Does Topological Ordering of Morphological Segments Reduce Morphological Modeling Complexity? A Preliminary Study on 13 Languages

Andreas Shcherbakov
The University of Melbourne
scherbakov.andreas@unimelb.edu.au

Kat Vylomova
The University of Melbourne
vylomovae@unimelb.edu.au

Abstract

Generalization to novel forms and feature combinations is the key to efficient learning. Recently, Goldman et al. (2022) demonstrated that contemporary neural approaches to morphological inflection still struggle to generalize to unseen words and feature combinations, even in agglutinative languages. In this paper, we argue that the use of morphological segmentation in inflection modeling allows decomposing the problem into sub-problems of substantially smaller search space. We suggest that morphological segments may be globally topologically sorted according to their grammatical categories within a given language. Our experiments demonstrate that such segmentation provides all the necessary information for better generalization, especially in agglutinative languages.

1 Introduction

Generalization is a form of abstraction where common patterns, or properties, that are observed across specific instances are then extended to a wider class of instances. This form of deductive inference allows humans to learn language more efficiently, form sophisticated concepts, and introduce semantic relations such as hypernymy. Still, computer systems are considered to be less successful in making generalizations from data (Lake and Baroni, 2018). Morphological inflection task is a popular playground to compare and evaluate systems' ability to generalize. The morphological inflection task is a type of language modelling that focuses on producing inflected forms from a given dictionary form (a *lemma*) and a set of morphosyntactic features (a *tagset*) that describes the word form to be produced, as in "$spider, (N; PL) \rightarrow spiders$". Table 1 provides a sample paradigm table for Czech and Turkish nouns for "cat". Annual contests on morphological inflection prediction were held since 2016, covering a variety of typologically diverse languages (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020; Pimentel et al., 2021). With the introduction of neural systems and the availability of large datasets, the task deemed to be solved with top performing systems achieving over 90% accuracy on most languages, even morphologically complex ones such as Uralic or Turkic. Most challenging cases were associated with under-resourced languages such as Chukchi or Evenki where majority of morphological paradigms were incomplete and sparse (Vylomova et al., 2020). However, a more fine-grained analysis from Pimentel et al. (2021) and Goldman et al. (2022) revealed that accuracy dropped substantially on unseen lemmas (i.e. in the condition where train, development, and test sets did not overlap lexically).

	Czech		Turkish	
Case	Singular	Plural	Singular	Plural
Nom	**kočka**	kočky	**kedi**	kediler
Gen	kočky	koček	kedinin	kedilerin
Dat	kočce	kočkám	kediye	kedilere
Acc	kočku	kočky	kediyi	kedileri
Ins	kočkou	kočkami	–	–
Ess	kočce	kočkách	kedide	kedilerde
Voc	kočko	kočky	–	–
Abl	kočko	kočky	kediden	kedilerden

Table 1: Sample paradigm tables for Czech and Turkish "cat" (its lemma form is in **bold**). The tags follow the UniMorph annotation schema (Sylak-Glassman, 2016). Turkish paradigm omits possessive and predicative forms.

This observation led to a significant reconsideration of the shared task design in 2022. The 2022 shared task (Kodner et al., 2022) focused on controlling the training, development, and test sets with respect to observed lemmas and tagsets. More specifically, the task organizers provided four conditions in which: 1) both the test lemma and tagset were observed in the training set (but separately!); 2) the test lemma was presented in the training set

but the test tagset was not included in the training set; 3) the test tagset was observed in the training set while the lemma was not; 4) (the most challenging where) both the lemma and the tagset appeared exclusively in the test set. The performance assessment and analysis were carried out separately for each of the four categories and revealed a notable lack of generalization ability in all submitted systems, the vast majority of which were neural sequence-to-sequence models. It is particularly striking that systems failed at modelling agglutinativity, the ability to compose novel combinations of morphemes that were previously observed in other combinations. Or, the opposite, deducing morphemes for a subset of a previously observed tagset. Many agglutination rules that seem to be simple to human learners, appear to be challenging when it comes to machines. This fact tells us that sequence-to-sequence models do not generalise well, and current approaches to morphology modelling should be reconsidered.

In this paper, we suggest that annotated morphological segmentation can significantly improve the generalization ability. We propose augmenting the inflection model with segmentation as an intermediate step. We aim to evaluate the claim that such task is easier to solve than the reinflection task in its classical setting, especially in agglutinative languages. We suggest that the reinflection task can be formalized as a classification task rather than a string-to-string transduction task. This approach dramatically reduces the search space during the inference phase as well as enhances the model's robustness to data sparsity.

