The MRL 2022 Shared Task on Multilingual Clause-level Morphology

Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Basmov, Shadrack Kirimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, et al.

The MRL 2022 Shared Task on Multilingual Clause-level Morphology

Omer Goldman[1] **Francesco Tinner**[2] **Hila Gonen**[5] **Benjamin Muller**[3]
Victoria Basmov[1] **Shadrack Kirimi**[4] **Lydia Nishimwe**[3] **Benoît Sagot**[3]
Djamé Seddah[3] **Reut Tsarfaty**[1] **Duygu Ataman**[6]
[1]Bar Ilan University [2]University of Amsterdam [3]Inria, Paris [4]Chuka University
[5]Paul G. Allen School of Computer Science & Engineering, University of Washington
[6]New York University

Abstract

The 2022 Multilingual Representation Learning (MRL) Shared Task was dedicated to clause-level morphology. As the first ever benchmark that defines and evaluates morphology outside its traditional lexical boundaries, the shared task on multilingual clause-level morphology sets the scene for competition across different approaches to morphological modeling, with 3 clause-level sub-tasks: *morphological inflection, reinflection* and *analysis*, where systems are required to generate, manipulate or analyze simple sentences centered around a single content lexeme and a set of morphological features characterizing its syntactic clause. This year's tasks covered eight typologically distinct languages: English, French, German, Hebrew, Russian, Spanish, Swahili and Turkish. The tasks has received submissions of four systems from three teams which were compared to two baselines implementing prominent multilingual learning methods. The results show that modern NLP models are effective in solving morphological tasks even at the clause level. However, there is still room for improvement, especially in the task of morphological analysis.

1 Introduction

Universality is an important premise of many morphological datasets and shared tasks. Recent shared tasks of SIGMORPHON have introduced the notion of comparative analysis in morphological studies by incorporating up to 100 languages (McCarthy et al., 2019; Vylomova et al., 2020) in their evaluation benchmark, by providing all of them with data that is annotated according to a single universal schema (Sylak-Glassman, 2016). Systems that succeed in these tasks ideally should boast in their ability to handle various morphological phenomena observed in almost any language family on earth (Peters and Martins, 2020, inter alia).

However, as pointed out recently by Goldman and Tsarfaty (2022), the perceived universality of morphological tasks is impaired by the lack of a working definition of *a morphosyntactic word* (Haspelmath, 2011). Without such definition, the boundary between morphology and syntax is blurred and the assignment of linguistic phenomena to either morphological or syntactic data results in inconsistency across languages. Thus, limiting the scope of morphological tasks to white-spaced words creates an undue advantage to some languages based on their grammarian traditions, typological characteristics, and some other arbitrary factors.

For example, some languages, like English, are considered isolating and have word-level inflection tables of tiny size, while other languages, like Turkish, are considered agglutinative and have huge inflection tables. However, isolating and agglutinative languages largely differ orthographically rather than linguistically, as both types concatenate pieces of text. The universal benchmark presented here allows testing of both models and theories while ignoring orthographic characteristics like white-spaces and treats equally languages with varying typological characteristics.

In this shared task, we operationalize a more universally applicable and comprehensive approach to morphology by liberating the evaluated tasks from the ill-defined formal restrictions dictated by white-spaces. We start with a fix universal set of inflectional features[1] and inflect lemmas in all languages to all possible combinations of features, disregarding the number of white-spaced words required to express them orthographically. The features define fully-saturated clauses and the result is a data set of clauses organized in inflection tables and tasks that go beyond the word-level and include phenomena considered syntactic, such as word order manipulation and the like. We can thus test the submitted systems' ability to cope with these phenomena.

[1]The set of features used in constructing our data is detailed in Appendix A.

The shared task includes 3 sub-tasks: *inflection*, where systems are to generate simple clauses from a lemma and a set of morphological features; *reinflection*, where systems should manipulate a source clause to a target clause; and *analysis*, where the task is to output a lemma and a set of features given a clause. See Table 1 for example annotations. Together, these tasks examine every aspect of the ability of systems to deal with clause-level morphological constructions, moving from abstract representation to concrete text and back.

All three sub-tasks include evaluation data annotated in the following eight languages: English, French, German, Hebrew, Russian, Spanish, Swahili and Turkish, from four different language families. The variety of languages induces a plethora of alternations to be modeled by the submitted systems, from pronoun incorporation in Swahili to verb-splitting German, from ablaut-extensive Semitic morphology to highly agglutinative Turkish. However, in terms of dimensions of meaning, the data is extremely uniform across languages as all morphological features are implemented in our data if they are implemented in the language.

The results included in this shared task compare 4 submitted systems and 2 baseline systems with various characteristics, from rule-based systems to systems based on a large pretrained language models. The best performing system outperforms the best baseline and reduces the error rates by 3 to 8 fold, depending on the sub-task. Future editions of the task are intended to further expand the number of languages and the scope of the data for better alignment with real-world phenomena and distributions.

