Well-Defined Morphology is Sentence-Level Morphology

Omer Goldman
Bar Ilan University
`omer.goldman@gmail.com`

Reut Tsarfaty
Bar Ilan University
`reut.tsarfaty@biu.ac.il`

Motivation and Background Morphological tasks have gained decent popularity in NLP in recent years, with many shared tasks in various flavours (Cotterell et al., 2016; McCarthy et al., 2019; Kann et al., 2020; inter alia), supported mainly by UniMorph (McCarthy et al., 2020) a large multi-lingual dataset providing morphological analysis of standalone words in the form of inflection tables. The computational task in these settings relies heavily on the notion of a '*word*' as these are the elements analyzed and inflected, while the relation between words is left for semantic operations. However, a detailed examination of the data reveals that no pre-determined criterion exists for *what constitutes a word* for the purpose of morphological inflection, and it is set in a per-language fashion relying mostly on speaker intuition.

Specifically, the decision of what is a *word* is clearly *not* based on delimitation of whitespaces. For example, the Finnish phrase *olen ajatellut* is considered a word and is annotated as V;ACT;PRS;PRF;POS;IND;1;SG, and likewise the Albanian phrase *do të mendosh* as V;1;PL;IND;FUT. In contrast, the English equivalents *have thought* and *will think* respectively (corresponding to the exact same features-bundles and meanings) are absent, and their construction is considered syntactic.

Moreover, there is no cross-linguistically consistent set of features and phenomena covered by UniMorph. Thus, *negation* is included in Latvian inflection tables, but not in German ones, while *interrogativity* is included in the Turkish dataset but not in Arabic. In fact, some features of meaning clearly expressed at word-level are absent from UniMorph altogether, most notably for languages that express object concords on their verbs, as Georgian and Bantu languages.UniMorph is skewed, including only forms that *lack* these morphemes.

This poses a problem for the development of truly multi-lingual morphological models since the models are examined on different dimensions of meaning and different sets of features. The current method of data construction induces a bias related to typological and orthographic characteristics of the languages included, so English is considered an isolating language with tiny inflection tables of size 5, while Turkish is considered an agglutinative language with inflection tables of hundreds of forms, although both languages exhibit a complex system of tense and aspect pronounced using linearly separable morphemes with the main difference that the English morphemes are separated by white-spaces.

The quagmire surrounding words and their demarcation is far from being unique to morphological NLP. In fact, the linguistic literature points to no coherent definition of word whatsoever (Zwicky and Pullum, 1983; Lieber and Scalise, 2006, *inter alia*) and the stance that no such cross-lingual definition even exists is also heard (Haspelmath, 2011). We suggest to bypass this linguistic debate while providing true universality to modelling of morphological tasks.

Proposal In order to make morphological modelling truly universal and define morphological tasks with no unduly advantage provided to white-space intensive languages, we propose to shift the focus of annotation from *words* to *features*. I.e., we propose to fix the set of inflectional features in all inflectional morphology tasks. Models will be required to inflect lemmas or forms to any bundle legal in a language, regardless of that language's expression of the features – be it in one word, a periphrastic construction or even by syntactic means as word order. The features included will be all those that are expressed clearly as an inflectional morpheme in *some* language. This will set all languages on equal footing. The current version of our annotation scheme includes all the aforementioned features of meaning: tense-aspect-mood (TAM), negation, interrogativity and pronominal arguments. Thus, for example, models will be required

	ENG		DEU		HEB		HEB$_{vocalized}$		TUR	
	word	sentence	word	sentence	word	sentence	word	sentence	word	sentence
LSTM	0.56	0.20	0.14	0.00	0.33	0.14	0.53	**0.33**	0.47	0.26
TRANSDUCE	0.88	**0.73**	0.78	0.34	0.80	**0.39**	0.47	0.01	0.80	**0.67**
MT5	NA	0.63	NA	**0.42**	NA	0.21	NA	NA	NA	0.32

Table 1: Word and sentence inflection results for all languages and systems. For every language the best performing system is marked in **bold**. Note that both word inflection models perform better in word inflection, confirming that sentence inflection is indeed a harder task.

