SIGMORPHON–UniMorph 2023 Shared Task 0: Typologically Diverse Morphological Inflection

Omer Goldman[1] **Khuyagbaatar Batsuren**[2] **Salam Khalifa**[3] **Aryaman Arora**[4]
Garrett Nicolai[5] **Reut Tsarfaty**[1] **Ekaterina Vylomova**[6]
[1]Bar-Ilan University [2]National University of Mongolia [3]Stony Brook University
[4]Georgetown University [5]University of British Columbia [6]University of Melbourne
omer.goldman@gmail.com vylomovae@unimelb.edu.au

Abstract

The 2023 SIGMORPHON–UniMorph shared task on typologically diverse morphological inflection included a wide range of languages: 26 languages from 9 primary language families. The data this year was all lemma-split, to allow testing models' generalization ability, and structured along the new hierarchical schema presented in (Batsuren et al., 2022). The systems submitted this year, 9 in number, showed ingenuity and innovativeness, including hard attention for explainability and bidirectional decoding. Special treatment was also given by many participants to the newly-introduced data in Japanese, due to the high abundance of unseen Kanji characters in its test set.[1]

1 Introduction

As a long-running shared task, the SIGMORPHON-UniMorph task on morphological inflection is a major engine behind the surging interest in computational morphology, as it facilitated both the building of UniMorph as a large multilingual morphological dataset, and the development and testing of morphological inflection models. In its first few installments (Cotterell et al., 2016, 2017, 2018; Vylomova et al., 2020) the focus of the task was first and foremost on generalization across languages, with the their number raising steadily from 10 languages in the task of 2016 to 90 languages in 2020.

Later studies, both in the 2021 shared task (Pimentel et al., 2021) and otherwise (Goldman et al., 2022a), discovered that the impressive results achieved by systems submitted to these tasks were in large part due the presence of test lemmas in the train set. As a result, the 2022 shared task (Kodner et al., 2022) focused on generalization to both unseen lemmas and unseen feature combinations.

[1]Data, evaluation scripts, and predictions are available at: https://github.com/sigmorphon/2023InflectionST

In this task we continue to test systems on the challenging lemma-split setting while circling back to the inclusivity objective that guided the task from its inception. To this end, we employ the hierarchical annotation schema of UniMorph 4.0 (Batsuren et al., 2022) that allows more natural annotation of languages with complex morphological structures such as case stacking and polypersonal agreement. This year we include 26 languages from 9 primary language families: Albanian, Amharic, Ancient Greek, Arabic (Egyptian and Gulf), Armenian, Belarusian, Danish, English, Finnish, French, Georgian, German, Hebrew, Hungarian, Italian, Japanese, Khaling, Macedonian, Navajo, Russian, Sámi, Sanskrit, Spanish, Swahili and Turkish. The inclusion of Japanese, written in Kanji characters that are rarely shared across lemmas, compelled all systems this year to find ways to deal with unseen characters in the test set.

In total, 9 systems were submitted by 3 teams, both neural and non-neural models, and they were compared against 2 baselines, neural and non-neural as well. The submitted systems experimented with innovative ideas for morphological inflection as well as for sequence-to-sequence modeling in general. Girrbach (2022) introduced an elaborate attention mechanism between static representations for explainability, and Canby and Hockenmaier (2023) experimented with a new type of decoder for transformer models that is able to decode from both left to right and vise versa simultaneously. Lastly, Kwak et al. (2023) improved the non-neural affixing system used as a baseline.

The results show that although on average systems achieve impressive results in inflecting unseen lemmas, some languages still present a substantial challenge, mostly extinct languages like Ancient Greek and Sanskrit or low resourced languages like Navajo and Sámi. In addition, the results point to a dependency on the writing system that could be further explored in future shared tasks.

Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 117–125
July 14, 2023

Family	Subfamily	ISO 639-2	Language	Source of Data	Annotators
Afro-Asiatic	Semitic	afb	Arabic, Gulf	Obeid et al. (2020)	Salam Khalifa
		arz	Arabic, Egyptian		Nizar Habash
		amh	Amharic	Gasser (2011)	Michael Gasser
		heb	Hebrew	Wiktionary	Omer Goldman
Indo-European	Albanian	sqi	Albanian	Wiktionary	Kirov et al. (2016)
	Armenian	hye	Eastern Armenian	Wiktionary	Hossep Dolatian
	Balto-Slavic	bel	Belarusian	Wiktionary	Ekaterina Vylomova
		mkd	Macedonian	Wiktionary	Ekaterina Vylomova
		rus	Russian	Wiktionary	Ekaterina Vylomova
	Germanic	dan	Danish	Wiktionary	Mans Hulden Khuyagbaatar Batsuren
		eng	English	Wiktionary	Mans Hulden Khuyagbaatar Batsuren
		deu	German	Wiktionary	Ryan Cotterell
	Helenic	grc	Ancient Greek	Wiktionary	Kirov et al. (2016)
	Indo-Aryan	san	Sanskrit	Huet's inflector	Aryaman Arora
	Romance	fra	French	Wiktionary	Géraldine Walther
		ita	Italian	Wiktionary	Géraldine Walther
		fra	Spanish	Wiktionary	Géraldine Walther
Japonic		jap	Japanese	Wiktionary	Khuyagbaatar Batsuren Omer Goldman
Kartvelian		kat	Georgian	Guriel et al. (2022)	David Guriel Simon Guriel Silvia Guriel-Agiashvili Nona Atanelov
Na-Dené	Southern Athabascan	nav	Navajo	Wiktionary	Mans Hulden Rob Malouf
Niger–Congo	Bantu	swa	Swahili	Goldman et al. (2022b)	Lydia Nishimwe Shadrak Kirimi Omer Goldman
Sino-Tibetan	Kiranti	klr	Khaling	Walther et al. (2013)	Géraldine Walther
Turkic	Oghuz	tur	Turkish	Wiktionary	Omer Goldman Duygu Ataman
Uralic	Finnic	fin	Finnish	Wiktionary	Mans Hulden
		sme	Sámi	Wiktionary	Mans Hulden
	Ugric	hun	Hungarian	Wiktionary	Judit Ács Khuyagbaatar Batsuren Gábor Bella, Ryan Cotterell Christo Kirov

Table 1: Languages presented in this year's shared task

2 Task Description

This year's task was organized in a very similar fashion 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 a cell in a morphological paradigm. They were provided with a training set of several thousands of examples, as well as a development set and test set for each language. The training data consisted of (`lemma`, `feature set`, `inflected form`) triples, while the inflected forms were held out from the test set. The development set was provided in both train- and test-like formats.

Data was made available to participants in two phases. In the first phase, the training and development sets were provided for most languages. In the second phase, training and development sets were released for some extra ("surprise") languages and the test sets were provided for all languages.[2]

Schema Differences The data this year followed the hierarchical annotation schema that was suggested by Guriel et al. (2022) and adopted in UniMorph 4.0 (Batsuren et al., 2022). The difference that was most pronounced in the data was the replacement of opaque tags that grouped several features such as AC3SM(a 3rd person singular masculine accusative argument) with the hierarchically combined features ACC(3,SG,MASC), i.e. without introducing a new tag for each feature combination in the cases of polypersonal agreement.

[2]The surprise languages were: Albanian, Belarusian, German, Gulf Arabic, Khaling, Navajo, Sámi and Sanskrit.

3 The Languages

The selection of languages used in this year's task is varied at almost any dimension. In terms of language genealogy we have representatives of 9 language families, some are widely used, like English and Spanish, and others are endangered or extinct, like Khaling and Sanskrit. The languages employ a wide variety of orthographic systems with varying degrees of transparency (Sproat and Gutkin, 2021): alphabets (e.g., German), abugidas (e.g., Sanskrit), abjads (e.g., Hebrew), and even one logographs using language (Japanese).

In light of the new annotation schema, many languages in this year's selection employ forms that refer to multiple arguments. Possessors are marked on nouns in 6 of the languages: Hebrew, Hungarian, Amharic, Turkish, Armenian and Finnish. In addition, polypersonal agreement appears in verbs of 5 of the languages: Georgian, Spanish, Hungarian, Khaling and Swahili.[3] Other notable morphological characteristics include, among others, the ablaut-extensive Semitic languages and prefix-inclined Navajo.

All in all, Table 1 enumerates the languages included in the shared task.