2 The Tasks

This shared task consists of three sub-tasks which test the ability of systems to deal with clause-level morphological data in multiple languages. In this Section we define and formalize the tasks, and in Table 1 we illustrate all three sub-tasks with concrete examples.

2.1 Tasks Description and formulation

Let l be a lemma, b be a feature bundle, and f an inflected form. Crucially, f may include zero or more white-space word delimiters. The *inflection* sub-task accepts a set of clause-level features and a verbal lemma as input $\langle l, b \rangle$, and requires the system to generate the desired output clause $\langle f \rangle$ that manifests these this lemma and inflectional features.

In the *reinflection* sub-task, each input item contains an example inflected form in a language accompanied by a set of morphological features that it realizes as well as a second set of features $\langle f_1, b_1, b_2 \rangle$. The system is required to generate the the respective form $\langle f_2 \rangle$ realizing this new set of features for the same lemma. It should do so without direct evidence of the lemma behind both forms.

Finally, the *analysis* sub-task evaluates the system performance in the opposite transformation of the inflection sub-task. That is, given a clause form $\langle f \rangle$ as input, the system needs to output its lemma and set of features being realized in this form $\langle l, b \rangle$.

The collection of the three sub-tasks aims to extensively assess the ability of a system to analyze and generate clause-level morphological data.

2.2 Evaluation

For all the tasks we provide the exact match accuracy between the predictions and the desired outputs.

However, systems' performance was ranked by another metric, varied by sub-task, that is more permissive and quantifies partial success. For the *inflection* and *reinflection* tasks we use an averaged edit distance between predictions and gold answers, a measure well-used in morphology to assess how close the predictions are to the ground truth on average (Cotterell et al., 2017, 2018, inter alia).

For the *analysis* task we used an F1 measure that takes into account the unordered nature of the outputs in this task. For each example we calculate the precision and recall of features in the prediction compared to the desired output, we then average the per-example F1 score over an entire set of examples.[2]

3 The Languages

Our selection of languages is diverse both typologically and genealogically. Most of the languages are Indo-European (English, French, German, Spanish and Russian), but we include languages from the Afro-Asiatic (Hebrew), Turkic (Turkish) and Atlantic-Congo (Swahili) families as well.

[2]For calculating this metric the lemma was treated as another feature but up-weighted and given an importance equal to 3 features.

Task	Model input		Reference output	
Inflection	take	IND;FUT;NOM(1,SG);ACC(3,SG,MASC)	I'll take him	
Reinflection	I'll take him	IND;FUT;NOM(1,SG);ACC(3,SG,MASC)		
		IND;PRS;NOM(1,PL);ACC(2);NEG	we don't take you	
Analysis	I'll take him		take	IND;FUT;NOM(1,SG);ACC(3,SG,MASC)

Table 1: Examples for the data format used for the evaluation of three sub-tasks. The inflection sub-task takes a lexeme and a set of tags in the given language and the model is required to produce the corresponding form. In the reinflection sub-task an inflected form accompanied by the sets of new features are input to the model, and a new form corresponding to the desired reinflection is produced. The analysis sub-task requires the model to discover the root and morphological features in a given sentence. In our annotations we use the Unimorph schema (Sylak-Glassman, 2016).

Language	Family	ISO 639-2	Annotators
English	Indo-European	eng	Omer Goldman
French	Indo-European	fra	Benjamin Muller, Djame Seddah & Benoît Sagot
German	Indo-European	deu	Omer Goldman
Hebrew	Afro-Asiatic	heb	Omer Goldman
Russian	Indo-European	rus	Victoria Basmov
Spanish	Indo-European	spa	Victoria Basmov
Swahili	Atlantic–Congo	swa	Omer Goldman, Shadrak Kirimi & Lydia Nishimwe
Turkish	Turkic	tur	Omer Goldman & Duygu Ataman

Table 2: The languages included in the benchmark.

The languages in our data exemplify almost any morpho-syntactic process that systems have to deal with in order to excel in clause-level morphological data. We have the pronoun incorporating Swahili, in which many clauses are expressed by a single word, and we have the isolating English, that makes an extensive use of multiple auxiliaries. Many of our languages concatenate words or morphemes in order to construct forms, but non-concatenative processes are also widely represented. For example, word/morpheme order is extensively used in German, especially with its infamous separable verb prefixes, and ablauts are used in inflecting almost any form in Hebrew due to its Semitic inflectional system. We have fusional languages, such as French, Russian and Spanish, in which a single morpheme corresponds to multiple features, and agglutinative languages like Turkish, in which the mapping is more one-to-one. The languages also vary in the prominence of phonological processes in them. Turkish provides an example for a language with high degree of morpho-phonological stem-affix interaction, expressed in vowel harmony, while French exemplifies post-lexical phonological processes that have effects beyond word boundaries, and in Swahili phonological interaction between inflectional morphemes is extremely rear.

Appendix B contains some additional linguistic characterization of the languages.

