Morphology Without Borders: Clause-Level Morphological Annotation

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

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

Abstract

Morphological tasks use large multi-lingual datasets that organize words into inflection tables, which then serve as training and evaluation data for various tasks. However, a closer inspection of these data reveals profound cross-linguistic inconsistencies, that arise from the lack of a clear linguistic and operational definition of what is a word, and that severely impair the universality of the derived tasks. To overcome this deficiency, we propose to view morphology as a clause-level phenomenon, rather than word-level. It is anchored in a fixed yet inclusive set of features homogeneous across languages, that encapsulates all functions realized in a saturated clause. We deliver MIGHTY MORPH, a novel dataset for clause-level morphology covering 4 typologically-different languages: English, German, Turkish and Hebrew. We use this dataset to derive 3 clause-level morphological tasks: inflection, reinflection and analysis. Our experiments show that the clause-level tasks are substantially harder than the respective word-level tasks, while having comparable complexity across languages. Furthermore, redefining morphology to the clause-level provides a neat interface with contextualized language models (LMs) and can be used to probe LMs capacity to encode complex morphology. Taken together, this work opens up new horizons in the study of computational morphology, leaving ample space for studying neural morphological modeling cross-linguistically.

1 Introduction

Morphology has long been viewed as a fundamental part of NLP, especially in cross-lingual settings — from translation (Minkov et al., 2007; Chahuneau et al., 2013) to sentiment analysis (Abdul-Mageed et al., 2011; Amram et al., 2018) — as languages vary wildly in the extent to which they use morphological marking as a means to realize meanings.

Recent years have seen a tremendous development in the data available for supervised morphological tasks, mostly via UniMorph (McCarthy et al., 2020), a large multi-lingual dataset that provides morphological analyses of standalone words, organized into inflection tables in over 170 languages. UniMorph was used in all of SIGMORPHON’s shared tasks in the last decade (Cotterell et al., 2016; Pimentel et al., 2021 inter alia).

Such labeled morphological data rely heavily on the notion of a ‘word’, as words are the elements occupying the cells of the inflection tables, and subsequently words are used as the input or output in the morphological tasks derived from these tables. However, a closer inspection of the data in UniMorph reveals that it is inherently inconsistent
with respect to how words are defined. For instance, it is inconsistent with regards to the inclusion or exclusion of auxiliary verbs such as "will" and "be" as part of the inflection tables, and it is inconsistent in the features words inflect for. A superficial attempt to fix this problem leads to the can of worms that is the theoretical linguistic debate regarding the definition of the morpho-syntactic word, where it seems that a coherent cross-lingual definition of words is nowhere to be found (Haspelmath, 2011).

Relying on a cross-linguistically ill-defined concept in NLP is not unheard of, but it does have its price here; it undermines the perceived universality of the morphological tasks, and skews annotation efforts as well as models’ accuracy in favor of those privileged languages in which morphology is not complex. To wit, even though English and Turkish exhibit comparably complex systems of tense and aspect marking, pronounced using linearly ordered morphemes, English is said to have a tiny verbal paradigm of 5 forms in UniMorph while Turkish has several hundred forms per verb.

Moreover, although inflection tables have a superficially similar structure across languages, they are in fact built upon an inherently language-specific set of features. As a result, models are tasked with arbitrarily different dimensions of meaning, guided by each language’s orthographic tradition (e.g., the abundance of white-spaces used) rather than the set of functions being realized. In this work we set out to remedy such cross-linguistic inconsistencies, by delimiting the realm of morphology by the set of functions realized, rather than the set of forms.

Concretely, in this work we propose to reintroduce universality into morphological tasks by sidestepping the issue of what is a word and giving up on any attempt to determine consistent word boundaries across languages. Instead, we anchor morphological tasks in a cross-linguistically consistent set of inflectional features, that is equivalent to a fully-saturated clause. Then, the lexemes in all languages are inflected to all legal feature combinations of this set, regardless of the amount of ‘words’ or ‘white spaces’ needed to realize its meaning. Under this revised definition, the inclusion of the Swahili form ‘siwapendi’ for the lexeme PENDA inflected to the following features: PRS;NEG;NOM(1,SG);ACC(3,PL), entails the inclusion of the English form ‘I don’t love them’, bearing the exact same lexeme and features.

