

SIGMORPHON–UniMorph 2022

Shared Task 0: Generalization and Typologically Diverse Morphological Inflection

Conference Paper**Author(s):**

Kodner, Jordan; Khalifa, Salam; Batsuren, Khuyagbaatar; Dolatian, Hossep; Cotterell, Ryan; Akkuş, Faruk; Anastasopoulos, Antonios; Andrushko, Taras; Arora, Aryaman; Atanalov, Nona; Bella, Gábor; Budianskaya, Elena; Ghanggo Ate, Yustinus; Goldman, Omer; Guriel, David; Guriel, Simon; Guriel-Agiashvili, Silvia; Kieraś, Witold; Krizhanovsky, Andrew; Krizhanovsky, Natalia; Marchenko, Igor; Markowska, Magdalena; Mashkovtseva, Polina; Nepomniashchaya, Maria; Rodionova, Daria; Sheifer, Karina; Serova, Alexandra; Yemelina, Anastasia; Young, Jeremiah; Vylomova, Ekaterina

Publication date:

2022-07

Permanent link:

<https://doi.org/10.3929/ethz-b-000588601>

Rights / license:

[Creative Commons Attribution 4.0 International](#)

Originally published in:

<https://doi.org/10.18653/v1/2022.sigmorphon-1.19>

SIGMORPHON–UniMorph 2022 Shared Task 0: Generalization and Typologically Diverse Morphological Inflection

Jordan Kodner¹ Salam Khalifa¹ Khuyagbaatar Batsuren² Hossep Dolatian¹
Ryan Cotterell³ Faruk Akkuş⁴ Antonios Anastasopoulos⁵ Taras Andrushko⁶
Aryaman Arora⁷ Nona Atanelov Gábor Bella⁸ Elena Budianskaya⁹
Yustinus Ghanggo Ate^{10,11} Omer Goldman¹² David Guriel¹² Simon Guriel
Silvia Guriel-Agiashvili Witold Kieraś¹³ Andrew Krizhanovsky¹⁴
Natalia Krizhanovsky¹⁴ Igor Marchenko⁶ Magdalena Markowska¹
Polina Mashkovtseva⁶ Maria Nepomniashchaya⁶ Daria Rodionova⁶ Karina Sheifer^{6,9}
Alexandra Serova⁶ Anastasia Yemelina⁶ Jeremiah Young¹⁵ Ekaterina Vylomova¹⁶
¹Stony Brook University ²National University of Mongolia ³ETH Zürich
⁴University of Massachusetts-Amherst ⁵George Mason University
⁶Higher School of Economics ⁷Georgetown University ⁸University of Trento
⁹Institute of Linguistics, Russian Academy of Sciences ¹⁰STKIP Weetebula
¹¹Australian National University ¹²Bar-Ilan University
¹³Institute of Computer Science, Polish Academy of Sciences
¹⁴Karelian Research Centre of the Russian Academy of Sciences
¹⁵University of Oregon ¹⁶University of Melbourne
jordan.kodner@stonybrook.edu vylomovae@unimelb.edu.au

Abstract

The 2022 SIGMORPHON–UniMorph shared task on large scale morphological inflection generation included a wide range of typologically diverse languages: 33 languages from 11 top-level language families: Arabic (Modern Standard), Assamese, Braj, Chukchi, Eastern Armenian, Evenki, Georgian, Gothic, Gujarati, Hebrew, Hungarian, Itelmen, Karelian, Kazakh, Ket, Khalkha Mongolian, Kholosi, Korean, Lamaholot, Low German, Ludic, Magahi, Middle Low German, Old English, Old High German, Old Norse, Polish, Pomak, Slovak, Turkish, Upper Sorbian, Veps, and Xibe. We emphasize generalization along different dimensions this year by evaluating test items with unseen lemmas and unseen features separately under small and large training conditions. Across the six submitted systems and two baselines, the prediction of inflections with unseen features proved challenging, with average performance decreased substantially from last year. This was true even for languages for which the forms were in principle predictable, which suggests that further work is needed in designing systems that capture the various types of generalization required for the world’s languages.¹

¹Data, evaluation scripts, and predictions are available at: <https://github.com/sigmorphon/2022InflectionST>

1 Introduction

Generalization, the ability to extend patterns from known to unknown items, is a critical part of morphological competence. Morphological systems, both human and machine, must be able to recognize and produce novel items as new words are encountered. Every learner, every speaker, and any system intended for general use constantly encounters new words, both new coinings and existing words that are new to them.

The centrality of generalization is emphasized by the morphological sparsity that pervades language use. Inflected forms, lemmas, and inflectional categories are all sparsely distributed and highly skewed in any input sample, following long-tailed, often Zipfian, frequency distributions (Chan, 2008). This has serious implications for learning, since the overwhelming majority of lemmas, if present at all in the input, will only be attested in a fraction of their possible forms. This is true even for a language like English, with only five inflected forms per verb and two per noun, and the problem only grows as a language’s paradigms increase in size and complexity.

The test paradigm that the SIGMORPHON inflection shared tasks have employed since 2016 (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021)

provides one test bed for generalization in morphological learning systems. The shared tasks leverage the UniMorph Database (Kirov et al., 2018; McCarthy et al., 2020; Batsuren et al., 2022), which provides data sets for an ever-growing range of typologically diverse morphologies.

In principle, there are at least two kinds of generalization which can be evaluated in our UniMorph-based test paradigm: generalization to unseen lemmas, and generalization to unseen inflectional categories (i.e., unseen feature sets). Contrasting seen and unseen lemmas and categories yields four different test conditions: 1) prediction of the form of a novel combination of a seen lemma and seen feature set, 2) prediction given a seen lemma but novel feature set, 3) prediction given a seen feature set but novel lemma, and 4) the prediction of a form when both the lemma and feature set are novel.

This year’s shared task include 33 languages from 11 top-level language families with a particular focus on Eastern Europe, Central Asia, and Siberia: Arabic (Modern Standard), Assamese, Braj, Chukchi, Eastern Armenian, Evenki, Georgian, Gothic, Gujarati, Hebrew, Hungarian, Itelmen, Karelian, Kazakh, Ket, Khalkha Mongolian, Kholosi, Korean, Lamahalot, Low German, Ludic, Magahi, Middle Low German, Old English, Old High German, Old Norse, Polish, Pomak, Slovak, Turkish, Upper Sorbian, Veps, and Xibe. Many of these were included last year, but we hoped that running them again would provide further insights into generalization.

1.1 Motivation for Generalization Task

Generalization to the unseen is a challenging task, the feasibility of which should be sensitive to the organization of a given language’s morphology. For a language with rampant unpredictable stem mutations or suppletion, it may not always be possible to generalize patterns accurately to unseen lemmas, but one would hope that a system could generalize well for a language with invariant stems or highly irregular stem changes. Similarly, it may not be possible for a system to generalize to unseen categories for a highly fusional language where forms cannot be predicted from their component features, but it should be possible for highly agglutinative languages where roughly each feature corresponds to its own morphological operation or for a language with a high degree of syncretism in which the expression of an unseen inflectional category is

| Feature Set | <i>guakamole</i> |
|----------------|------------------------|
| N;ACC;SG | ? |
| N;ACC;PL | <i>guakamoleleri</i> |
| N;DAT;SG | <i>guakamoleye</i> |
| N;DAT;PL | ? |
| N;ACC;PL;PSS3S | <i>guakamolelerini</i> |
| N;DAT;PL;PSS3S | <i>guakamolelerine</i> |
| ... | ... |

Table 1: A partial paradigm for Turkish *guakamole* ‘guacamole,’ illustrating inference for novel feature sets in an agglutinative language.

likely the same as one that has already been learned. This was shown to be feasible in practice for Nen, a Papuan language with a large degree of syncretism (Muradoglu et al., 2020).

Previous iterations of this shared task have looked at some aspects of this problem, but none made this a focus. Last year’s task (Pimentel et al., 2021) reported separate performance numbers for seen and unseen lemmas, but did not control for seen/unseen feature overlap. The 2018 task (Cotterell et al., 2018), sampled train and test sets with frequency weighting from Wikipedia, which made for a more naturalistic sparse sampling setting, but did not control for either kind overlap. In preparation for this year’s iteration, we found that the proportion of test items with seen feature sets varied greatly across languages in the 2018 task and may have been a major driver of performance.

For example, the best performing system on Turkish, consistently scored just under the proportion of test items with seen feature sets at each training size (Table 2), even though Turkish is a agglutinative language for which generalization to unseen categories should be possible. Table 1 provides a partial noun paradigm from Turkish UniMorph which illustrates why this type of generalization should be possible. Say the feature sets N;ACC;SG and N;DAT;PL were never attested in training, but the lemma *guakamole* was. It should be possible to deduce their forms anyway – this would be a fair homework problem for an undergraduate course.

Looking at the table, *-ler-* corresponds to PL here, \emptyset to SG, and *-in-* to PSS3S. Both forms with ACC end in *-i*, while DAT seems to correspond to *-ye* in the singular and *-e* in the plural. From this alone, one can correctly infer that N;DAT;PL should be *guakamole-ler-e*, while N;ACC;SG should be *guakamole-yi* or maybe *guakamole-i*. The former is indeed correct: *y*-insertion is well attested elsewhere in the language and would certainly be

present with other lemmas and with other feature sets containing ACC. While unseen Turkish inflectional categories are not completely predictable, since they also contain some morphological eccentricities which obscure predictability, “could an undergraduate solve it?” is a good rule of thumb for whether generalization to unseen feature sets is a feasible task.

Performance was divergent on closely related languages whose test sets’ feature set overlaps differed. Turkish and Azeri are closely related Oghuz Turkic languages with some mutual intelligibility (Salehi and Neysani, 2017) and very similar morphological paradigms, nevertheless, scores for Azeri during the 2018 task were much higher than for Turkish. Table 3 shows feature overlap and performance for Azeri. It is tempting to propose that Azeri scores were higher than Turkish scores because overlap proportions were higher.

Taken together, this suggests two things. First, the proportion of test items with feature sets attested in training is an uncontrolled factor in the data that could be driving performance in a way that obscures language-internal factors. Second, this could suggest that the systems of the day were not able to generalize across inflectional categories,² but a more focused evaluation would be needed to investigate these hypotheses. We perform such an investigation this year.

| Turkish | Overlap% | Best Acc% | Δ |
|---------|----------|-----------|----------|
| Low | 39.600 | 39.500 | -0.1 |
| Medium | 94.100 | 90.700 | -3.4 |
| High | 100 | 98.500 | -1.5 |

Table 2: Comparison of best 2018 system accuracy on Turkish low-, medium-, and high-train conditions and percent of test items with feature sets attested during training.

| Azeri | Overlap% | Best Acc% | Δ |
|--------|----------|-----------|----------|
| Low | 71.000 | 65.000 | -6.0 |
| Medium | 99.000 | 96.000 | -3.0 |
| High | 100 | 100 | 0 |

Table 3: Comparison of best 2018 system accuracy on Azeri low-, medium-, and high-train conditions and percent of test items with feature sets attested during training.

²Recent work has shown that lemma overlap is also an important predictor of performance (Goldman et al., 2022), but an analysis of 2018 results suggests that feature set overlap is an even better predictor (see Appendix A).

2 Task Description

From the participants’ perspective, this task was organized very similarly to previous iterations. Participants were asked to design supervised learning systems which could predict an inflected form given a lemma and a morphological feature set corresponding to an inflectional category or cell in a morphological paradigm. They were provided with a small, and data permitting, large training set, as well as a development set and test set for each language. The train and dev sets consisted of (lemma, inflected, feature set) triples, while the inflected forms were held out from the test set.

Data was made available to participants in two phases. In the first phase, train and dev sets were provided, with the expectation that model development and tuning be carried out primarily on these languages. In the second phase, test sets were released for all languages during the evaluation phase. Teams produced predicted inflected forms for each test set. They were given the opportunity to submit two sets of predictions from two separate models, one trained on the small training sets and one trained on the large training sets, with the latter being a super set of the former.

3 Description of Languages

This section provides brief descriptions of each language that was newly included or newly updated for this year’s task. Further information about returning languages can be found in previous years’ papers (Vylomova et al., 2020; Pimentel et al., 2021). Table 4 summarizes the list of languages and provides citation and attribution information.

3.1 Armenian (Indo-European)

Armenian is an independent branch of the Indo-European family. Its oldest attested form is Old Armenian or Classical Armenian (~5th century). It has two modern standardized varieties: Western Armenian and **Eastern Armenian**. Western Armenian is a diasporic language that developed in the Ottoman Empire, while Eastern Armenian is the official language of the Republic of Armenia (Dum-Tragut, 2009). Inflection is largely agglutinative, with some residues of Indo-European fusional morphology. For verb morphology, verbs fall into different conjugation classes. Most tenses are formed via periphrasis via a non-finite converb and a finite auxiliary, though some tenses are synthetic. Nouns

| Family | Subfamily | ISO 639-2 | Language | Source of Data | Annotators |
|---------------------|-------------------|-----------|------------------------|---|---|
| Afro-Asiatic | Semitic | ara | Modern Standard Arabic | Taji et al. (2018) | Salam Khalifa Nizar Habash |
| | | heb | Hebrew | Wiktionary | Omer Goldman |
| Austronesian | Malayo-Polynesian | slp | Lamahalot | Nagaya (2012) | Yustinus Ghanggo Ate |
| Chukotko-Kamchatkan | Northern | ckt | Chukchi | Chuklang; Tyers and Mishchenkova (2020) | Karina Sheifer Maria Ryskina |
| | Southern | itl | Itelmen | | Karina Sheifer Sofya Ganieva Matvey Plugaryov |
| Indo-European | Armenian | hye | Eastern Armenian | Wiktionary | Hossep Dolatian |
| | Germanic | got | Gothic | Wiktionary | Khuyagbaatar Batsuren |
| | | nds | Low German | Wiktionary | Jeremiah Young |
| | | gml | Middle Low German | Wiktionary | " |
| | | ang | Old English | Wiktionary | Khuyagbaatar Batsuren |
| | | goh | Old High German | Wiktionary | Jeremiah Young |
| | | non | Old Norse | Wiktionary | " |
| | Indic | asm | Assamese | Wiktionary | Khuyagbaatar Batsuren Aryaman Arora Shyam Ratan |
| | | bra | Braj | Kumar et al. (2018) | Ritesh Kumar Aryaman Arora |
| | | guj | Gujarati | Wiktionary | Khuyaagbaatar Batsuren |
| | | hsi | Kholosi | Arora and Etebari (2021) | Aryaman Arora |
| | Slavic | mag | Magahi | Kumar et al. (2014) | Mohit Raj, Ritesh Kumar |
| | | pol | Polish | Woliński et al. (2020); Woliński and Kieraś (2016) | Witold Kieraś Marcin Woliński |
| | | poma | Pomak | Jusuf Karahodža et al. (2022) | Ritvan Karahodja |
| | | slo | Slovak | Hajič and Hric (2017) | Antonios Anastasopoulos |
| | | hsb | Upper Sorbian | Fraser (2020) | Witold Kieraś Taras Andrushko Igor Marchenko |
| Kartvelian | | kat | Georgian | Guriel et al. (2022) | David Guriel Simon Guriel Silvia Guriel-Agiashvili |
| | | | | | Nona Atanelov |
| Koreanic | | kor | Korean | Wiktionary | Maria Nepomniashchaya |
| Mongolic | Central | khk | Khalkha Mongolian | Munkhjargal et al. (2016); Batsuren et al. (2019) | Daria Rodionova Anastasia Yemelina |
| Tungusic | Northern | evn | Evenki | Kazakevich and Klyachko (2013) | Khuyagbaatar Batsuren |
| | Southern | hsb | Xibe | Zhou et al. (2020) | Elena Klyachko |
| Turkic | Kipchak | kaz | Kazakh | (Nabiye, 2015; Turkicum, 2019), Polish Wiktionary | Eleanor Chodroff Khuyagbaatar Batsuren |
| | Oghuz | tur | Turkish | Wiktionary | Omer Goldman Duygu Ataman |
| Uralic | Ugric | hun | Hungarian | Wiktionary | Judit Ács Khuyagbaatar Batsuren Gábor Bella, Ryan Cotterell |
| | | | | | Christo Kirov Andrew Krizhanovsky |
| | Finnic | kr1 | Karelian | Boyko et al. (2021, VepKar) | Natalia Krizhanovsky |
| | | lud | Ludic | Boyko et al. (2021, VepKar) | Elizabeth Salesky |
| Yeniseian | Northern | vep | Veps | Boyko et al. (2021, VepKar) | " " " |
| | | ket | Ket | Ket corpus | " " " |
| Yeniseian | Northern | ket | Ket | Ket corpus | Elena Budianskaya Polina Mashkovtseva |
| | | | | | Alexandra Serova |

Table 4: Languages presented in this year’s shared task

fall into different declension classes, based on the choice of plural and case suffixes.

3.2 Finno-Ugric (Uralic)

Finno-Ugric is a branch of Uralic, a language family with around 25 million native speakers spread between Northern Russia, Scandinavia, and Hungary. The majority of them are agglutinating and extensively use suffixes. They are also known for a relatively rich grammatical case system. Verbs are inflected for number, person, tense, and mood. Phonologically, these languages often present vowel harmony and palatalization.