The diversity in the languages included in our data forces models to be flexible and powerful enough to be able to deal with all the different strategies chosen by speakers to construct inflected forms. Thus, a model that is successful on our selection is likely to succeed if supplied with data in other languages as well.

4 The Data

The data included in this task is based on the MIGHTYMORPH data set presented by Goldman and Tsarfaty (2022). The data for four of the languages was prepared in prior work, and in this shared task we have doubled the number of languages to include eight languages in total from four language families.

For most languages the data was created by expanding the UniMorph (Batsuren et al., 2022) word-level inflection tables into respective clauses that saturate all the required arguments of the verbal lemma.

This was done in two phases. Initially, we used a language-specific rule-based grammar that included the inflection tables of any relevant auxiliaries in order to construct all possible periphrastic constructions of the inflected verb. For example, when constructing the future perfect form for the English verb *receive*, equivalent to the features IND;FUT;PRF, we used the past participle from the UniMorph inflection table *received* and the auxiliaries *will* and *have* to construct *will have received*.

lexeme=LOVE	IND		IND;PERF		COND	
PRS;DECL;NOM(2,SG)	POS	NEG	POS	NEG	POS	NEG
ACC(1,SG)	you love me	you don't love me	you have loved me	you haven't loved me	you would love me	you wouldn't love me
ACC(1,PL)	you love us	you don't love us	you have loved us	you haven't loved us	you would love us	you wouldn't love us
ACC(2,SG,RFLX)	you love yourself	you don't love yourself	you have loved yourself	you haven't loved yourself	you would love yourself	you wouldn't love yourself
ACC(3,SG)	you love him	you don't love him	you have loved him	you haven't loved him	you would love him	you wouldn't love him
ACC(3,PL)	you love them	you don't love them	you have loved them	you haven't loved them	you would love them	you wouldn't love them

Table 3: A fraction of a clause-level inflection table in English.

We then manually determined which arguments each verb can take in order to generate a fully-saturated clause. To retain the tasks with a single lemma, all arguments are realized as pronominal features. For example, the English verb *receive* has 2 possible argument combinations: {NOM, ACC} and {NOM, ACC, ABL}, equivalent to sentences like "I received it" and "I received it from you", respectively. For each argument combination we exhaustively generated all suitably cased pronouns without regarding the semantic plausibility of the resulted clause.

Turkish and Swahili are somewhat exceptional to the process described above in the sense that the clause-level tables were constructed solely by grammars of morphemes without relying on the UniMorph word-level tables.

In addition to using UniMorph we generated the French data based on the Lefff (Sagot, 2010), which is a large-coverage and freely available morphological and syntactic lexicon for French. In contrast with the other languages, the types of arguments and their combinations for each verb was not determined manually but automatically with the Lefff. The auxiliary allowed for each verb was also decided using the Lefff.

The result is a fully-populated clause-level inflection table, where each entry in the table is structured as $(lemma, features, form)$. See Table 3 for a fraction of an English inflection table, and Appendix A for a glossary of all features used in our data. In this shared task we limited generation of example sentences to ones composed of a single main clause with a verbal head.

4.1 Sampling and Splitting

To prepare splits for the tasks we sampled 500 inflection tables per language. From the tables we sample 12,000 examples per task. For inflection and analysis, every example is one entry in the inflection table with the input being the $lemma$ and the $features$ and the $form$ constituting the output, or the other way around. The examples for the reinflection task are composed of two entries in the inflection table without use of the shared lemma, such that the input is $features1, form1, features2$ and the output is $form2$.

The data is split such that lemmas do not overlap between splits, thus the train set contains 10,000 examples from 400 lemmas and the test and dev sets each include 1,000 examples from 100 lemmas.[3]

5 Systems

5.1 Baseline Systems

We provide two baselines for the share task: (a) A text-to-text transformer (Raffel et al., 2020) that is trained using our training data; (b) A model based on the already pretrained mT5 model (Xue et al., 2021), fine-tuned using our training data. In both cases we train a separate monolingual model for each language. More details for each baseline are listed below.

We use the same format as provided in the training data. The morphological features are added to the vocabulary as special tokens, randomly initialized, and trained with the rest of the parameters of the models. When the input or output are separated into two parts (e.g. lemma/features), we use a separator token. Finally, we use 50 epochs across models with a learning rate of $5e-5$, and take the final checkpoint as the final trained model.[4]

The dimensions of the models were selected via hyper-parameter tuning.

[3]All data is available at https://github.com/omagolda/MRL_shared-task_2022.

[4]The scripts used to build and train the baseline models are available at https://github.com/omagolda/MRL_shared-task_2022.

Transformer Baseline We experiment with 6 configurations of different sizes, tuning on the development sets of English and Hebrew. According to the tuning process, we choose a transformer with a single-layer encoder and a single-layer decoder, with 3 self-attention heads, and with 128 as the dimension of the self-attention layers, and 256 as the dimension of the feed-forward layers.

mT5 Baseline We experiment with the base and large architectures, tune them on the development sets of English and Hebrew, and choose to use the large model. As mentioned above, we use the pre-trained model and only fine-tune it with our data.