Languages new to UniMorph A couple of languages, namely Swahili and Sanskrit, have seen their respective UniMorph data increased substantially in size for this task. The Swahili data, that previously had partial inflection tables, was expanded using the clause morphology data of Goldman et al. (2022b), so a Swahili verbal inflection table includes more that 14,000 forms rather than mere 180. The Sanskrit data was massively expanded, mostly in terms of the number of lemmas, by incorporating data from Gérard Huet's Sanskrit inflector.[4]

In addition, one previously unrepresented language was introduced to UniMorph — Japanese. The data was crawled from Wiktionary and canonicalized to match the UniMorph 4.0 format. The usage of Kanji characters, logograms of Chinese origin that are completely unrepresentative of the pronunciation and almost uniquely used per lemma, can pose an interesting challenge to inflection systems that will have to deal with many unseen characters.

[3]Nouns in Arabic also mark their possessor and Verbs in Navajo also agree with multiple arguments, but the UniMorph data includes partial inflection tables for these languages.

[4]`https://sanskrit.inria.fr/index.fr.html`

# Inflection Tables	Languages
500	`fin, fra, grc, heb, hun, hye, ita kat, klr, nav, san, sme, spa, sqi swa, tur`
1000	`amh, bel, deu, jap, mkd, rus`
2000	`dan`
3000	`afb, arz, eng`

Table 2: Results of all the systems, submitted and baselines over the test sets in all languages. the best system(s) per language in marked in **bold**. The systems are ordered by the averaged success.

4 Data Preparation

All data for this task is provided in standard UniMorph format, with training items consisting of (`lemma, morphosyntactic features, inflected form`) triples. Since the goal of the task is to predict inflected forms, the test set was presented as (`lemma, features`) pairs. The data for all languages was lemma-split (Goldman et al., 2022a).

For each language, a number of inflection tables (i.e., lemmas) were sampled from the entire UniMorph dataset. 80% of the tables were used for the train set, and the rest were split between the validation and the test sets, then 10,000 forms were sampled from the inflection tables of the train set, and 1,000 forms were sampled for the validation and test sets from the respective tables. The number of inflection tables used was capped at 500, in cases where the tables were too small to generate enough data more tables were added until it was sufficient. Table 2 details the amount of tables used for each language.

5 The systems

5.1 Baseline Systems

The baseline systems provided this year are a recurrent appearance of the baselines of yesteryears: a **neural** character-level transformer (Wu et al., 2021, details in Appendix A), and a **non-neural** statistical application of affixing rules firstly used by Cotterell et al. (2017).

5.2 Submitted Systems

University of Arizona Kwak et al. (2023) submitted several non-neural models. Their first system (**AZ1**) is a re-implementation of the non-neural baseline, while another system of theirs (**AZ2**) uses the same framework but improves the rules used for both processing of the training data and making the predictions over the test set. In addition, they experimented with a weighted finite-state transducers (WFST; **AZ3**), and they provided an ensemble of the WFST with AZ2 (**AZ4**).

University of Tübingen Girrbach (2023) focused on explainability of the predictions of a neural inflection model. They did not get into debate on whether soft attention between model's hidden states is a good explanation (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), but rather applied a hard attention mechanism directly over static character representations. The models complexity comes solely from the attention module itself, that includes a LSTMs that run over the example's source and target.

University of Illinois Canby and Hockenmaier (2023) provided the most extensive set of experiments with transformer-based neural models. The ultimate focus of their work was the directionality of the decoder. Rather than decoding left-to-right, their first system (**IL1**) used two unidirectional models and chose a prediction that got the higher probability assessed by its respective model. In addition, they experimented with a model capable of deciding whether to decode from left or right at each step separately and used it either to select between unidirectional predictions (**IL2**) or as a standalone model (**IL3**). Lastly, they equipped IL3 with a beam re-ranker (**IL4**).

Common system characteristics The Japanese data, with its high abundance of unseen characters, posed a major problem to the neural submitted systems. Thus, they all gave the Japanese data special treatment and replaced the unseen characters with special place holders that were filled in with the lemma characters as a post-processing step.[5]

None of the systems submitted made explicit use of the hierarchy of the features. The teams opted for flattening the structure and letting the models understand the relations between the features from the order. Thus, for example, the feature bundle V;PRS;NOM(1,SG);ACC(2,PL) was treated as V;PRS;NOM;1;SG;ACC;2;PL, with multiple person and number features on the same level.

6 Results and analysis

Table 3 summarizes the accuracy results of all systems over all languages based on the exact match between the prediction and gold outputs. In addition, we also provide macro-averaged score over languages.