Hungarian, with its 13 million native speakers, is the most widely spoken Uralic language. Hungarian is an agglutinative language with a rich set of affixes expressing derivation or inflection, such as in the verb. Another feature of Hungarian morphology, adding to its complexity from a computational perspective, is vowel harmony: the vowels of certain affixes adapt to those of the stem (Rounds, 2009). Compounding in Hungarian is frequent and productive, leading to further complexity in its morphological analysis (Kiefer and Nemeth, 2019).

Karelian and Ludic are two closely related Finnic varieties spoken in Russian and Finnish Karelia and the regions around Lakes Onega and Ladoga. The data for both languages, along with Veps returning from last year, has been collected as part of the VepKar project (Boyko et al., 2021) and includes multiple dialects. Typical of Finnic, these languages are highly agglutinative, present vowel harmony processes, and overtly express well over ten cases on nominals and adjectives. Ludic is often described as a dialect of Karelian, although it has certain unique features such as the presence of a reflexive conjugation (Novak et al., 2019) and the use of the full temporal paradigm of the conditional. It is seriously endangered, with about 150 remaining speakers.³

3.3 Georgian (Kartvelian)

Kartvelian, or South Caucasian, languages are primarily spoken in the South Caucasus with no demonstrable genetic relation to other languages in the region. Georgian, an official language of Georgia, has about four million speakers worldwide. Georgian morphology is mostly agglutinative. Nouns have number (singular/plural), but no grammatical gender. Its grammatical case system is relatively rich, having seven cases. Nouns are declined for number and case. Verbs exhibit polyper-

sonal agreement (incorporating the number and the person of both subject and objects). In addition, verbs are divided into 4 classes: transitive, intransitive, indirect, and medial, and present many irregularities.

3.4 Germanic (Indo-European)

The Germanic family constitutes one of the primary branches of Indo-European. It in turn contains three sub-branches. The West Germanic sub-branch includes English, Dutch, and German, among others. The North Germanic sub-branch contains the Germanic languages of Scandinavia. The East Germanic sub-branch is extinct and contained Gothic. At a high level, Germanic morphology is similar to that of other Indo-European branches, but it does diverge in some key ways (Ringe, 2017). Germanic languages, particularly in the past, had an inherited three-way gender distinction, an inherited three-way number system, and overt inflectional case systems, all reduced to some degree from Indo-European. Nominals fall into several inflectional classes with different case/number expressions.⁴ The 2020 shared task revealed some major inconsistencies in the data (Vylomova et al., 2020). In this iteration, the data has been re-extracted and checked.

Gothic is an extinct East Germanic language. Nearly the entire extant Gothic corpus comes from a partial translation of the Christian Bible by bishop Wulfila. Gothic is in many ways more conservative than other Germanic languages. It lacks Umlaut, which is a type of vowel alternation on nouns and verbs present in the rest of the family, but it retains reduplicated perfects, and it sometimes uses the accusative as a vocative case. Data for Gothic was sourced from Wiktionary and contains both Gothic script and Latin transcriptions.

Old English, Old High German, and Old Norse were three closely related West and North Germanic languages and early attested ancestors of modern English, High German varieties, and liv-

⁴Five of the six Germanic languages presented this year are historical. They no longer have living speakers, and their corpora are of a fixed size. Paradigms were initially extracted from Wiktionary. Given the highly skewed long-tailed distributions of inflected forms, lemmas, and inflected categories (Chan, 2008), which do not differ in historical corpora (Kodner, 2019), the large majority of potential inflected forms, even for known lemmas, are not attested in the historical record. As such, most of the forms in the full paradigms available on Wiktionary are generated and not actually attested. This is likely not a major concern for the purpose of this task, but the caveat must be expressed.

³<https://lyydi.net/>

ing North Germanic languages today. Inflectional classes in these languages are often less transparent than in Gothic due to successive sound changes obscuring their basis.

Middle Low German was a collection of West Germanic dialects spoken along the southern North Sea coast. It was a major trade language, the *lingua franca* of the Hanseatic League during the European Medieval period. The language retains overt case distinctions on nominals, but it shows a greater degree of syncretism than earlier Germanic languages. This trend of increased syncretism extends to the verbal system as well (Lasch, 1914).

Low German is a collection of West Germanic varieties descended from Middle Low German occupying an intermediate space in a dialect continuum between Dutch and High German. Varieties exist in a state of diglossia, mostly with Standard German, a High German variety. Several million native speakers remain in the 21st century, though numbers are declining. Outside of Europe, Low German is spoken in some diaspora communities including Mennonite groups in the Americas.

3.5 Hebrew (Semitic)

The Semitic languages, a branch of the larger Afro-Asiatic family, are spoken by over 300 million people across North Africa and Southwest Asia. Hebrew is a Northwest Semitic language with around 5 million native speakers, spoken mainly in Israel. Typically of Semitic, Hebrew makes heavy use of *templatic* non-concatenative morphology (Coffin and Bolozky, 2005). Verbs are expressed through triliteral consonant roots which occupy slots in a template of vowels. Verbs occupy inflectional classes called *binyanim* in Hebrew. Person, number, and tense marking is indicated primarily with affixation. Both prefixation and suffixation are applied depending on the tense. Nouns and adjectives indicate gender and number through suffixation, sometimes with stem mutations. Verbs, nouns, and propositions may take possessive or pronominal object clitics. In the current shared task we introduce a vocalized version of Hebrew that has been recently added to the UniMorph.

3.6 Indic, or Indo-Aryan (Indo-European)

Indic is a branch of Indo-Iranian, itself a primary branch of Indo-European. The family has a long history, with a large attested corpus of Vedic and Classical Sanskrit. It currently has over 800 million speakers extending through all countries in South

Asia. Morphologically conservative languages express a three-way gender distinction and case on nouns, tense, aspect, mood, number, and person on verbs. Inflectional morphology is primarily suffixing. Some languages possess overt formality distinctions on verbs.

Assamese is mainly spoken in the northeast Indian state of Assam, with over 20 million native speakers. While gender is not grammatically marked, Assamese presents a rich system of noun classifiers. The Assamese data has been extracted from the English edition of Wiktionary. **Gujarati** (Baxi et al., 2021) is spoken predominantly in the Indian state of Gujarat, with over 50 million native speakers. **Kholosi** is an under-documented Indo-Aryan language spoken in two villages (Kholus and Gotav) in Hormozgan Province, Iran. The data has been collected during field work (Arora and Etebari, 2021).

3.7 Ket (Yeniseian)

Yeniseian languages were historically spoken along the Yenisei River region of central Siberia. Ket, the only living member, is critically endangered, with only about 60 remaining speakers at any level of linguistic competence. The language presents mainly agglutinative morphology, with extensive use of suffixes, prefixes, and infixes. Although verbal conjugation and noun declension systems are well-developed, the boundaries between word classes are fuzzy (Verner, 1997). Noun classes differentiate between masculine and non-masculine in the singular, animate and inanimate in the plural. The grammatical case system contains between 8 and 10 cases depending on the analysis. Ket verbs express polypersonal agreement, with the case and number of all arguments reflected on the verb.

The data for Ket was sourced from a text collection compiled during the field work of the Laboratory for Computational Lexicography of the Moscow State University, that took place between 2004 and 2009. It contains word forms from twelve categories, seven of which (ADJ, NUM, ADV, INTJ, ADP, PART, CONJ) are invariable.

3.8 Khalkha Mongolian (Mongolic)

The Mongolic language family has 5.200 million active speakers of 14 language varieties, which are actively spoken in Mongolia, Russia, China, and Afghanistan. The Khalkha Mongolian is de facto the official national language of Mongolia and both the most widely spoken and most-known

member of the Mongolic language family. Khalkha Mongolian is an agglutinative language with a rich set of suffixes, but no prefixes. It also expresses complex vowel harmony patterns (Jaimai et al., 2005).

3.9 Korean (Koreanic)

Korean, spoken by about 80 million people, is often described as a language isolate. However, the Jeju dialect, spoken on the southern island of Jeju is highly divergent and often considered its own language. The language expresses limited inflectional morphology on nominals. Verbs express valency, tense, aspect, mood, and various dimensions of formality through suffixation. The current dataset consists of mostly predicates, so the resulting lemmas are mainly verbs and a smaller number of adjectives.

3.10 Lamaholot (Austronesian)

Lamaholot, or Solor, is one of the Central-Malayo-Polynesian languages, a proposed branch in the Malayo-Polynesian within Austronesian. As of 2010, it had about 200,000 native speakers, primarily on the eastern part of Flores Island, and neighboring islands of Flores (Solor, Adonara, Lembata, and Alor). Nearby Papuan languages have had a significant influence on this language phonologically and syntactically (Nagaya, 2011; Arka, 2007; Klamer, 2002, 2009). The language has several dialects. We use data mainly from the Lewotobi dialect (Nagaya, 2011) spoken by about 6,000 people in Kecamatan Ile Bura, East Flores. Morphologically, Lamaholot is a nearly isolating language (each word typically has one morpheme) with a small inventory of affixes (mostly prefixes and a handful of suffixes) and clitics (mainly enclitics). This language has two salient morphological features, namely agreement and nominalization.

3.11 Slavic (Indo-European)

Slavic, another primary branch of Indo-European, contains approximately 20 languages, with half of them having over 1 million speakers. The languages are spoken in Central and Eastern Europe, the Balkans, and Russia. They are traditionally divided into three branches: East Slavic (incl. Belarusian, Russian, Rusyn, and Ukrainian), West Slavic (incl. Czech, Kashubian, Polish, Silesian, Slovak, and Upper and Lower Sorbian, among others), and South Slavic (incl. varieties of Bosnian-Croatian-Montenegrin-Serbian, varieties of Macedonian and

Bulgarian including Pomak, and Slovenian).

Slavic morphology is generally typical of Indo-European, with several inflectional classes for both verbs and nouns, nominal inflection by case, number, and three genders. It elaborates Indo-European verbal inflectional paradigms marking aspect, tense, number, person, and sometimes gender.

Slovak (Mistrík, 1988), and **Upper Sorbian** are two closely related West Slavic languages. Masculine nouns additionally mark animacy, which is often described as a part of the gender system of these languages. The case systems of both languages are fairly similar, however in Slovak, vocative is usually syncretic with nominative. Upper Sorbian retains a dual number and has a greater variety of verbal past forms than other West Slavic languages. The Slovak data was obtained by automatic conversion of extensive inflectional dictionaries used for morphological analysis to the UniMorph scheme.⁵ The data for Upper Sorbian was combined from WMT and online grammars.⁶

Pomak is a South Slavic language, a dialect of Southeastern Bulgarian spoken in Greece and European Turkey. It has around 30,000 speakers as of 2021 but lacks standardized orthography (Jusuf Karahóga et al., 2022). Bulgarian and Macedonian varieties are unusual among Slavic for having mostly lost case marking on nouns and for marking voice synthetically on verbs.

4 Data Preparation

All data for this task is provided in standard UniMorph format, with training items consisting of (lemma, inflected form, morphosyntactic features) triples. Since the goal of the task is to predict inflected forms, the test set was presented as (lemma, features) pairs. Data was canonicalized as in previous years using <https://github.com/unimorph/um-canonicalize>, which ensures consistent ordering of the features in the feature sets.

4.1 Training-Test Overlap

As always, we ensured that there are no lemma-feature set pairs that occur in both the training and test sets. However, since test items contain both lemmas and features, other overlaps between training and test are possible. This year’s data splitting algorithm aimed to control for the four logically

⁵<https://github.com/unimorph/slk>

⁶https://www.statmt.org/wmt20/unsup_and_very_low_res/, <https://baltoslav.eu/hsb/>

possible licit types of lemma and feature overlap, which define four kinds of test items:

Both Overlap: Both the lemma and feature set of a training pair are attested in the training set (but not together in the same triple)

Lemma Overlap: A test pair’s lemma is attested in training, but its feature set is novel

Feature Overlap: A test pair’s feature set is attested in training, but its lemma is novel

Neither Overlap: A test pair is entirely unattested in training. Both its lemma and features are novel.

For illustration, consider the sample training and test sets provided in (1)-(2). In this example, each test pair exhibits a different kind of overlap.

(1) **Example Training Set**

```
eat  eating  V;V.PTCP;PRS
run  ran      V;PST
```

(2) **Example Test Set**

```
eat  V;PST      <-- both
run  V;NFIN     <-- lemma
see  V;PST      <-- feature
go   V;PRS;3;SG <-- neither
```

4.2 Data Splits

The data set for each language was split into training, development, and test sets. For languages with sufficiently large corpora, both large and small training sets were produced with the small set being a subset of the large one. We aimed for 7,000/1,000/2,000-item large train/dev/test splits and a 700-item small train split when possible, but splits for most languages were somewhat smaller in practice. Chukchi, Kholosi, Lamahlot, and Xibe in particular were too small to extract even full small training sets, while Braj, Gujarati, Itelmen, Ket, Low German, Magahi, Middle Low German, Old High German, Upper Sorbian were too small to extract large training sets. Split sizes are summarized in Table 5.⁷

⁷Triples which shared their lemma and feature set with another item in the data were removed after splitting, which is why some languages fall short of 7,000/1,000/2,000 splits.

4.3 Motivation for Data Splitting

The sampling script attempts to control the size of each overlap category in the test set. The challenge here is controlling for both *lemma overlap* and *feature overlap* simultaneously. Since no frequency information is provided in the UniMorph annotation scheme, any uniform sampling over triples, controlling for *lemma overlap* or otherwise, will tend to drive *feature overlap* to near 100%. This is unnatural. Since both lemmas and inflectional categories tend to follow long-tailed sparse frequency distributions in real language (Chan, 2008, ch. 3), a naturalistic split weighted by token frequencies of individual items will tend to oversample high frequency lemmas and inflectional categories (i.e., feature sets), and undersample most others. This skewed sampling should yield a mix of overlap types in the test set. This is what was achieved in 2018, though the ratios of overlap types were uncontrolled. In contrast, this year’s data splitting achieves a controlled mixture of overlap types even in the absence of frequency information.

4.4 Splitting Process

The algorithm began by randomly partitioning a language’s feature sets into OVERLAPPABLE and NON-OVERLAPPABLE sets and uniformly sampling the large training set from only those triples that contain feature sets in OVERLAPPABLE. If there were not enough triples with with feature sets in OVERLAPPABLE for a given language, then the OVERLAPPABLE partition was increased incrementally until enough training triples could be sampled. If there was insufficient data to create the large training set, then the small training set was sampled this way instead. If there was enough data, then the small training set was down-sampled uniformly from the large training set.

The test set was sampled from the remaining items, with half drawn from triples with feature sets in OVERLAPPABLE and half from triples with feature sets in NON-OVERLAPPABLE features. The development set was drawn from the remainder in the same fashion.

As summarized in Table 5, this approach resulted in a much more even mixtures of overlapping pairs at both training sizes than is achieved by sampling that does not take *feature overlap* into account, though the actual ratios varied by language due to corpus-specific and language-specific factors. In controlling for *feature overlap*, a good mixture of

lemma overlap items is achieved simultaneously. Since most languages provide ample attestation of each overlap type, we could evaluate on each overlap type individually to gauge models’ generalization abilities across both the lemma and inflectional category dimensions. Additionally, in aiming for a more uniform ratio of overlap types across languages, overall performance on each language is more directly comparable.

5 Baseline Systems

The organizers provided one neural and one non-neural baseline system. The neural system, Neural, is a character-level transformer (Wu et al., 2021). It is identical to the system CHR-TRM which was used in the 2021 task. The non-neural system, NonNeur, is identical to the non-neural baseline made available in 2020 and 2021.⁸

6 Submitted Systems

CLUZH (Silvan Wehrli and Makarov, 2022): The CLUZH team adapted their earlier model, character-level neural transducer, to work on large datasets (Makarov and Clematide, 2020). The model has previously shown superior performance, especially in low-resource scenarios. This year, the team optimized the training procedure using mini-batches. They only relied on the teacher-forcing approach, i.e., using gold labels rather than what was predicted during the training phase. Morphosyntactic features were treated individually, and their embeddings were summed. The team explored performance of the model across various task settings and demonstrated its ability to capture feature behaviour better than other team’s models, especially in the small training condition. The system is identical to the one submitted to this year’s acquisition-inspired subtask (Kodner and Khalifa, 2022).

OSU (Elsner and Court, 2022): OSU’s system is identical to the one submitted to this year’s acquisition-inspired subtask. This inflection system is a transformer whose input is augmented with an analogical exemplar model showing how to inflect a different word into the target cell. In addition, alignment-based heuristic features indicate how well the exemplar is likely to match the output. The system works only when examples of the target cell are present in the training set and can serve as exemplars; otherwise, it outputs the

lemma as a placeholder. Thus, the system’s scores are expected to be higher for the *feature overlap* and *both overlap* evaluation categories and very low when the target cell is unknown.

TüMorph-Main (Merzhevich et al., 2022): TüMorph’s neural system is a modification of the character-level adaptation of transformer to morphology from Wu et al. (2021). In particular, the team trained the transformer to predict a distribution over states of FST (whose states are characters) rather than character sequences themselves. The model is scored third on both the small and large training settings.

