoretically be possible. To sum up, the Evenki case presents the chal-

lenges of oral non-standardized speech.

Turkic: This family was represented by nine languages. Mean accuracy across systems was relatively high at 93%, and relative to other families, variance across systems was low. Accuracy ranged from 51.5% to 100%. Accuracy was lower for Azerbaijani and Turkmen, which after closer inspection revealed some slight contamination in the 'gold' files. There was very marginal variation in the accuracy for these languages across systems. Besides these two, accuracies were predominantly above 98%. A few systems struggled with the choice and inflection of the postverbal auxiliary in various languages (e.g., Kyrgyz, Kazakh, and Uzbek).

Uralic: This family was represented by 16 languages. Mean accuracy across systems was average at 81.5%, but the variance across systems and languages was very high. Accuracy ranged from 0% to 99.8%. Languages from the Uralic family were included in the Development Phase.

Uto-Aztecan: This family was represented by one language, O'odham. Mean accuracy across systems was slightly below average at 76.4%, but the variance across systems and languages was fairly low. Accuracy ranged from 54.8% to 82.5%. The systems with higher accuracy may have benefited from better recall of suppletive forms relative to lower accuracy systems.

10 Conclusion

This years's shared task on morphological reinflection focused on building models that could generalize across an extremely typologically diverse set of languages, many from understudied language families and with limited available text resources. As in previous years, neural models performed well, even in relatively low-resource cases. Submissions were able to make productive use of multilingual training to take advantage of commonalities across languages in the dataset. Data augmentation techniques such as hallucination helped fill in the gaps and allowed networks to generalize to unseen inputs. These techniques, combined with architecture tweaks like sparsemax, resulted in excellent overall performance on many languages (over 90% accuracy on average). However, the task's focus on typological diversity revealed that some morphology types and language families (Tungusic, Oto-Manguean, Southern Daly) remain a challenge for even the best systems. These families are extremely low-resource, represented in this dataset by few or a single language. This makes cross-linguistic transfer of similarities by multilanguage training less viable. They may also have morphological properties and rules (e.g., Evenki is agglutinating with many possible forms for each lemma) that are particularly difficult for machine learners to induce automatically from sparse data. For some languages (Ingrian, Tajik, Tagalog, Zarma, and Lingala), optimal performance was only achieved in this shared task by hand-encoding linguist knowledge in finite state grammars. It is up to future research to imbue models with the right kinds of linguistic inductive biases to overcome these challenges.

Acknowledgements

We would like to thank each of the participants for their time and effort in developing their task systems. We also thank Jason Eisner for organization and guidance. We thank Vitalij Chernyavskij for his help with Võro and Umida Boltaeva and Bahriddin Abdiev for their contribution in Uzbek data annotation.

References

Murat Abdulin. 2016. *Turkmen Verbs: 100 Turkmen Verbs Conjugated in All Tenses*. CreateSpace Independent Publishing Platform, Online.

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

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. 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 983–995.

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

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

Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, Piyush Mishra, Saliha Muradoğlu, Lan Sang, Tyler Short, Sagarika Shreevastava, Elizabeth Spaulding, Tetsumichi Umada, Beilei Xiang, Changbing Yang, and Mans Hulden. 2020. Linguist vs. machine: Rapid development of finite-state morphological grammars. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Emily M. Bender. 2009. Linguistically naïve!= language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32.

Emily M. Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Jean Berko. 1958. The child's learning of english morphology. *Word*, 14(2-3):150–177.

Balthasar Bickel and Johanna Nichols. 2013a. Exponence of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Balthasar Bickel and Johanna Nichols. 2013b. Fusion of selected inflectional formatives. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Balthasar Bickel and Johanna Nichols. 2013c. Inflectional synthesis of the verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Eric Campbell. 2016. Tone and inflection in zenzontepec chatino. *Tone and inflection*, pages 141–162.

Marc Canby, Aidana Karipbayeva, Bryan Lunt, Sahand Mozaffari, Charlotte Yoder, and Julia Hockenmaier. 2020. University of illinois 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*.

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, Sebastian J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 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.

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.

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

Hilaria Cruz. 2014. *Linguistic Poetic and Rhetoric of Eastern Chatino of San Juan Quiahije*. Ph.D. thesis.

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

Qianji Di, Ekaterina Vylomova, and Timothy Baldwin. 2019. Modelling Tibetan verbal morphology. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 35–40.

Matthew S. Dryer. 2013. Position of case affixes. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Renate Egli-Wildi. 2007. *Züritüütsch verstaa - Züritüütsch rede*. Küsnacht.

Timothy Feist and Enrique L. Palancar. 2015. *Oto-Manguean Inflectional Class Database*. University of Surrey, Online.

Raphael Finkel and Gregory Stump. 2007. Principal parts and morphological typology. *Morphology*, 17(1):39–75.

Martina Forster and Clara Meister. 2020. SIGMORPHON 2020 task 0 system description: ETH Zürich team. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Zygmunt Frajzyngier. 2018. Afroasiatic Languages. In *Oxford Research Encyclopedia of Linguistics*.

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.

Joseph Harold Greenberg. 1963. Universals of language.

Mohammad Nazmul Haque, Nasimul Noman, Regina Berretta, and Pablo Moscato. 2016. Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification. *PloS one*, 11(1).

Ray Harlow. 2007. *Maori: A Linguistic Introduction*. Cambridge University Press.

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

Jeffrey Heath. 2014. Grammar of humburi senni (songhay of hombori, mali).

Mans Hulden, Markus Forsberg, and Malin Ahlberg. 2014. Semi-supervised learning of morphological paradigms and lexicons. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 569–578.

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

Paa Kwesi Imbeah. 2012. *102 Akan Verbs*. CreateSpace Independent Publishing Platform, Online.

Sulev Iva. 2007. *Võru kirjakeele sõnamuutmissüsteem*. Ph.D. thesis.

Pratik Jayarao, Siddhanth Pillay, Pranav Thombre, and Aditi Chaudhary. 2020. Exploring neural architectures and techniques for typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

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

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

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

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

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

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

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

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

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

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

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

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

Elena Klyachko, Alexey Sorokin, Natalia Krizhanovskaya, Andrew Krizhanovsky, and Galina Ryazanskaya. 2020. LowResourceEval-2019: a shared task on morphological analysis for low-resource languages. *arXiv preprint arXiv:2001.11285*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Harlan LaFontaine and Neil McKay. 2005. *550 Dakota Verbs*. Minnesota Historical Society Press, Online.

Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon: A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2393–2400.

Ling Liu and Mans Hulden. 2020. Leveraging principal parts for morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

John Mansfield. 2019. *Murrinhpatha Morphology and Phonology*, volume 653. Walter de Gruyter.

André Martins and Ramon Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623.

John C. Moorfield. 2019. *Te Aka Online Māori Dictionary*. Online.

Nikitha Murikinati and Antonios Anastasopoulos. 2020. The CMU-LTI submission to the SIGMORPHON 2020 shared task 0: Language-specific cross-lingual transfer. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Temir Nabiyev. 2015. *Kazakh Language: 101 Kazakh Verbs*. Preceptor Language Guides, Online.

Mirembe Namono. 2018. *Luganda Language: 101 Luganda Verbs*. CreateSpace Independent Publishing Platform, Online.

Idai Nandoro. 2018. *Shona Language: 101 Shona Verbs*. CreateSpace Independent Publishing Platform, Online.

Center for Southeast Asian Studies NIU. 2017. *Table of Tagalog Verbs*. CreateSpace Independent Publishing Platform, Online.

Enrique L Palancar and Jean Léo Léonard. 2016. *Tone and inflection: New facts and new perspectives*, volume 296. Walter de Gruyter GmbH & Co KG.

Ben Peters and André F. T Martins. 2020. One-size-fits-all multilingual models. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Ben Peters and André FT Martins. 2019. It–ist at the sigmorphon 2019 shared task: Sparse two-headed models for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 50–56.

Ben Peters, Vlad Niculae, and André FT Martins. 2019. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1504–1519.

Robert L Rankin, John Boyle, Randolph Graczyk, and John E Koontz. 2003. Synchronic and diachronic perspective on'word'in siouan. *Word: a cross-linguistic typology*, pages 180–204.

Dakila Reyes. 2015. *Cebuano Language: 101 Cebuano Verbs*. CreateSpace Independent Publishing Platform, Online.

Anj Santos. 2018. *Hiligaynon Language. 101 Hiligaynon Verbs*. CreateSpace Independent Publishing Platform, Online.

Andei 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*.

Andrei Shcherbakov, Ekaterina Vylomova, and Nick Thieberger. 2016. Phonotactic modeling of extremely low resource languages. In *Proceedings of the Australasian Language Technology Association Workshop 2016*, pages 84–93.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99, Vancouver. Association for Computational Linguistics.

Assaf Singer and Katharina Kann. 2020. The NYU-CUBoulder systems for SIGMORPHON 2020 task 0 and task 2. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Morris Swadesh. 1950. Salish internal relationships. *International Journal of American Linguistics*, 16(4):157–167.

