SIGMORPHON 2022 Task 0 Submission Description: Modelling Morphological Inflection with Data-Driven and Rule-Based Approaches

Tatiana Merzhevich[1], Nkonye Gbadegoye[1], Leander Girrbach[1], Jingwen Li[1], Ryan Soh-Eun Shim[2]
[1]Department of Linguistics, University of Tübingen
{firstname.lastname}@student.uni-tuebingen.de
[2]Institute for Natural Language Processing, University of Stuttgart
soh-eun.shim@ims.uni-stuttgart.de

Abstract

This paper describes our participation in the 2022 SIGMORPHON-UniMorph Shared Task on Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation. We present two approaches: one being a modification of the neural baseline encoder-decoder model, the other being hand-coded morphological analyzers using finite-state tools (FST) and outside linguistic knowledge. While our proposed modification of the baseline encoder-decoder model underperforms the baseline for almost all languages, the FST methods outperform other systems in the respective languages by a large margin. This confirms that purely data-driven approaches have not yet reached the maturity to replace trained linguists for documentation and analysis especially considering low-resource and endangered languages.

1 Introduction

There are two tracks of the task of language Inflection Generation: Typologically Diverse Morphological (Re-)Inflection and (Automatic) Morphological Acquisition Trajectories. We only participate in the first track, Typologically Diverse Morphological (Re-)Inflection.

Here, the main goal is to predict inflected forms of a word by given lemmas and sets of morphological tags. In total, the task features 32 languages, for several of which both a low-resource scenario and a high resource scenario are proposed.

Our participation was split into two systems: One is a modification of the encoder-decoder baseline described in Wu et al. (2021), which is applied to all languages and resource settings.[1] The other system is based on hand-coded finite-state transducers for Chukchi (ckt), Upper Sorbian (hsb), and Kholosi (hsi).

The modification of the encoder-decoder baselines aims for better interpretability of predictions, but underperforms the baseline on almost all languages. The finite-state approaches yield very strong performance on the respective languages, however, their creation may have accidentally violated the train set / test set separation by usage of publicly available data UniMorph provided by Kirov et al. (2018) while constructing the transducers.

2 Methodology

2.1 Data-Driven Approach

In order to enable more explicit control of predicted forms and better interpretability, we propose a modification of the encoder-decoder baseline as in Wu et al. (2021). The main idea is as follows: We provide a directed graph whose states represent generated characters. Edges represent allowed transitions. This graph could be a full FST, or a simpler structure. Then, at each time-step, the encoder-decoder model predicts a distribution over states instead of characters as in the baseline model. While the difference may seem negligible, we argue there are several reasons why formulating the morphological prediction task in this way is useful: The provided graph can be used to control which sequences can be generated by disallowing illegal transitions during decoding. Also, the graph can be created and edited automatically or manually, which allows to inject expert knowledge. Here, different states may generate the same character, but in this way disambiguate possible trajectories through the graph. Finally, since each prediction can be directly mapped to a certain location in the graph topology, the model predictions can be interpreted relative to the given graph. If the graph is designed in a sufficiently informative way, this may allow better interpretation of predictions and also errors.

For training, each target form is converted to a path through the given graph and the characters are

[1] Unfortunately, we failed to submit results for Middle Low German (glm).

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 204 - 211
July 14, 2022 ©2022 Association for Computational Linguistics

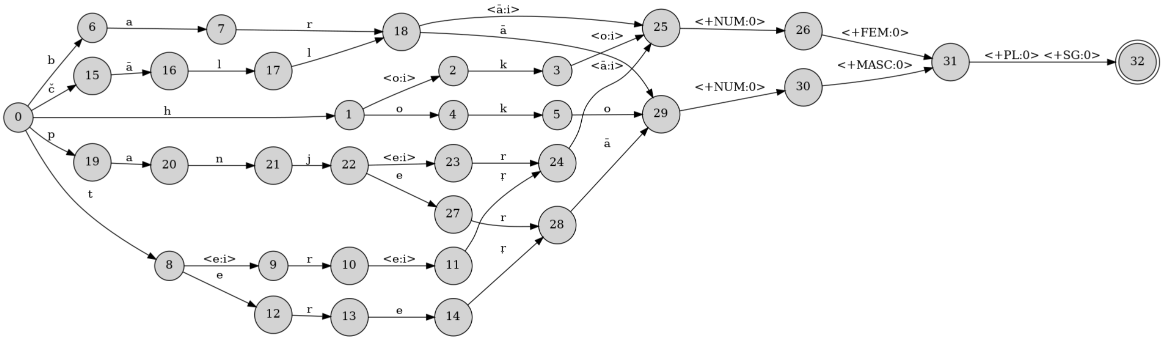

Fig. 1: Finite-state transducer for Kholosi numerals 1-5.