5.2 Submitted Systems

UBC Jaidi et al. (2022) submitted a transformer-based system with four attention heads over four encoder and decoder layers. The innovation of their system is the introduction of byte-pair encoding (BPE) (Sennrich et al., 2016) to morphological tasks in order to shorten the lengths of sequences. In addition they augmented the data to bias the model more strongly towards copying and found that it helps to improve results only for the inflection and analysis tasks.

KUIS-AI Acikgoz et al. (2022) sent multiple systems:

- **KUIS-AI-1** is a transformer with four encoder and four decoder layers. Data perturbation using hallucinated data (Anastasopoulos and Neubig, 2019) was optionally added to the training set to support system capacity, with varied amount depending on the development set performance. This system participated only in the inflection and reinflection tasks.
- **KUIS-AI-2** is based on the pre-trained mGPT by Wolf et al. (2019) with additional prefix of fine-tuned vectors. This system participated only in the analysis task.

Göttingen Dönicke (2022)'s system is a rule based system that participated only in the analysis task. The system uses rules to map word-level features that are themselves either from UniMorph or from SpaCy[5] model trained over the Universal Dependencies data set (De Marneffe et al., 2021).

6 Results

Tables 4, 5 and 6 summarize the results per system for the inflection, reinflection and analysis tasks, respectively, while averaging over all languages in our selection. Results broken down by language can be found in Appendix C.

Team	ED	EM
KUIS-AI-1	**0.292**	0.919
UBC	0.496	0.855
Base-mT5	2.577	0.530
Base-transformer	3.278	0.392

Table 4: Results for task 1: inflection, for all submitted and baseline systems, averaged across languages. Edit distance (ED) is the main evaluation metric, and Exact match accuracy (EM) is given for reference.

Team	ED	EM
KUIS-AI-1	**0.705**	0.747
UBC	0.983	0.670
Base-mT5	2.826	0.481
Base-transformer	4.642	0.156

Table 5: Results for task 2: reinflection, for all submitted and baseline systems, averaged across languages. Edit distance (ED) is the main evaluation metric, and Exact match accuracy (EM) is given for reference.

All systems significantly outperformed the fine-tuned mT5 which is the strongest baseline. The systems submitted by the KUIS-AI team rank first in all tasks, both in terms of the main evaluation metric used in each task(edit distance or F1) and in terms of the exact match accuracy.

Comparing the performance of all systems over all tasks in terms of exact match accuracy, it is clearly shown that inflection is the easier task of the three. Since reinflection can be conceptually and practically decomposed to an analysis followed by an inflection operation, one can hypothesize that the under-performance in this task stems from the difficulty of the analysis operation.

Figures 1 and 2 average the performance over all systems to gain some insights into the relative difficulty of the tasks in the various languages. The trends in the different tasks point to different languages as being more or less difficult. For example, Swahili was one the toughest languages in the analysis task but one of the easiest in inflection and reinflection, while the opposite is true for Russian.

Systems also tended to under-perform in vocalized Hebrew in both inflection and reinflection, pointing to the complexity of the Semitic inflectional system. However, in the analysis task, performance over vocalized Hebrew was actually better

[5] https://spacy.io/

Team	F1	EM
KUIS-AI-2	**0.950**	0.778
Göttingen	0.940	0.658
UBC	0.914	0.680
Base-mT5	0.845	0.368
Base-transformer	0.800	0.278

Table 6: Results for task 3: analysis, for all submitted and baseline systems, averaged across languages. Weighted F1 is the main evaluation metric, and Exact match accuracy (EM) is given for reference.

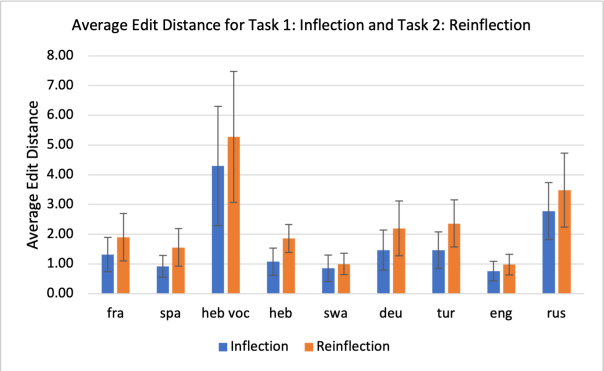

Figure 1: Average Levensthein edit distance for tasks *inflection* and *reinflection* by languages. Error bars are one standard deviation, n=4 for *inflection*, n=4 for *reinflection*

than that over the unvocalized version, probably due to the ease of disambiguation when vowels are written.

Interestingly, the inclusion of the Swahili language drove down the overall result of the Göttingen system in the analysis task, potentially depriving it from leading the table. This points both to the importance of inclusion of low-resourced languages in multilingual tasks, and also to the limits of rule-based systems that may be dependent on the knowledge of their designers of the languages at hand.