TüMorph-FST (Merzhevich et al., 2022): As their second submission, the team manually developed FSTs using grammars and corresponding UnMorph repositories. Since that requires more human labour and linguistic competence, the team focused only on three languages: Chukchi, Kholosi, and Upper Sorbian. The resulting FST models outperformed all other submitted systems on two of three languages. The authors confirm earlier observations from Beemer et al. (2020) that such systems are able to reach superior results compared to neural ones, especially in low-resource scenarios and high morphological complexity, but require substantially more human working hours.

UBC (Yang et al., 2022): The UBC team proposed enriching the character-level transformer of Wu et al. (2021) with reverse positional embeddings to better account for suffixing, one of the most common word formation processes. In addition, the team explored a synthetic data augmentation technique proposed by Anastasopoulos and Neubig (2019) and student-forcing (Nicolai and Silfverberg, 2020), a training strategy where the model outputs are replaced with gold labels for some percentage of samples to alleviate exposure bias. Data augmentation leads to significant improvements, especially in the small training condition, confirming its utility. The student forcing training also provides a certain accuracy gain but presents mixed results when used together with data hallucination.

Flexica (Scherbakov and Vylomova, 2022) is a modified version of the non-neural system submitted to the SIGMORPHON 2020 Shared Task on morphological reinflection (Scherbakov, 2020). The system is based on refined alignment patterns between lemmas and inflected forms. In this year’s submission, grammatical tag interchangeability learning was added to address smaller fea-

⁸Available here: <https://github.com/sigmorphon/2022InflectionST/tree/main/baselines/nonneural>

| Language | Train/Dev/Test Split Sizes | | | | Test/Small Train Overlaps | | | | Test/Large Train Overlaps | | | |
|----------|----------------------------|--------|------|-------|---------------------------|--------|--------|----------|---------------------------|-------|--------|----------|
| | #Small | #Large | #Dev | #Test | #Both | #Lemma | #Feats | #Neither | #Both | #Feat | #Lemma | #Neither |
| ang | 700 | 7000 | 866 | 1969 | 158 | 217 | 815 | 779 | 697 | 821 | 278 | 173 |
| ara | 700 | 7000 | 988 | 1995 | 84 | 93 | 843 | 975 | 549 | 529 | 447 | 470 |
| asm | 700 | 7000 | 996 | 1990 | 416 | 498 | 558 | 518 | 979 | 990 | 12 | 9 |
| bra | 700 | — | 365 | 734 | 64 | 161 | 146 | 363 | — | — | — | — |
| ckt | 167 | — | 22 | 46 | 0 | 16 | 1 | 29 | — | — | — | — |
| evn | 700 | 7000 | 959 | 1743 | 1 | 519 | 2 | 1221 | 3 | 1065 | 0 | 675 |
| gml | 700 | — | 229 | 358 | 42 | 316 | 0 | 0 | — | — | — | — |
| goh | 700 | — | 986 | 1877 | 713 | 800 | 199 | 165 | — | — | — | — |
| got | 700 | 7000 | 994 | 1994 | 146 | 174 | 836 | 838 | 825 | 795 | 169 | 205 |
| guj | 700 | — | 994 | 1941 | 764 | 823 | 204 | 150 | — | — | — | — |
| heb | 700 | 7000 | 1000 | 2000 | 419 | 454 | 581 | 546 | 1000 | 1000 | 0 | 0 |
| hsb | 240 | — | 40 | 80 | 0 | 13 | 3 | 64 | — | — | — | — |
| hsi | 70 | — | 15 | 30 | 1 | 18 | 0 | 11 | — | — | — | — |
| hun | 700 | 7000 | 1000 | 2000 | 40 | 40 | 949 | 971 | 308 | 315 | 692 | 685 |
| hye | 700 | 7000 | 1000 | 2000 | 145 | 158 | 838 | 859 | 678 | 715 | 322 | 285 |
| itl | 700 | — | 572 | 1083 | 85 | 191 | 449 | 358 | — | — | — | — |
| kat | 630 | 7000 | 1000 | 2000 | 162 | 406 | 721 | 711 | 816 | 832 | 184 | 168 |
| kaz | 700 | 7000 | 998 | 1994 | 375 | 510 | 609 | 500 | 966 | 992 | 28 | 8 |
| ket | 700 | — | 85 | 137 | 13 | 48 | 14 | 62 | — | — | — | — |
| khk | 700 | 7000 | 996 | 1980 | 205 | 284 | 788 | 703 | 976 | 985 | 17 | 2 |
| kor | 700 | 7000 | 987 | 1964 | 221 | 245 | 748 | 750 | 886 | 925 | 83 | 70 |
| kr1 | 700 | 7000 | 998 | 1996 | 148 | 174 | 844 | 830 | 804 | 816 | 192 | 184 |
| lud | 700 | 7000 | 991 | 1976 | 87 | 105 | 880 | 904 | 775 | 297 | 212 | 692 |
| mag | 700 | — | 215 | 430 | 45 | 107 | 105 | 173 | — | — | — | — |
| nds | 700 | — | 963 | 1900 | 813 | 936 | 106 | 45 | — | — | — | — |
| non | 700 | 7000 | 992 | 1991 | 362 | 442 | 609 | 578 | 931 | 964 | 61 | 35 |
| pol | 700 | 7000 | 1000 | 2000 | 8 | 11 | 847 | 1134 | 61 | 70 | 939 | 930 |
| poma | 700 | 7000 | 921 | 1999 | 17 | 14 | 980 | 988 | 169 | 172 | 830 | 828 |
| sjo | 700 | — | 350 | 1857 | 184 | 286 | 754 | 633 | — | — | — | — |
| slk | 700 | 7000 | 1000 | 2000 | 4 | 5 | 869 | 1122 | 56 | 47 | 944 | 953 |
| slp | 240 | — | 40 | 79 | 2 | 56 | 3 | 18 | — | — | — | — |
| tur | 700 | 7000 | 1000 | 2000 | 333 | 575 | 469 | 623 | 874 | 869 | 126 | 131 |
| vep | 700 | 7000 | 995 | 1993 | 42 | 58 | 936 | 957 | 412 | 428 | 583 | 570 |

Table 5: Training, development, and test data sizes along with overlap sizes between small training and test and between large training and test. Items were excluded post-hoc from dev and test if there were multiple triples with the same lemma and features.

ture overlap. The system learns transformation patterns based on maximal continuous matches between lemma and inflected forms. The extraction of a pattern from an inflection sample starts with finding the longest common substring and then recurrently continues to the remaining parts until no more common characters can be found. Then, each of such extracted patterns is augmented with a set of more concrete patterns. Concrete patterns are produced from abstract ones by replacing some ‘wildcard’ characters back with concrete characters observed in a training sample. At prediction time, an inflected form is inferred by choosing a pattern that matches the respective lemma and yields a maximum score.

7 Results and Evaluation

Performance was evaluated by exact match accuracy. Macro-averages across languages on the entire test set and partitioned over the four overlap types are provided in Table 6. Results by language for both small and large training conditions are provided in Tables 14–18 in Appendix B.

A few points stand out immediately. First, overall performance is much lower this year compared to last year’s similar task. During the 2021 iteration,

all systems achieved over 90% accuracy on most of languages, while this year, no system achieves over 72% average in either training condition. This task was designed to be particularly challenging because the test set required systems to make predictions with only partial information. The results bore out this expectation.

Flexica, the only general non-neural submitted system, surpasses the non-neural baseline, but does not surpass 40% overall accuracy in either training condition. Being a hand-built system, TüMorph-FST outperformed all other systems on two of three languages that it was developed for.

As expected, all systems that submitted full or nearly full predictions for both the small and large training conditions performed substantially better with more training data. CLUZH, TüMorph-Main, UBC, and the neural baseline each improved by over ten points, while Flexica and the non-neural baseline showed smaller gains of around four points.

UBC achieved the highest performance of any system in either training condition. To understand why this is, it is necessary to look at a breakdown of performance by overlap type. The system is more resilient to novel feature sets than any other except for the hand-built FSTs.

| System | Small Training Condition | | | | | Large Training Condition | | | | |
|----------|--------------------------|---------------|---------------|---------------|---------------|--------------------------|---------------|---------------|---------------|---------------|
| | Overall | Both | Lemma | Feature | Neither | Overall | Both | Lemma | Feature | Neither |
| CLUZH | 56.871 | 77.308 | 31.269 | 77.966 | 43.255 | 67.853 | 90.991 | 41.425 | 87.171 | 60.300 |
| Flexica | 34.406 | 59.503 | 6.390 | 61.616 | 14.562 | 38.243 | 66.846 | 4.985 | 73.007 | 21.337 |
| OSU | <i>47.688</i> | <i>79.310</i> | <i>8.565</i> | <i>82.308</i> | <i>44.133</i> | 46.734 | 89.565 | 4.843 | 85.308 | 16.768 |
| TüM-FST | <i>67.308</i> | <i>100.00</i> | <i>55.319</i> | <i>75.000</i> | <i>72.115</i> | — | — | — | — | — |
| TüM-Main | <i>41.591</i> | <i>58.907</i> | <i>18.597</i> | <i>62.469</i> | <i>27.613</i> | 57.627 | 77.995 | 34.916 | 76.009 | 48.720 |
| UBC | 57.234 | 75.963 | 35.519 | 74.201 | 46.060 | 71.259 | 89.503 | 50.583 | 85.063 | 66.224 |
| Neural | 47.626 | 65.027 | 24.929 | 66.539 | 35.601 | 62.391 | 80.462 | 42.166 | 77.627 | 55.563 |
| NonNeur | 33.321 | 58.475 | 5.566 | 59.969 | 14.431 | 37.583 | 67.434 | 4.843 | 72.283 | 16.768 |

Table 6: Macro-average accuracy for each system. Three systems (OSU, TüMorph-Main, and TüMorph-FST) only submitted predictions for a subset of languages in the small training condition, so their numbers (italicized) are not directly comparable to the others. Flexica and NonNeur are non-neural.

7.1 Analysis by Overlap Partition

A breakdown by overlap partition reveals some consistent trends. As expected, *neither overlap* items proved challenging, since systems had to infer the forms for simultaneously novel lemmas and novel feature sets. Surprisingly, all systems performed better on *neither overlap* items than *lemma overlap* items. It is not clear why this would be, since it is observed on average for many but not all of the tested systems. It may be an artifact of the data splitting algorithm favoring balancing feature overlap over lemma overlap. However, the results are consistent with the observation over the 2018 data that systems struggle generalizing across feature sets more so than generalizing over lemmas.

They perform better on generalizations across lemmas to such an extent that the proportion of items with feature overlap in the test set washes out the effect of seen and unseen lemmas. Tables 7-8 illustrate this point quantitatively. Table 7 compares average performance on test items with feature sets attested in training (*both overlap* \cup *feature overlap* items) with test items with novel feature sets (*neither overlap* \cup *lemma overlap* items). All systems perform better on items with attested feature sets, but the gap in performance varies greatly from UBC’s 32 points in the small training condition to OSU’s 79 points in the large training condition. OSU’s drop in performance is expected because it outputs the lemma when the feature set is unknown. In these cases it makes correct predictions exactly when the inflected form is identical to the lemma, pointing to a degree of syncretism in the data.

Table 8 shows the same, but for test items with lemmas attested during training *both overlap* \cup *lemma overlap* items) and test items with novel feature sets (*neither overlap* \cup *feature overlap* items). Every system actually performs *worse* on the attested lemma items than the novel lemma items.

The penalty of novel feature sets overpowers gains incurred by attested lemmas.

| Features System | Small Train | | Large Train | |
|-----------------|---------------|---------------|-------------|--------|
| | Seen | Novel | Seen | Novel |
| CLUZH | 77.790 | 39.417 | 89.753 | 47.874 |
| OSU | <i>80.573</i> | <i>21.174</i> | 88.186 | 8.918 |
| TüM-FST | <i>80.000</i> | <i>66.887</i> | — | — |
| TüM-Main | <i>61.521</i> | <i>24.797</i> | 77.351 | 39.633 |
| UBC | 74.672 | 42.684 | 88.064 | 55.928 |
| Flexica | 60.916 | 12.894 | 68.757 | 10.614 |

Table 7: Macro-Average performance for submitted systems on test items with attested feature sets (*both overlap* and *feature overlap*) and items with novel feature sets (*lemma overlap* and *neither overlap* types). Italicized small training results were calculated over partial submissions.

| Lemma System | Small Train | | Large Train | |
|--------------|---------------|---------------|-------------|--------|
| | Seen | Novel | Seen | Novel |
| CLUZH | 50.175 | 59.690 | 65.399 | 72.764 |
| OSU | <i>38.248</i> | <i>62.811</i> | 45.821 | 48.560 |
| TüM-FST | <i>56.250</i> | <i>72.222</i> | — | — |
| TüM-Main | <i>35.442</i> | <i>44.116</i> | 55.752 | 61.378 |
| UBC | 52.128 | 59.384 | 69.407 | 74.962 |
| Flexica | 28.629 | 37.309 | 35.378 | 44.300 |

Table 8: Macro-Average performance for submitted systems on test items with attested lemmas (*both overlap* and *lemma overlap*) and items with novel lemmas (*feature overlap* and *neither overlap* types). Italicized small training results were calculated over partial submissions.

Tables 6-7 together elucidate a clear difference between CLUZH and UBC. While the former outperforms the latter on items with seen feature sets, the latter outperforms the former on items with novel feature sets. This means that UBC outperformed CLUZH on this data set because it is better suited for generalization to unseen features, something that would likely been hidden if tested on previous years’ data.

However, there is a sense in which testing on items with novel feature sets is not entirely fair for

all languages. In highly fusional languages in particular, it may not actually be possible to predict the mapping from a set of semantic features to a particular inflection given what is known about the member features. On the other hand, it should be solvable for a canonically agglutinative language where each member feature contributes one piece of the inflected form like “beads on a string.” Thus, it could be possible that the lower aggregate performance observed on novel feature test items is not due to a failure of generalization in the systems but rather the impossible nature of the task.

Table 9 tests this hypothesis. It shows average performance only on languages considered to be primarily agglutinative: Chukchi, Evenki, Georgian, Hungarian, Itelmen, Karelian, Kazakh, Ket, Korean, Ludic, Mongolian, Turkish, Veps, and Xibe. Further information can be gleaned from performance on each language individually as reported in Tables 14-18 in Appendix B.

In principle, a system should be able to infer the appropriate morphological operations for unseen feature sets in these languages, as was illustrated for Turkish in Table 1. While this is not a perfect test, since real agglutinative languages also contain some morphological eccentricities which obscure predictability, “could an undergraduate solve it?” does apply. It provides a clear result: the gap between performance on test items attested and novel features does not generally improve even for these languages where it should, if the unfairness of the task were driving decreased performance on fusional languages. This shows that generalization to novel feature sets, that is, to previously unattested inflectional categories, remains a legitimate concern for nearly all the systems.

7.2 Results by Part-of-Speech

As in previous years, the data employed for this task contains items from several parts-of-speech. Languages vary considerably in how much inflection they apply to different POS categories. As such, collapsing over POS categories can obscure interesting patterns. Tables 19-26 provide results for test items tagged with the four most common part-of-speech features in this year’s data: verb (V), noun (N), adjective (ADJ), and participle (V.PTCP). Given the overall challenging nature of this year’s task, performance across POS categories is generally weaker than what was reported for last year.

| Features System | Small Train | | Large Train | |
|-----------------|---------------|---------------|-------------|--------|
| | Seen | Novel | Seen | Novel |
| CLUZH | 78.837 | 34.118 | 90.198 | 40.657 |
| OSU | 77.800 | 30.376 | 88.497 | 13.456 |
| TÜM-FST | <i>100.00</i> | <i>17.778</i> | — | — |
| TÜM-Main | <i>61.730</i> | <i>14.816</i> | 74.667 | 29.433 |
| UBC | 75.994 | 39.232 | 89.213 | 49.799 |
| Flexica | 60.885 | 11.386 | 69.173 | 10.094 |

| Lemma System | Small Train | | Large Train | |
|--------------|---------------|---------------|-------------|--------|
| | Seen | Novel | Seen | Novel |
| CLUZH | 44.850 | 56.649 | 62.082 | 66.201 |
| OSU | <i>30.012</i> | <i>61.435</i> | 45.315 | 53.753 |
| TÜM-FST | <i>6.250</i> | <i>26.667</i> | — | — |
| TÜM-Main | <i>28.956</i> | <i>37.569</i> | 48.871 | 53.093 |
| UBC | 50.439 | 57.022 | 67.471 | 68.427 |
| Flexica | 22.361 | 36.604 | 35.123 | 41.965 |

Table 9: Macro-Average performance for submitted systems on seen and unseen feature and lemma items *for agglutinative languages only*. Compare to Tables 7-8. Italicized small training accuracies were calculated over partial submissions.

8 Error Analysis by Language

This section contains qualitative error analysis for six languages from five different top-level families.

8.1 Arabic

As shown in Table 17, none of the systems outperformed either of the baselines in the *overall* partition in the large training setting.