replaced by state identifiers. However, in order to simplify the provided graph, we may also define special states that allow the prediction of arbitrary characters, for example to predict base morphemes or copy them from the input lemma. In this case, we do not replace the respective characters with state identifiers. In any case, we train the baseline encoder-decoder model in the standard way but with modified targets.

The proposed idea is similar to learning weights of a FST as described in Rastogi et al. (2016). However, in our case, the encoder-decoder model does not have explicit access to the graph, but has to implicitly learn the possible transitions and their weights.

For this shared task, we were only able to test a simple but automatic way of constructing the proposed graph as auxiliary data structure for decoding: First, we align each paradigm in the train set, i.e. each set of forms with the same lemma, by iteratively aligning forms to the already aligned forms using the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) with column-sum as scoring metric. We replace all aligned substrings that are present in the lemma and all its forms by a placeholder symbol. This approach is similar to the method suggested in Forsberg and Hulden (2016). Finally, we use the same procedure to align the resulting forms of all paradigms.

Having obtained such alignments, each position in the multiple sequence alignment becomes a state in the graph. So far, we do not consider constraints on the edges, i.e. we effectively treat the graph as fully connected. However, in the future we would like to generate and evaluate more expressive auxiliary graphs.

2.2 Rule Based Approach

The morphological analyzers for three manually annotated languages were built using a finite-state compiler Foma (Hulden, 2009), which is based on lexicon and rules. The lexicon stores a list of words to which morphological analysis is applied. The rule transducers are established from regular expressions and applied to the list of identified word forms. For the system to perform better it is necessary to have a large lexical dataset to obtain higher accuracy of the morphological analysis performance. Therefore, we used the wordsets provided by the Universal Morphology project (UniMorph) (Kirov et al., 2018), which offers lists with lemmas, word forms and universal feature schemas with morphological categories.

Language	ISO	Speakers	Status
Chukchi	ckt	5.100	Threatened
Kholosi[2]	hsi	1.800	Unknown
Upper Sorbian	hsb	13.300	Threatened

Table 1: Manually annotated languages with their respective number of speakers and status (According to Eberhard et al. (2021)).

Chukchi Chukchi is a polysynthetic language spoken on the Chukotka Peninsula, in the northern part of the Russian Federation. It is composed of a rich inflectional and derivational morphology with progressive and regressive vowel harmony, productive incorporation, and extensive circumfixing across all its parts of speech described in Andriyanets and Tyers (2018). Chukchi is an ergative absolutive language with a highly complex system of verbal agreement constituting prefixal and suffixal components as stated in Bobaljik (1998).

[2] Data from Anonby and Bahmani (2016).

These components are commonly described as having some form of "split" ergativity such that prefixes show a nominative-accusative alignment, while suffixes show an absolutive-ergative bias (Wexler, 1982; Spencer, 2000).[3]

Chukchi also displays various types of perfective aspect as described in Volkov and Pupinina (2016). Examples of such are provided below.[4]

(1)	*etˀəm*	*Wełwəne*	*ɣetułˀetłinet*
	etˀəm	Wełwə-ne	ɣe-tułˀet-łinet
	apparently	Wełwə-ERG	PF-steal-3PL.PFV

’Apparently, Welwe stole them (deer)’

(2)	*ɣəm*	*tˀəłɣˀi*	*ɣətkaɣtə*
	ɣəm	tˀəł-ɣˀi	ɣətka-ɣtə
	I.ABS	hurt-AOR.3SG	leg-ALL

’I hurt my leg’

(3)	*ɣənin*	*əneqej*		
	ɣənin	əneqej		
	your	old.brother.ABS.SG		
	ɣekełitkułin	*kałetkorak*	*?*	
	ɣe-kełitku-łin	kałetkora-k	?	
	PFV-study-3SG.PFV	school-LOC		

’Did your older brother go to school ?’