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). *Johns Hopkins University*.

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

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

Turkmenistan US Embassy. 2018. *501 Turkmen verbs*. Peace Corps, Online.

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*, pages 5998–6008.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction.

Xiang Yu, Ngoc Thang Vu, and Jonas Kuhns. 2020. Ensemble self-training for low-resource languages:grapheme-to-phoneme conversion and morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

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.

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

Long Zhou, Jiajun Zhang, and Chengqing Zong. 2019. Synchronous bidirectional neural machine translation. *Transactions of the Association for Computational Linguistics*, 7:91–105.

A Language data statistics

Lang	Total			Inconsistency (%)			Contradiction (%)		In Vocabulary (%)	
	Train	Dev	Test	Train	Dev	Test	Dev	Test	Dev	Test
aka	2793	380	763	0.0	0.0	0.0	0.0	0.0	24.7	12.5
ang	29270	4122	8197	11.8	1.8	3.4	21.6	21.9	35.1	21.3
ast	5096	728	1457	0.0	0.0	0.0	0.0	0.0	23.9	12.4
aze	5602	801	1601	11.9	1.9	4.0	22.3	20.9	31.5	20.2
azg	8482	1188	2396	0.8	0.0	0.0	1.3	1.1	26.9	13.8
bak	8517	1217	2434	0.0	0.0	0.0	0.0	0.0	59.8	40.1
ben	2816	402	805	0.0	0.0	0.0	0.0	0.0	29.9	16.0
bod	3428	466	936	1.0	0.2	0.3	2.4	1.9	80.0	73.4
cat	51944	7421	14842	0.0	0.0	0.0	0.0	0.0	20.8	10.4
ceb	420	58	111	1.0	0.0	0.0	0.0	2.7	72.4	62.2
cly	3301	471	944	0.0	0.0	0.0	0.0	0.0	37.4	19.3
cpa	5298	727	1431	3.4	0.6	0.8	6.6	4.3	60.2	39.8
cre	4571	584	1174	18.5	2.1	4.9	29.8	29.6	5.5	2.7
crh	5215	745	1490	0.0	0.0	0.0	0.0	0.0	77.4	60.7
ctp	2397	313	598	15.9	1.6	3.0	22.0	21.7	52.7	34.1
czn	1088	154	305	0.2	0.0	0.0	1.3	0.0	86.4	74.8
dak	2636	376	750	0.0	0.0	0.0	0.0	0.0	75.5	55.7
dan	17852	2550	5101	16.5	2.5	5.0	34.5	32.9	71.4	51.8
deu	99405	14201	28402	0.0	0.0	0.0	0.0	0.0	55.8	37.8
dje	56	9	16	0.0	0.0	0.0	0.0	0.0	100.0	87.5
eng	80865	11553	23105	1.1	0.2	0.4	2.1	1.9	80.3	66.2
est	26728	3820	7637	2.7	0.4	0.8	6.1	5.1	22.4	11.6
evn	5413	774	1547	9.6	2.8	4.3	8.9	10.0	38.9	32.5
fas	25225	3603	7208	0.0	0.0	0.0	0.0	0.0	7.6	3.8
fin	99403	14201	28401	0.0	0.0	0.0	0.0	0.0	32.6	17.2
frm	24612	3516	7033	0.0	0.0	0.0	0.0	0.0	17.1	8.6
frr	1902	224	477	4.0	0.0	1.7	9.8	6.1	22.8	10.7
fur	5408	772	1546	0.0	0.0	0.0	0.0	0.0	21.6	10.9
gaa	607	79	169	0.0	0.0	0.0	0.0	0.0	74.7	47.3
glg	24087	3441	6882	0.0	0.0	0.0	0.0	0.0	14.1	7.1
gmh	496	71	141	1.2	0.0	0.0	5.6	2.8	38.0	20.6
gml	890	127	255	17.3	3.1	5.5	22.8	27.8	39.4	20.4
gsw	1345	192	385	0.0	0.0	0.0	0.0	0.0	55.7	35.6
hil	859	116	238	0.0	0.0	0.0	0.0	0.0	59.5	36.6
hin	36300	5186	10372	0.0	0.0	0.0	0.0	0.0	5.0	2.5
isl	53841	7690	15384	1.0	0.1	0.3	1.9	2.0	48.8	29.5
izh	763	112	224	0.0	0.0	0.0	0.0	0.0	42.9	22.3
kan	3670	524	1049	13.2	2.7	4.7	18.7	20.7	21.9	14.0
kaz	7852	1063	2113	1.1	0.2	0.4	1.9	1.8	10.6	5.3
kir	3855	547	1089	0.0	0.0	0.0	0.0	0.0	17.9	9.0
kjh	840	120	240	0.0	0.0	0.0	0.0	0.0	50.8	30.4
kon	568	76	156	0.0	0.0	0.0	0.0	0.0	78.9	71.8
kpv	57919	8263	16526	0.0	0.0	0.0	0.0	0.0	48.8	35.0
krl	80216	11225	22290	0.2	0.0	0.0	0.3	0.3	19.7	10.3
lin	159	23	46	0.0	0.0	0.0	0.0	0.0	100.0	73.9
liv	2787	398	802	0.0	0.0	0.0	0.0	0.0	40.7	24.1

Table 6: Number of samples in training, development, test sets, as well as statistics on systematic errors (inconsistency) and percentage of samples with lemmata observed in the training set.

Lang	Total			Inconsistency (%)			Contradiction (%)		In Vocabulary (%)	
	Train	Dev	Test	Train	Dev	Test	Dev	Test	Dev	Test
lld	5073	725	1450	0.0	0.0	0.0	0.0	0.0	24.3	12.3
lud	294	41	82	7.8	0.0	3.7	9.8	11.0	31.7	20.7
lug	3420	489	977	4.0	0.6	0.8	5.1	7.6	18.2	9.1
mao	145	21	42	0.0	0.0	0.0	0.0	0.0	61.9	81.0
mdf	46362	6633	13255	1.6	0.2	0.5	3.1	3.3	49.0	35.1
mhr	71143	10081	20233	0.3	0.0	0.0	0.4	0.5	48.8	34.3
mlg	447	62	127	0.0	0.0	0.0	0.0	0.0	90.3	74.0
mlt	1233	176	353	0.1	0.0	0.0	0.6	0.0	52.3	30.6
mwf	777	111	222	2.6	0.0	0.9	2.7	4.5	25.2	13.1
myv	74928	10738	21498	1.7	0.3	0.5	3.1	3.1	45.5	32.7
nld	38826	5547	11094	0.0	0.0	0.0	0.0	0.0	58.2	38.4
nno	10101	1443	2887	3.4	0.4	1.0	6.0	6.8	80.0	70.2
nob	13263	1929	3830	10.5	1.8	3.1	18.5	19.7	80.5	70.5
nya	3031	429	853	0.0	0.0	0.0	0.0	0.0	46.4	26.5
olo	43936	6260	12515	1.4	0.3	0.5	3.3	2.9	83.0	70.8
ood	1123	160	314	0.4	0.0	0.0	1.9	1.0	70.0	58.0
orm	1424	203	405	0.2	0.0	0.2	0.5	0.7	41.9	22.7
ote	22962	3231	6437	0.4	0.1	0.1	0.5	0.8	48.4	29.5
otm	21533	3020	5997	0.9	0.1	0.3	1.8	1.7	49.4	29.4
pei	10017	1349	2636	15.8	2.6	4.9	21.5	21.4	9.1	4.7
pus	4861	695	1389	3.9	0.6	1.6	9.9	7.7	34.2	23.0
san	22968	3188	6272	3.1	0.5	0.9	4.5	5.5	26.9	14.6
sme	43877	6273	12527	0.0	0.0	0.0	0.0	0.0	28.2	16.3
sna	1897	246	456	0.0	0.0	0.0	0.0	0.0	31.3	18.0
sot	345	50	99	0.0	0.0	0.0	0.0	0.0	48.0	25.3
swa	3374	469	910	0.0	0.0	0.0	0.0	0.0	20.7	10.5
swe	54888	7840	15683	0.0	0.0	0.0	0.0	0.0	70.6	51.9
syc	1917	275	548	3.5	1.5	0.4	7.6	8.6	47.3	28.1
tel	952	136	273	1.4	0.0	1.1	0.7	2.6	62.5	39.6
tgk	53	8	16	0.0	0.0	0.0	0.0	0.0	0.0	0.0
tgl	1870	236	478	7.6	1.3	1.0	11.9	10.0	74.2	55.6
tuk	20963	2992	5979	9.5	1.5	3.2	16.8	16.0	16.7	8.3
udm	88774	12665	25333	0.0	0.0	0.0	0.0	0.0	38.1	24.8
uig	5372	750	1476	0.3	0.0	0.0	0.3	0.5	12.0	6.1
urd	8486	1213	2425	0.0	0.0	0.0	0.0	0.0	9.4	6.0
uzb	25199	3596	7191	0.0	0.0	0.0	0.0	0.0	11.9	6.0
vec	12203	1743	3487	0.0	0.0	0.0	0.0	0.0	20.8	10.6
vep	94395	13320	26422	10.9	1.8	3.3	19.3	19.8	25.1	12.9
vot	1003	146	281	0.0	0.0	0.0	0.0	0.0	35.6	19.6
vro	357	51	103	1.1	0.0	0.0	2.0	1.0	70.6	50.5
xno	178	26	51	0.0	0.0	0.0	0.0	0.0	19.2	9.8
xty	2110	299	600	0.1	0.3	0.0	0.3	1.3	78.6	65.8
zpv	805	113	228	0.0	0.0	0.4	2.7	0.9	78.8	78.9
zul	322	42	78	1.9	0.0	0.0	2.4	0.0	83.3	66.7
TOTAL	1574004	223649	446580	2.0	0.3	0.6	3.6	3.6	41.1	27.9

Table 7: Number of samples in training, development, test sets, as well as statistics on systematic errors (inconsistency) and percentage of samples with lemmata observed in the training set.