2 The Dataset

In our experiments we used datasets for inflectional paradigms and segmentation for Catalan (`cat`), Czech (`ces`), German (`deu`), English (`eng`), Finnish (`fin`), French (`fra`), Hungarian (`hun`), Italian (`ita`), Mongolian (`mon`), Portuguese (`por`), Russian (`rus`), Spanish (`spa`), and Swedish (`swe`) provided in MorphyNet resource (Batsuren et al., 2021).

3 Learning the Order of Segments

We hypothesise that the order of morphological segments[1] *within a language* is defined by the order of their corresponding grammatical categories (such as grammatical number, person, case). For instance, Turkish nouns would first specify the number and then the case (as shown on Table 1).

In the dataset described above, each word form w^j stands for a sequence of $[(s_i, t_i)]^j$, where s_i is i-segment in word form j, t_i is a tagset describing the segment (*segmental tagset*; such as "$GEN;PL$" for fusional or "GEN" for agglunitative languages). Let us illustrate this notation by the following example from Catalan, taken from MorphyNet dataset.

ossificar	ossificaven
V|IND;PST;IPFV|3;PL	ossificar|ava|en

Here, an inflected form is expressed as a sequence of three segments: s_0 = "ossificar", s_1 = "ava" and s_2 = "en". Each segment bears its respective tagset. In such a way, a *whole* word's Unimorph tagset "$V;IND;PST;IPFV;3;PL$" associated with the word form is represented as a sequence of three segmental tagsets $t_0 \dots t_2$, where t_0 = V, t_1 = IND;PST;IPFV and t_2 = 3;PL.

As we mentioned, we suggest that segment tagsets are strictly ordered globally withing a given language. More formally, we claim that it is possible to sort all unique tag combinations $t = (t_i)_{i=0\dots i_{max}(j)}$ topologically, i.e. to associate each *unique* t with a number $ord(t^j)$ in such a way that for each w^j we have:

$$k > i \Rightarrow ord(t^j_k) > ord(t^j_i) \qquad (1)$$

To test the hypothesis, we propose the following learning algorithm. First, we initialize $ord(t^j_i) := 0$ for all segment-wise tag combinations t^j_i. Then, in each epoch, for each w^j observed in the dataset we check whether the equation 1 has already been satisfied for all i, k. If not, we add $(i - \tilde{i})$ to $ord(t^j_i)$ for each i, where $\tilde{i}$ is mean i value (half the number of segments in w^j). This way, we attempt to either learn the global segmental tagset order or disprove existence of it. We repeat the procedure until the number of forms in which segmentation was compliant to equation (1), stops to increase. A simplified pseudocode which implements such a process is given below.

```
RATE = 0.01                          ▷ A tunable hyperparameter
function FITTAGORDER(tagsets, update)
    mixed := false
    last := LEFTPAD                  ▷ A dummy tagset
    S = |tagsets|
    for i ∈ 0 ... S − 1 do
        if update then
```

[1] We will use morphological segments and morphemes interchangeably.

```
            increment L[tagsets[i]] by RATE
×(2i − S + 1), default = 0
            if L[tagsets[i]] ≤ L[last] then
                mixed = true
            last := max_{set→L[set]} (last, tagset)
        return mixed
    procedure EPOCH(samples)
        for sample ∈ samples do
            ts = segment tagsets in sample
            if FITTAGORDER(ts, false) then
                FITTAGORDER(ts, true)
                report sample as outlier
```

Indeed, we find that the global order of segmental tagsets *does* exist in all languages represented in MorphyNet. Swedish is the only language where a few (only two) exceptions were found; however, even those exceptions may be attributed to fuzziness of segment tagging rules. This result suggests that for a morphological inflection system it should be sufficient to produce a set of segments and use their global topological order to properly sort them rather than deal with segmentation order for every sample individually. Therefore, a "full scale" character-level sequence-to-sequence model can be replaced by a simpler classifier model to carry out the segmentation process. This important finding allows to reduce the model decision space without any loss in accuracy while enabling better generalization, especially in agglutinative languages (and higher robustness to training data sparsity).

4 Decomposing Tagsets

As grammatical feature combinations are often complex, one might expect that there should be numerous ways to decompose those corresponding to morphological segments, thus, making decomposition a separate complex subtask. In this section, we refute it by demonstrating the statistics on decomposition variety per distinct segmental tagset.