15% of the lemmas in the test set were not inflected correctly by all the systems. Nouns (N) made up the majority of those errors (47.8%). Focusing on the noun majority, errors included inaccurate plurals, minor orthographic errors, and “reasonable” confusion of different state and possession features. The plural inflection errors follow a similar pattern to those in this year’s acquisition-inspired subtask. See Kodner and Khalifa (2022) for more in-depth analysis. Orthographic errors include minor common mistakes resulting from missing orthotactic operations or an alternative spelling in the gold form. Lastly, there seems to be some confusion between SPEC, DEF, PSSD tags⁹ in the dual and masculine plural forms since both those suffixes inflect for case and state. This confusion is mainly due to the existence of possible different forms of the same lemma sharing the same feature set or vice versa in the training data.

On the other hand, all systems correctly inflected 29% of the lemmas. In this case, 55% of those

⁹For more details about the state, case, and possession tags, please see the mapping description here: https://github.com/unimorph/ara#ara_atb

cases are adjectives (ADJ). This is not very surprising since adjectives in Arabic are more regular than nouns in pluralization in particular. Most of the plurals in this set are those ending with the feminine plural suffix, which does not inflect for case and state the same way the masculine plural suffix does. On the other hand, most of the masculine adjectives are singular and therefore the case and state inflections are easier.

In the small training setting, systems follow a similar trend, shown in Table 14. However, there is a higher percentage of verbs (V) among the lemmas that all systems inflected incorrectly. This is expected since verbal paradigms in Modern Standard Arabic tend to be very large in size, therefore, more sparsity in smaller training sets.

8.2 Armenian

Armenian orthography is quite close to the pronunciation of words. But all four models had issues when the triggers for inflectional allomorphy were from phonology, semantics, or morphological classes.¹⁰

The different learning models had problems in respecting the rather close correspondence between the orthography and phonology. For example, given a word with a final orthographic <a> like <anjnya> ‘personable’, adding a vowel-initial suffix sequence like *-i-s* (-GEN-POSS2SG) triggers a glide in both the orthography and pronunciation: <anjny^{ay}is>. All four models incorrectly generated a glideless form for this word <*anjnyais>.

There were also cases of transparent phonological-conditioned allomorphy that caused errors. The definite suffix is <-n> after vowels, but <-ə> after consonants. Given a vowel-final word like <mořeni> ‘raspberry,’ the definite form should thus be <mořeniⁿ>, yet all four models made some type of error. The Flexica model used an entirely different ablative suffix *-ic*’, while the other three models used the wrong definite allomorph *-ə*. This allomorphy rule is exceptionless and is fully transparent from the reformed Armenian orthography. These errors suggest that the models didn’t fully exploit the phonological properties that are reflected in the orthography. It is possible that such errors would reduce if the models incorporated some level of

phonological information, such as by making the input forms be transcribed forms, and by having the models have a priori knowledge of cross-linguistic phonological feature systems.

Some errors were unavoidable and are due to phonology-semantics interactions. The plural suffix is <-er> after monosyllabic words, but <-ner> after polysyllabic words. For example, the monosyllabic word <nyut> ‘material’ takes the plural <nyut’-er>. But if a word is an endocentric compound, then the plural suffix must count the number of syllables in the second stem of the word (the head). For example, the word <řparanyut> ‘makeup’ is an endocentric compound of <řpar> ‘makeup’ and <nyut>. Its plural unambiguously takes *-er* because of the transparent semantic connections between the compound and the monosyllabic second stem. But all four models incorrectly generated the polysyllabic-selecting suffix *-ner*. It is not surprising that all four models made errors of this type. To avoid such errors, the models would need access to semantic information of the compound, and to also access the semantics of other words in the lexicon (the stems).

Some errors were due to purely morphological under-learning. Armenian has many different declension and conjugation classes. The different models made over-regularization mistakes, whereby they used regular inflectional suffixes over irregular ones. Sometimes the use of a suffix triggers morphological alternations in the stem. The models however preferred to keep the shape of the stem constant. Such ‘mistakes’ are common in colloquial speech, but they are absent in the prescriptive declension patterns that the Wiktionary data uses.

8.3 Hungarian

The richness of the Hungarian inflection system made prediction hard for all systems. While most errors show failures of generalization, many are attributable to genuinely hard, i.e., irregular or weakly systematic, forms of inflection. Mistakes due to vowel harmony are very frequent, as the vowels to be used in inflections are often unpredictable and can only be judged in terms of frequency in everyday use. Thus, **megtilt+enėlek* is clearly ungrammatical (it should be *megtilt+anėlak*), but forms such as *szellős+ök* or *objektív+től*, not present in the gold standard, are actually used. Another recurrent mistake is the presence or ab-

¹⁰Transliteration is the Hübschmann-Meillet-Benveniste (HMB) system: https://en.wiktionary.org/wiki/Wiktionary:Armenian_transliteration. Forms in <angled brackets> are transliterations.

sence of the *-j-* in possessives where, again, systematicity is weak: in *siketfajd+(j)a*, the form without the *-j-* is not acceptable, but in other cases (*hangár+(j)aitok*, *tranziszor+(j)a*) native speakers may accept either form. Unsurprisingly, all systems tended to fail over irregular inflections, such as hard-to-predict (but frequently used) inflectional classes, such as *low vowel nouns* (singular *út* but plural *utak*) or *v-stems* (singular *ló* but plural *lovak*). Finally, homonymy can also explain apparent mistakes, such as *szél* that means both *wind* and *edge*: in the first case its plural is *szelek* while in the second case it is *szélek*.

8.4 Khalkha Mongolian

Mongolian inflectional suffixes are highly unambiguous given a lemma’s POS feature. Every inflectional suffix often belongs to only one morphological feature (Denwood, 2011; Munkhjargal et al., 2016). For example, Mongolian *-iin* belongs only to the genitive case while German *-s* suffix has two meanings by making the inflectional forms of either the genitive case or plural nouns. In this sense of low ambiguity, it is not surprising to see that the all participating systems have zero accuracy over the *lemma overlap* settings in Tables 15 and 18.

8.5 Polish

Performance on Polish was decent overall. In the small training condition, CLUZH managed to achieve nearly 91% on the *lemma overlap* items. While number decreased to 84% in the large training condition, which likely suggests that the *lemma overlap* test partitions contained coincidentally easy items, it does demonstrate generalization. Not all systems succeeded on the *lemma overlap* items. OSU, Flexica, and the non-neural baseline showed the usual performance drop.

Masculine genitive singular inflection proved challenging. There are two possible endings, *-u* and *-a*, but their distribution is unpredictable. As a classic example of paradigmatic gaps, native speakers themselves frequently disagree on which ending to apply (Dąbrowska, 2001). Then it is unsurprisingly that systems sometimes predict the wrong ending. For example CLUZH produced **przystępa* for *przystępu* as the genitive singular of *przystęp*. It also produced *filungu* instead of *filunga* as the genitive singular of *filung*, which is a known variant form in the language, but not the one presented in the gold standard data.

Systems also confuse masculine and feminine forms or inflect the wrong case. They also misapply *yers*, or palatalization, a pervasive process in Polish and in Slavic more generally. These types of errors were also identified in an error analysis of the 2017 task in Gorman et al. (2019). See that paper for more information.

8.6 Turkish

Turkish exhibits both front/back and rounding harmony. Harmony mismatches are a major source of errors on the language. For example, Flexica produces **dokumalisin*, a front/back violation for expected *dokumalısın*, and CLUZH produces a rounding violation **yoldurtmuşım* for *yoldurtmuşum*. Flexica, the only non-neural submitted system particularly struggled in this area.

Voicing assimilation, which can occur intervocalically and at some morpheme boundaries, also proved to be challenging. For example, Flexica and CLUZH, the stem *çıldirt-* ends in voiceless stop, therefore the consonant of the following past tense suffix should be devoiced and realized as [t], however, in these three systems it remains [d], thus resulting in forms like **çıldirttım mı* for expected *çıldirttım mı*. CLUZH and Flexica do not perform intervocalic voicing for *akrebinizi* from *akrep* and instead produce **akrepinizi*. Similarly no system except for TüMorph-Main correctly produces *asidi* from *asit*. They instead produce **asiti*. Related to this, systems sometimes fail to insert epenthetic glides between vowels in hiatus.

Sometimes systems produce commission errors, substituting a morpheme with one absent in the feature set. For example, for CLUZH in the small training condition, the case marking is wrong for the lemma *balta*: instead of producing the genitive *-ın*, it adds the ablative *-dan* even though the GEN feature is present. The same issue holds in quite a few lines as well. For example, for Flexica, the features contain GEN, but the system generates it with dative case (along with a vowel harmony error as in Hungarian), thus producing **havai fişeklara* instead of the expected form *havai fişeklerin*. All systems struggle significantly on items with unseen feature sets. This is interesting, because Turkish should have been one of the languages most conducive to generalization over unseen feature sets. The systems may not be associating the features in a set with their corresponding agglutinative realizations.

9 Discussion

This year’s shared task investigated two dimensions of generalization in morphological inflection: generalization over lemmas and generalization over inflectional categories. Test items with lemmas or feature sets that were attested in training were evaluated separately from those with novel lemmas or feature sets to gain a better understanding of generalization. This proved to be a challenging version of the task, as performance is substantially lower across systems compared to previous years.

We carried on the tradition of including a range of typologically diverse languages in the task. From the perspective of the two dimensions of generalization, different morphological paradigms could prove more or less challenging. In particular, it is more reasonable to expect a system to generalize to an unseen feature set if the form of the corresponding inflectional category is in some way derived from forms associated with each of the member features. Similarly, a language with relatively invariant stem forms and little unpredictable stem-conditioned realization of inflectional categories should be conducive to generalization across lemmas, while a language with more stem changes or lexically arbitrary inflectional classes should prove more challenging.

Two major patterns emerged which held across systems. First, overall averages were lower than previous years in which overlaps between lemmas and features in training and test were left uncontrolled. The task was challenging. Second, performance test items with novel feature sets was almost uniformly weaker than performance on test items with novel lemmas. This was true for all systems and still held true for agglutinative languages which stood the best chance of generalization across feature sets.

9.1 Implications for Future Work

The results of this year’s shared tasks have some implications for future systems and future shared tasks. First, since overlap type has a major effect on performance, cross-linguistic differences in performance in morphological inflection tasks may sometimes be driven by these distributions rather language-internal. Since these overlaps were hardly evaluated in previous years, a reanalysis of prior years’ shared tasks along these lines may uncover interesting results. Related to this, train/test/dev splits created by uniform sampling

of UniMorph will not only lead to uncontrolled overlap ratios, but will tend to drive feature overlap unrealistically high when training sets are large. This year’s shared task provided an algorithm to make splits more uniformly with respect to overlap types, and it is recommended that future tasks also control for and separately analyze overlap types.

Second, both lemmas and inflectional categories are sparsely distributed in natural language use. As a result, systems in use in the real world will likely be asked to produce inflections for which lemmas or feature sets were not previously attested in their training. As focus grows on low-resource languages and language revitalization, a wide range of morphological typologies, including polysynthetic systems, will have to be reckoned with. The ability to generalize to unseen feature sets will become increasingly critical. Yet, there is a general weakness in generalization across inflectional categories in today’s systems. Every system showed serious performance degradation. This was even true for agglutinative languages. Nevertheless, systems do appear to have generalized to unseen feature sets to a significant degree, and CLUZH and UBC, which showed similar overall performance, differed in their ability to handle unseen feature sets in particular. Thus, we believe there is reason for optimism and that there are real-world performance gains to be had by further developing this type of generalization.

Acknowledgements

We would like to thank Garrett Nicolai, Maria Ryskina, Ben Ambridge, Jeff Heinz, and all those who provided valuable advice and logistical support in the early stages of this project, Judit Ács, Duygu Ataman, Zigniew Bronk, Eleanor Chodroff, Sofya Ganieva, Włodzimierz Gruszczyński, Nizar Habash, Jan Hajič, Jan Hric, Ritvan Karahodja, Christo Kirov, Elena Klyachko, Ritesh Kumar, Vahagn Petrosyan, Matvey Plugaryov, Mohit Raj, Maria Ryskina, Elizabeth Salesky, Zygmunt Saloni, Danuta Skowrońska, Marcin Woliński, and Robert Wołosz, who prepared and authored lexicons used in this project, as well as Jeff Heinz, Sarah Payne, and Charles Yang, who provided feedback on this overview paper. The neural baseline system was trained on the SeaWulf HPC cluster maintained by RCC, and IACS at Stony Brook University and made possible by NSF grant #1531492.

References

- Antonios Anastasopoulos and Graham Neubig. 2019. [Pushing the limits of low-resource morphological inflection](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.
- I Wayan Arka. 2007. Creole genesis and extreme analyticity in Flores languages. In *the 5th International East Nusantara Conference on Language and Culture (ENUS)*, Kupang.
- Aryaman Arora and Ahmed Etebari. 2021. [Kholosi dictionary](#).
- Khuyagbaatar Batsuren, Amarsanaa Ganbold, Altangerel Chagnaa, and Fausto Giunchiglia. 2019. [Building the Mongolian WordNet](#). In *Proceedings of the 10th Global Wordnet Conference*, pages 238–244, Wrocław, Poland. Global Wordnet Association.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina J. Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Gurriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud’hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. [Unimorph 4.0: Universal morphology](#).
- Jatayu Baxi, Dr Bhatt, et al. 2021. Morpheme boundary detection & grammatical feature prediction for Gujarati: Dataset & model. *arXiv preprint arXiv:2112.09860*.
- Sarah Beemer, Zak Boston, April Bukoski, Daniel Chen, Princess Dickens, Andrew Gerlach, Torin Hopkins, Parth Anand Jawale, Chris Koski, Akanksha Malhotra, Piyush Mishra, Saliha Muradoglu, Lan Sang, Tyler Short, Sagarika Shreevastava, Elizabeth Spaulding, Testumichi 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*, pages 162–170, Online. Association for Computational Linguistics.
- Tatyana Boyko, Nina Zaitseva, Natalya Krizhanovskaya, Andrew Krizhanovsky, Irina Novak, Natalia Pellinen, Alexandra Rodionova, and Elizaveta Trubina. 2021. [The linguistic corpus VepKar is a language refuge for the Baltic-Finnish languages of Karelia](#). *Transactions of the Karelian Research Centre of the Russian Academy of Sciences*, (7):100–115.
- Erwin Chan. 2008. *Structures and distributions in morphological learning*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA.
- Edna Amir Coffin and Shmuel Bolozky. 2005. *A reference grammar of Modern Hebrew*. Cambridge University Press.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. [The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection](#). In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. [CoNLL–SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages](#). In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.
- Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. [The SIGMORPHON 2016 shared Task—Morphological reinflection](#). In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.
- Ewa Dąbrowska. 2001. Learning a morphological system without a default: The Polish genitive. *Journal of child language*, 28(3):545–574.

- Ann Denwood. 2011. Template and morphology in Khalkha Mongolian—and beyond? In *Living on the Edge*, pages 543–562. De Gruyter Mouton.
- Jasmine Dum-Tragut. 2009. *Armenian: Modern Eastern Armenian*. Number 14 in London Oriental and African Language Library. John Benjamins Publishing Company, Amsterdam/Philadelphia.
- Micha Elsner and Sara K. Court. 2022. OSU at SIGMORPHON 2022: Analogical Inflection With Rule Features. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.
- Alexander Fraser. 2020. *Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT*. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771. Association for Computational Linguistics.
- Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models’ performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational 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.
- David Guriel, Omer Goldman, and Reut Tsarfaty. 2022. Morphological reinflection with multiple arguments: An extended annotation schema and a Georgian case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.
- Jan Hajič and Jan Hric. 2017. *Morfflex SK 170914*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Purev Jaimai, Tsolmon Zundui, Altangerel Chagnaa, and Cheol-Young Ock. 2005. PC-KIMMO-based description of Mongolian morphology. *Journal of Information Processing Systems*, 1(1):41–48.
- Ritván Jusúf Karahóga, Panagiotis G. Krimpas, Vivian Stamou, Vasileios Arampatzakis, Dimitrios Karatskos, Vasileios Sevetlidis, Nikolaos Constantinides, Nikolaos Kokkas, George Pavlidis, and Stella Markantonatou. 2022. *Morphologically annotated corpora of Pomak*. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 179–186, Dublin, Ireland. Association for Computational Linguistics.
- Olga Kazakevich and Elena Klyachko. 2013. Creating a multimedia annotated text corpus: a research task (sozdaniye multimedijnogo annotirovannogo kor-pusa tekstov kak issledovatel’skaya protsedura). In *Proceedings of International Conference Computational linguistics*, pages 292–300.
- Ferenc Kiefer and Boglarka Nemeth. 2019. *Compounds and multi-word expressions in Hungarian: Compounds and Multi-Word Expressions*, pages 337–358. De Gruyter.
- Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqi, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. *UniMorph 2.0: Universal Morphology*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Marian Klamer. 2002. Typical features of Austronesian languages in Central/Eastern Indonesia. *Oceanic Linguistics*, 41:250–263.
- Marian Klamer. 2009. The use of language data in comparative research: A note on Blust (2008) and Onvlee (1984). *Oceanic Linguistics*, 250–263.
- Jordan Kodner. 2019. Estimating child linguistic experience from historical corpora. *Glossa: a journal of general linguistics*, 4(1).
- Jordan Kodner and Salam Khalifa. 2022. SIGMORPHON-UniMorph 2022 Shared Task 0: Modeling Inflection in Language Acquisition. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.
- Ritesh Kumar, Bornini Lahiri, and Deepak Alok. 2014. *Developing LRs for Non-scheduled Indian Languages: A Case of Magahi*. In *Human Language Technology Challenges for Computer Science and Linguistics*, Lecture Notes in Computer Science, pages 491–501. Springer International Publishing, Switzerland. Original-date: 2014.
- Ritesh Kumar, Bornini Lahiri, Deepak Alok Atul Kr. Ojha, Mayank Jain, Abdul Basit, and Yogesh Dawar. 2018. Automatic identification of closely-related Indian languages: Resources and experiments. In *Proceedings of the 4th Workshop on Indian Language Data Resource and Evaluation (WILDRE-4)*, Paris, France. European Language Resources Association (ELRA).
- Agathe Lasch. 1914. *Mittelniederdeutsche Grammatik*. Max Niemeyer Verlag.
- Peter Makarov and Simon Clematide. 2020. *CLUZH at SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion*. In *Proceedings*