B Accuracy trends

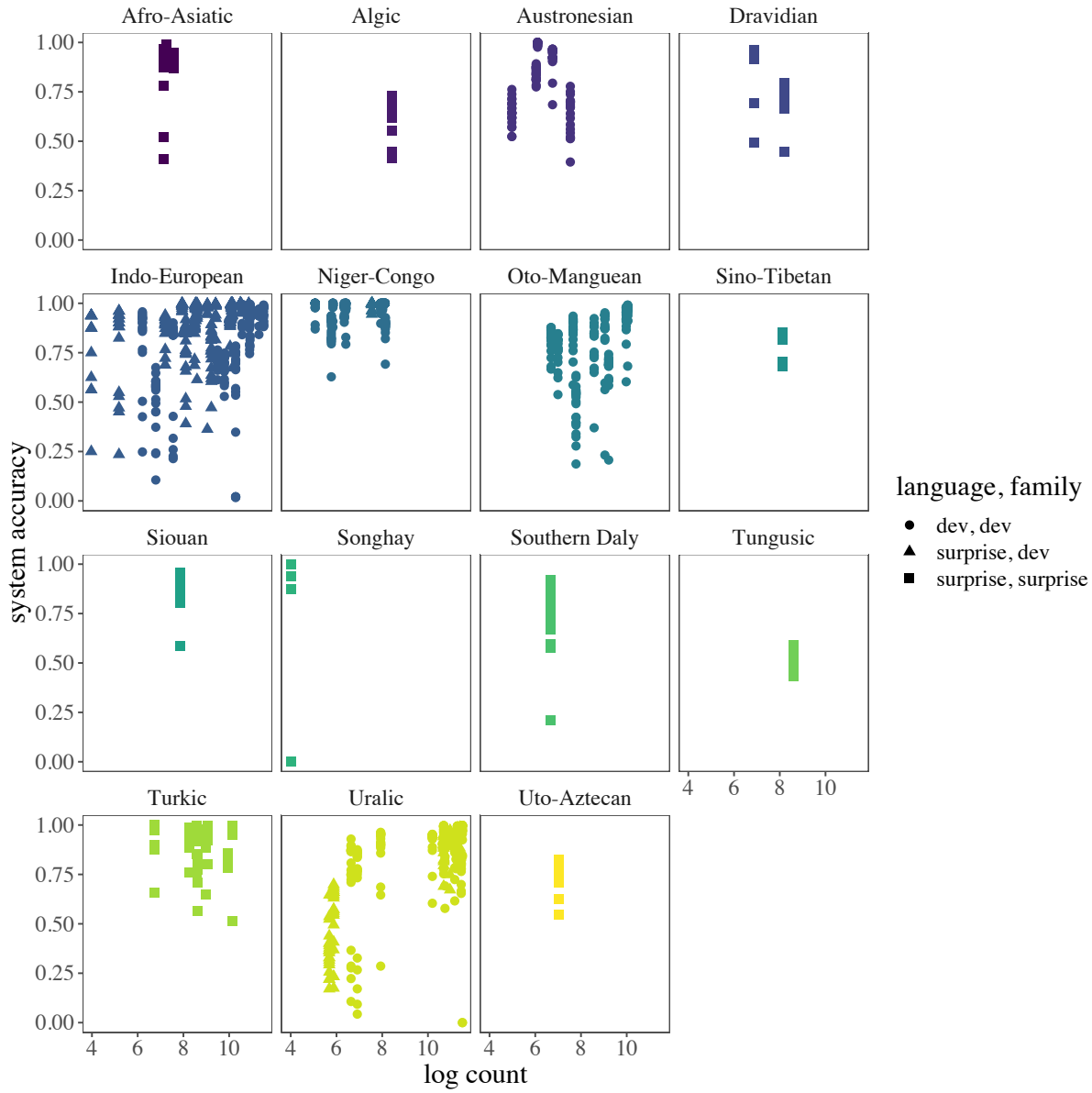

Figure 7: Accuracy for each system and language by the log size of the dataset, **grouped by language family**. Points are color-coded according to language family, and shape-coded according to language type: development language – development family, surprise language – development family, surprise language – surprise family.

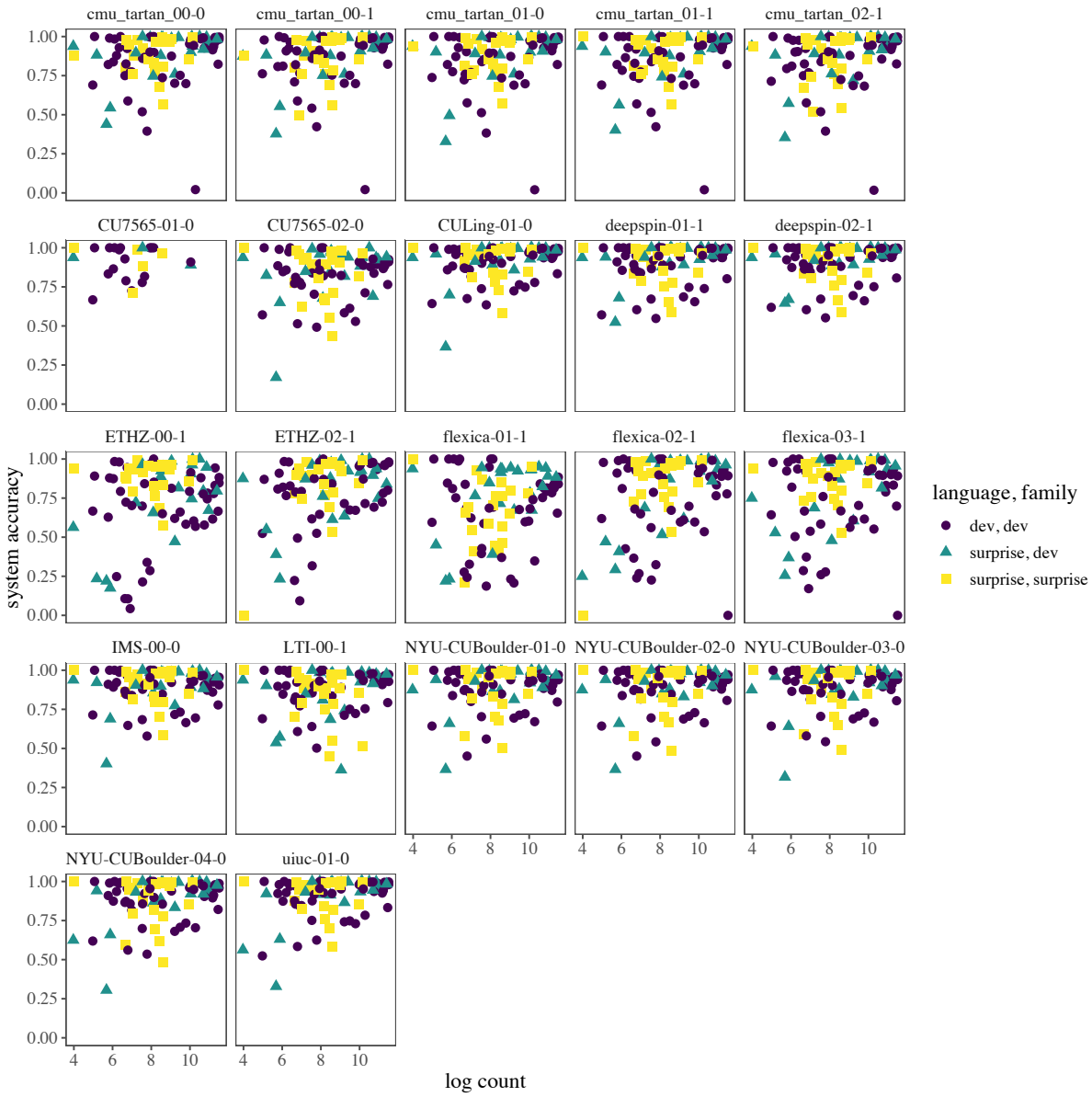

Figure 8: Accuracy for each language by the log size of the dataset, **grouped by submitted system**. Points are color- and shape-coded according to language type: development language – development family, surprise language – development family, surprise language – surprise family.