We thus present MIGHTYMORPH, a novel dataset for clause-level inflectional morphology, covering 4 typologically different languages: English, German, Turkish and Hebrew. We sample data from MIGHTYMORPH for 3 clause-level morphological tasks: inflection, reinflection and analysis. We experiment with standard and state-of-the-art models for word-level morphological tasks (Silfverberg and Hulden, 2018; Makarov and Clematide, 2018; Peters and Martins, 2020) and show that clause-level tasks are substantially harder compared to their word-level counterparts, while exhibiting comparable cross-linguistic complexity.

Operating on the clause level also neatly interfaces morphology with general-purpose pre-trained language models, such as T5 (Raffel et al., 2020) and BART (Lewis et al., 2020), to harness them for morphological tasks which were so far considered non-contextualized. Using the multilingual pre-trained model mT5 (Xue et al., 2021) on our data shows that complex morphology is still genuinely challenging for such LMs. We conclude that our redefinition of morphological tasks is more theoretically sound, crosslingually more consistent, and lends itself to more sophisticated modelling, leaving ample space to probe the ability of LMs to encode complex morphological phenomena.

The contributions of this paper are manifold. First, we uncover a major inconsistency in the current setting of supervised morphological tasks in NLP (§2). Second, we redefine morphological inflection to the clause level (§3) and deliver MIGHTYMORPH, a novel clause-level morphological dataset reflecting the revised definition (§4). We then present data for 3 clause-level morphological tasks with strong baseline results that demonstrate the profound challenge posed by our new approach to contemporary models (§5).

2 Morphological Essential Preliminaries

2.1 Morphological Tasks

Morphological tasks in NLP are typically devided into generation and analysis tasks. In both cases, the basic morphological structure assumed is an inflection table. The dimensions of an inflection table are defined by a set of attributes (e.g., gender, number, case, etc.) and their possible values (e.g., gender:{masculine,feminine,neutral}). A specific attribute:value pair defines an inflectional feature (henceforth, a feature) and a specific combination of features is called an inflectional feature bundle.
A paradigm (here, a feature bundle). An inflection table includes, for a given lexeme \( l_i \), an exhaustive list of \( m \) inflected word-forms \( \{ w_{lj}^i \}_{j=0}^m \), corresponding to all available feature bundles \( \{ t_j \}_{j=0}^n \). See Table 1a for a fraction of an inflection table in Swahili. A paradigm in a language (verbal, nominal, adjectival, etc.) is a set of inflection tables. The set of inflection tables for a given language can be used to derive labeled data for (at least) 3 different tasks, inflection, reinflection and analysis.¹

In morphological inflection (1a), the input is a lemma \( l_i \) and a feature bundle \( t_j \) that specifies the target word-form. The output is that inflected word-form \( w_{lj}^i \) realizing the features in the bundle. (1b) is an example for inflection in the French verbal paradigm for the lemma finir (complete), inflected to an indicative IND future tense FUT with a 1st person singular subject 1; SG.

\[
(1) \quad a. \quad \langle l_i, t_j \rangle \mapsto w_{lj}^i \\
(2) \quad b. \quad \langle \text{finir}, \text{IND}; \text{FUT}; 1; \text{SG} \rangle \mapsto \text{finirai}
\]

The morphological inflection task is in fact a specific version of a more general task which is called morphological reinflection. In the general case, the source of inflection can be any form rather than only the lemma. Specifically, a source word-form \( w_{lj}^i \) from some lexeme \( l_i \) is given as input accompanied by its own feature bundle \( t_j \), and the model reinflects it to a different feature bundle \( t_k \), resulting in the word \( w_{lk}^i \) (2a). In (2b) we illustrate for the same French lemma finir, a reinflection from the indicative present tense with first person singular ‘finis’ to the subjunctive past and second person singular ‘finisses’.