- of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 171–176, Online. Association for Computational Linguistics.
- Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. [UniMorph 3.0: Universal Morphology](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.
- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. [The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection](#). In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.
- Tatiana Merzhevich, Nkonye Gbadegoye, Leander Gierbach, Jingwen Li, and Ryan Soh-Eun Shim. 2022. Modelling Morphological Inflection with Data-Driven and Rule-Based Approaches. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.
- J. Mistrík. 1988. *A Grammar of Contemporary Slovak*. Slovenské pedagogické nakladateľstvo.
- Zoljargal Munkhjargal, Altangerel Chagnaa, and Purev Jaimai. 2016. Morphological transducer for Mongolian. In *International Conference on Computational Collective Intelligence*, pages 546–554. Springer.
- Saliha Muradoglu, Nicholas Evans, and Ekaterina Vylomova. 2020. [Modelling verbal morphology in Nen](#). In *Proceedings of the The 18th Annual Workshop of the Australasian Language Technology Association*, pages 43–53, Virtual Workshop. Australasian Language Technology Association.
- Temir Nabiyev. 2015. *Kazakh Language: 101 Kazakh Verbs*. Preceptor Language Guides, Online.
- Naonori Nagaya. 2011. *The Lamaholot language of Eastern Indonesia*. Ph.D. thesis, Rice University, Houston, TX.
- Naonori Nagaya. 2012. *The Lamaholot language of eastern Indonesia*. Ph.D. thesis, Rice University.
- Garrett Nicolai and Miikka Silfverberg. 2020. [Noise isn't always negative: Countering exposure bias in sequence-to-sequence inflection models](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2837–2846, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- I Novak, M Penttonen, A Ruuskanen, and L Siilin. 2019. Karel'skiy yazyk v grammatikakh (Karelian in grammars). *Sravnitel'noe issledovanie foneticheskoy i morfologicheskoy sistem–Petrozavodsk: KarRC RAS*, page 22.
- Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. [SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages](#). In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.
- Donald A Ringe. 2017. *From Proto-Indo-European to Proto-Germanic*, volume 1. Oxford University Press.
- Carol Rounds. 2009. *Hungarian: An essential grammar*. Routledge.
- Mohammad Salehi and Aydin Neysani. 2017. Receptive intelligibility of Turkish to Iranian-Azerbaijani speakers. *Cogent Education*, 4(1):1326653.
- Andreas Scherbakov. 2020. The UniMelb submission to the SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 177–183.
- Andrey Scherbakov and Ekaterina Vylomova. 2022. Morphology is not just a Naïve Bayes! In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.
- Simon Clematide Silvan Wehrli and Peter Makarov. 2022. CLUZH at SIGMORPHON 2022 Shared

- Tasks on Morpheme Segmentation and Inflection Generation. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.
- Dima Taji, Salam Khalifa, Ossama Obeid, Fadhil Eryani, and Nizar Habash. 2018. *An Arabic morphological analyzer and generator with copious features*. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 140–150, Brussels, Belgium. Association for Computational Linguistics.
- Turkicm. 2019. *The Kazakh Verbs: Review Guide*. Preceptor Language Guides, Online.
- Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204.
- G.K. Verner. 1997. Jeniseiskije jazyki. *Jazyki mira. Paleosiasiskije jazyki.*, pages 169–177.
- Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. *SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection*. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.
- Marcin Woliński and Witold Kieraś. 2016. *The online version of grammatical dictionary of Polish*. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2589–2594, Portorož, Slovenia. European Language Resources Association (ELRA).
- Marcin Woliński, Zygmunt Saloni, Robert Wołosz, Włodzimierz Gruszczyński, Danuta Skowrońska, and Zbigniew Bronk. 2020. *Słownik gramatyczny języka polskiego*, 4th edition. Warsaw. <http://sgjp.pl>.
- Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. *Applying the transformer to character-level transduction*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.
- Changbing Yang, Ruixin Yang, Garrett Nicolai, and Miikka Silfverberg. 2022. Generalizing Morphological Inflection Systems to Unseen Lemmas. In *Proceedings of the SIGMORPHON 2022 Shared Task: Morphological Inflection*, Seattle. North American Chapter of the Association for Computational Linguistics.
- He Zhou, Juyeon Chung, Sandra Kübler, and Francis Tyers. 2020. Universal dependency treebank for Xibe. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 205–215.

A Lemma and Feature Overlap in 2018

Under the hypothesis that systems struggle at generalization to novel lemmas or feature sets, the proportion of test items which are novel should serve as a performance ceiling. Tables 2-3 show an apparent ceiling effect for two closely related highly agglutinative languages, Turkish and Azeri. This appendix provides performance and ceiling numbers for both lemma and feature overlap for the best performing system on each language on the low training size condition in the 2018 inflection task (Cotterell et al., 2018). This condition was chosen for illustration because it showed the most language-to-language variation in overlaps.

Tables 10-11 show a ceiling effect for feature overlap in the low training condition in 2018 task. The best systems manage to surpass the hypothesized ceiling for only 17 of 104 languages, most of which are agglutinative. In contrast, lemma overlap, shown in Tables 12-13, does not seem to produce a ceiling effect. The best systems surpass it for 74 of 104 languages, which can only be possible if the systems possess a significant ability to generalize to unseen lemmas.

| Language | F Overlap% | Acc% | Δ |
|-------------------------|-------------|-------------|-------------|
| Adyghe | 98.3 | 90.6 | -7.7 |
| Albanian | 54.8 | 36.4 | -18.4 |
| Arabic | 54.2 | 45.2 | -9.0 |
| Armenian | 55.3 | 64.9 | 9.6 |
| Asturian | 65.2 | 74.6 | 9.4 |
| Azeri | 71.0 | 65.0 | -6.0 |
| Bashkir | 98.0 | 77.8 | -20.2 |
| Basque | 5.6 | 13.3 | 7.7 |
| Belarusian | 86.3 | 33.4 | -52.9 |
| Bengali | 83.0 | 72.0 | -11.0 |
| Breton | 74.0 | 72.0 | -2.0 |
| Bulgarian | 66.1 | 62.9 | -3.2 |
| Catalan | 86.9 | 72.5 | -14.4 |
| Classical Syriac | 95.0 | 96.0 | 1.0 |
| Cornish | 68.0 | 40.0 | -28.0 |
| Crimean Tatar | 98.0 | 91.0 | -7.0 |
| Czech | 56.7 | 46.5 | -10.2 |
| Danish | 96.2 | 87.7 | -8.5 |
| Dutch | 95.2 | 69.3 | -25.9 |
| English | 100. | 91.8 | -8.2 |
| Estonian | 70.3 | 35.2 | -35.1 |
| Faroese | 85.7 | 49.8 | -35.9 |
| Finnish | 58.1 | 25.7 | -32.4 |
| French | 85.5 | 66.6 | -18.9 |
| Friulian | 89.0 | 79.0 | -10.0 |
| Galician | 73.0 | 61.1 | -11.9 |
| Georgian | 93.8 | 88.2 | -5.6 |
| German | 79.6 | 67.1 | -12.5 |
| Greek | 57.7 | 32.3 | -25.4 |
| Greenlandic | 100. | 80.0 | -20.0 |
| Haida | 45.0 | 63.0 | 18.0 |
| Hebrew | 82.4 | 56.7 | -25.7 |
| Hindi | 38.8 | 78.0 | 39.2 |
| Hungarian | 78.9 | 48.2 | -30.7 |
| Icelandic | 92.2 | 56.2 | -36.0 |
| Ingrian | 100. | 46.0 | -54.0 |
| Irish | 82.7 | 37.7 | -45.0 |
| Italian | 82.8 | 57.4 | -25.4 |
| Kabardian | 99.0 | 92.0 | -7.0 |
| Kannada | 74.0 | 61.0 | -13.0 |
| Karelian | 88.0 | 94.0 | 6.0 |
| Kashubian | 100. | 68.0 | -32.0 |
| Kazakh | 100. | 86.0 | -14.0 |
| Khakas | 100. | 86.0 | -14.0 |
| Khaling | 22.0 | 33.8 | 11.8 |
| Kurmanji | 90.2 | 87.4 | -2.8 |
| Ladin | 77.0 | 72.0 | -5.0 |
| Latin | 52.3 | 33.1 | -19.2 |
| Latvian | 80.1 | 57.3 | -22.8 |
| Lithuanian | 65.4 | 32.6 | -32.8 |
| Livonian | 73.0 | 35.0 | -38.0 |
| Lower Sorbian | 75.9 | 54.3 | -21.6 |

Table 10: Difference between proportion of 2018 test set items with *feature overlap* and best performance in the low training condition (Adyghe-Lower Sorbian). Bolded rows indicate better percent correct than overlap.

B Full Results by Language

This section provides performance breakdowns by overlap type for each individual language for both small training (Tables 14-16) and large training (17-18) conditions. Data partition sizes can be found in Table 5.

| Language | F Overlap% | Acc% | Δ |
|-----------------------|-------------|-------------|-------------|
| Macedonian | 79.2 | 68.8 | -10.4 |
| Maltese | 99.0 | 49.0 | -50.0 |
| Mapudungun | 88.0 | 86.0 | -2.0 |
| Middle French | 86.7 | 84.5 | -2.2 |
| Middle High German | 94.0 | 84.0 | -10.0 |
| Middle Low German | 92.0 | 54.0 | -38.0 |
| Murrinhpatha | 98.0 | 38.0 | -60.0 |
| Navajo | 88.9 | 20.8 | -68.1 |
| Neapolitan | 90.0 | 89.0 | -1.0 |
| Norman | 88.0 | 66.0 | -22.0 |
| Northern Sami | 69.1 | 35.8 | -33.3 |
| North Frisian | 85.0 | 45.0 | -40.0 |
| Norwegian Bokmaal | 99.3 | 90.1 | -9.2 |
| Norwegian Nynorsk | 98.3 | 83.6 | -14.7 |
| Occitan | 91.0 | 77.0 | -14.0 |
| Old Armenian | 47.4 | 42.0 | -5.4 |
| Old Church Slavonic | 97.0 | 53.0 | -44.0 |
| Old English | 81.0 | 46.5 | -34.5 |
| Old French | 65.8 | 46.2 | -19.6 |
| Old Irish | 46.0 | 8.0 | -38.0 |
| Old Saxon | 68.3 | 46.6 | -21.7 |
| Pashto | 59.0 | 48.0 | -11.0 |
| Persian | 54.7 | 67.6 | 12.9 |
| Polish | 75.9 | 49.4 | -26.5 |
| Portuguese | 73.7 | 75.8 | 2.1 |
| Quechua | 21.4 | 70.2 | 48.8 |
| Romanian | 79.4 | 46.2 | -33.2 |
| Russian | 80.2 | 53.5 | -26.7 |
| Sanskrit | 68.9 | 58.0 | -10.9 |
| Scottish Gaelic | 100. | 74.0 | -26.0 |
| Serbo Croatian | 34.5 | 44.8 | 10.3 |
| Slovak | 90.0 | 51.8 | -38.2 |
| Slovene | 70.8 | 58.0 | -12.8 |
| Sorani | 38.2 | 40.1 | 1.9 |
| Spanish | 82.7 | 73.2 | -9.5 |
| Swahili | 39.0 | 72.0 | 33.0 |
| Swedish | 95.0 | 79.0 | -16.0 |
| Tatar | 98.0 | 90.0 | -8.0 |
| Telugu | 86.0 | 96.0 | 10.0 |
| Tibetan | 100. | 58.0 | -42.0 |
| Turkish | 39.6 | 39.5 | -0.1 |
| Turkmen | 100. | 90.0 | -10.0 |
| Ukrainian | 85.4 | 57.1 | -28.3 |
| Urdu | 41.3 | 72.5 | 31.2 |
| Uzbek | 75.0 | 92.0 | 17.0 |
| Venetian | 88.5 | 78.8 | -9.7 |
| Votic | 94.0 | 34.0 | -60.0 |
| Welsh | 88.0 | 55.0 | -33.0 |
| West Frisian | 100. | 56.0 | -44.0 |
| Yiddish | 100. | 87.0 | -13.0 |
| Zulu | 43.5 | 33.0 | -10.5 |

Table 11: Difference between proportion of 2018 test set items with *feature overlap* and best performance in the low training condition (Macedonian-Zulu). Bolded rows indicate better percent correct than percent overlap.

C Performance by Part-of-Speech

This section provides performance breakdowns by part-of-speech for both small training (Tables 19-22) and large training (Tables 23-26) conditions. Information on the four most common parts-of-speech in the data overall: verbs V, nouns N, adjectives ADJ, and participles V.PTCP is provided. Results for TüMorph-FST are provided separately in Table 27.

| Language | L Overlap% | Acc% | Δ |
|------------------|------------|------|----------|
| Adyghe | 4.6 | 90.6 | 86.0 |
| Albanian | 26.3 | 36.4 | 10.1 |
| Arabic | 3.4 | 45.2 | 41.8 |
| Armenian | 2.2 | 64.9 | 62.7 |
| Asturian | 22.0 | 74.6 | 52.6 |
| Azeri | 36.0 | 65.0 | 29.0 |
| Bashkir | 8.7 | 77.8 | 69.1 |
| Basque | 87.8 | 13.3 | -74.5 |
| Belarusian | 10.2 | 33.4 | 23.2 |
| Bengali | 53.0 | 72.0 | 19.0 |
| Breton | 86.0 | 72.0 | -14.0 |
| Bulgarian | 5.4 | 62.9 | 57.5 |
| Catalan | 5.5 | 72.5 | 67.0 |
| Classical Syriac | 47.0 | 96.0 | 49.0 |
| Cornish | 100. | 40.0 | -60.0 |
| Crimean Tatar | 4.0 | 91.0 | 87.0 |
| Czech | 3.4 | 46.5 | 43.1 |
| Danish | 3.2 | 87.7 | 84.5 |
| Dutch | 1.4 | 69.3 | 67.9 |
| English | 0.5 | 91.8 | 91.3 |
| Estonian | 12.8 | 35.2 | 22.4 |
| Faroese | 3.0 | 49.8 | 46.8 |
| Finnish | 0.2 | 25.7 | 25.5 |
| French | 1.6 | 66.6 | 65.0 |
| Friulian | 42.0 | 79.0 | 37.0 |
| Galician | 17.8 | 61.1 | 43.3 |
| Georgian | 3.0 | 88.2 | 85.2 |
| German | 0.8 | 67.1 | 66.3 |
| Greek | 2.1 | 32.3 | 30.2 |
| Greenlandic | 100. | 80.0 | -20.0 |
| Haida | 100. | 63.0 | -37.0 |
| Hebrew | 17.4 | 56.7 | 39.3 |
| Hindi | 33.1 | 78.0 | 44.9 |
| Hungarian | 0.6 | 48.2 | 47.6 |
| Icelandic | 2.2 | 56.2 | 54.0 |
| Ingrian | 94.0 | 46.0 | -48.0 |
| Irish | 2.7 | 37.7 | 35.0 |
| Italian | 1.5 | 57.4 | 55.9 |
| Kabardian | 33.0 | 92.0 | 59.0 |
| Kannada | 51.0 | 61.0 | 10.0 |
| Karelian | 100. | 94.0 | -6.0 |
| Kashubian | 88.0 | 68.0 | -20.0 |
| Kazakh | 100. | 86.0 | -14.0 |
| Khakas | 76.0 | 86.0 | 10.0 |
| Khaling | 18.1 | 33.8 | 15.7 |
| Kurmanji | 1.1 | 87.4 | 86.3 |
| Ladin | 47.0 | 72.0 | 25.0 |
| Latin | 0.9 | 33.1 | 32.2 |
| Latvian | 1.4 | 57.3 | 55.9 |
| Lithuanian | 9.3 | 32.6 | 23.3 |
| Livonian | 40.0 | 35.0 | -5.0 |
| Lower Sorbian | 10.3 | 54.3 | 44.0 |

Table 12: Difference between proportion of 2018 test set items with *lemma overlap* and best performance in the low training condition (Adyghe-Lower Sorbian). Bolded rows indicate better percent correct than overlap.