A very critical set of rules incorporated into the FST were circumfixation and vowel harmony. Vowels in Chukchi are divided into two groups based on vowel height in addition to a schwa sound. The first group are the dominant vowels, which consist of letters э, о, а. The second group are the recessive vowels which are и, у, э (Andriyanets and Tyers, 2018). Both groups contain "э", which in both cases, are distinguished based on vowel harmony. Vowel harmony occurs progressively and regressively, influencing the entire word, thus morphological and phonological features can cause vowel changes in the stem and vice versa. For example, the verb "тэлпык", in the "V;PFV;IND;SG;3;PST" context becomes "гэтэлпылин".[5] While on the other hand, the verb "панрэватык" in the same context, becomes "гапанрэбатлен", thus changing the prefix-suffix combination "гэ..лин" into "га..лен", as a result of the dominant vowel "а" in the stem.

Chukchi also has morphological processes that on many occasions, result in the mutation or elision of letters. For example, the word "итык" changes to "титгъэк" in the "V;PFV;IND;SG;1;PST" context, thus resulting in the elision of the last two letters "ы" and "к".

The morphological and phonological analyzer accounts for some of the morphological and phonological processes in Chukchi. The finite state transducer for Chukchi adjectives can be seen in Figure 2.

Kholosi The Kholosi FST is additionally based on preliminary descriptions of the language's morphology. Since a systematic account of Kholosi morphology is yet to be published, we work exclusively with the work of Arora (2020), which is based on elicitation from a single native Kholosi speaker.

One interesting phenomenon is gender alternation with vowel harmony. Kholosi has two grammatical genders, which can be reflected by morphemes, -o for masculine and -i for feminine (Arora, 2020). In numerals, for instance, the feminine form always ends with -i. Hence we have a rule that transform the last character to an i when the FEM tag is expected. From the given five pairs of MASC/FEM numerals in the training data, we observe a change of non-a/ā vowels to i.

We also note the (notational) discrepancies among different data sources:[6] In the training data provided by the shared task, the numeral lemmas ends with an ā instead of an o which is different from what was proposed by Arora (2020).[7] We also found glossed sentences in Kholosi where instead of *baro* (or *barā*, depending on the data source), *bahro* is used for the masculine (lemma) form of the numeral *two*.[8]

The resulting numeral FST is shown in Figure 1. Adjectives can also inflect with respect to gender ac-

[3] Absolutive, as it is used traditionally, refers to the grouping of an intransitive subject and direct object of a transitive verb. Nominative here is reserved to indicate the grouping of the intransitive subject and transitive subject.

[4] This paper follows the Leipzig Glossing Rules (Can be accessed from: https://www.eva.mpg.de/lingua/resources/glossing-rules.php), with additionally AOR = aorist.

[5] The "л" sound is used as a substitute for the Cyrillic letter "El with hook".

[6] With possible errors inherited from UniMorph data: V;IPFV;IND;SG;3;PRS form of the verb *karen* is attested as *keraw* in glossed sentences but provided as *kerav* in train data, while the same forms for other verbs all observe an *-aw* suffix.

[7] Except the case of *hoko*, meaning *one*.

[8] Can be accessed from https://aryamanarora.github.io/kholosi/sentences.html

cording to Arora (2020), so similar rules are added to the adjective FST, although there are no feminine adjective forms provided in the training data at all.

Upper Sorbian Sorbian is a West Slavic language spoken in eastern Germany, in Saxony and Brandenburg (also called Lusatia). Sorbian demonstrate closeness with Czech and Polish, and at the same time shares certain features with South Slavic languages, such as the use of the double grammatical number with nouns, adjectives and verbs, as well as the use of specific forms of past tense. Unfortunately, due to the constant contact with German, Sorbian includes a large number of German loanwords in its standardized lexicon (Glaser, 2007).

According to Eberhard et al. (2021), the number of Upper Sorbian speakers estimated as no more than 13.000. Their community is fully bilingual, which means that if the rule of thumb proposed by Payne and Payne (1997) is applied, the Upper Sorbian might become extinct by the year 2070. However, the actual number of Sorbian speakers is based on estimations. According to the principles of minority law applicable in the Federal Republic of Germany, the commitment to a minority is free and not registered officially, as reported by Marti (2007).