\[
(2) \quad a. \quad \langle t_j, w_{lj}^i, \langle t_k, \_ \_ \rangle \rangle \mapsto w_{lk}^i \\
(3) \quad b. \quad \langle \text{IND}; \text{PRS}; 1; \text{SG}, \text{finis} \rangle , \langle \text{SBJV}; \text{PST}; 2; \text{SG}, \_ \_ \rangle \mapsto \text{finisses}
\]

Morphological inflection and reinflection are generation tasks, in which word forms are generated from feature specifications. In the opposite direction, morphological analysis is a task where word-forms are the input, and models map them to their lemmas and feature bundles (3a). This task is in fact an inverted version of inflection, as can be seen in (3), which are the exact inverses of (1).

\[
(3) \quad a. \quad w_{lj}^i \mapsto \langle l_i, t_j \rangle \\
(4) \quad b. \quad \text{finirai} \mapsto \langle \text{finir}, \text{IND}; \text{FUT}; 1; \text{SG} \rangle
\]

²UniMorph

The most significant source of inflection tables for training and evaluating all of the aforementioned tasks is UniMorph² (Sylak-Glassman et al., 2015; McCarthy et al., 2020), a large inflectional-morphology dataset covering over 170 languages. For each language the data contains a list of lexemes with all their associated feature bundles and the words realizing them. Formally, every entry in UniMorph is a triplet \( \langle l, t, f \rangle \) with lemma \( l \), a feature bundle \( t \), and a word-form \( f \). The tables in UniMorph are exhaustive, that is, the data generally does not contain partial tables; their structure is fixed for all lexemes of the same paradigm, and each cell is filled in with a single form, unless that form doesn’t exist in that language.³ The data is usually crawled from Wiktionary⁴ or from some preexisting finite-state automaton. The features for all languages are standardized to be from a shared inventory of features, but every language makes use of a different subset of that inventory.

So far, the formal definition of UniMorph seems cross-linguistically consistent. However, a closer inspection of UniMorph reveals an inconsistent definition of words, which then influences the dimensions included in the inflection tables in different languages. For example, the Finnish phrase ‘olen ajatellut’ is considered a single word, even though it contains a white-space. It is included in the relevant inflection table and annotated as ACT;PRS;PRF;POS;IND;1;SG. Likewise, the Albanian phrase ‘do të mendosh’ is also considered a single word, labeled as IND;FUT;1;PL. In contrast, the English equivalents have thought and will think, corresponding to the exact same feature-bundles and meanings, are absent from UniMorph, and their construction is considered purely syntactic.

This overall inconsistency encompasses the inclusion or exclusion of various auxiliary verbs as well as the inclusion of particles, clitics, light verb constructions and more. The decision on what or how much phenomena to include is done in a per-language fashion that is inherited from the specific language’s grammatical traditions and sources, and in practice it is quite arbitrary and taken without any consideration of universality. In fact, the definition of inflected words can be inconsistent even in

¹The list of tasks mentioned above is of course not exhaustive; other tasks may be derived from labeled inflection tables, e.g. the Paradigm Cell Completion Problem (Ackerman et al., 2009; Cotterell et al., 2017).

²https://unimorph.github.io

³In cases of overabundance, i.e., availability of more than one form per cell, only one canonical form occupies the cell.

⁴https://www.wiktionary.org
closely related languages in the same language family, e.g., the Arabic definite article is included in the Arabic nominal paradigm, while the equivalent definite article is excluded for Hebrew nouns.

One possible attempted solution could be to define words by white-spaces and strictly exclude any forms with more than one space-delimited word. However, this kind of solution will severely impede the universality of any morphological task as it would give a tremendous weight to the orthographic tradition of a language and would be completely inapplicable for languages that do not use a word-delimiting sign like Mandarin Chinese and Thai. In addition, a decades-long debate about space-agnostic word definition have failed to result in any workable solution (see Section 6).