| Language | L Overlap% | Acc% | Δ |
|---------------------|------------|------|----------|
| Macedonian | 0.8 | 68.8 | 68.0 |
| Maltese | 54.0 | 49.0 | -5.0 |
| Mapudungun | 100. | 86.0 | -14.0 |
| Middle French | 17.5 | 84.5 | 67.0 |
| Middle High German | 98.0 | 84.0 | -14.0 |
| Middle Low German | 78.0 | 54.0 | -24.0 |
| Murrinhpatha | 98.0 | 38.0 | -60.0 |
| Navajo | 17.9 | 20.8 | 2.9 |
| Neapolitan | 96.0 | 89.0 | -7.0 |
| Norman | 100. | 66.0 | -34.0 |
| North Frisian | 88.0 | 45.0 | -43.0 |
| Northern Sami | 6.3 | 35.8 | 29.5 |
| Norwegian Bokmaal | 2.1 | 90.1 | 88.0 |
| Norwegian Nynorsk | 1.5 | 83.6 | 82.1 |
| Occitan | 43.0 | 77.0 | 34.0 |
| Old Armenian | 3.7 | 42.0 | 38.3 |
| Old Church Slavonic | 53.0 | 53.0 | 0.0 |
| Old English | 10.3 | 46.5 | 36.2 |
| Old French | 5.9 | 46.2 | 40.3 |
| Old Irish | 90.0 | 8.0 | -82.0 |
| Old Saxon | 18.4 | 46.6 | 28.2 |
| Pashto | 35.0 | 48.0 | 13.0 |
| Persian | 30.3 | 67.6 | 37.3 |
| Polish | 1.6 | 49.4 | 47.8 |
| Portuguese | 2.2 | 75.8 | 73.6 |
| Quechua | 17.0 | 70.2 | 53.2 |
| Romanian | 4.0 | 46.2 | 42.2 |
| Russian | 0.4 | 53.5 | 53.1 |
| Sanskrit | 13.3 | 58.0 | 44.7 |
| Scottish Gaelic | 80.0 | 74.0 | -6.0 |
| Serbo Croatian | 0.9 | 44.8 | 43.9 |
| Slovak | 10.4 | 51.8 | 41.4 |
| Slovene | 5.3 | 58.0 | 52.7 |
| Sorani | 52.5 | 40.1 | -12.4 |
| Spanish | 2.5 | 73.2 | 70.7 |
| Swahili | 78.0 | 72.0 | -6.0 |
| Swedish | 1.0 | 79.0 | 78.0 |
| Tatar | 5.0 | 90.0 | 85.0 |
| Telugu | 100. | 96.0 | -4.0 |
| Tibetan | 80.0 | 58.0 | -22.0 |
| Turkish | 2.6 | 39.5 | 36.9 |
| Turkmen | 84.0 | 90.0 | 6.0 |
| Ukrainian | 5.9 | 57.1 | 51.2 |
| Urdu | 76.9 | 72.5 | -4.4 |
| Uzbek | 100. | 92.0 | -8.0 |
| Venetian | 24.3 | 78.8 | 54.5 |
| Votic | 92.0 | 34.0 | -58.0 |
| Welsh | 39.0 | 55.0 | 16.0 |
| West Frisian | 61.0 | 56.0 | -5.0 |
| Yiddish | 7.0 | 87.0 | 80.0 |
| Zulu | 18.9 | 33.0 | 14.1 |

Table 13: Difference between proportion of 2018 test set items with *lemma overlap* and best performance in the low training condition (Macedonian-Zulu). Bolded rows indicate better percent correct than percent overlap.

| Lang | Partition | CLUZH | Flexica | OSU | Tüm FST | Tüm Main | UBC | Neural | NonNeur |
|------|-----------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| ang | overall | 54.241 | 37.075 | — | — | 45.962 | 51.346 | 49.822 | 33.215 |
| | both | 70.253 | 58.861 | — | — | 66.456 | 72.785 | 68.354 | 43.671 |
| | lemma | 38.710 | 17.512 | — | — | 34.562 | 38.710 | 42.396 | 8.756 |
| | features | 70.307 | 61.350 | — | — | 58.282 | 66.503 | 61.350 | 58.037 |
| | neither | 38.511 | 12.709 | — | — | 32.092 | 34.660 | 36.072 | 11.938 |
| ara | overall | 66.566 | 32.581 | — | — | 62.857 | 47.870 | 65.965 | 22.757 |
| | both | 71.429 | 50.000 | — | — | 67.857 | 60.714 | 75.000 | 39.286 |
| | lemma | 63.441 | 9.677 | — | — | 61.290 | 54.839 | 65.591 | 0 |
| | features | 74.614 | 58.719 | — | — | 71.530 | 52.313 | 70.700 | 47.568 |
| | neither | 59.487 | 10.667 | — | — | 55.077 | 42.256 | 61.128 | 2.051 |
| asm | overall | 57.286 | 30.452 | — | — | 38.995 | 55.025 | 54.673 | 26.231 |
| | both | 74.760 | 57.692 | — | — | 63.702 | 68.029 | 70.192 | 47.115 |
| | lemma | 40.562 | 0 | — | — | 23.494 | 44.177 | 47.189 | 1.807 |
| | features | 72.043 | 65.591 | — | — | 56.093 | 65.771 | 61.649 | 56.452 |
| | neither | 43.436 | 0 | — | — | 15.637 | 43.436 | 41.892 | 0.386 |
| bra | overall | 60.354 | 58.856 | 57.902 | — | 53.134 | 56.131 | 55.041 | 57.902 |
| | both | 26.562 | 26.562 | 25.000 | — | 21.875 | 25.000 | 28.125 | 21.875 |
| | lemma | 21.739 | 17.391 | 18.012 | — | 16.770 | 22.360 | 20.497 | 18.012 |
| | features | 74.658 | 76.027 | 71.233 | — | 67.808 | 68.493 | 66.438 | 72.603 |
| | neither | 77.686 | 76.033 | 76.033 | — | 68.871 | 71.625 | 70.523 | 76.033 |
| ckt | overall | 13.043 | 10.870 | 10.870 | 19.565 | 8.696 | 21.739 | 6.522 | 13.043 |
| | both | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | lemma | 0 | 0 | 0 | 6.250 | 12.500 | 18.750 | 12.500 | 0 |
| | features | 100.00 | 100.00 | 0 | 100.00 | 0 | 100.00 | 0 | 100.00 |
| | neither | 17.241 | 13.793 | 17.241 | 24.138 | 6.897 | 20.690 | 3.448 | 17.241 |
| evn | overall | 28.514 | 3.328 | — | — | 23.867 | 34.481 | 29.260 | 25.014 |
| | both | 100.00 | 100.00 | — | — | 100.00 | 100.00 | 100.00 | 100.00 |
| | lemma | 14.258 | 2.312 | — | — | 15.992 | 22.736 | 21.580 | 9.441 |
| | features | 50.000 | 50.000 | — | — | 0 | 50.000 | 50.000 | 0 |
| | neither | 34.480 | 3.604 | — | — | 27.191 | 39.394 | 32.432 | 31.613 |
| gml | overall | 56.704 | 26.257 | 20.950 | — | — | 44.693 | 42.737 | 22.067 |
| | both | 88.095 | 71.429 | 33.333 | — | — | 88.095 | 97.619 | 40.476 |
| | lemma | 52.532 | 20.253 | 19.304 | — | — | 38.924 | 35.443 | 19.620 |
| | features | — | — | — | — | — | — | — | — |
| | neither | — | — | — | — | — | — | — | — |
| goh | overall | 60.629 | 40.224 | 52.637 | — | 52.158 | 59.03 | 56.420 | 42.568 |
| | both | 84.853 | 66.620 | 87.237 | — | 76.578 | 81.487 | 84.151 | 63.114 |
| | lemma | 32.500 | 8.875 | 15.125 | — | 26.500 | 31.875 | 30.125 | 15.125 |
| | features | 93.970 | 90.955 | 95.980 | — | 69.849 | 91.457 | 75.377 | 87.437 |
| | neither | 52.121 | 16.970 | 32.727 | — | 49.697 | 54.545 | 41.212 | 32.727 |
| got | overall | 51.204 | 18.154 | — | — | 47.693 | 61.384 | 60.582 | 38.816 |
| | both | 78.082 | 36.301 | — | — | 81.507 | 89.041 | 86.986 | 72.603 |
| | lemma | 26.437 | 5.747 | — | — | 34.483 | 52.299 | 50.575 | 4.023 |
| | features | 76.196 | 32.057 | — | — | 68.660 | 78.349 | 76.675 | 71.292 |
| | neither | 26.730 | 3.699 | — | — | 23.628 | 41.527 | 42.005 | 7.757 |
| guj | overall | 66.924 | 47.141 | 49.253 | — | 40.855 | 63.112 | 39.979 | 48.429 |
| | both | 96.728 | 86.518 | 96.073 | — | 64.136 | 94.895 | 63.743 | 93.717 |
| | lemma | 34.143 | 5.468 | 1.580 | — | 17.861 | 30.741 | 12.272 | 1.580 |
| | features | 94.118 | 90.686 | 91.667 | — | 59.804 | 91.667 | 69.118 | 92.647 |
| | neither | 58.000 | 16.000 | 14.667 | — | 22.667 | 40.000 | 31.333 | 14.667 |
| heb | overall | 40.850 | 19.250 | — | — | 31.150 | 35.150 | 39.650 | 14.750 |
| | both | 77.804 | 44.630 | — | — | 66.826 | 71.838 | 81.862 | 28.640 |
| | lemma | 5.066 | 0.220 | — | — | 0.881 | 0.441 | 1.322 | 6.167 |
| | features | 74.182 | 33.907 | — | — | 57.487 | 68.675 | 75.904 | 20.482 |
| | neither | 6.777 | 0 | — | — | 0.916 | 0.183 | 0.549 | 5.128 |
| hsb | overall | 15.000 | 13.750 | 8.750 | 83.750 | 7.500 | 3.750 | 5.000 | 10.000 |
| | both | — | — | — | — | — | — | — | — |
| | lemma | 7.692 | 0 | 0 | 61.538 | 0 | 0 | 0 | 0 |
| | features | 100.00 | 66.667 | 66.667 | 66.667 | 66.667 | 0 | 33.333 | 100.00 |
| | neither | 12.500 | 14.062 | 7.812 | 89.062 | 6.250 | 4.688 | 4.688 | 7.812 |

Table 14: Partitioned test performance in the small training condition (ang-hsb). No *feature overlap* or *neither overlap* items for gml and no *both overlap* items for hsb were included in the test set.

| Lang | Partition | CLUZH | Flexica | OSU | Tüm FST | Tüm Main | UBC | Neural | NonNeur |
|------|-----------|---------------|---------|---------------|---------------|---------------|---------------|---------------|---------------|
| hsi | overall | 16.667 | 13.333 | 20.000 | 96.667 | 0 | 13.333 | 0 | 20.000 |
| | both | 0 | 0 | 0 | 100.00 | 0 | 0 | 0 | 0 |
| | lemma | 11.111 | 5.556 | 16.667 | 94.444 | 0 | 16.667 | 0 | 16.667 |
| | features | — | — | — | — | — | — | — | — |
| | neither | 27.273 | 27.273 | 27.273 | 100.00 | 0 | 9.091 | 0 | 27.273 |
| hun | overall | 60.000 | 25.900 | — | — | 51.850 | 61.750 | 65.000 | 23.900 |
| | both | 85.000 | 60.000 | — | — | 85.000 | 85.000 | 90.000 | 52.500 |
| | lemma | 40.000 | 0 | — | — | 27.500 | 45.000 | 65.000 | 0 |
| | features | 80.295 | 51.423 | — | — | 71.338 | 80.400 | 78.925 | 47.313 |
| | neither | 39.959 | 0.618 | — | — | 32.441 | 43.254 | 50.360 | 0.824 |
| hye | overall | 82.350 | 39.250 | — | — | 61.450 | 86.250 | 64.750 | 38.750 |
| | both | 95.862 | 80.690 | — | — | 52.414 | 95.172 | 51.724 | 82.759 |
| | lemma | 67.722 | 0 | — | — | 43.038 | 74.684 | 45.570 | 3.165 |
| | features | 91.050 | 79.714 | — | — | 68.377 | 89.737 | 69.093 | 76.611 |
| | neither | 74.272 | 0 | — | — | 59.604 | 83.469 | 66.240 | 0.931 |
| itl | overall | 33.333 | 31.210 | 31.487 | — | 33.056 | 34.441 | 34.257 | 28.163 |
| | both | 42.353 | 41.176 | 43.529 | — | 47.059 | 43.529 | 48.235 | 28.235 |
| | lemma | 3.141 | 0 | 0 | — | 3.665 | 6.283 | 6.283 | 0 |
| | features | 65.702 | 65.702 | 66.370 | — | 60.802 | 62.138 | 59.465 | 61.247 |
| | neither | 6.704 | 2.235 | 1.676 | — | 10.615 | 12.570 | 14.246 | 1.676 |
| kat | overall | 59.200 | 34.350 | — | — | 47.800 | 51.800 | 60.200 | 43.600 |
| | both | 51.852 | 43.210 | — | — | 51.852 | 48.148 | 57.407 | 53.704 |
| | lemma | 16.995 | 3.695 | — | — | 7.389 | 14.532 | 23.399 | 6.404 |
| | features | 95.284 | 73.925 | — | — | 92.372 | 90.430 | 93.620 | 94.730 |
| | neither | 48.383 | 9.705 | — | — | 24.754 | 34.740 | 47.961 | 10.689 |
| kaz | overall | 61.735 | 34.203 | — | — | 55.165 | 65.747 | 55.667 | 42.879 |
| | both | 96.800 | 64.800 | — | — | 83.467 | 96.800 | 83.467 | 85.611 |
| | lemma | 36.471 | 1.569 | — | — | 30.392 | 45.098 | 31.373 | 0 |
| | features | 98.686 | 70.115 | — | — | 94.745 | 97.701 | 95.567 | 100.00 |
| | neither | 16.200 | 0.800 | — | — | 11.000 | 24.600 | 11.000 | 0 |
| ket | overall | 33.577 | 18.978 | 35.036 | — | 13.139 | 26.277 | 10.949 | 32.847 |
| | both | 23.077 | 30.769 | 30.769 | — | 38.462 | 30.769 | 30.769 | 23.077 |
| | lemma | 12.500 | 0 | 12.500 | — | 2.083 | 2.083 | 0 | 12.500 |
| | features | 50.000 | 50.000 | 57.143 | — | 57.143 | 57.143 | 35.714 | 42.857 |
| | neither | 48.387 | 24.194 | 48.387 | — | 6.452 | 37.097 | 9.677 | 48.387 |
| khk | overall | 41.768 | 22.374 | — | — | 39.495 | 29.899 | 41.616 | 28.182 |
| | both | 83.902 | 48.293 | — | — | 89.268 | 61.951 | 92.195 | 56.098 |
| | lemma | 0 | 0 | — | — | 0 | 0.352 | 0 | 0.352 |
| | features | 83.122 | 43.655 | — | — | 76.015 | 58.629 | 80.584 | 55.584 |
| | neither | 0 | 0 | — | — | 0 | 0.284 | 0 | 0.569 |
| kor | overall | 50.509 | 30.957 | — | — | 17.821 | 44.348 | 23.523 | 28.870 |
| | both | 70.588 | 59.276 | — | — | 41.176 | 57.466 | 54.299 | 55.656 |
| | lemma | 33.061 | 0.408 | — | — | 18.776 | 33.061 | 28.163 | 0 |
| | features | 71.658 | 62.433 | — | — | 20.989 | 62.968 | 25.134 | 59.358 |
| | neither | 29.200 | 1.200 | — | — | 7.467 | 25.600 | 11.333 | 0 |
| krl | overall | 41.333 | 23.497 | — | — | 10.421 | 45.842 | 16.182 | 5.411 |
| | both | 68.919 | 37.838 | — | — | 16.216 | 68.919 | 22.297 | 1.351 |
| | lemma | 19.540 | 1.149 | — | — | 2.299 | 27.011 | 9.195 | 0.575 |
| | features | 63.389 | 45.735 | — | — | 16.588 | 63.744 | 22.986 | 8.886 |
| | neither | 18.554 | 3.012 | — | — | 4.819 | 27.470 | 9.639 | 3.614 |
| lud | overall | 87.702 | 88.006 | — | — | 46.559 | 84.565 | 46.609 | 88.715 |
| | both | 91.954 | 95.402 | — | — | 93.103 | 93.103 | 91.954 | 96.552 |
| | lemma | 18.095 | 16.190 | — | — | 2.857 | 17.143 | 3.810 | 18.095 |
| | features | 94.091 | 95.227 | — | — | 93.977 | 95.114 | 93.409 | 95.909 |
| | neither | 89.159 | 88.606 | — | — | 0.996 | 81.305 | 1.659 | 89.159 |
| mag | overall | 64.419 | 58.140 | 57.209 | — | 51.163 | 56.744 | 51.163 | 55.349 |
| | both | 53.333 | 44.444 | 37.778 | — | 31.111 | 51.111 | 40.000 | 31.111 |
| | lemma | 15.888 | 4.673 | 4.673 | — | 5.607 | 7.477 | 3.738 | 4.673 |
| | features | 86.667 | 83.810 | 83.810 | — | 76.190 | 80.952 | 79.048 | 79.048 |
| | neither | 83.815 | 79.191 | 78.613 | — | 69.364 | 73.988 | 66.474 | 78.613 |