3 Results

The data-driven approach earned third place for both small and large languages in part one of the shared task, although under-performing the neural baseline. The official preliminary results are available in Table 3.[9]

The rule based approach for three languages with relatively small datasets outperformed all other systems. However, the analyzers were not only built by the provided train data, but also with help of linguistic knowledge and UniMorph schemas, which in large encompassed the test set. The performance results are shown in Table 2.

Language	Result
Chukchi	19.565
Upper Sorbian	83.750
Kholosi	96.667

Table 2: Results (overall test scores) of the finite-state approach.

4 Discussion

The findings of our study follow up on the work of Beemer et al. (2020), where it was concluded that "it is very difficult in many cases to outperform a state-of-the-art neural network model without significant development effort and attention to nuanced morphophonological patterns". The finite-state grammars in Beemer et al. (2020) outperformed the seq2seq results only in languages with high morphophonological complexity such as Tagalog, and came at the cost of 5.5 manual working hours on average per week, over the course of 5 weeks.

Our work similarly required a high number of working hours, but was able to outperform other systems in low-resource scenarios precisely due to the reliance on the linguistic expertise of the FST creators. The trade-off we observe in our submission is therefore how much interpretability and intuition-guided modifications of a model is desired, where for sufficiently well-documented languages the benefit of FSTs may not be as obvious, but for scenarios where sufficient data may not be able to be collected, our submission would indicate that FSTs still maintain an edge over neural approaches.

Beemer et al. (2020) note that for certain languages the amount of inconsistencies makes it unlikely for a hand-written grammar to surpass neural systems, where certain rules were deemed irregular enough to not warrant treatment by their FSTs. We believe our study provides a partial defense for FSTs with precisely the same point: in cases where the amount of data is insufficient for neural models to infer the rules of low-resource languages, it is unlikely that the neural models can perform well without further data; for rule-based approaches, even with limited amount of data (e.g. due to a lack of orthography or access to native speakers), the models can always rely on linguistic knowledge to provide working solutions.

Our usage of data outside of the training set is also based on this concern: it is unlikely that there will be enough human resources for most of the world's languages to have enough data collected, but for the practical situation where a morphological analyzer is nevertheless needed, our results indicate that this approach still remains to be the most practical solution.

[9] Taken from `https://github.com/sigmorphon/2022InflectionST/blob/main/results/preliminary.md`

For our other submission where we experimented with a data-driven approach, we believe that it constitutes a step towards more interpretable encoder-decoder predictions, which in light of the above, may also stand as a future research direction, which could be beneficial for practical scenarios.

5 Conclusion

We presented two different approaches to morphological inflection, a data preprocessing method to be used in conjunction with standard encoder-decoder models and hand-coded finite-state methods. Despite the problems with both approaches, i.e. insufficient performance of the data-driven approach and large amounts of effort needed to engineer FSTs, we think that both have their benefits, as discussed in Section 4.

In particular we would like to note that the effort invested into creating FSTs expands computational resources for under-researched and low-resource languages and can be considered as a collaborative part in language revitalization as proposed in Pine and Turin (2017). Also, both approaches allow for future extensions, e.g. a big improvement of finite-state analyzers would be expansion of current lexicons with guessers for assigning possible stems and part-of-speech tags.

References

Vasilisa Andriyanets and Francis Tyers. 2018. A prototype finite-state morphological analyser for Chukchi. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 31–40, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Erik Anonby and Hassan Mohebbi Bahmani. 2016. Shipwrecked and landlocked: Discovery of Kholosi, an Indo-Aryan language in south-west Iran. In Jila Ghomeshi, Carina Jahani, and Agnes Lenepveu-Hotz, editors, *Further Topics in Iranian Linguistics. Proceedings of the 5th International Conference on Iranian Linguistics, held in Bamberg on 24-26 August 2013*, volume 58 of *Cahiers de Studia Iranica*, pages 13–36. Peeters, Louvain.

Aryaman Arora. 2020. Historical Phonology and other Observations on Kholosi.

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.