7 Conclusion and Future Directions

The first shared task on Multi-lingual Clause-level Morphology proposed novel means for modeling the evaluation of morphosyntactic representations in a more universally inclusive setting. The multilingual and typologically diverse nature of the data used in the construction of the benchmark allows its usage in comparative studies from different fields and schools. Apart from including more languages, future shared tasks should take into account the overall good, even if not perfect, performance of

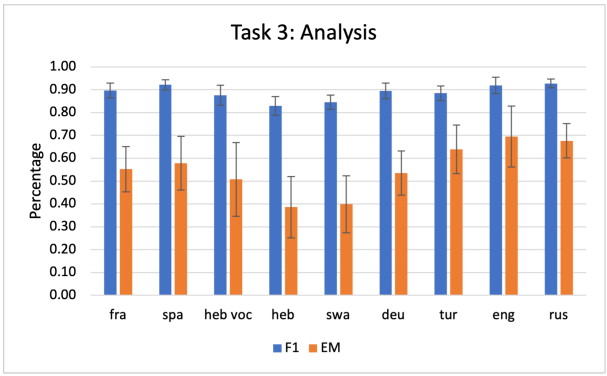

Figure 2: Average F1 and exact match accuracy for task *analysis* by languages. Error bars are one standard deviation, n=5

the systems and try to tease apart the characteristics that make morphological tasks easier in order to figure out whether they are justified.

An example for that may be the invariability in morphological data, compared to other NLP tasks such as translation. Forster et al. (2021) pointed to the remarkably different behavior of models in decoding language from morphological data, specifically to the sufficiency of greedy decoding. This is not surprising due to the conceptualization of morphological data as containing a single inflected form for every bundle of inflectional features. However, on the clause-level such one-to-one mapping is less justified, as speakers can vary the word order of a sentence or the grammatical construction chosen to pronounce the same meaning. Hence, future shared tasks could allow multiple realizations of feature bundles, making the decoding more complicated.

Semantic plausibility is another factor that was largely ignored in creating the data for this shared task. This path was chosen in order to test the systems' ability to recreate the human grammar that is well able to produce implausible sentences. However, different settings can take this factor into account so systems will not be punished for failures to predict sentences are not used in practice.

Finally, while this task included only clauses with verbal head, future tasks may include nominal and adjectival clauses as well. However, different languages use different means to express tenses in this kind of clauses, so this requires a careful linguistic treatment of copulas in comparison to (partially) zero-copula languages like Turkish and Hebrew.

To conclude, the shared task showed that modern

NLP models, whether relying on pretrained models or not, are capable of solving clause-level morphological tasks to a large extent. Still, there is room for improvement, both in the systems' ability to analyze data and in terms of the data included in these tasks.

References

Emre Can Acikgoz, Tilek Chubakov, Müge Kural, Gözde Gül Sahin, and Deniz Yuret. 2022. Transformers on multilingual clause-level morphology. In *Proceedings of the 2nd Workshop on Multilingual Representation Learning*, Abu Dhabi. Association for Computational Linguistics.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

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 Karahóǧa, 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.

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.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Tillmann Dönicke. 2022. Rule-based clause-level morphology for multiple languages. In *Proceedings of the 2nd Workshop on Multilingual Representation Learning*, Abu Dhabi. Association for Computational Linguistics.

Martina Forster, Clara Meister, and Ryan Cotterell. 2021. Searching for search errors in neural morphological inflection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1388–1394, Online. Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2022. Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10.

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Badr Jaidi, Utkarsh Saboo, Xihan Wu, Garrett Nicolai, and Miikka Silfverberg. 2022. Impact of sequence length and copying on clause-level inflection. In *Proceedings of the 2nd Workshop on Multilingual Representation Learning*, Abu Dhabi. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J.

Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Ben Peters and André F. T. Martins. 2020. One-size-fits-all multilingual models. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 63–69, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Benoît Sagot. 2010. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

A Features Glossary

Table 7 enumerates all features used in our data together with their meaning. Most features are taken from UniMorph annotation guidelines (Sylak-Glassman, 2016), with accidental gaps filled with new features. Some language specific features (LGSPEC1, LGSPEC2, etc.) were used to distinguish different constructions with the same meaning.

B The Languages' Linguistic Characteristics

English is the most widely spoken language, if counting L2 speakers, according to Ethnologue,[6] and by far the language that enjoys the most attention in the NLP literature. Morphologically, it is considered mostly an isolating language, with tense, aspect and mood being regularly expressed using white-space-separated auxiliaries. As a Germanic language, its verbs are classified into *weak verbs* that use a morpheme to form the past tense form and the past participle, and *strong verbs* that form the same forms with an ablaut in the stem.

English's word order is usually SVO, although some remnants of the Germanic V2 order do exist. Word order is used to form yes-no questions, with the auxiliary or a supporting *do* appearing in the beginning of the sentence. A supporting *do* is also added to negated sentences with no auxiliary. An array of phonological post-lexical contractions are also optionally used and affect almost all auxiliaries and the negation clitic *n't*.