Table 8: **Results per Language Family: Afro-Asiatic and Algic**

System	Rank	Acc
uiuc-01-0	1.0	96.4
CULing-01-0	1.0	96.3
deepspin-02-1	3.7	95.2
BASE: trm-single	4.0	95.5
BASE: trm-aug-single	4.0	95.0
deepspin-01-1	4.0	94.7
NYU-CUBoulder-01-0	4.0	94.4
NYU-CUBoulder-02-0	4.0	94.4
NYU-CUBoulder-04-0	9.7	94.3
BASE: mono-single	6.3	92.8
cmu_tartan_00-0	6.3	92.7
cmu_tartan_01-0	9.3	89.6
cmu_tartan_01-1	9.3	89.4
cmu_tartan_02-1	10.0	80.9
ETHZ-00-1	6.7	94.7
BASE: trm-shared	6.7	94.2
BASE: trm-aug-shared	6.7	94.0
IMS-00-0	6.7	93.6
BASE: mono-aug-single	6.7	93.5
NYU-CUBoulder-03-0	12.3	93.7
flexica-02-1	9.3	92.9
ETHZ-02-1	9.3	92.3
flexica-03-1	9.3	92.1
BASE: mono-shared	9.3	91.5
**CU7565-01-0*	*19.3*	*93.7*
BASE: mono-aug-shared	16.0	89.8
CU7565-02-0	15.0	91.6
cmu_tartan_00-1	17.7	91.7
LTI-00-1	17.7	91.3
flexica-01-1	28.3	73.4
Oracle (Baselines)		98.7
Oracle (Submissions)		99.7
Oracle (All)		99.8

(a) Results on the Afro-Asiatic family (3 languages)

System	Rank	Acc
CULing-01-0	1.0	73.0
flexica-03-1	1.0	70.4
IMS-00-0	1.0	70.3
uiuc-01-0	1.0	70.3
ETHZ-02-1	1.0	69.4
cmu_tartan_02-1	1.0	69.4
flexica-02-1	1.0	69.4
cmu_tartan_00-1	8.0	69.2
BASE: mono-aug-shared	8.0	68.5
BASE: mono-aug-single	8.0	68.5
ETHZ-00-1	8.0	68.4
BASE: trm-aug-shared	8.0	68.0
BASE: trm-aug-single	8.0	68.0
cmu_tartan_01-1	8.0	68.0
NYU-CUBoulder-01-0	8.0	67.9
BASE: trm-shared	8.0	67.7
BASE: trm-single	8.0	67.7
cmu_tartan_00-0	8.0	67.6
cmu_tartan_01-0	8.0	67.6
BASE: mono-shared	8.0	66.8
BASE: mono-single	8.0	66.8
NYU-CUBoulder-02-0	8.0	66.5
deepspin-02-1	8.0	66.5
deepspin-01-1	24.0	65.1
NYU-CUBoulder-03-0	24.0	64.7
NYU-CUBoulder-04-0	26.0	61.8
CU7565-02-0	27.0	55.5
LTI-00-1	28.0	44.9
flexica-01-1	28.0	41.5
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		86.9
Oracle (Submissions)		98.7
Oracle (All)		98.8

(b) Results on the Algic family (1 language)

Table 9: **Results per Language Family: Austronesian and Dravidian**

System	Rank	Acc
CULing-01-0	1.0	84.4
IMS-00-0	1.6	85.1
NYU-CUBoulder-03-0	1.6	83.6
ETHZ-00-1	1.6	83.4
NYU-CUBoulder-01-0	1.6	82.9
NYU-CUBoulder-04-0	1.6	82.9
BASE: trm-shared	1.6	82.8
NYU-CUBoulder-02-0	1.6	82.7
deepspin-02-1	3.2	82.4
BASE: trm-aug-single	3.2	81.6
**CU7565-01-0*	*6.8*	*82.7*
uiuc-01-0	5.4	82.3
BASE: trm-single	6.0	81.2
BASE: mono-aug-shared	6.0	82.9
LTI-00-1	6.0	82.0
BASE: mono-aug-single	7.8	81.3
deepspin-01-1	7.6	81.0
BASE: trm-aug-shared	7.6	79.8
flexica-03-1	7.6	79.3
cmu_tartan_00-0	8.2	79.1
BASE: mono-shared	10.4	79.2
BASE: mono-single	10.4	77.6
cmu_tartan_00-1	12.8	80.3
cmu_tartan_02-1	12.8	78.9
cmu_tartan_01-0	12.8	78.6
flexica-02-1	12.8	78.3
cmu_tartan_01-1	12.8	78.2
ETHZ-02-1	12.0	77.4
**CU7565-02-0*	*22.4*	*73.7*
flexica-01-1	21.2	69.7
Oracle (Baselines)		89.1
Oracle (Submissions)		93.5
Oracle (All)		93.7

(a) Results on the Austronesian family (5 languages)

System	Rank	Acc
IMS-00-0	1.0	87.6
CULing-01-0	1.0	87.0
BASE: trm-aug-shared	1.0	86.8
cmu_tartan_00-0	1.0	86.3
cmu_tartan_01-1	1.0	86.3
BASE: trm-aug-single	1.0	85.9
BASE: trm-shared	1.0	85.8
ETHZ-02-1	1.0	85.5
cmu_tartan_01-0	5.0	85.7
deepspin-02-1	5.0	85.6
cmu_tartan_02-1	5.0	85.5
BASE: trm-single	5.0	85.4
uiuc-01-0	5.0	85.3
deepspin-01-1	5.0	85.2
LTI-00-1	5.0	85.0
ETHZ-00-1	5.0	84.9
BASE: mono-single	5.0	84.8
BASE: mono-aug-single	5.0	84.1
NYU-CUBoulder-02-0	12.0	82.2
NYU-CUBoulder-01-0	12.0	82.2
NYU-CUBoulder-03-0	12.0	82.1
NYU-CUBoulder-04-0	12.0	81.9
CU7565-02-0	14.5	81.4
flexica-02-1	16.5	83.7
BASE: mono-shared	16.5	83.7
flexica-03-1	16.5	83.0
cmu_tartan_00-1	19.0	62.6
BASE: mono-aug-shared	23.5	79.7
flexica-01-1	28.5	56.9
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		95.9
Oracle (Submissions)		98.2
Oracle (All)		98.6

(b) Results on the Dravidian family (2 languages)

Table 10: **Results per Language Family: Indo-European and Niger-Congo**

System	Rank	Acc
deepspin-02-1	2.3	92.9
uiuc-01-0	3.1	91.6
deepspin-01-1	2.9	92.9
BASE: trm-single	2.9	91.7
CULing-01-0	3.9	93.5
BASE: trm-aug-single	3.4	92.9
NYU-CUBoulder-04-0	7.3	90.7
BASE: trm-shared	12.0	86.9
cmu_tartan_00-1	8.1	88.6
BASE: mono-shared	8.9	90.3
NYU-CUBoulder-03-0	10.0	91.2
cmu_tartan_00-0	8.9	88.5
NYU-CUBoulder-02-0	11.4	90.6
BASE: mono-aug-shared	12.9	90.5
NYU-CUBoulder-01-0	12.4	90.4
BASE: mono-single	8.1	88.0
BASE: mono-aug-single	7.9	91.9
cmu_tartan_01-0	10.5	88.6
cmu_tartan_01-1	9.9	88.5
IMS-00-0	15.9	90.4
cmu_tartan_02-1	10.7	88.4
BASE: trm-aug-shared	15.0	88.6
LTI-00-1	15.8	87.5
CU7565-02-0	20.3	86.3
flexica-03-1	19.4	80.7
ETHZ-02-1	18.1	83.8
ETHZ-00-1	23.5	73.7
flexica-02-1	21.8	77.5
**CU7565-01-0*	*28.8*	*91.4*
flexica-01-1	26.0	76.7
Oracle (Baselines)		98.0
Oracle (Submissions)		98.8
Oracle (All)		99.1

(a) Results on the Indo-European family (28 languages)