System performance In terms of averaged performance, all neural systems outperformed the non-neural systems, with IL4 having the best performance. When examining the results per language, the neural baseline and three of the Illinois-submitted systems take the lead in about 6 languages each. The exceptions to this are English, Danish and French, in which the non-neural baseline is the best performing system. Partial explanation may be the small size of the inflection tables in Danish and English that necessitated inclusion of many lemmas in the training set and may facilitated better generalization ability of the non-neural baseline. Admittedly, this explanation is not valid for French, but this language was proven difficult in previous shared tasks (Cotterell et al., 2017, 2018) and in other works (Silfverberg and Hulden, 2018; Goldman and Tsarfaty, 2021).

The neural baseline system was significantly hampered by the lack of a special mechanism for the unseen characters in Japanese. When discarding the Japanese performance for all systems, the neural baseline is second in averaged performance. That is to say that devising a strategy to deal with unseen characters is highly necessary when inflecting lemma-split data in general, and logographic languages in particular.

Being the neural system with the lowest averaged accuracy, TÜB seem to trade some predictive power in favor of having more explainable outputs, as exemplified in Figure 1.

Although the WFST system that is AZ3 is the system with the lowest scores, including it as part of an ensemble resulted in some advantages and helped producing the best non-neural system — AZ4.

Language performance The performance of the per-language best system over most languages is quite impressive, and in some cases like Swahili and Khaling even exceptionally impressive. How-

[5] Another possible solution to this bind could have been to introduce a copy mechanism in the model itself, such as the one used by Makarov and Clematide (2018). However, no team chose this path.

Language	AZ3	AZ1	Baseline Non-neural	AZ2	AZ4	TÜB	Baseline Neural	IL1	IL2	IL3	IL4
macro average	56.1	67.2	69.6	71.7	72.4	76.9	81.6	82.6	84.0	84.1	**84.3**
afb	34.5	30.8	30.8	52.7	52.7	75.8	80.1	80.7	82.2	84.1	**84.6**
amh	59.9	65.4	65.4	74.0	74.0	83.8	82.2	88.9	**90.6**	88.9	88.6
arz	75.7	77.2	77.9	80.8	80.8	87.6	**89.6**	89.2	88.7	89.1	88.7
bel	46.2	68.1	68.1	64.5	64.5	56.3	74.5	73.5	**74.7**	72.9	72.9
dan	64.8	**89.5**	**89.5**	87.4	87.4	85.7	88.8	88.8	**89.5**	86.5	87.5
deu	59.9	79.8	79.8	77.9	77.9	74.5	**83.7**	79.7	79.7	80.2	79.7
eng	67.0	**96.6**	**96.6**	96.2	96.2	96.0	95.1	95.6	95.9	94.6	95.0
fin	48.2	80.8	80.8	80.6	80.6	67.6	85.4	79.2	80.6	85.7	**86.1**
fra	76.7	**77.7**	**77.7**	76.3	76.3	67.9	73.3	69.3	74.7	71.7	72.9
grc	40.4	52.6	52.6	54.8	54.8	36.7	54.0	48.9	53.7	**56.0**	**56.0**
heb	51.6	64.5	64.5	76.7	76.7	81.3	83.2	77.3	79.3	**83.7**	83.6
heb$_{\text{voc}}$	34.7	30.9	30.9	65.3	65.3	82.7	92.0	**92.9**	92.6	90.9	91.0
hun	45.9	74.7	74.7	74.7	74.7	75.9	80.5	76.3	79.8	84.3	**85.0**
hye	88.9	86.3	86.3	86.2	88.9	85.9	91.0	88.4	91.5	**94.4**	94.3
ita	78.0	75.0	75.0	63.6	78.0	84.7	94.1	95.8	**97.2**	92.1	92.2
jap	67.0	64.1	64.1	64.1	67.0	**95.3**	26.3	92.8	94.2	94.9	94.9
kat	71.7	82.0	82.0	82.1	82.1	70.5	84.5	84.1	**84.7**	81.3	82.9
klr	27.8	54.5	54.5	53.1	53.1	96.4	**99.5**	99.4	99.4	99.4	99.4
mkd	64.9	91.6	91.6	90.8	90.8	86.7	**93.8**	91.9	92.4	92.1	92.4
nav	23.7	35.8	35.8	41.8	41.8	53.6	52.1	54.0	55.1	55.1	**55.6**
rus	66.8	86.0	86.0	85.6	85.6	82.1	**90.5**	87.4	87.3	84.2	85.5
san	47.0	62.2	62.2	62.1	62.1	54.5	66.3	63.3	**69.1**	67.7	65.9
sme	30.1	56.0	56.0	49.7	49.7	58.5	**74.8**	69.9	71.8	67.4	67.3
spa	86.3	87.8	87.8	87.4	87.4	88.7	93.6	90.9	91.4	**93.8**	93.1
sqi	73.8	19.3	83.4	78.1	78.1	71.5	85.9	87.6	88.9	**92.0**	91.6
swa	56.2	60.5	60.5	65.0	65.0	94.7	93.7	93.1	93.1	**96.6**	**96.6**
tur	28.1	64.6	64.6	64.6	64.6	81.8	**95.0**	90.9	90.8	90.3	92.0