As both segments and their corresponding tagsets are listed for each word form in MorphyNet, it may appear that a "natural" way of segmentation modelling would look as follows. First, decompose the initial tagset into segment-wise sub-combinations and, second, map each sub-combination into a distinct morphological segment. However, as we discovered, this technique does not work well because the assignment of tag combinations to segments appears to be highly ambiguous in MorphyNet. In many cases, it is due to the tags that represent an inherent property of a lemma. These tags, therefore, are not realized as a segment (e.g., animacy in nouns). The lack of consistent rules governing tag-to-segment annotation is another source of ambiguity as it frequently leads to different tagging across similar samples. Fortunately, there is an *alternative technique* that works better. Namely, we should consider *unique combinations* of resulting moprhological segments rather than focus on the variants of tagset decomposition. Our experiments demonstrate that the number of distinct morphological segments per tagset is less than 4 for the majority of tag combinations, and only in 5% cases reaches 15 (in total, approximately 2,400 tag combinations were considered, as counted separately for each language). The stem segment of any word was replaced by a wildcard symbol matching with other word stem segments.[2] Figure 1 shows the distribution of the number of segment variants per a distinct tag combination. It is worth mentioning that more than a half of tag combinations that are realized by "15 or more segment sequences" each were Russian verb forms. The first letter of suffixes in those verbs may depend on the adjacent ending of the verb's stem. This dependency results in either copying of a stem trailing consonant or a consonant mutation. Thus, it is necessary to take adjacent letter into account in order to predict the segment correctly.

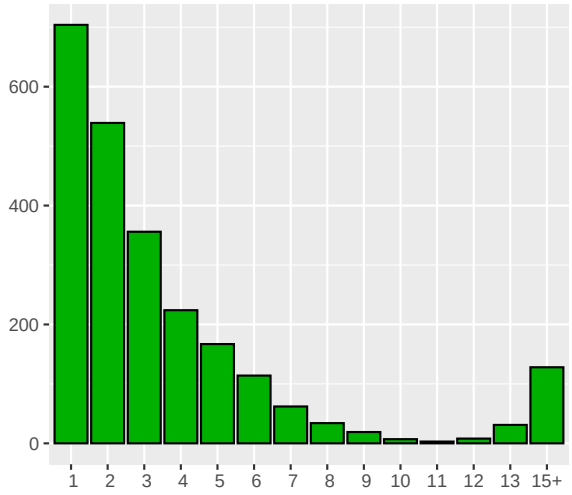

Figure 1: A frequency distribution for the number of different morphological segments per tagset. Here we consider distinct (language, tagset) pairs.

[2] To keep the setting simple, we excluded inflected forms of German compounds, in which the order of two stems was swapping.

5 Segment Composability

In Section 3 we have demonstrated that the order of segments is deterministic. Still, in the condition when the data is sparse an inflection system should be able to retrieve relevant segments from training samples, especially in agglutinative languages. Typically, the observed tagsets are different from the one that needs to be predicted. We define a *"segment composability"* measure over a segmentation dataset as a percentage of tagsets T with the following property: *The segment has ever been seen in at least two data samples, one with tagset t and one, with tagset $t' \neq t$.* While evaluating this percentage, we prune all tagsets that contain tags that only occur once, i.e. in that particular tagset (which means the tagset cannot be reconstructed from the rest of the data). A "segment composability" is a probability for a segmentation corresponding to the tagset to be reconstructed from segments observed in other tagsets, given that the predictor uses a "perfect" oracle over segments observed in a training set. The composability values measured over MorphyNet are provided in Table 2. They appear to be close to 100% for languages with high agglutinativity,[3] demonstrating a notable usability of MorphyNet segmentation datasets for the inference of unseen word forms. Here is a pseudocode explaining our approach to computation of composability.

function COMPOSABILITYRATE
  for $sample \in samples$ **do**
    $(segments, tagsets) = sample$
    $T = \{\forall tag \in \forall set \in tagsets\}$
    for $(seg, set) \in sample^{\mathsf{T}}$ **do**
      for $\tau \in set$ **do** ▷ single tags
        $uses_t[\tau] := uses_t[\tau] \cup \{T\}$
        $uses_s[seg] := uses_s[seg] \cup \{T\}$

$$combined = \left\{ \begin{array}{c} T: \\ \exists \tau : \{T\} \subset uses_t[\tau] \\ \neg\exists \tau : \{T\} = uses_t[\tau] \end{array} \right\}$$

▷ Word tag sets without exclusive tags

$$compos = \left\{ \begin{array}{c} T \in combined: \\ \neg \text{ISSTEM}(seg) \wedge \\ \neg\exists s : \{T\} = uses_s[seg] \end{array} \right\}$$

▷ "Composable" word tag sets that share all representing segments to some other tag sets

return $|compos|/|combined|$

$\mathscr{L}$	Interc., %	$\mathscr{L}$	Interc., %
cat	85	hun	88
ces	100	ita	55
deu	96	por	55
eng	50	rus	98
fin	100	spa	96
fra	52	swe	97

Table 2: "Segment composability" as measured over MorphyNet datasets.