French is a Romance language in the Indo-European language family. Influenced by Germanic and Celtic languages, it has evolved more drastically from Latin than other romance languages like Spanish and Portuguese. For instance, French requires the use of the subject pronouns, hence it is classified as a non-pro-drop language. It has four main moods, and about 21 distinct tenses which can be simple or compounded with one of the auxiliaries, *être* and *avoir*. French has 3 persons and 2 numbers. French's basic word order is SVO language, but it can be altered for grammatical reasons. For instance, interrogative form are typically constructed by inverting the subject and the verb. Additionally, when taking a pronominal form, the object is inverted with the verb leading to a SOV order (e.g. *je te dis*, literally *I you tell*).

[6] https://www.ethnologue.com/language/eng

French contains multiple phonetic contractions. For instance, the negative particle *ne* becomes *n'* when followed by a vowel. Similar phenomenon is also applied to the 1st person subject pronoun and to many object pronouns. French also contains a few phonetic-based insertions. For instance, the *-t-* in *a-t-il dit* — did he say — is added for phonetic purposes.

German Another representative of the Germanic branch of the Indo-European language family is German. It shares many characteristics with its close relative English, most prominently the concatenation of auxiliaries to express complex inflections and the division of verbs into *strong* and *weak* classes. However, it has some characteristics that are unique in our selection of languages, mostly in the realm of syntax. The German word order is V2 with the first auxiliary or verb-part appearing as the second constituent while the rest are at the end of the sentence. Some verbs also consists of a separable prefix that appear at the end of the sentence but only in some inflections, thus making German a hard language to learn for humans and machines alike. Nouns and pronouns take one of 4 possible cases, but verbs' arguments can be also introduced with a wide array of prepositions that interact with the cases to specify some fine-grained dimensions of meaning.

Hebrew As a member of the Semitic branch of the Afro-Asiatic language family, Hebrew exhibits the typical ablaut-extensive Semitic inflectional system, where lexemes are expressed via roots that are mostly tri-consonantal and an array of interwoven vowels as well as suffixes are used to inflect the verbs. Hebrew verbs belong to 7 major classes (*Binyanim*) with many subclasses depending on the phonological features of the root's consonants. Verbs inflect for number, gender, and tense-mood.

In terms of syntax, Hebrew's word order is SVO and yes-no questions are typically expressed using intonation only, although an introduction word, האם, is optionally available. Hebrew displays a partial pro-drop where non-third-person subjects are dropped in non-present tenses. Some of the prepositions used to express nominal arguments are fused prepositions, i.e., written without a white-space before the noun. But all prepositions are fused when introducing a pronoun that appears in a clitic form.

As a typical Semitic languages, Hebrew is writ-

Attribute		Value
Tense		PST(past),PRS(present),FUT(future) IMMED(immediate)
Mood		IND(indicative) IMP(imperative) SBJV(subjunctive) INFR†(inferential) NEC†(necessitative) COND(conditional) QUOT(quotative)
Aspect		HAB(habitual) PROG(progressive) PRF(perfect) PRSP(prospective) PRV(perfective) IPRV(imperfective)
Non-locative Cases		NOM(nominative) ACC(accusative) DAT(dative) GEN(genitive) INS(instrumental) COM(comitative) BEN(benefactive) PRIM(primary)† SEC(secondary)†
Locative Cases		LOC†(general locative) ABL(ablative) ALL(allative) ESS(essive) APUD(apudessive) PERL†(perlative) CIRC(near) ANTE(in front) CONTR†(against) AT(at, general vicinity) ON(on) IN(in) VON†(about) ONVR(vertical on) SUB(under) PROL(prolative) VERS(versative) TERM(terminative) INTER(among) POST(behind) REM(distal) PROXM(proximal)
Sentence Features		NEG(negative) Q(interrogative)
Argument Features	Person	1(1st person) 2(2nd person) 3(3rd person)
	Number	SG(singular) PL(plural)
	Gender	MASC(masculine) FEM(feminine) NEUT(neuter)
	Swa classes	M-WA† M-MI† JI-MA† KI-VI† N† U† KU†
	Misc.	FORM(formal) INFM(informal) RFLX†(reflexive)

Table 7: A list of all features used in constructing the data for all 8 languages. Features not taken from Sylak-Glassman (2016) are marked with †.

ten using an abjad where the vowels are sparsely marked in unvocalized text. This style of writing somewhat waters down the complexity of the Semitic morphology as the alternating vowels are largely not written. For this reason we include data in vocalized Hebrew in addition to the commonly-used unvocalized data.

Russian is an East Slavic Language. It belongs to the Balto-Slavic branch of the Indo-European language family. Russian has a rich, fusional, highly synthetic morphology, typical of most Slavic languages.