We therefore suggest to proceed in the opposite, far more inclusive, direction. We propose not to try to delineate ‘words’, but rather to delineate a consistent feature set that lexemes can be inflected for, regardless of the number of ‘words’ and white spaces needed to realize it.

3 The Proposal: Word-Free Morphology

In this work we extend inflectional morphology, data and tasks, to the clause level. We define an inclusive cross-lingual set of inflectional features \( \{t_j\} \) and inflect lemmas in all languages to the same set, no matter how many white-spaces have to be used in the realized form. By doing so, we reintroduce universality into morphology, equating the treatment of languages in which clauses are frequently expressed with a single word with those that use several of them. Figure 1 exemplifies how this approach induces universal treatment for typologically different languages, as lexemes are inflected to the same feature bundles in all of them.

The Inflectional Features Our guiding principle in defining an inclusive set of features is the inclusion of all features types expressed at the word level in some language. This set essentially defines a saturated clause, with feature types including sentence-level features and various of argument markings.

Concretely, our universal feature set contains the obvious tense, aspect and mood (TAM) features, as well as negation, interrogativity and all argument-marking features such as: person, number, gender, case, formality and reflexivity. TAM features are obviously included as the hallmark of almost any inflectional system, particularly of most European languages, negation is expressed at the word level in many Bantu languages (Wilkes and Nkosi, 2012; Mpiranya, 2014), and interrogativity — in, e.g., Inuit (Webster, 1968) and to a lesser degree in Turkish.

Perhaps more important (and less familiar) is the fact that in many languages multiple arguments can be marked on a single verb. For example, agglutinating languages like Georgian and Basque show poly-personal agreement, where the verb morphologically indicates features of multiple arguments, above and beyond the subject. For example:

\[
\begin{align*}
\text{a. Georgian: } & \text{გაგიწვებათ} \\
& \text{Trans: “we will let you go”} \\
& \text{IND;FUT;NOM(1,PL);ACC(2,SG)}
\end{align*}
\]

\[
\begin{align*}
\text{b. Spanish: } & \text{dimelo} \\
& \text{Trans: “tell it to me”} \\
& \text{IMP;NOM(2,SG);ACC(3,SG,NEUT);DAT(1,SG)}
\end{align*}
\]

\[
\begin{align*}
\text{c. Basque: } & \text{dakarkiogu} \\
& \text{Trans: “we bring it to him/her”} \\
& \text{IND;PRS;ERG(1,PL);ABS(3,SG);DAT(3,SG)}
\end{align*}
\]

Following Anderson (1992)’s feature layering approach, we propose the annotation of arguments to be done as complex features, i.e. features that allow a feature set as their value.\(^5\) So, the Spanish verb form dimelo, (translated: ‘tell it to me’), for example, will be tagged as IMP;NOM(2,SG);ACC(3,SG,NEUT);DAT(1,SG).

For languages that do not mark the verb’s arguments by morphemes, we used personal pronouns to realize the relevant feature-bundles, e.g. the bold elements in the English translations in (4).

Treating pronouns as feature realizations keeps the clauses single-lexemed for all languages, whether argument incorporating or not. To keep the inflected clauses single-lexemed we also limited the forms to main clauses, avoiding subordination.

Although we collected the inflectional features empirically and bottom-up, the list we ended up with corresponds to Anderson (1992, p. 219)’s suggestion for clausal inflections: “[for VP:] auxiliaries, tense markers, and pronominal elements representing the arguments of the clause; and determiners and possessive markers in NP”\(^6\). Thus, our suggested feature set is not only diverse and inclusive in practice, it is also theoretically sound.

---

\(^5\)This is reminiscent of feature structures in Unification Grammars (Shieber, 2003) as GPSG, HPSG and LFG (Gazdar et al., 1989; Pollard and Sag, 1994; Bresnan et al., 2015).