	ENG	DEU	HEB	TUR
UniMorph table size	5	29	29	702
intransitive table size	450	512	132	702
# inflection tables	486	200	779	200

Table 2: Some statistics over the proposed MIGHTY-MORPH data. The inclusion of periphrastic constructions expands the size of the inflection tables comparing to UniMorph even without additional arguments.

to inflect the lemma equivalent to LOVE into the features IND;PRS;NEG;INTR;SUBJ(2;SG);OBJ(1;SG) in all languages to resulting in Swahili *hunipendi?*, Turkish *beni sevmez misin?* and English *Don't you love me?*, regardless of the number of white-spaces.

Task and Data The specific morphological task in this paper is modelled after morphological reinflection (Cotterell et al., 2016) where given a word-form of a certain lemma and it's feature bundle models are asked to realize the form equivalent to another bundle of features. As mentioned, we consider simple sentences as forms, rather than words, to allow inflection for all included features in all languages. To keep sentences single-lemmaed verbs' arguments are filled in as pronouns.

We annotated a dataset for this task, MIGHTY-MORPH, covering 4 languages: English, German, Turkish and Hebrew.[1] For each language we sampled verbal lemmas from UniMorph, giving priority to frequently used verbs, and for each lemma we exhaustively generated a full table of all simple sentences with their respective features. For every lemma, we manually determined all possible arguments using a monolingual dictionary, including both cased arguments and arguments licensed by an adposition, where arguments' features include person, number, gender and reflexivity. Some statistics of the generated data are provided in Table 2.

The data for our sentence reinflection task was sampled from MIGHTYMORPH. For each language, we sampled 10, 000 pairs of sentences with their morphological features from 200 inflection tables, such that each pair shares a lemma. We split the data 90% for the train set and the rest are a test set. The lemmas in the train- and test-set are disjoint (Goldman et al., 2021).

Models As an initial attempt at sentence inflection we apply SOTA models for word-inflection to both the word- and sentence-level inflection tasks:
• LSTM: by Silfverberg and Hulden (2018)
• TRANSDUCE: by Makarov and Clematide (2018)
Both models handle characters as input and output, treating white-space as yet another character rather than a special word-delimiter. In addition, moving to sentence inflection allows the use of contextualized pretrained language models. We used the MT5 (Xue et al., 2021) finetuned on each language separately. The morphological features were added to the input as new tokens with randomly initialized embeddings, and the rest of the input and output was tokenized using MT5's own tokenizer.

Results Table 1 compares the results for word- and sentence-inflection for the LSTM and TRANSDUCE models. It is clear that sentence inflection is a harder task. It is not surprising as it involves longer character sequences, and more sophisticated edits such as manipulating word order (e.g. S-V inversions English and German). On sentence inflection, the TRANSDUCE performs better on average, although variation across languages does occur.

Conclusion We suggest a paradigm shift for morphological NLP, from inflection table defined by the ill-defined concept of *word* to tables defined by a cross linguistically fixed set of *features*. We argue that this shift requires a mode from words to simple sentences. We believe our framework brings better universality to morphology-related NLP tasks and we formulated a sentence-reinflection task accompanied by suitable data. We showed that while this task is significantly harder, it also provides an opportunity to interface with contextualized LMs, thus it is a thread of research worth developing.

---

[1]For Hebrew we have 2 versions: vocalized/unvocalized.

References

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.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2021. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance.

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

Katharina Kann, Arya D. McCarthy, Garrett Nicolai, and Mans Hulden. 2020. The SIGMORPHON 2020 shared task on unsupervised morphological paradigm completion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 51–62, Online. Association for Computational Linguistics.

Rochelle Lieber and Sergio Scalise. 2006. The lexical integrity hypothesis in a new theoretical universe. *Lingue e linguaggio*, (1):7–32.

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.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

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.

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, Online. Association for Computational Linguistics.

Arnold M Zwicky and Geoffrey K Pullum. 1983. Cliticization vs. inflection: English n't. *Language*, 59(3):502–513.