Jonathan David Bobaljik. 1998. Pseudo-Ergativity in Chukotko-Kamchatkan Agreement Systems. *Recherches Linguistiques de Vincennes*, 27:21–44.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition*, volume 16. SIL international, Dallas, TX.

Markus Forsberg and Mans Hulden. 2016. Learning Transducer Models for Morphological Analysis from Example Inflections. In *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata*, pages 42–50, Berlin, Germany. Association for Computational Linguistics.

Konstanze Glaser. 2007. Minority languages and cultural diversity in Europe. In *Minority Languages and Cultural Diversity in Europe*. Multilingual matters.

Mans Hulden. 2009. Foma: a Finite-State Compiler and Library. In *EACL*.

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

Roland Marti. 2007. Lower Sorbian — twice a minority language. *International journal of the sociology of language*, 2007(183):31–51.

Saul B Needleman and Christian D Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Thomas E Payne and Thomas Edward Payne. 1997. *Describing morphosyntax: A guide for field linguists*. Cambridge University Press.

Aidan Pine and Mark Turin. 2017. *Language revitalization*. Oxford University Press.

Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting Finite-State Transductions With Neural Context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 623–633, San Diego, California. Association for Computational Linguistics.

Andrew Spencer. 2000. Agreement morphology in Chukotkan. *Amsterdam studies in the theory and history of linguistic science*, pages 191–222.

Oleg Volkov and Maria Pupinina. 2016. The category of perfect in chukotko-kamchatkan languages. *Acta Linguistica Petropolitana*, pages 535–568.

Paul Wexler. 1982. Bernard Comrie The Languages of the Soviet Union. *Language Problems and Language Planning*, 6(2):166–175.

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.

A Appendix

Language	Small	Large
ang	45.962	60.945
ara	62.857	75.338
asm	38.995	63.065
bra	53.134	-
ckt	8.696	-
evn	23.867	52.037
gml	*	-
goh	52.158	-
got	47.693	65.346
guj	40.855	-
heb	31.15	47.9
hsb	7.5	-
hsi	0.0	-
hun	51.85	68.15
hye	61.45	66.7
itl	33.056	-
kat	47.8	78.85
kaz	55.165	53.611
ket	13.139	-
khk	39.495	47.727
kor	17.821	47.556
krl	10.421	24.098
lud	46.559	50.506
mag	51.163	-
nds	21.947	-
non	47.313	79.759
pol	53.85	67.7
poma	45.873	58.829
sjo	54.496	-
slk	56.05	65.75
slp	12.658	-
tur	19.25	33.6
vep	27.446	44.104

Table 3: Results (overall accuracy of test set predictions) of data-driven approach for all languages. “-” means not part of this shared task. “*”: We accidentally did not submit results for gml

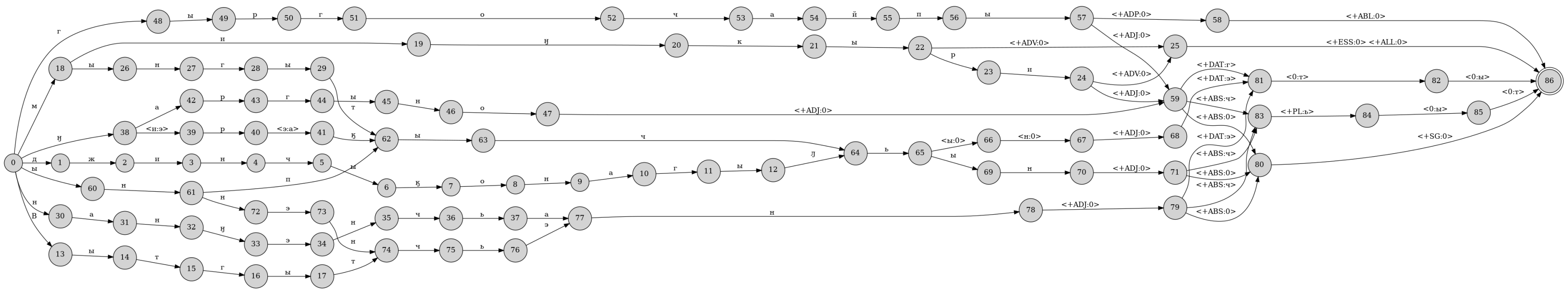

Fig. 2: Finite-state transducer for Chukchi adjectives.