System	Rank	Acc
IMS-00-0	1.0	98.1
uiuc-01-0	1.0	97.9
NYU-CUBoulder-01-0	1.3	98.1
NYU-CUBoulder-02-0	1.3	98.1
deepspin-02-1	1.3	98.0
NYU-CUBoulder-03-0	1.3	98.0
BASE: mono-aug-single	1.3	97.9
deepspin-01-1	1.3	97.9
NYU-CUBoulder-04-0	1.3	97.8
LTI-00-1	1.3	97.7
BASE: trm-shared	1.3	97.7
BASE: trm-single	1.3	97.7
BASE: mono-single	1.3	97.7
BASE: mono-shared	1.3	97.6
BASE: trm-aug-single	1.3	97.5
BASE: trm-aug-shared	1.3	97.4
BASE: mono-aug-shared	1.3	97.2
**CU7565-01-0*	*3.9*	*98.0*
CULing-01-0	3.4	97.1
flexica-03-1	3.1	96.9
flexica-02-1	3.1	96.9
cmu_tartan_01-1	3.6	96.4
cmu_tartan_00-0	3.6	96.3
cmu_tartan_01-0	3.6	96.3
CU7565-02-0	6.5	95.6
cmu_tartan_00-1	7.8	95.4
flexica-01-1	9.2	94.2
cmu_tartan_02-1	11.2	94.4
ETHZ-02-1	18.9	91.7
ETHZ-00-1	20.3	89.3
Oracle (Baselines)		99.2
Oracle (Submissions)		99.4
Oracle (All)		99.6

(b) Results on the Niger-Congo family (10 languages)

Table 11: **Results per Language Family: Oto-Manguean and Sino-Tibetan**

System	Rank	Acc
uiuc-01-0	1.0	87.5
BASE: trm-single	2.0	86.2
CULing-01-0	3.1	86.7
deepspin-02-1	3.4	85.4
deepspin-01-1	3.4	85.3
NYU-CUBoulder-04-0	6.4	84.2
BASE: mono-single	7.9	82.4
NYU-CUBoulder-03-0	8.4	83.5
BASE: mono-aug-single	6.1	83.5
BASE: mono-shared	8.2	82.9
NYU-CUBoulder-02-0	9.1	83.5
IMS-00-0	10.3	83.3
LTI-00-1	9.4	82.4
NYU-CUBoulder-01-0	9.4	83.6
BASE: mono-aug-shared	9.8	82.0
cmu_tartan_00-0	13.9	78.5
cmu_tartan_01-1	14.9	78.5
cmu_tartan_02-1	15.2	78.2
BASE: trm-shared	14.5	80.2
BASE: trm-aug-shared	20.3	73.8
flexica-01-1	26.3	47.2
BASE: trm-aug-single	7.4	84.3
cmu_tartan_00-1	14.1	79.0
ETHZ-02-1	14.0	81.4
CU7565-02-0	20.9	75.1
cmu_tartan_01-0	18.3	76.5
**CU7565-01-0*	*27.8*	*81.0*
ETHZ-00-1	25.4	70.5
flexica-02-1	25.6	67.0
flexica-03-1	26.1	64.2
Oracle (Baselines)		94.1
Oracle (Submissions)		96.2
Oracle (All)		96.7

(a) Results on the Oto-Manguean family (10 languages)

System	Rank	Acc
deepspin-01-1	1.0	85.1
deepspin-02-1	1.0	85.0
LTI-00-1	1.0	84.7
uiuc-01-0	1.0	84.4
BASE: trm-single	1.0	84.4
BASE: trm-shared	1.0	84.4
CULing-01-0	1.0	84.1
ETHZ-02-1	1.0	83.8
flexica-02-1	1.0	83.7
cmu_tartan_01-1	1.0	83.4
BASE: mono-aug-shared	1.0	83.4
BASE: mono-aug-single	1.0	83.4
NYU-CUBoulder-01-0	1.0	83.4
IMS-00-0	1.0	83.3
BASE: trm-aug-single	1.0	83.3
BASE: trm-aug-shared	1.0	83.3
BASE: mono-shared	1.0	83.2
BASE: mono-single	1.0	83.2
cmu_tartan_00-0	1.0	83.1
cmu_tartan_02-1	1.0	83.1
cmu_tartan_00-1	1.0	83.0
NYU-CUBoulder-03-0	22.0	82.8
ETHZ-00-1	22.0	82.8
cmu_tartan_01-0	22.0	82.7
NYU-CUBoulder-02-0	22.0	82.6
flexica-03-1	22.0	82.5
NYU-CUBoulder-04-0	22.0	81.7
flexica-01-1	28.0	70.6
CU7565-02-0	28.0	67.9
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		91.3
Oracle (Submissions)		96.0
Oracle (All)		96.2

(b) Results on the Sino-Tibetan family (1 language)

Table 12: **Results per Language Family: Siouan and Songhay**

System	Rank	Acc
NYU-CUBoulder-01-0	1.0	95.7
BASE: trm-single	1.0	95.6
CULing-01-0	1.0	95.6
BASE: trm-shared	1.0	95.6
ETHZ-00-1	1.0	95.5
uiuc-01-0	1.0	94.9
deepspin-01-1	1.0	94.8
NYU-CUBoulder-02-0	1.0	94.8
NYU-CUBoulder-03-0	1.0	94.7
deepspin-02-1	1.0	94.5
BASE: mono-aug-shared	1.0	94.4
BASE: mono-aug-single	1.0	94.4
NYU-CUBoulder-04-0	1.0	94.3
ETHZ-02-1	14.0	93.3
BASE: mono-single	14.0	92.9
BASE: mono-shared	14.0	92.9
BASE: trm-aug-single	14.0	92.5
BASE: trm-aug-shared	14.0	92.5
flexica-02-1	14.0	91.5
IMS-00-0	14.0	90.9
LTI-00-1	21.0	89.7
flexica-03-1	21.0	89.3
cmu_tartan_01-0	23.0	85.7
cmu_tartan_01-1	23.0	85.7
cmu_tartan_02-1	23.0	85.7
cmu_tartan_00-0	23.0	85.5
cmu_tartan_00-1	23.0	85.5
CU7565-02-0	28.0	80.5
flexica-01-1	29.0	58.4
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		97.3
Oracle (Submissions)		98.1
Oracle (All)		98.1

(a) Results on the Siouan family (1 language)

System	Rank	Acc
BASE: mono-aug-single	1.0	100.0
BASE: trm-aug-single	1.0	100.0
CU7565-02-0	1.0	100.0
CU7565-01-0	1.0	100.0
uiuc-01-0	1.0	100.0
NYU-CUBoulder-02-0	1.0	100.0
NYU-CUBoulder-03-0	1.0	100.0
BASE: mono-aug-shared	1.0	100.0
NYU-CUBoulder-01-0	1.0	100.0
LTI-00-1	1.0	100.0
IMS-00-0	1.0	100.0
flexica-01-1	1.0	100.0
deepspin-02-1	1.0	100.0
deepspin-01-1	1.0	100.0
CULing-01-0	1.0	100.0
cmu_tartan_01-1	1.0	100.0
NYU-CUBoulder-04-0	1.0	100.0
BASE: trm-aug-shared	1.0	100.0
flexica-03-1	1.0	93.8
ETHZ-00-1	1.0	93.8
cmu_tartan_02-1	1.0	93.8
cmu_tartan_01-0	1.0	93.8
cmu_tartan_00-0	1.0	87.5
cmu_tartan_00-1	1.0	87.5
BASE: trm-shared	1.0	87.5
BASE: trm-single	1.0	87.5
flexica-02-1	27.0	0.0
BASE: mono-shared	27.0	0.0
BASE: mono-single	27.0	0.0
ETHZ-02-1	27.0	0.0
Oracle (Baselines)		100.0
Oracle (Submissions)		100.0
Oracle (All)		100.0

(b) Results on the Songhay family/genus (1 language)

Table 13: **Results per Language Family: Southern Daly and Tungusic**

System	Rank	Acc
CULing-01-0	1.0	91.9
BASE: trm-single	1.0	89.6
BASE: trm-shared	1.0	89.6
ETHZ-00-1	1.0	88.7
uiuc-01-0	1.0	87.8
BASE: trm-aug-single	1.0	86.9
BASE: trm-aug-shared	1.0	86.9
IMS-00-0	1.0	86.0
deepspin-01-1	9.0	83.8
deepspin-02-1	9.0	83.3
cmu_tartan_01-1	9.0	81.1
cmu_tartan_01-0	9.0	81.1
cmu_tartan_00-0	9.0	80.2
cmu_tartan_00-1	9.0	80.2
ETHZ-02-1	15.0	77.9
CU7565-02-0	15.0	77.5
flexica-03-1	15.0	73.4
flexica-02-1	15.0	72.5
LTI-00-1	15.0	70.3
cmu_tartan_02-1	20.0	67.1
BASE: mono-shared	20.0	60.8
BASE: mono-single	20.0	60.8
NYU-CUBoulder-04-0	20.0	59.5
NYU-CUBoulder-03-0	20.0	59.0
NYU-CUBoulder-02-0	20.0	57.7
NYU-CUBoulder-01-0	20.0	57.7
BASE: mono-aug-single	27.0	44.6
BASE: mono-aug-shared	27.0	44.6
flexica-01-1	29.0	21.2
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		91.4
Oracle (Submissions)		96.4
Oracle (All)		96.4

(a) Results on the Southern Daly family (1 language)