One peculiarity of the Russian verbal system is that its 2 aspects: perfective and imperfective. are assigned in the lexemic level, so each verb is either perfective or imperfective. Most verbs come in pairs (e.g. делать/сделать - to do/to have done). This system of aspects is characteristic of Slavic languages in general. In addition, verbs can be reflexive (using the reflexive suffix -ся/-сь).

In terms of inflectional morphology, Russian verbs have 3 tenses and 3 moods. Verbs agree with the subject in person and number in non-past tenses, and in gender and number in past forms. The vast majority of verb forms are synthetic, while future tense of imperfective verbs and the subjunctive are analytic and formed with auxiliaries.

Nouns and pronouns take one of the 6 possible cases, but, similarly to German, verbs' arguments can be also introduced with a wide variety of prepositions that interact with the cases to specify fine-grained relationships.

The basic word order in Russian is SVO, but since grammatical relationships are marked by inflection, a considerable freedom of word order is allowed. Changes in word order are mainly used to express logical stress. Similarly to Hebrew, yes-no questions are typically expressed using intonation only, but optionally the interrogative particle ли can be used.

Spanish is a Romance language of the Indo-European language family. It belongs to the Ibero-Romance group of languages. Most grammatical characteristics of Spanish are typical of Romance languages in general.

Spanish is a fusional language with a rich morphology. It has a very rich verb conjugation with about 50 forms per verb (not counting periphrastic forms). The Spanish verb paradigm has 16 distinct tense, aspect and mood combinations, 8 simple and 8 compound. Other verb forms include infinitive, imperative, gerund, and past participle. Each of the 16 tenses has 3 persons and 2 numbers. In both singular and plural, different persons are used for formal and informal addressees. Also, the sets of second-person verb forms can differ by dialect (i.e., voseo vs. tuteo).

Spanish nouns belong to either the masculine or the feminine gender and have 2 numbers. Nouns don't inflect by case. Instead, grammatical relations are expressed with prepositions. Personal pronouns are inflected by person, number, gender and (in a very reduced manner) by case.

The basic word order is SVO, but considerable variations are possible, so that VSO, VOS and OVS are also relatively common. Interestingly, in the

OVS order, the direct object noun is supplemented with the corresponding direct object pronoun, e.g. *La cena* ***la*** *preparo yo* (literally, *"The dinner* ***it*** *will make I"*).

A very characteristic feature of Spanish are clitics, or weak personal pronouns. They are used enclitically (after the verb) or proclitically (before the verb) depending on the verb form. Enclitic pronouns are written as part of the verb (e.g. *comprármelo* - to buy it for me). Clitics can be also attached to one another forming arrays, but these arrays obey strict ordering rules (e.g. *comprármelo* is grammatical while **comprárlome* is not).

Swahili is the only representative of the Atlantic-Congo language family in our selection and the most low-resourced language, lacking even a Universal Dependencies dataset. Being the most agglutinative in our data, Swahili inflects verbs mostly by concatenating non-interacting morphemes, although some may express several dimensions of meaning like the combined morphemes for nominative agreement and polarity. In addition, an auxiliary verb *kuwa* is also used to express some compound tense-aspect-mood combinations.

Swahili uses a secundative alignment of verbs' arguments, meaning that the direct object of monotransitive verbs is treat similarly to the indirect object of di-transitive verbs and this category is referred to as the *primary object*, while the direct object of di-transitive verbs is a separate *secondary* category. In main clauses, verbs agree with the nominative and the primary arguments, while secondary objects appear only as a separate word. In addition, prepositions and coverbs are used sparsely to introduce arguments of some verbs. Swahili is a pro-drop language, omitting pronouns to any argument that is expressed on the verb. The word order is SVO.

Turkish The other agglutinative language in our selection is Turkish, of the Oghuz branch of the Turkic languages. Characterized by Turkic vowel harmony, most morphemes have either 2 or 4 allomorphs, and they are used to express tense, mood and agreement with the nominative argument as well compoundable aspects and dimensions of meaning that are usually considered syntactic in other languages, like morphemes for subordination and conjunction. Some tense-aspect-mood combinations require the usage of the auxiliary *olmak*. Yes-no questions are formed using the *mi* particle that takes the nominative agreement instead of the verbs in many inflections.

Turkish typical word order is SOV and nouns take one of 6-7 cases. They can also be introduced by postpositions, mostly the beneficiary *için*.

C Detailed Results

Tables 8, 9 and 10 provide the full results for all systems and languages for the inflection, reinflection and analysis tasks, respectively.

Tables 11, 12 and 13 show the results per language, averaged over the systems.