Table 15: Partitioned test performance in the small training condition (hsi-mag). No *feature overlap* items were included in the hsi test set.

| Lang | Partition | CLUZH | Flexica | OSU | Tüm FST | Tüm Main | UBC | Neural | NonNeur |
|------|-----------|---------------|---------------|---------------|---------|---------------|---------------|---------------|---------------|
| nds | overall | 47.789 | 31.316 | 34.947 | – | 21.947 | 50.421 | 25.789 | 16.053 |
| | both | 65.560 | 46.863 | 72.079 | – | 38.376 | 67.897 | 43.665 | 32.226 |
| | lemma | 32.799 | 16.239 | 1.603 | – | 7.906 | 36.859 | 10.256 | 1.603 |
| | features | 57.547 | 48.113 | 59.434 | – | 29.245 | 52.830 | 34.906 | 26.415 |
| | neither | 15.556 | 24.444 | 0 | – | 0 | 11.111 | 4.444 | 0 |
| non | overall | 48.820 | 39.126 | – | – | 47.313 | 52.436 | 55.902 | 30.638 |
| | both | 61.602 | 50.276 | – | – | 56.630 | 62.431 | 69.613 | 47.238 |
| | lemma | 37.330 | 22.851 | – | – | 47.738 | 49.548 | 58.824 | 5.430 |
| | features | 63.054 | 61.248 | – | – | 49.918 | 61.741 | 56.322 | 60.755 |
| | neither | 34.602 | 21.280 | – | – | 38.408 | 38.581 | 44.637 | 7.785 |
| pol | overall | 71.800 | 43.300 | – | – | 53.850 | 78.350 | 59.250 | 30.100 |
| | both | 75.000 | 87.500 | – | – | 100.00 | 100.00 | 100.00 | 87.500 |
| | lemma | 90.909 | 9.091 | – | – | 72.727 | 90.909 | 72.727 | 0 |
| | features | 85.596 | 70.130 | – | – | 61.393 | 86.423 | 65.289 | 68.123 |
| | neither | 61.287 | 23.280 | – | – | 47.707 | 72.046 | 54.321 | 1.587 |
| poma | overall | 50.975 | 29.315 | – | – | 45.873 | 46.023 | 51.426 | 22.311 |
| | both | 70.588 | 64.706 | – | – | 58.824 | 47.059 | 70.588 | 52.941 |
| | lemma | 42.857 | 21.429 | – | – | 42.857 | 35.714 | 50.000 | 0 |
| | features | 61.020 | 44.694 | – | – | 55.816 | 54.388 | 57.041 | 42.245 |
| | neither | 40.789 | 13.563 | – | – | 35.830 | 37.854 | 45.547 | 2.328 |
| sjo | overall | 71.998 | 65.751 | 68.174 | – | 54.496 | 76.737 | 58.643 | 67.905 |
| | both | 71.739 | 73.370 | 70.652 | – | 70.652 | 75.543 | 76.087 | 68.478 |
| | lemma | 36.014 | 20.280 | 24.476 | – | 27.273 | 50.699 | 36.713 | 24.476 |
| | features | 93.103 | 91.512 | 91.512 | – | 89.257 | 92.971 | 89.125 | 91.379 |
| | neither | 63.191 | 53.397 | 59.400 | – | 20.695 | 69.510 | 27.172 | 59.400 |
| slk | overall | 74.500 | 51.600 | – | – | 56.05 | 84.100 | 61.000 | 38.450 |
| | both | 75.000 | 75.000 | – | – | 50.000 | 75.000 | 50.000 | 75.000 |
| | lemma | 80.000 | 60.000 | – | – | 80.000 | 80.000 | 80.000 | 20.000 |
| | features | 87.457 | 83.774 | – | – | 65.823 | 89.413 | 67.664 | 82.739 |
| | neither | 64.439 | 26.560 | – | – | 48.396 | 80.036 | 55.793 | 4.100 |
| slp | overall | 29.114 | 8.861 | 6.329 | – | 12.658 | 30.380 | 15.190 | 5.063 |
| | both | 100.00 | 100.00 | 100.00 | – | 100.00 | 100.00 | 100.00 | 100.00 |
| | lemma | 25.000 | 3.571 | 0 | – | 10.714 | 28.571 | 16.071 | 0 |
| | features | 66.667 | 33.333 | 66.667 | – | 33.333 | 33.333 | 33.333 | 33.333 |
| | neither | 27.778 | 11.111 | 5.556 | – | 5.556 | 27.778 | 0 | 5.556 |
| tur | overall | 61.250 | 18.350 | – | – | 19.250 | 85.800 | 34.600 | 16.600 |
| | both | 80.18 | 54.655 | – | – | 17.718 | 95.796 | 28.228 | 51.952 |
| | lemma | 58.957 | 0 | – | – | 10.087 | 89.391 | 24.000 | 0 |
| | features | 72.068 | 39.446 | – | – | 37.740 | 85.501 | 51.173 | 31.983 |
| | neither | 45.104 | 0 | – | – | 14.607 | 77.368 | 35.313 | 1.445 |
| vep | overall | 40.291 | 20.622 | – | – | 27.446 | 42.097 | 35.575 | 21.325 |
| | both | 54.762 | 47.619 | – | – | 42.857 | 52.381 | 45.238 | 40.476 |
| | lemma | 25.862 | 1.724 | – | – | 15.517 | 32.759 | 24.138 | 1.724 |
| | features | 56.624 | 40.598 | – | – | 39.850 | 53.632 | 46.154 | 40.385 |
| | neither | 24.556 | 1.045 | – | – | 15.361 | 30.930 | 25.496 | 3.03 |

Table 16: Partitioned test performance in the small training condition (nds-vep).

| Lang | Partition | CLUZH | Flexica | OSU | TüM Main | UBC | Neural | NonNeur |
|------|-----------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| ang | overall | 64.855 | 41.138 | 44.540 | 60.945 | 59.980 | 61.097 | 43.118 |
| | both | 82.496 | 73.171 | 80.488 | 82.066 | 80.918 | 83.070 | 78.479 |
| | lemma | 48.356 | 11.693 | 10.840 | 42.509 | 41.778 | 41.048 | 10.840 |
| | features | 76.619 | 64.388 | 73.741 | 71.942 | 74.101 | 73.381 | 68.705 |
| | neither | 53.179 | 14.451 | 12.717 | 45.665 | 39.306 | 47.977 | 12.717 |
| ara | overall | 75.890 | 37.544 | 40.902 | 75.338 | 67.218 | 78.546 | 26.917 |
| | both | 79.964 | 66.302 | 80.874 | 81.603 | 74.317 | 81.239 | 52.823 |
| | lemma | 73.913 | 10.397 | 1.323 | 71.834 | 71.078 | 77.316 | 1.323 |
| | features | 81.655 | 65.548 | 78.747 | 78.523 | 65.548 | 81.879 | 50.783 |
| | neither | 67.872 | 7.872 | 2.766 | 68.936 | 56.170 | 73.617 | 2.766 |
| asm | overall | 70.653 | 34.271 | 43.467 | 63.065 | 75.628 | 76.784 | 31.859 |
| | both | 90.807 | 68.744 | 86.313 | 77.222 | 85.393 | 83.861 | 62.615 |
| | lemma | 50.909 | 0 | 1.111 | 49.091 | 65.758 | 69.697 | 1.111 |
| | features | 83.333 | 75.000 | 75.000 | 91.667 | 83.333 | 83.333 | 83.333 |
| | neither | 33.333 | 0 | 0 | 22.222 | 88.889 | 77.778 | 0 |
| evn | overall | 48.939 | 3.844 | 24.957 | 52.037 | 57.487 | 57.717 | 25.072 |
| | both | 66.667 | 66.667 | 0 | 66.667 | 66.667 | 66.667 | 66.667 |
| | lemma | 40.376 | 1.878 | 12.582 | 45.634 | 52.394 | 53.427 | 12.582 |
| | features | — | — | — | — | — | — | — |
| | neither | 62.370 | 6.667 | 44.593 | 62.074 | 65.481 | 64.444 | 44.593 |
| got | overall | 65.747 | 21.264 | 51.254 | 65.346 | 73.370 | 72.166 | 46.038 |
| | both | 95.515 | 38.182 | 95.879 | 93.333 | 95.758 | 95.758 | 84.606 |
| | lemma | 35.723 | 3.522 | 4.654 | 38.239 | 52.201 | 49.560 | 4.654 |
| | features | 92.899 | 41.420 | 94.083 | 91.716 | 91.716 | 93.491 | 87.574 |
| | neither | 40.000 | 5.366 | 17.073 | 36.098 | 50.244 | 47.317 | 17.073 |
| heb | overall | 51.750 | 28.000 | 50.000 | 47.900 | 43.950 | 48.450 | 20.350 |
| | both | 94.100 | 55.900 | 94.400 | 94.400 | 86.500 | 96.600 | 35.100 |
| | lemma | 9.400 | 0.100 | 5.600 | 1.400 | 1.400 | 0.300 | 5.600 |
| | features | — | — | — | — | — | — | — |
| | neither | — | — | — | — | — | — | — |
| hun | overall | 72.350 | 32.950 | 47.100 | 68.150 | 74.900 | 77.200 | 37.250 |
| | both | 94.805 | 64.286 | 94.156 | 94.481 | 93.831 | 94.805 | 75.000 |
| | lemma | 54.603 | 2.540 | 1.270 | 45.397 | 60.000 | 61.905 | 1.270 |
| | features | 93.497 | 62.861 | 93.064 | 92.775 | 91.474 | 94.364 | 73.121 |
| | neither | 49.051 | 2.628 | 0.584 | 41.898 | 56.496 | 58.978 | 0.584 |
| hye | overall | 86.05 | 42.750 | 48.900 | 66.700 | 93.400 | 69.800 | 44.850 |
| | both | 97.935 | 85.841 | 97.640 | 61.357 | 98.083 | 61.947 | 90.708 |
| | lemma | 72.448 | 0 | 1.818 | 55.105 | 88.671 | 60.280 | 1.818 |
| | features | 94.410 | 84.783 | 94.099 | 91.304 | 94.720 | 90.062 | 83.540 |
| | neither | 82.456 | 0 | 0 | 80.702 | 92.632 | 89.474 | 0 |
| kat | overall | 74.350 | 45.100 | 52.400 | 78.850 | 83.200 | 87.250 | 45.500 |
| | both | 95.098 | 79.289 | 94.608 | 95.956 | 98.284 | 97.426 | 77.696 |
| | lemma | 53.005 | 7.572 | 9.255 | 61.779 | 68.990 | 77.163 | 9.255 |
| | features | 96.739 | 95.652 | 96.739 | 96.739 | 96.739 | 97.283 | 96.739 |
| | neither | 54.762 | 9.524 | 12.500 | 60.714 | 65.476 | 76.786 | 12.500 |
| kaz | overall | 58.375 | 34.203 | 49.198 | 53.611 | 65.747 | 55.667 | 42.879 |
| | both | 96.170 | 67.702 | 98.758 | 89.959 | 97.516 | 90.683 | 85.611 |
| | lemma | 20.867 | 0.806 | 0 | 17.44 | 34.375 | 20.867 | 0 |
| | features | 100.00 | 71.429 | 96.429 | 96.429 | 92.857 | 96.429 | 100.00 |
| | neither | 0 | 0 | 0 | 0 | 25.000 | 0 | 0 |

Table 17: Partitioned results on large training (ang-kaz). No *feature overlap* evn items and no *feature overlap* or *both overlap* heb items were included in the test set.

| Lang | Partition | CLUZH | Flexica | OSU | Tüm Main | UBC | Neural | NonNeur |
|------|-----------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| khk | overall | 47.879 | 23.384 | 49.242 | 47.727 | 46.263 | 49.141 | 38.03 |
| | both | 95.492 | 46.619 | 97.746 | 95.184 | 92.316 | 98.053 | 75.102 |
| | lemma | 0 | 0 | 0.508 | 0 | 0 | 0 | 0.508 |
| | features | 94.118 | 47.059 | 94.118 | 94.118 | 88.235 | 94.118 | 88.235 |
| | neither | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| kor | overall | 51.833 | 33.198 | 29.990 | 47.556 | 54.684 | 56.161 | 32.332 |
| | both | 79.007 | 67.494 | 61.738 | 69.300 | 76.185 | 78.668 | 66.140 |
| | lemma | 25.946 | 0.865 | 0 | 28.000 | 35.351 | 36.865 | 0 |
| | features | 71.084 | 55.422 | 50.602 | 56.627 | 60.241 | 62.651 | 59.036 |
| | neither | 27.143 | 0 | 0 | 20.000 | 31.429 | 18.571 | 0 |
| krl | overall | 58.367 | 37.876 | 45.190 | 24.098 | 64.429 | 27.104 | 5.361 |
| | both | 88.557 | 72.264 | 87.811 | 29.975 | 88.06 | 31.468 | 4.478 |
| | lemma | 27.328 | 2.083 | 0.858 | 8.578 | 39.828 | 13.725 | 0.858 |
| | features | 87.500 | 69.792 | 85.938 | 57.812 | 85.417 | 57.812 | 20.833 |
| | neither | 33.696 | 13.043 | 13.043 | 32.065 | 48.370 | 35.326 | 13.043 |
| lud | overall | 73.077 | 89.221 | 89.676 | 50.506 | 72.419 | 52.986 | 89.372 |
| | both | 94.839 | 95.871 | 96.774 | 96.000 | 94.710 | 96.516 | 95.871 |
| | lemma | 21.212 | 51.515 | 51.515 | 11.111 | 39.057 | 20.202 | 51.515 |
| | features | 87.264 | 91.981 | 92.925 | 93.396 | 88.208 | 94.340 | 93.396 |
| | neither | 66.618 | 97.110 | 97.110 | 3.324 | 56.936 | 5.636 | 97.110 |
| non | overall | 76.896 | 47.162 | 48.016 | 79.759 | 87.243 | 84.982 | 37.318 |
| | both | 90.763 | 68.743 | 90.548 | 89.796 | 93.340 | 92.374 | 67.991 |
| | lemma | 63.900 | 25.207 | 5.705 | 70.851 | 82.054 | 78.838 | 5.705 |
| | features | 85.246 | 77.049 | 85.246 | 80.328 | 90.164 | 88.525 | 80.328 |
| | neither | 51.429 | 25.714 | 17.143 | 57.143 | 62.857 | 51.429 | 17.143 |
| pol | overall | 86.500 | 52.850 | 47.800 | 67.700 | 90.950 | 69.450 | 43.600 |
| | both | 91.803 | 78.689 | 90.164 | 77.049 | 95.082 | 78.689 | 85.246 |
| | lemma | 84.286 | 15.714 | 0 | 71.429 | 87.143 | 68.571 | 0 |
| | features | 96.060 | 85.942 | 94.888 | 74.015 | 95.740 | 74.441 | 86.262 |
| | neither | 76.667 | 20.538 | 1.075 | 60.430 | 86.129 | 63.871 | 1.075 |
| poma | overall | 60.430 | 33.867 | 36.568 | 58.829 | 61.481 | 63.882 | 24.462 |
| | both | 73.373 | 48.521 | 74.556 | 69.231 | 69.822 | 75.148 | 40.828 |
| | lemma | 46.512 | 12.791 | 1.744 | 47.674 | 50.581 | 59.884 | 1.744 |
| | features | 76.145 | 54.458 | 70.120 | 69.398 | 73.253 | 74.096 | 47.831 |
| | neither | 44.928 | 14.614 | 2.415 | 48.430 | 50.242 | 52.174 | 2.415 |
| slk | overall | 85.550 | 58.250 | 47.400 | 65.750 | 93.950 | 70.100 | 47.450 |
| | both | 87.500 | 87.500 | 89.286 | 57.143 | 89.286 | 57.143 | 87.500 |
| | lemma | 89.362 | 44.681 | 2.128 | 51.064 | 95.745 | 57.447 | 2.128 |
| | features | 93.538 | 90.042 | 92.161 | 70.445 | 95.657 | 71.081 | 92.373 |
| | neither | 77.335 | 25.708 | 2.833 | 62.329 | 92.445 | 70.514 | 2.833 |
| tur | overall | 87.200 | 35.600 | 48.500 | 33.600 | 94.150 | 39.650 | 36.400 |
| | both | 97.941 | 72.654 | 96.224 | 36.041 | 98.398 | 37.414 | 72.654 |
| | lemma | 80.667 | 0.345 | 0.230 | 23.360 | 93.326 | 31.415 | 0.230 |
| | features | 93.651 | 57.937 | 95.238 | 80.159 | 92.857 | 79.365 | 66.667 |
| | neither | 52.672 | 0.763 | 5.344 | 40.458 | 72.519 | 70.992 | 5.344 |
| vep | overall | 57.451 | 30.457 | 36.929 | 44.104 | 62.268 | 48.821 | 32.413 |
| | both | 75.485 | 58.01 | 72.330 | 55.825 | 70.146 | 57.039 | 64.078 |
| | lemma | 42.757 | 1.402 | 1.402 | 25.935 | 54.907 | 33.879 | 1.402 |
| | features | 71.527 | 58.834 | 69.983 | 57.461 | 68.782 | 59.177 | 60.377 |
| | neither | 41.053 | 3.333 | 4.211 | 35.614 | 55.439 | 43.509 | 4.211 |

Table 18: Partitioned results on large training (khk-vep).