System	Rank	Acc
deepspin-02-1	1.0	59.0
deepspin-01-1	1.0	58.8
uiuc-01-0	1.0	58.3
IMS-00-0	1.0	58.2
CULing-01-0	1.0	58.0
BASE: trm-aug-single	1.0	57.7
BASE: trm-aug-shared	1.0	57.7
ETHZ-00-1	1.0	57.2
BASE: trm-single	1.0	57.1
cmu_tartan_01-0	1.0	57.1
BASE: trm-shared	1.0	57.1
cmu_tartan_00-0	12.0	56.8
cmu_tartan_01-1	12.0	56.5
cmu_tartan_00-1	12.0	55.9
LTI-00-1	12.0	55.0
cmu_tartan_02-1	16.0	54.1
BASE: mono-single	16.0	54.0
BASE: mono-shared	16.0	54.0
ETHZ-02-1	16.0	53.6
BASE: mono-aug-single	16.0	53.5
BASE: mono-aug-shared	16.0	53.5
flexica-02-1	16.0	53.1
flexica-03-1	16.0	52.7
NYU-CUBoulder-01-0	24.0	50.0
NYU-CUBoulder-03-0	24.0	48.8
NYU-CUBoulder-02-0	24.0	48.6
NYU-CUBoulder-04-0	24.0	48.2
flexica-01-1	28.0	46.5
CU7565-02-0	29.0	43.5
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		67.7
Oracle (Submissions)		75.9
Oracle (All)		76.3

(b) Results on the Tungusic family (1 language)

Table 14: **Results per Language Family: Turkic and Uralic**

System	Rank	Acc
BASE: trm-single	1.0	91.8
BASE: trm-aug-single	1.0	91.8
uiuc-01-0	1.8	92.0
CULing-01-0	3.5	91.9
deepspin-02-1	6.7	91.3
deepspin-01-1	6.7	91.1
NYU-CUBoulder-04-0	5.5	90.4
BASE: mono-single	5.1	90.9
NYU-CUBoulder-02-0	6.8	90.6
NYU-CUBoulder-03-0	6.8	90.5
cmu_tartan_01-1	7.2	91.0
cmu_tartan_00-1	6.6	90.8
BASE: mono-aug-single	7.3	90.7
BASE: trm-shared	7.7	91.3
cmu_tartan_02-1	7.4	90.8
NYU-CUBoulder-01-0	8.9	90.5
BASE: trm-aug-shared	9.3	91.1
cmu_tartan_00-0	9.7	90.9
cmu_tartan_01-0	11.8	90.7
ETHZ-00-1	16.6	88.9
IMS-00-0	11.2	91.0
BASE: mono-shared	15.1	88.9
flexica-02-1	13.1	89.7
LTI-00-1	17.1	83.3
flexica-03-1	17.0	88.6
BASE: mono-aug-shared	19.5	86.3
CU7565-02-0	21.6	85.9
ETHZ-02-1	17.5	88.6
**CU7565-01-0*	*29.1*	*96.4*
flexica-01-1	28.9	72.4
Oracle (Baselines)		95.8
Oracle (Submissions)		97.4
Oracle (All)		97.5

(a) Results on the Turkic family (10 languages)

System	Rank	Acc
deepspin-02-1	1.8	90.7
deepspin-01-1	3.1	89.7
uiuc-01-0	2.8	88.2
CULing-01-0	3.9	88.9
BASE: trm-single	3.8	88.1
BASE: trm-aug-single	4.3	88.5
NYU-CUBoulder-04-0	10.6	86.8
NYU-CUBoulder-02-0	13.4	86.4
NYU-CUBoulder-03-0	13.4	86.0
IMS-00-0	14.8	86.1
NYU-CUBoulder-01-0	15.4	85.9
cmu_tartan_00-1	7.7	85.8
cmu_tartan_02-1	9.8	84.8
LTI-00-1	12.3	86.7
cmu_tartan_01-1	7.6	86.0
cmu_tartan_00-0	8.7	86.2
BASE: trm-aug-shared	18.8	82.6
**CU7565-02-0*	*22.2*	*79.4*
**CU7565-01-0*	*28.2*	*92.9*
BASE: mono-single	10.8	83.0
cmu_tartan_01-0	10.6	84.8
BASE: mono-shared	17.6	81.1
BASE: mono-aug-shared	19.4	81.9
BASE: trm-shared	19.5	76.8
ETHZ-02-1	22.6	67.9
BASE: mono-aug-single	11.4	85.9
flexica-02-1	19.5	70.7
flexica-03-1	20.5	67.8
flexica-01-1	26.8	66.0
ETHZ-00-1	28.3	54.9
Oracle (Baselines)		95.5
Oracle (Submissions)		96.8
Oracle (All)		97.2

(b) Results on the Uralic family (16 languages)

Table 15: **Results per Language Family (Uto-Aztecan) and Semitic Genus (Afro-Asiatic Family)**

System	Rank	Acc
uiuc-01-0	1.0	82.5
NYU-CUBoulder-01-0	1.0	82.2
NYU-CUBoulder-02-0	1.0	81.8
NYU-CUBoulder-03-0	1.0	81.5
IMS-00-0	1.0	81.5
BASE: trm-single	1.0	80.9
CULing-01-0	1.0	80.9
BASE: trm-shared	1.0	80.9
deepspin-02-1	1.0	80.6
NYU-CUBoulder-04-0	1.0	79.6
ETHZ-00-1	1.0	79.3
LTI-00-1	1.0	79.0
deepspin-01-1	1.0	79.0
BASE: trm-aug-single	14.0	78.0
BASE: trm-aug-shared	14.0	78.0
flexica-02-1	14.0	77.7
BASE: mono-aug-single	14.0	77.4
BASE: mono-aug-shared	14.0	77.4
cmu_tartan_00-0	14.0	76.1
cmu_tartan_00-1	14.0	76.1
cmu_tartan_01-0	14.0	75.8
cmu_tartan_01-1	14.0	75.8
BASE: mono-shared	14.0	75.8
BASE: mono-single	14.0	75.8
flexica-03-1	14.0	75.5
ETHZ-02-1	14.0	74.5
cmu_tartan_02-1	14.0	74.2
CU7565-01-0	28.0	71.0
CU7565-02-0	29.0	62.4
flexica-01-1	30.0	54.8
Oracle (Baselines)		87.2
Oracle (Submissions)		92.0
Oracle (All)		92.3

(a) Results on the Uto-Aztecan family (1 language)

System	Rank	Acc
uiuc-01-0	1.0	95.6
CULing-01-0	1.0	94.9
deepspin-02-1	5.0	93.3
BASE: trm-single	5.5	93.9
BASE: trm-aug-single	5.5	93.1
deepspin-01-1	5.5	92.5
NYU-CUBoulder-01-0	5.5	92.4
NYU-CUBoulder-02-0	5.5	92.3
NYU-CUBoulder-04-0	14.0	92.0
BASE: mono-aug-shared	9.0	91.3
BASE: mono-single	9.0	90.2
cmu_tartan_00-0	9.0	90.0
cmu_tartan_01-1	13.5	85.4
cmu_tartan_01-0	13.5	85.2
cmu_tartan_02-1	14.5	72.3
ETHZ-00-1	9.5	92.5
BASE: trm-aug-shared	9.5	91.8
BASE: trm-shared	9.5	91.7
IMS-00-0	9.5	91.7
BASE: mono-aug-single	9.5	90.9
CU7565-02-0	9.5	90.6
NYU-CUBoulder-03-0	18.0	91.2
LTI-00-1	13.5	90.1
flexica-02-1	13.5	90.1
ETHZ-02-1	13.5	89.5
flexica-03-1	13.5	89.2
cmu_tartan_00-1	13.5	89.0
BASE: mono-shared	13.5	88.5
**CU7565-01-0*	*28.5*	*88.3*
flexica-01-1	28.0	63.9
Oracle (Baselines)		98.4
Oracle (Submissions)		99.6
Oracle (All)		99.7

(b) Results on the Semitic genus (2 languages)

Table 16: **Results per Language Genus (in Indo-European family)**

System	Rank	Acc
deepspin-02-1	3.4	87.1
deepspin-01-1	4.6	87.0
uiuc-01-0	3.5	87.4
BASE: trm-single	3.1	87.5
CULing-01-0	3.5	88.3
BASE: trm-aug-single	4.9	87.4
IMS-00-0	15.1	83.1
BASE: mono-single	5.3	86.3
BASE: mono-aug-single	6.8	86.3
NYU-CUBoulder-04-0	10.2	85.2
NYU-CUBoulder-02-0	13.1	83.3
NYU-CUBoulder-03-0	12.0	84.4
LTI-00-1	11.1	84.3
cmu_tartan_00-1	9.8	79.5
NYU-CUBoulder-01-0	14.5	83.0
BASE: mono-aug-shared	13.2	84.4
cmu_tartan_01-0	11.1	78.9
cmu_tartan_01-1	11.1	78.8
cmu_tartan_00-0	10.8	79.3
BASE: trm-shared	19.5	77.7
BASE: trm-aug-shared	19.5	79.1
BASE: mono-shared	11.7	83.7
cmu_tartan_02-1	13.2	78.5
CU7565-02-0	19.4	78.6
ETHZ-02-1	18.9	76.4
flexica-01-1	26.2	66.6
flexica-03-1	25.5	66.5
flexica-02-1	25.9	64.2
ETHZ-00-1	27.1	60.1
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		97.0
Oracle (Submissions)		98.4
Oracle (All)		98.9