\(^6\)Our resulted set may still be incomplete, but the principle
We present MIGHTY, the first multilingual clause-level morphological dataset. Like UniMorph, MIGHTY contains inflection tables with entries of the form of lemma, features, form. The data can be used to elicit training sets for any clause-level morphological task. The data covers 4 languages from 3 language families: English, German, Turkish and Hebrew. Our selection covers languages classified as isolating, agglutinative and fusional. The languages vary also in the extent they utilize morpho-syntactic processes: from the ablaut extensive Hebrew to no ablauts in Turkish; from fixed word order in Turkish to the meaning-conveying word-order in German. Our data for each language contains 500 inflection tables.

To illustrate, Table 1 shows a fragment of a clause-level inflection table in Swahili and its English equivalent. It shows that while the Swahili forms are expressed with one word, their English equivalents express the same feature bundles with several words. Including the exact same feature combinations, while allowing for multiple ‘word’ expressions in the inflections, finally makes the comparison between the two languages straightforward, and showcases the comparable complexity of clause-level morphology across languages.

**The Tasks** To formally complement out proposal, we amend the task definitions in Section 2 to refer to forms in general $f_i^l$ rather than words:

5. Clause-Level Morphological Tasks

a. inflection $\langle t_i, t_j \rangle \rightarrow f_i^l$

b. reinflection $\langle t_j, f_i^l, \langle t_k, ? \rangle \rightarrow f_i^l$

c. analysis $f_i^l \rightarrow \langle t_i, t_j \rangle$

| lexeme—PENDA | IND | POS | NEG | IND.PERF | POS | NEG | COND | POS | NEG |
|--------------|-----|-----|-----|----------|-----|-----|------|-----|-----|
| ACC(1,SG)   | unanienda | hunipendi | umenienda | humenienda | unenienda | usenienda | usingenienda | usingenienda | usingenienda |
| ACC(1,PL)   | unatuenda | hutupendi | umetuenda | hometuenda | ungetuenda | usingetuenda | usingejipenda | usingetienda | usingewapenda |
| ACC(2,SG,FLX) | unajienda | hujipendi | umejipenda | hujipenda | unjejipenda | usingejipenda | usingewapenda | usingewapenda | usingewapenda |
| ACC(2,PL)   | unawapenda | huwapendi | unawapenda | unawapenda | ungempenda | usingempenda | usingawapenda | usingawapenda | usingawapenda |
| ACC(3,SG)   | unampenda | humpendi | umempenda | humpenda | unampenda | usingampenda | usingampenda | usingampenda | usingampenda |
| ACC(3,PL)   | unapenda | hupendi | umempenda | humpenda | unapenda | usingapenda | usingapenda | usingapenda | usingapenda |

(a) Swahili inflection table

| lexeme—LOVE | IND | POS | NEG | IND.PERF | POS | NEG | COND | POS | NEG |
|-------------|-----|-----|-----|----------|-----|-----|------|-----|-----|
| ACC(1,SG)   | you love | you don’t | you have | you haven’t | you would | you wouldn’t |
| ACC(1,PL)   | you love | you don’t | you have | you haven’t | you would | you wouldn’t |
| ACC(2,SG,FLX) | you love | you don’t | you have | you haven’t | you would | you wouldn’t |
| ACC(2,PL)   | you love | you don’t | you have | you haven’t | you would | you wouldn’t |
| ACC(3,SG)   | you love | you don’t | you have | you haven’t | you would | you wouldn’t |
| ACC(3,PL)   | you love | you don’t | you have | you haven’t | you would | you wouldn’t |

(b) English inflection table

Table 1: A fraction of a clause-level inflection table, in both English and Swahili. The tables are completely aligned in terms of meaning, but differ in the amount of words needed to realize each cell. In practice we did not use the ‘y’all pronoun and it is given here for the illustration.

4 The MIGHTY Benchmark

We present MIGHTY, the first multilingual clause-level morphological dataset. Like UniMorph, MIGHTY contains inflection tables holds: when adding a new language with new word-level features, these features will be realized for all languages.