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|------|--------|---------|--------|--------|--------|
| ang | 483 | 43.478 | 36.025 | – | 32.091 | 58.799 |
| ara | 341 | 32.551 | 11.730 | – | 27.859 | 52.199 |
| asm | 809 | 38.072 | 14.339 | – | 31.397 | 51.545 |
| bra | 208 | 21.635 | 17.308 | – | 17.788 | 18.269 |
| ckt | 20 | 10.000 | 10.000 | 5.000 | 0 | 5.000 |
| evn | 504 | 5.952 | 0.794 | – | 12.103 | 30.556 |
| gml | 229 | 63.755 | 23.581 | – | – | 93.886 |
| goh | 976 | 48.053 | 29.611 | – | 36.578 | 83.402 |
| got | 1003 | 41.874 | 19.541 | – | 36.491 | 80.160 |
| guj | 1214 | 57.908 | 36.903 | – | 28.007 | 93.987 |
| heb | 1729 | 43.493 | 18.681 | – | 33.372 | 79.294 |
| hsb | 21 | 4.762 | 0 | 71.429 | 0 | 0 |
| hsi | 19 | 10.526 | 5.263 | 100.00 | 0 | 15.789 |
| hun | 370 | 33.243 | 23.784 | – | 27.027 | 60.000 |
| hye | 764 | 74.215 | 37.696 | – | 26.702 | 96.859 |
| itl | 401 | 6.484 | 3.741 | – | 6.484 | 9.975 |
| kat | 453 | 5.519 | 5.298 | – | 6.402 | 47.461 |
| kaz | 576 | 72.222 | 23.438 | – | 56.771 | 92.188 |
| ket | 38 | 5.263 | 2.632 | – | 0 | 2.632 |
| khk | 78 | 1.282 | 6.410 | – | 1.282 | 30.769 |
| kor | 918 | 64.488 | 34.423 | – | 23.203 | 63.834 |
| krl | 1595 | 41.944 | 22.696 | – | 6.959 | 72.727 |
| lud | 903 | 85.050 | 86.157 | – | 1.661 | 91.251 |
| mag | 175 | 28.571 | 14.286 | – | 13.714 | 33.143 |
| nds | 880 | 43.523 | 37.273 | – | 28.636 | 75.114 |
| non | 585 | 41.709 | 34.188 | – | 39.316 | 60.000 |
| pol | 501 | 66.267 | 46.307 | – | 30.539 | 81.238 |
| poma | 747 | 54.217 | 27.711 | – | 49.665 | 63.989 |
| sjo | 297 | 29.630 | 10.438 | – | 35.690 | 56.902 |
| slk | 660 | 75.455 | 54.848 | – | 17.121 | 87.121 |
| slp | 63 | 26.984 | 6.349 | – | 12.698 | 57.143 |
| tur | 1446 | 65.698 | 19.018 | – | 11.549 | 91.978 |
| vep | 740 | 31.081 | 20.000 | – | 12.027 | 62.703 |

Table 19: Performance on verbs (V) in the small training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|------|--------|---------|--------|--------|--------|
| ang | 342 | 68.421 | 49.708 | – | 52.632 | 57.895 |
| ara | 833 | 67.827 | 31.933 | – | 65.306 | 49.340 |
| asm | 1103 | 73.255 | 44.152 | – | 44.968 | 70.898 |
| bra | 368 | 79.076 | 79.620 | – | 71.739 | 73.913 |
| ckt | 14 | 0 | 0 | 21.429 | 14.286 | 21.429 |
| evn | 867 | 45.559 | 0.231 | – | 35.409 | 47.174 |
| gml | 16 | 50.000 | 31.250 | – | – | 62.500 |
| goh | 839 | 77.116 | 54.470 | – | 71.514 | 75.924 |
| got | 206 | 34.466 | 12.136 | – | 28.155 | 43.204 |
| guj | 700 | 81.286 | 66.714 | – | 61.143 | 74.857 |
| heb | 226 | 28.761 | 27.434 | – | 20.354 | 28.761 |
| hsb | 37 | 16.216 | 10.811 | 91.892 | 8.108 | 2.703 |
| hsi | 5 | 40.000 | 40.000 | 100.00 | 0 | 20.000 |
| hun | 1287 | 64.180 | 28.127 | – | 55.245 | 63.403 |
| hye | 884 | 86.991 | 40.611 | – | 81.787 | 85.747 |
| itl | 217 | 49.309 | 50.230 | – | 54.839 | 54.378 |
| kat | 1505 | 74.684 | 42.724 | – | 59.801 | 64.518 |
| kaz | 1418 | 57.475 | 38.575 | – | 54.513 | 59.520 |
| ket | 44 | 18.182 | 15.909 | – | 18.182 | 20.455 |
| khk | 1847 | 44.721 | 23.714 | – | 42.285 | 32.052 |
| krl | 285 | 40.000 | 29.474 | – | 23.860 | 37.895 |
| lud | 878 | 91.230 | 90.774 | – | 90.319 | 91.344 |
| mag | 77 | 84.416 | 83.117 | – | 72.727 | 76.623 |
| non | 541 | 53.420 | 45.841 | – | 42.884 | 52.680 |
| pol | 259 | 63.707 | 69.884 | – | 55.212 | 65.251 |
| poma | 133 | 61.654 | 60.902 | – | 61.654 | 60.902 |
| sjo | 447 | 94.183 | 95.973 | – | 92.841 | 95.526 |
| slk | 111 | 65.766 | 63.964 | – | 54.955 | 63.063 |
| slp | 1 | 100.00 | 0 | – | 100.00 | 100.00 |
| tur | 538 | 50.929 | 16.543 | – | 40.335 | 73.978 |
| vep | 971 | 44.490 | 19.876 | – | 35.015 | 43.151 |

Table 20: Performance on verbs (N) in the small training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|------|--------|---------|--------|--------|--------|
| ang | 1085 | 57.512 | 35.484 | – | 52.535 | 57.419 |
| ara | 821 | 79.415 | 41.900 | – | 74.909 | 59.440 |
| bra | 69 | 57.971 | 55.072 | – | 55.072 | 59.420 |
| ckt | 1 | 0 | 0 | 100.00 | 0 | 0 |
| evn | 49 | 24.490 | 8.163 | – | 38.776 | 59.184 |
| gml | 78 | 57.692 | 39.744 | – | – | 50.000 |
| got | 309 | 59.547 | 12.945 | – | 58.900 | 67.961 |
| hsb | 18 | 22.222 | 27.778 | 77.778 | 16.667 | 5.556 |
| hsi | 4 | 0 | 0 | 75.000 | 0 | 0 |
| hun | 343 | 73.178 | 19.825 | – | 65.889 | 79.300 |
| hye | 315 | 94.603 | 38.730 | – | 89.206 | 93.651 |
| itl | 30 | 66.667 | 66.667 | – | 63.333 | 63.333 |
| kat | 42 | 83.333 | 47.619 | – | 64.286 | 73.810 |
| ket | 1 | 0 | 0 | – | 0 | 0 |
| kor | 221 | 69.683 | 42.534 | – | 21.267 | 57.466 |
| krl | 50 | 38.000 | 14.000 | – | 22.000 | 36.000 |
| lud | 105 | 92.381 | 92.381 | – | 92.381 | 90.476 |
| mag | 3 | 100.00 | 100.00 | – | 66.667 | 100.00 |
| nds | 887 | 49.605 | 21.082 | – | 13.191 | 55.919 |
| non | 652 | 50.920 | 35.583 | – | 56.748 | 58.896 |
| pol | 428 | 70.327 | 54.907 | – | 82.710 | 91.822 |
| poma | 242 | 66.529 | 54.959 | – | 58.678 | 60.744 |
| sjo | 2 | 0 | 0 | – | 0 | 50.000 |
| slk | 1142 | 75.569 | 48.862 | – | 77.758 | 87.916 |
| tur | 16 | 6.250 | 18.750 | – | 6.250 | 18.750 |
| vep | 233 | 54.506 | 26.180 | – | 45.064 | 57.082 |

Table 21: Performance on verbs (ADJ) in the small training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|-----|--------|---------|--------|--------|--------|
| ang | 59 | 0 | 1.695 | – | 0 | 0 |
| asm | 78 | 30.769 | 3.846 | – | 33.333 | 23.077 |
| ckt | 2 | 50.000 | 50.000 | 50.000 | 0 | 50.000 |
| evn | 30 | 0 | 0 | – | 0 | 0 |
| gml | 31 | 12.903 | 6.452 | – | – | 12.903 |
| goh | 62 | 35.484 | 14.516 | – | 35.484 | 22.581 |
| got | 476 | 72.689 | 21.218 | – | 72.479 | 82.563 |
| hun | 12 | 25.000 | 33.333 | – | 33.333 | 8.333 |
| hye | 19 | 73.684 | 73.684 | – | 89.474 | 73.684 |
| kor | 127 | 63.780 | 45.669 | – | 15.748 | 48.031 |
| krl | 55 | 41.818 | 29.091 | – | 32.727 | 40.000 |
| nds | 133 | 63.910 | 60.150 | – | 36.090 | 55.639 |
| non | 213 | 50.235 | 46.479 | – | 51.643 | 53.521 |
| pol | 615 | 81.138 | 25.203 | – | 58.374 | 79.512 |
| poma | 875 | 42.400 | 18.857 | – | 36.800 | 37.371 |
| sjo | 246 | 33.333 | 15.041 | – | 25.203 | 40.244 |
| slk | 62 | 74.194 | 51.613 | – | 83.871 | 88.710 |
| slp | 8 | 50.000 | 25.000 | – | 12.500 | 50.000 |
| vep | 25 | 36.000 | 28.000 | – | 32.000 | 36.000 |

Table 22: Performance on verbs (V.PTCP) in the small training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|------|--------|---------|--------|--------|--------|
| ang | 483 | 62.733 | 45.963 | 52.174 | 51.967 | 51.346 |
| ara | 341 | 65.396 | 21.408 | 37.243 | 65.982 | 58.651 |
| asm | 809 | 65.760 | 20.643 | 38.072 | 54.141 | 67.367 |
| evn | 504 | 33.532 | 0.595 | 0 | 36.706 | 37.698 |
| got | 1003 | 52.044 | 20.538 | 47.856 | 54.337 | 64.806 |
| heb | 1729 | 52.747 | 27.357 | 50.781 | 48.120 | 44.997 |
| hun | 370 | 56.757 | 31.622 | 48.919 | 53.514 | 58.378 |
| hye | 764 | 80.628 | 42.670 | 50.131 | 32.068 | 92.016 |
| kat | 453 | 71.082 | 16.777 | 39.735 | 84.768 | 91.391 |
| kaz | 576 | 62.674 | 23.438 | 40.625 | 54.688 | 81.076 |
| khk | 78 | 11.538 | 8.974 | 14.103 | 12.821 | 11.538 |
| kor | 918 | 64.815 | 38.562 | 37.255 | 61.547 | 70.044 |
| krl | 1595 | 58.621 | 37.743 | 44.389 | 14.420 | 63.699 |
| lud | 903 | 53.045 | 87.375 | 88.372 | 3.765 | 51.717 |
| non | 585 | 69.231 | 38.632 | 38.120 | 72.308 | 80.513 |
| pol | 501 | 79.641 | 51.497 | 41.717 | 36.926 | 81.637 |
| poma | 747 | 60.241 | 34.806 | 42.704 | 65.060 | 63.989 |
| slk | 660 | 87.273 | 60.606 | 46.515 | 23.788 | 93.030 |
| tur | 1446 | 92.600 | 38.036 | 48.548 | 22.407 | 97.994 |
| vep | 740 | 54.730 | 29.459 | 31.486 | 17.432 | 59.865 |

Table 23: Performance on verbs (V) in the large training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|------|--------|---------|--------|--------|--------|
| ang | 342 | 80.117 | 54.094 | 58.772 | 73.977 | 71.053 |
| ara | 833 | 72.389 | 37.935 | 34.454 | 71.909 | 61.825 |
| asm | 1103 | 76.156 | 45.603 | 47.235 | 71.079 | 83.409 |
| evn | 867 | 65.052 | 0.231 | 43.599 | 68.166 | 73.818 |
| got | 206 | 61.165 | 20.874 | 56.796 | 52.427 | 58.738 |
| heb | 226 | 54.425 | 38.496 | 53.982 | 55.752 | 44.690 |
| hun | 1287 | 73.660 | 34.266 | 49.728 | 69.852 | 75.913 |
| hye | 884 | 90.498 | 43.439 | 50.113 | 88.575 | 94.796 |
| kat | 1505 | 75.083 | 53.223 | 55.880 | 77.010 | 80.731 |
| kaz | 1418 | 56.629 | 38.575 | 52.680 | 53.173 | 59.520 |
| khk | 1847 | 50.731 | 24.689 | 52.084 | 50.514 | 49.053 |
| krl | 285 | 58.246 | 38.947 | 48.772 | 64.912 | 71.228 |
| lud | 878 | 91.230 | 92.027 | 92.141 | 93.508 | 92.141 |
| non | 541 | 78.373 | 51.386 | 59.704 | 73.752 | 83.549 |
| pol | 259 | 79.923 | 74.903 | 62.934 | 81.467 | 84.942 |
| poma | 133 | 74.436 | 70.677 | 60.902 | 73.684 | 80.451 |
| slk | 111 | 76.577 | 74.775 | 72.973 | 80.180 | 78.378 |
| tur | 538 | 72.862 | 29.182 | 48.513 | 63.941 | 83.829 |
| vep | 971 | 59.423 | 29.763 | 40.886 | 58.805 | 62.925 |

Table 24: Performance on verbs (N) in the large training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|------|--------|---------|--------|--------|--------|
| ang | 1085 | 64.332 | 37.143 | 39.078 | 64.147 | 63.594 |
| ara | 821 | 83.800 | 43.849 | 48.965 | 82.704 | 76.248 |
| evn | 49 | 69.388 | 8.163 | 8.163 | 71.429 | 71.429 |
| got | 309 | 84.790 | 19.094 | 59.223 | 82.524 | 89.644 |
| hun | 343 | 84.257 | 29.446 | 35.277 | 77.551 | 88.921 |
| hye | 315 | 91.429 | 40.635 | 41.270 | 90.476 | 96.190 |
| kat | 42 | 83.333 | 59.524 | 64.286 | 80.952 | 83.333 |
| kor | 221 | 77.828 | 41.629 | 35.747 | 65.611 | 72.851 |
| krl | 50 | 64.000 | 36.000 | 54.000 | 68.000 | 70.000 |
| lud | 105 | 92.381 | 92.381 | 92.381 | 92.381 | 90.476 |
| non | 652 | 82.822 | 46.319 | 48.006 | 91.411 | 96.626 |
| pol | 428 | 83.645 | 58.645 | 55.140 | 96.028 | 99.065 |
| poma | 242 | 77.686 | 59.091 | 48.760 | 69.835 | 72.314 |
| slk | 1142 | 85.639 | 54.991 | 45.184 | 87.653 | 96.848 |
| tur | 16 | 81.250 | 31.250 | 43.750 | 25.000 | 93.750 |
| vep | 233 | 61.803 | 35.193 | 37.339 | 68.240 | 69.528 |

Table 25: Performance on verbs (ADJ) in the large training condition

| Lang | # | CLUZH | Flexica | OSU | Tüm-M | UBC |
|------|-----|--------|---------|--------|--------|--------|
| ang | 59 | 3.390 | 0 | 0 | 0 | 0 |
| asm | 78 | 43.590 | 15.385 | 46.154 | 42.308 | 51.282 |
| evn | 30 | 3.333 | 0 | 0 | 10.000 | 16.667 |
| got | 476 | 84.244 | 24.370 | 50.840 | 82.983 | 87.185 |
| hun | 12 | 41.667 | 25.000 | 33.333 | 41.667 | 33.333 |
| hye | 19 | 78.947 | 78.947 | 78.947 | 94.737 | 78.947 |
| kor | 127 | 70.079 | 52.756 | 41.732 | 60.630 | 62.205 |
| krl | 55 | 50.909 | 41.818 | 50.909 | 52.727 | 49.091 |
| non | 213 | 76.056 | 62.441 | 45.540 | 79.812 | 86.385 |
| pol | 615 | 94.959 | 42.764 | 42.764 | 72.846 | 94.797 |
| poma | 875 | 53.829 | 20.571 | 24.343 | 48.343 | 53.600 |
| slk | 62 | 95.161 | 70.968 | 54.839 | 95.161 | 96.774 |
| vep | 25 | 52.000 | 44.000 | 48.000 | 52.000 | 64.000 |

Table 26: Performance on verbs (V.PTCP) in the large training condition

| Lang | V | N | ADJ | V.PTCP |
|------|--------|--------|--------|--------|
| ckt | 5.000 | 21.429 | 100.00 | 50.000 |
| hsb | 71.429 | 91.892 | 77.778 | – |
| hsi | 100.00 | 100.00 | 75.000 | – |

Table 27: TüMorph-FST results by POS. TüMorph-FST was only run on three languages, all in the small training condition.