(a) Results on the Germanic genus (13 languages)

System	Rank	Acc
uiuc-01-0	1.0	98.2
deepspin-02-1	1.5	98.1
deepspin-01-1	1.5	98.0
BASE: trm-single	1.5	97.9
BASE: trm-aug-single	1.5	97.8
BASE: trm-shared	2.8	97.9
CULing-01-0	7.5	98.0
BASE: mono-single	6.0	97.6
NYU-CUBoulder-04-0	5.0	97.7
cmu_tartan_02-1	7.8	97.4
cmu_tartan_00-1	7.0	97.4
BASE: mono-shared	7.0	97.3
cmu_tartan_01-1	7.8	97.3
cmu_tartan_00-0	8.8	97.1
NYU-CUBoulder-03-0	8.5	97.4
NYU-CUBoulder-02-0	9.2	97.4
NYU-CUBoulder-01-0	9.2	97.3
BASE: trm-aug-shared	11.0	97.7
BASE: mono-aug-single	9.5	97.2
flexica-03-1	9.5	97.1
flexica-02-1	11.0	96.8
ETHZ-02-1	11.5	97.4
ETHZ-00-1	13.8	96.4
BASE: mono-aug-shared	15.8	94.2
cmu_tartan_01-0	17.2	96.9
IMS-00-0	17.0	96.6
CU7565-02-0	19.8	94.8
LTI-00-1	19.8	81.5
**CU7565-01-0*	*29.0*	*89.0*
flexica-01-1	28.8	88.1
Oracle (Baselines)		99.2
Oracle (Submissions)		99.6
Oracle (All)		99.7

(b) Results on the Indic genus (4 languages)

Table 17: **Results per Language Genus (in Indo-European family)**

System	Rank	Acc
CULing-01-0	1.0	95.3
deepspin-01-1	2.0	94.6
deepspin-02-1	2.0	94.6
BASE: trm-aug-shared	2.0	94.5
BASE: trm-aug-single	2.0	94.5
BASE: trm-shared	2.0	86.2
cmu_tartan_02-1	4.3	94.0
BASE: mono-aug-single	4.3	93.8
BASE: mono-shared	4.3	92.0
NYU-CUBoulder-03-0	4.3	91.8
cmu_tartan_00-1	4.3	91.8
ETHZ-02-1	4.3	91.8
NYU-CUBoulder-04-0	4.3	83.7
uiuc-01-0	9.3	82.5
BASE: trm-single	9.3	82.2
IMS-00-0	9.3	94.3
ETHZ-00-1	10.3	81.7
cmu_tartan_01-0	10.0	94.0
cmu_tartan_01-1	10.0	93.8
cmu_tartan_00-0	13.0	91.9
NYU-CUBoulder-02-0	10.0	91.8
flexica-03-1	11.7	87.2
BASE: mono-single	14.0	62.7
**CU7565-01-0*	*20.3*	*93.8*
BASE: mono-aug-shared	14.7	93.3
NYU-CUBoulder-01-0	14.7	91.4
CU7565-02-0	17.7	90.9
LTI-00-1	18.3	86.2
flexica-01-1	19.3	77.5
flexica-02-1	20.0	70.6
Oracle (Baselines)		97.3
Oracle (Submissions)		97.5
Oracle (All)		97.7

(a) Results on the Iranian genus (3 languages)

System	Rank	Acc
deepspin-02-1	1.0	99.3
BASE: trm-single	1.0	99.2
deepspin-01-1	1.0	99.1
uiuc-01-0	1.0	98.7
BASE: trm-aug-single	2.5	98.7
CULing-01-0	3.8	99.1
cmu_tartan_00-0	4.4	98.0
BASE: mono-shared	7.1	97.0
NYU-CUBoulder-04-0	4.9	98.8
cmu_tartan_01-0	6.4	98.2
BASE: trm-shared	8.0	96.6
BASE: mono-aug-shared	10.4	97.6
cmu_tartan_00-1	7.4	97.9
cmu_tartan_01-1	9.0	98.1
NYU-CUBoulder-03-0	9.8	98.9
NYU-CUBoulder-01-0	9.8	98.6
NYU-CUBoulder-02-0	10.2	98.5
BASE: mono-aug-single	10.2	97.5
BASE: mono-single	11.5	95.5
cmu_tartan_02-1	10.5	97.8
BASE: trm-aug-shared	14.5	97.2
flexica-03-1	17.2	93.1
IMS-00-0	19.0	97.6
LTI-00-1	20.4	96.3
CU7565-02-0	23.1	92.9
flexica-02-1	21.2	92.0
flexica-01-1	26.9	87.1
ETHZ-02-1	25.1	86.1
ETHZ-00-1	27.5	81.4
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		99.4
Oracle (Submissions)		99.7
Oracle (All)		99.7

(b) Results on the Romance genus (8 languages)

Table 18: **Results per Language Genus (in Niger-Congo family)**

System	Rank	Acc
uiuc-01-0	1.0	97.7
IMS-00-0	1.0	97.6
CULing-01-0	1.0	96.9
NYU-CUBoulder-01-0	1.4	97.9
NYU-CUBoulder-02-0	1.4	97.9
NYU-CUBoulder-03-0	1.4	97.9
deepspin-02-1	1.4	97.6
BASE: mono-aug-single	1.4	97.5
BASE: trm-single	1.4	97.4
deepspin-01-1	1.4	97.3
NYU-CUBoulder-04-0	1.4	97.3
LTI-00-1	1.4	97.3
BASE: trm-shared	1.4	97.2
BASE: mono-single	1.4	97.1
BASE: trm-aug-single	1.4	97.0
BASE: mono-shared	1.4	97.0
BASE: trm-aug-shared	1.4	96.7
BASE: mono-aug-shared	1.4	96.6
**CU7565-01-0*	*4.6*	*97.4*
flexica-02-1	3.6	96.2
flexica-03-1	3.6	96.2
CU7565-02-0	4.2	95.8
cmu_tartan_01-1	4.2	95.6
cmu_tartan_01-0	4.2	95.5
cmu_tartan_00-0	4.2	95.5
cmu_tartan_00-1	6.5	94.9
flexica-01-1	7.9	93.4
cmu_tartan_02-1	13.8	93.3
ETHZ-02-1	16.9	92.0
ETHZ-00-1	18.2	89.6
Oracle (Baselines)		98.9
Oracle (Submissions)		99.3
Oracle (All)		99.5

(a) Results on the Bantoid genus (8 languages)

System	Rank	Acc
BASE: mono-shared	1.0	100.0
BASE: mono-single	1.0	100.0
CU7565-01-0	1.0	100.0
IMS-00-0	1.0	100.0
deepspin-02-1	1.0	100.0
deepspin-01-1	1.0	100.0
flexica-03-1	1.0	99.9
BASE: trm-shared	1.0	99.9
BASE: mono-aug-single	1.0	99.9
cmu_tartan_00-0	1.0	99.9
BASE: trm-aug-shared	1.0	99.9
BASE: trm-aug-single	1.0	99.7
cmu_tartan_01-1	1.0	99.7
BASE: mono-aug-shared	1.0	99.6
NYU-CUBoulder-04-0	1.0	99.6
LTI-00-1	1.0	99.5
flexica-02-1	1.0	99.3
cmu_tartan_01-0	1.0	99.3
BASE: trm-single	1.0	98.8
NYU-CUBoulder-01-0	1.0	98.8
NYU-CUBoulder-02-0	1.0	98.8
NYU-CUBoulder-03-0	1.0	98.7
cmu_tartan_02-1	1.0	98.7
uiuc-01-0	1.0	98.5
CULing-01-0	13.0	98.0
cmu_tartan_00-1	13.0	97.7
flexica-01-1	14.5	97.4
CU7565-02-0	15.5	94.9
ETHZ-02-1	27.0	90.4
ETHZ-00-1	28.5	87.9
Oracle (Baselines)		100.0
Oracle (Submissions)		100.0
Oracle (All)		100.0

(b) Results on the Kwa genus (2 languages)