Table 3: Results of all the systems, submitted and baselines over the test sets in all languages. the best system(s) per language in marked in **bold**. The systems are ordered by the averaged success.

ever, there are still some languages over which no system achieves over 80% accuracy. These are: Navajo, Ancient Greek, Sanskrit, Belarusian, Sami and French. While there is no one characteristic shared between all of these languages, it is worth noting that this list includes the only two extinct languages tested in this task, and the only mostly prefixing language. Perhaps further development of tailored models could close this gap.

Language	AZ4	IL4
afb	52.7	84.6
afb no diacr.	80.8	89.2
arz	80.8	88.7

Table 4: Results of all the best neural and non-neural systems over Gulf Arabic, with and without omission of diacritics. Results over Egyptian Arabic are provided for reference. Further evaluations and results for all systems appear in Appendix B.

The orthography's influence As in previous years, the Hebrew data was provided in two formats: the standard unvocalized abjad where vowels are largely omitted from the text, and the rarely used fully vocalized form that is computationally equivalent to an alphabet.

For most systems, the difference in performance between the two variants is stark. In general, the non-neural systems succeeded better over the unvocalized variant, presumably because omitting the vowels masks the non-concatenative ablauts. However, the neural systems fared better over the vocalized data, potentially due to the far lower level of ambiguity it exhibits.

However, the Arabic data complicates this picture. Although Egyptian and Gulf Arabic are closely related dialects with marginal differences in the inflectional system, most systems' success rates differ significantly between these two Arabic varieties. Error analysis revealed that inconsistent diacritization in the Gulf data is the main driving factor in this discrepancy in performance. Unlike the Egyptian Arabic data, not all forms in the Gulf data are diacritized. While all lemmas are diacritized in Gulf, only a subset of the verbal inflected forms are diacritized and the rest are not. In total, around 46% of the training data is diacritized.

The result is that the non-neural systems failed to generate vowel diacritics in the same somewhat arbitrary pattern unlike the neural systems, which

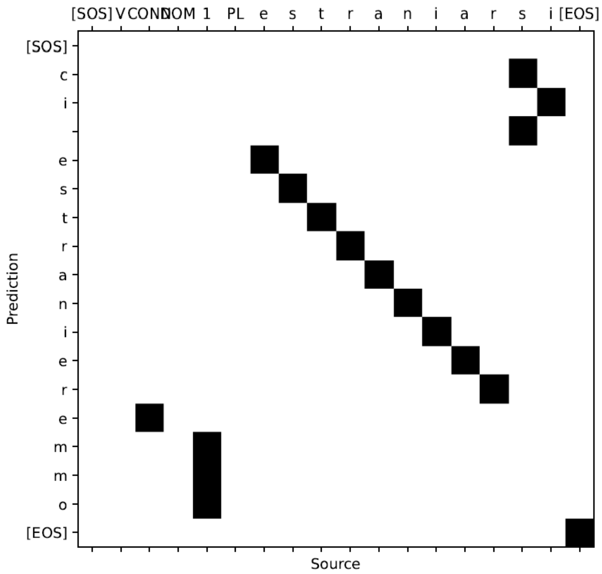

Figure 1: An example of explained inflection by TÜB. Each predicted character is anchored in one input symbol, other conditioning symbols omitted and can be found in Girrbach (2023).

managed to deal well with the inconsistency in the data. The exact match accuracy for Gulf Arabic for the best neural and non-neural systems, which was calculated after omission of all diacritics, is presented in Table 4 and detailed for all systems in Appendix B. It shows that without this source of inconsistency, the performance of Gulf Arabic is in line with the performance of Egyptian.