6 From Segments to Surface Forms

Even when all morphological segments are predicted, a conversion into a surface form is yet to be done. Luckily, in most cases, such a conversion only requires to remove segment separators and concatenate the substrings. However, to account for phonotactics, additional string edit operations may be necessary. Our analysis discovered the following major cases when they are needed: (1) removal or modification of *affixes* that are relevant only to the lemma form and are not separated from the stem into a different segment. This mostly concerns verbs. For example, deletion of `-ar` and insertion of `-u-` in Spanish (`catalogar` → `cataloguem V|IND;PRS;1;PL catalogar|em`); (2) removal of adjacent duplicate letters in some languages; (3) replacement of certain adjacent letter combinations at segment boundaries as in the following Czech example: `čtverec` → `čtvercem N;SG|INST;MASC;INAN čtverec|em`.

Predicting such transformations is generally a sequence-to-sequence task. Still, it is rather specific sub-task in which source and target sequences are aligned, and only local character modifications are to be learnt. In our experiments, a hard attention model (Aharoni and Goldberg, 2017) yields nearly perfect prediction of segments "gluing" into a word.[4] German was the only exception due to compounding.

[3] High composability figures, besides a language's agglutinativity, may result from a large size of the corresponding dataset or high variety of word forms presented there. As shown in Table 2, values for closely related language may differ significantly. A high composability is particularly important for agglutinative morphology modelling. However, it shouldn't be preceived as a *measure* of a language's agglutinativity.

[4] Grammatical tags were ignored (set to some constant value).

$\mathscr{L}$	Accuracy	$\mathscr{L}$	Accuracy
cat	0.99	hun	0.98
ces	0.98	ita	0.99
deu	0.89	mon	1.00
eng	0.99	por	1.00
fra	0.99	swe	0.98

Table 3: Segments-to-form conversion accuracy achieved with a hard attention model

7 Discussion

The experiment results suggest that the usage of morphological segmentation dataset enables principal reduction of the complexity of the morphological inflection task. This allows breaking the inflection task into two consecutive stages, (1) producing segments for a given (lemma, tagset) pair, and (2) concatenating segments into a surface word form. As our experiments suggest, prediction of segments in stage (1) is a classification task with a relatively limited feature set, while stage (2) translates into a (minor) string edit task. Here, we have just outlined this perspective direction; a detailed performance exploration is yet to be done. Still, the statistics we collected in our experiments allows us to be optimistic about filling two major gaps in the state-of-the-art systems' performance on these tasks: (1) the ability to generalize to unseen grammatical tag combinations (Kodner et al., 2022), and (2) to better account for phonotactics, as described in Section 6. Also, the proposed reduction of search space should be beneficial for smaller training sets and is crucial for under-resourced languages.

Although morphological segmentation allows a decent amount of fuzziness, it facilitates the discovery of important latent variables that participate in inflection processes. We hypothesize that it would be sufficient to allow an inflection system consider the latent variables within its architecture and fit them during the training process. While the above is the only option for the languages not yet represented in MorphyNet and similar resources, the usage of annotated segmentation datasets should significantly increase generalization ability in the inflection task.

8 Conclusion

We conducted a series of experiments with morphological segmentation and demonstrated that annotated segment sequences may significantly simplify the prediction of inflected forms. We outlined that inflection task can be transformed from sequence-to-sequence into a classification task, with better capacities to address language agglutinativity challenges.

References

Roee Aharoni and Yoav Goldberg. 2017. Morphological inflection generation with hard monotonic attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. Morphynet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th sigmorphon workshop on computational research in phonetics, phonology, and morphology*, pages 39–48.

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.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkuş, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanelov, Gábor Bella, Elena

Budianskaya, Yustinus Ghanggo Ato, 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 Sheifer, Alexandra Serova, 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*.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

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.

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.

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.