Table 19: **Results per Language Genus (in Oto-Manguean Family)**

System	Rank	Acc
CULing-01-0	1.0	93.9
uiuc-01-0	1.0	93.5
BASE: trm-single	1.0	92.8
deepspin-01-1	2.5	93.1
NYU-CUBoulder-04-0	2.5	93.1
deepspin-02-1	2.5	92.6
NYU-CUBoulder-03-0	2.5	92.5
NYU-CUBoulder-02-0	6.0	92.3
BASE: mono-single	6.0	92.1
NYU-CUBoulder-01-0	6.0	92.0
BASE: mono-aug-single	6.0	91.6
BASE: trm-aug-single	6.0	91.4
IMS-00-0	10.5	91.4
BASE: mono-aug-shared	10.5	90.0
BASE: mono-shared	10.5	89.9
LTI-00-1	13.0	89.6
cmu_tartan_00-1	13.0	87.9
ETHZ-02-1	15.5	89.7
BASE: trm-shared	15.5	89.5
cmu_tartan_02-1	18.0	87.3
cmu_tartan_00-0	18.0	87.1
cmu_tartan_01-1	20.5	86.7
cmu_tartan_01-0	18.0	86.3
BASE: trm-aug-shared	21.0	84.2
ETHZ-00-1	22.0	82.7
**CU7565-01-0*	*28.0*	*81.7*
CU7565-02-0	26.5	76.3
flexica-02-1	26.5	69.2
flexica-03-1	28.0	66.1
flexica-01-1	29.5	40.9
Oracle (Baselines)		96.4
Oracle (Submissions)		97.1
Oracle (All)		97.4

(a) Results on the Amuzgo-Mixtecan genus (2 languages)

System	Rank	Acc
uiuc-01-0	1.0	81.1
CULing-01-0	1.5	80.3
BASE: trm-single	3.5	78.9
deepspin-02-1	2.2	78.7
deepspin-01-1	2.2	78.3
NYU-CUBoulder-04-0	2.2	77.2
IMS-00-0	3.8	78.0
NYU-CUBoulder-02-0	3.8	77.1
NYU-CUBoulder-03-0	3.8	77.0
LTI-00-1	6.8	73.9
NYU-CUBoulder-01-0	4.8	77.5
BASE: mono-aug-single	8.2	73.8
BASE: mono-aug-shared	9.2	72.9
cmu_tartan_01-1	12.0	69.2
cmu_tartan_00-0	13.0	68.5
cmu_tartan_02-1	13.0	68.5
BASE: trm-aug-shared	15.2	65.9
BASE: mono-shared	11.2	73.5
flexica-01-1	21.8	51.0
BASE: trm-aug-single	9.2	75.7
ETHZ-02-1	15.0	71.7
CU7565-02-0	16.5	68.5
BASE: trm-shared	15.2	71.0
BASE: mono-single	15.2	70.4
cmu_tartan_00-1	16.5	68.9
cmu_tartan_01-0	17.5	66.5
**CU7565-01-0*	*26.2*	*75.7*
ETHZ-00-1	26.2	60.5
flexica-02-1	27.0	54.3
flexica-03-1	28.2	49.0
Oracle (Baselines)		89.9
Oracle (Submissions)		93.7
Oracle (All)		94.3

(b) Results on the Zapotecan genus (4 languages)

Table 20: **Results per Language Genus (in Oto-Manguean and Uralic Families)**

System	Rank	Acc
BASE: mono-shared	1.0	98.6
uiuc-01-0	1.0	98.6
deepspin-02-1	1.0	98.5
BASE: trm-single	1.0	98.4
BASE: mono-single	1.0	98.4
BASE: mono-aug-single	1.0	98.4
deepspin-01-1	1.0	98.4
BASE: mono-aug-shared	8.0	98.2
BASE: trm-aug-single	8.0	98.1
CULing-01-0	9.5	97.7
LTI-00-1	11.5	97.2
cmu_tartan_01-1	12.0	96.2
cmu_tartan_00-1	12.0	96.8
cmu_tartan_00-0	12.0	96.7
NYU-CUBoulder-04-0	13.5	96.5
cmu_tartan_02-1	14.0	96.3
ETHZ-02-1	15.5	95.9
BASE: trm-shared	16.5	94.2
NYU-CUBoulder-03-0	18.5	94.1
NYU-CUBoulder-02-0	18.5	94.1
NYU-CUBoulder-01-0	20.0	93.7
flexica-03-1	21.0	93.1
flexica-02-1	22.5	93.1
cmu_tartan_01-0	20.5	91.9
CU7565-02-0	25.0	91.1
IMS-00-0	24.5	91.0
**CU7565-01-0*	*28.5*	*90.9*
BASE: trm-aug-shared	25.5	87.3
ETHZ-00-1	27.5	85.3
flexica-01-1	29.5	64.2
Oracle (Baselines)		99.7
Oracle (Submissions)		99.9
Oracle (All)		99.9

(a) Results on the Otomian genus (2 languages)

System	Rank	Acc
deepspin-02-1	2.2	87.4
uiuc-01-0	2.6	83.5
deepspin-01-1	3.8	85.8
BASE: trm-aug-single	4.0	84.1
BASE: trm-single	4.3	83.4
CULing-01-0	5.2	84.6
NYU-CUBoulder-04-0	7.0	83.0
NYU-CUBoulder-02-0	10.0	82.8
NYU-CUBoulder-03-0	9.8	82.2
IMS-00-0	12.3	82.2
NYU-CUBoulder-01-0	12.0	82.4
cmu_tartan_00-1	8.3	80.0
cmu_tartan_02-1	8.3	80.2
LTI-00-1	12.3	81.9
cmu_tartan_01-1	8.0	80.3
cmu_tartan_00-0	9.4	80.8
BASE: trm-aug-shared	18.9	76.9
CU7565-02-0	20.3	74.0
**CU7565-01-0*	*27.1*	*92.9*
BASE: mono-single	12.6	75.5
cmu_tartan_01-0	11.7	78.6
BASE: mono-shared	15.8	74.8
BASE: mono-aug-shared	16.9	77.4
BASE: trm-shared	21.2	67.3
ETHZ-02-1	20.6	61.0
BASE: mono-aug-single	11.2	80.7
flexica-02-1	21.2	57.3
flexica-03-1	23.0	52.5
flexica-01-1	26.6	56.1
ETHZ-00-1	28.2	45.7
Oracle (Baselines)		93.9
Oracle (Submissions)		95.8
Oracle (All)		96.3

(b) Results on the Finnic genus (10 languages)

Table 21: **Results per Language Genus (in Uralic Family)**

System	Rank	Acc
deepspin-01-1	1.0	97.9
deepspin-02-1	1.0	97.9
CULing-01-0	2.0	97.8
BASE: trm-single	3.0	97.7
cmu_tartan_00-1	5.0	97.4
uiuc-01-0	5.0	97.6
BASE: trm-aug-single	5.0	97.6
cmu_tartan_00-0	6.0	97.4
cmu_tartan_01-1	6.0	97.3
cmu_tartan_02-1	12.5	95.6
cmu_tartan_01-0	9.0	97.1
BASE: mono-single	9.5	97.0
BASE: mono-aug-single	11.0	96.7
NYU-CUBoulder-04-0	14.0	95.6
LTI-00-1	13.5	96.7
BASE: trm-shared	14.5	95.7
BASE: trm-aug-shared	17.0	95.6
flexica-02-1	18.5	95.0
NYU-CUBoulder-02-0	18.5	94.8
IMS-00-0	19.0	94.8
NYU-CUBoulder-03-0	18.5	94.8
NYU-CUBoulder-01-0	18.5	94.7
flexica-03-1	19.0	94.6
BASE: mono-shared	21.0	94.5
CU7565-02-0	23.5	93.3
BASE: mono-aug-shared	26.0	91.5
flexica-01-1	27.0	88.7
ETHZ-02-1	28.0	79.4
ETHZ-00-1	29.0	73.4
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		98.6
Oracle (Submissions)		99.0
Oracle (All)		99.2

(a) Results on the Permic genus (2 languages)

System	Rank	Acc
deepspin-02-1	1.0	94.0
CULing-01-0	1.0	93.9
BASE: trm-single	1.0	93.9
uiuc-01-0	1.0	93.8
BASE: trm-aug-single	3.5	93.7
deepspin-01-1	3.5	93.6
cmu_tartan_02-1	6.5	93.3
cmu_tartan_00-1	6.5	93.2
cmu_tartan_01-1	6.5	93.2
cmu_tartan_01-0	6.5	93.2
cmu_tartan_00-0	6.5	93.2
BASE: mono-single	9.5	93.0
LTI-00-1	9.5	92.8
BASE: trm-shared	13.5	92.0
BASE: mono-aug-single	14.5	92.3
BASE: trm-aug-shared	15.0	91.9
IMS-00-0	17.0	91.5
NYU-CUBoulder-04-0	18.5	90.8
flexica-03-1	18.5	90.5
flexica-02-1	18.5	90.5
NYU-CUBoulder-03-0	19.5	90.2
NYU-CUBoulder-02-0	19.5	90.2
NYU-CUBoulder-01-0	23.5	89.5
BASE: mono-shared	21.5	88.9
BASE: mono-aug-shared	24.5	87.2
CU7565-02-0	25.5	85.2
flexica-01-1	27.0	82.1
ETHZ-02-1	28.0	73.7
ETHZ-00-1	28.5	67.9
**CU7565-01-0*	*30.0*	*0.0*
Oracle (Baselines)		97.0
Oracle (Submissions)		97.6
Oracle (All)		98.0

(b) Results on the Mordvin genus (2 languages)