All in all, it seems like a consistent indication of vowels does not have the same effects in Hebrew and Arabic, despite their typological and orthographic similarity. The results over Arabic dialects are similar regardless of whether diacritics were omitted, while in Hebrew the vocalization played a greater role. This conundrum may point to a need to investigate further the role of the orthographic system in the success rate of inflection models, both neural and non-neural.

7 Conclusions

This year's shared task further promoted the goals of the recurring UniMorph inflection task: we tested innovative inflection systems on a challenging lemma-split data, and did so in an inclusive fashion both in terms of typological diversity of the languages included and the annotation schema that allows treatment of more complex morphological phenomena.

We received 9 submitted systems and tested them on 16 typologically diverse languages. The most interesting pattern arising from our results is the greatly varied performance between languages, with the best performing system ranging from 55.6 to 99.4 accuracy percentage. We thus conclude that further research is needed to close this gap.

Moreover, this year's task gave a prominent role to the orthographic systems of the languages selected, both by including for the first time a logographically written language and by analysing the role of abjad-vocalization in Semitic languages. We believe that this direction is a promising lead for promoting the understanding of the factors influencing the performance of inflection models.

Acknowledgements

The research of Omer Goldman and Reut Tsarfaty was funded by a grant from the European Research Council, ERC-StG grant number 677352, and a grant from the Israeli Ministry of Science and Technology (MOST), grant number 3-17992, for which they are grateful. All of Salam Khalifa's contributions were supported by the department of Linguistics and the Institute of Advanced Computational Science (IACS) at Stony Brook University.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott 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 Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóğa, 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. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Marc E. Canby and Julia Hockenmaier. 2023. A framework for bidirectional decoding: Case study in morphological inflection.

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.

Michael Gasser. 2011. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*, pages 94–99.

Leander Girrbach. 2022. SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 124–130, Seattle, Washington. Association for Computational Linguistics.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Toronto, Canada. Association for Computational Linguistics.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022a. (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.

Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Basmov, Shadrack Kirimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, and Duygu Ataman. 2022b. The MRL 2022 shared task on multilingual clause-level morphology. In *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 134–146, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2021. Minimal supervision for morphological inflection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2088, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 196–202, Dublin, Ireland. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Alice Kwak, Michael Hammond, and Cheyenne Wing. 2023. Morphological reinflection with weighted finite-state transducers. In *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Toronto, Canada. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

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.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889, Brussels, Belgium. Association for Computational Linguistics.

Richard Sproat and Alexander Gutkin. 2021. The taxonomy of writing systems: How to measure how logographic a system is. *Computational Linguistics*, 47(3):477–528.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Géraldine Walther, Guillaume Jacques, and Benoît Sagot. 2013. Uncovering the inner architecture of khaling verbal morphology. In *3rd workshop on Sino-Tibetan languages of Sichuan*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

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 Hyper Parameters of the Neural Baseline

For the neural baseline models we used the standard hyper parameters of Wu et al. (2021). These are:

- 4 transformer layers
- 4 attention heads
- 256 dimensions in the embeddings
- 1024 dimensions in the hidden feed forward layers
- 0.3 dropout chance
- 400 examples per batch
- 20,000 training steps at max
- Inverse square root scheduler with 4,000 worm up steps
- Adam optimizer with β of 0.98
- learning rate of 0.001
- label smoothing with α of 0.1

Language	AZ3	AZ1	Baseline Non-neural	AZ2	AZ4	TÜB	Baseline Neural	IL1	IL2	IL3	IL4
afb original	34.5	30.8	30.8	52.7	52.7	75.8	80.1	80.7	82.2	84.1	84.6
afb mixed	66.9	70.7	70.7	70.3	70.3	77.4	82.2	83.1	84.5	86.0	86.5
afb no diacr.	74.4	77.4	77.4	80.8	80.8	81.9	87.9	87.8	89.2	89.0	89.2
arz	75.7	77.2	77.9	80.8	80.8	87.6	89.6	89.2	88.7	89.1	88.7

Table 5: Results of all the systems over Gulf Arabic with different considerations for inconsistent diacritization of the original data. Results over Egyptian Arabic are provided for reference.

B Detailed evaluations for Gulf Arabic

Table 5 details several evaluations done over Gulf Arabic, with the results of Egyptian Arabic provided for reference. Specifically:

- *original* is the evaluation done over the inconsistently diacritized data, as it appears in Table 3.
- *mixed* is the evaluation done after removing diacritics only the predictions whose respective gold contains no diacritics
- *no diacr.* is the evaluation done after removing all diacritics from both predictions and gold outputs.