Team	Metric	fra	spa	heb$_{voc}$	heb	swa	deu	tur	eng	rus
KUIS-AI-1	ED	0.124	0.199	0.550	0.113	0.019	0.241	0.333	0.221	0.828
	EM	0.932	0.920	0.898	0.942	0.996	0.918	0.898	0.889	0.877
UBC	ED	0.276	0.210	0.724	0.347	0.103	0.630	0.281	0.339	1.558
	EM	0.864	0.883	0.846	0.852	0.918	0.771	0.914	0.803	0.847
Base-transformer	ED	2.839	1.803	5.671	2.390	2.202	3.705	3.187	1.874	5.834
	EM	0.485	0.516	0.252	0.496	0.262	0.191	0.429	0.508	0.389
Base-mT5	ED	2.032	1.467	10.240	1.472	1.093	1.303	2.074	0.619	2.889
	EM	0.449	0.587	0.258	0.395	0.524	0.673	0.517	0.794	0.574

Table 8: Detailed results for task 1: inflection, for all submitted and baseline systems, both in terms of edit distance and exact match accuracy.

Team	Metric	fra	spa	heb$_{voc}$	heb	swa	deu	tur	eng	rus
KUIS-AI-1	ED	0.758	0.480	0.796	1.002	0.182	0.788	1.011	0.477	0.854
	EM	0.683	0.776	0.833	0.577	0.845	0.665	0.774	0.723	0.849
UBC	ED	0.641	0.593	1.072	1.093	0.471	1.430	0.781	0.648	2.114
	EM	0.693	0.757	0.792	0.536	0.701	0.476	0.762	0.611	0.704
Base-transformer	ED	4.584	3.628	8.531	3.347	2.004	5.360	4.653	2.170	7.502
	EM	0.197	0.163	0.043	0.050	0.211	0.044	0.197	0.288	0.213
Base-mT5	ED	1.595	1.531	10.686	1.993	1.343	1.198	3.005	0.614	3.468
	EM	0.539	0.566	0.243	0.239	0.465	0.675	0.320	0.788	0.497

Table 9: Detailed results for task 2: reinflection, for all submitted and baseline systems, both in terms of edit distance and exact match accuracy.

Team	Metric	fra	spa	heb$_{voc}$	heb	swa	deu	tur	eng	rus
KUIS-AI-2	F1	0.956	0.981	0.928	0.821	0.905	0.959	0.954	0.996	0.975
	EM	0.819	0.894	0.735	0.362	0.626	0.834	0.847	0.985	0.886
UBC	F1	0.892	0.940	0.949	0.863	0.936	0.891	0.925	0.878	0.955
	EM	0.597	0.727	0.820	0.513	0.743	0.594	0.768	0.552	0.810
Göttingen	F1	0.977	0.943	0.955	0.965	0.789	0.974	0.929	0.993	0.931
	EM	0.693	0.637	0.748	0.827	0.067	0.550	0.816	0.974	0.609
Base-transformer	F1	0.799	0.874	0.735	0.744	0.808	0.779	0.796	0.804	0.866
	EM	0.291	0.407	0.050	0.098	0.300	0.238	0.365	0.282	0.474
Base-mT5	F1	0.855	0.868	0.814	0.754	0.789	0.872	0.822	0.923	0.908
	EM	0.363	0.229	0.183	0.130	0.258	0.458	0.400	0.683	0.604

Table 10: Detailed results for task 3: analysis, for all submitted and baseline systems, both in terms of weighted F1 and exact match accuracy.

Language	Exact Match	Edit Dist.	F1
fra	0.683	1.318	0.926
spa	0.727	0.920	0.958
heb	0.564	4.296	0.879
heb_{voc}	0.671	1.081	0.900
swa	0.675	0.855	0.812
deu	0.638	1.470	0.917
tur	0.690	1.469	0.884
eng	0.749	0.763	0.955
rus	0.672	2.777	0.931
all lang.	0.674	1.661	0.907

Table 11: Results per language for the *inflection* sub-task. Edit distance is the most important metric, as it quantifies the difference between the correct and predicted clause. n=4

Language	Exact Match	Edit Dist.	F1
fra	0.528	1.895	0.889
spa	0.566	1.558	0.931
heb	0.351	1.859	0.791
heb_{voc}	0.478	5.271	0.838
swa	0.555	1.000	0.757
deu	0.465	2.194	0.874
tur	0.513	2.363	0.808
eng	0.603	0.977	0.934
rus	0.566	3.485	0.910
all lang.	0.514	2.289	0.859

Table 12: Results per language for the *reinflection* sub-task. Edit distance is the most important metric for this task, as it quantifies the difference between the correct and predicted clause. n=4

Language	Exact Match	Edit Dist.	F1
fra	0.553	2.111	0.896
spa	0.579	3.112	0.921
heb	0.507	3.802	0.876
heb_{voc}	0.386	2.088	0.829
swa	0.399	5.799	0.845
deu	0.535	2.311	0.895
tur	0.639	2.069	0.885
eng	0.695	0.699	0.919
rus	0.677	2.568	0.927
all lang.	0.552	2.729	0.888

Table 13: Results per language for the *analysis* sub-task. Accuracy quantifies the amount of perfectly predicted features (lemma and morphological structure). F1-score considers each morphological feature and sub-feature equally, but assigns the lemma more importance (by assigning the lemma feature weight three). n=5

.