---

8For Hebrew we annotated a vocalized version in addition to the commonly-used unvocalized forms.

9With the exception of Hebrew, where 780 tables were annotated for each of the vocalized and unvocalized variants.

We will openly release our annotation infrastructure to facilitate the addition of more languages into MIGHTY.
Table 2: Comparison of statistics over the 4 languages common to UniMorph (UM) and MIGHTY-MORPH (MM): size of an inflection table, size of the feature set used, and average amount of features per form. For all indicators, the values for MIGHTY-MORPH are more uniform across languages.

determine valency, we consulted a native speaker or a monolingual dictionary.\textsuperscript{10} We treated all argument types equally and annotated them with a feature, whether expressed with an affix or an adposition. We exhaustively generated all possible combinations of verb inflections and arguments without regarding the semantic plausibility of the resulted clause. So, the clause \textit{I read him to them} is expected to appear in the inflection table of \textsc{Read}, even though it may sound odd.

**Data Analysis** The MIGHTY-MORPH benchmark represents inflectional morphology in 4 typologically very-different languages, yet, as shall be seen shortly, the data is both more uniform across languages and more diverse in the features realized for each language, compared to the de-facto standard word-level morphological annotations.

Table 2 compares aggregated values between UniMorph and MIGHTY-MORPH across languages: the size of an inflection table,\textsuperscript{11} the amount of unique features used, and the average number of features used per form. We see that MIGHTY-MORPH is more consistent cross-lingually on all 3 comparisons: the size of tables is less varied, so English no longer has extraordinarily small tables; the fact that the set of potential features is fixed for all languages makes the sets of features that were used per language highly similar; and finally, forms in all languages are now described by feature bundles with comparable averaged length. The residual variation in all of these values arises only from true linguistic variation, like the fact that Hebrew does not express verbal aspects at all, or that German uses a wider array of adpositions for verbal arguments. This is a strong empirical indication that applying morphological annotation to clauses reintroduces universality into morphological data.

In addition, the bigger inflection tables in MIGHTY-MORPH include more diverse phenomena, like word-order changes in English, lexeme-dependent perfective auxiliary in German, and partial pro-drop in Hebrew. Thus, Models trying to tackle clause-level morphology will need to address these newly added phenomena. We conclude that our proposed data and tasks are both more challenging and more universal than previously-studied word-level tasks.

5 Experiments

**Goal** We set out to assess the challenges and opportunities in clause-level morphological tasks. To this end we experimented with the 3 tasks defined in Section 3: inflection, reinflection and analysis, all executed both at the word-level and clause-level.

**Splits** For each task we sampled 10,000 examples or pairs of examples in the suitable format from 500 inflection tables. We used 80% of the examples for training and the rest was divided between the validation and test sets. We sampled the same amount of examples from each table and, following Goldman et al. (2021), we split the data such that the lexemes in each set are disjoint. So, 400 lexemes are used in the train set, and 50 are for each of the validation and test sets.

**Models** As solid baselines, we applied contemporary models designed for word-level morphological tasks. The word-level inflection models handle characters as input and output, and they were applied to the clause-level task straightforwardly by treating white-space as yet another character rather than a special word-delimiter.\textsuperscript{12} We set out to assess the challenges and opportunities in clause-level morphological tasks. To this end we experimented with the 3 tasks defined in Section 3: inflection, reinflection and analysis, all executed both at the word-level and clause-level.

- LSTM: A char-based sequence-to-sequence architecture of LSTM encoder-decoder with attention by Silfverberg and Hulden (2018)
- TRANSduce: A neural transducer predicting actions between the input and output strings by Makarov and Clematide (2018).
- DEEPSpIn: an RNN-based system by Peters and Martins (2020) using sparsemax instead of softmax.

\textsuperscript{10}For German we used Duden dictionary, and for Turkish we used the Türk Dil Kurumu’s dictionary.
\textsuperscript{11}At the clause level we compared to an intransitive table, as the table size is dependent on the transitivity of the verb.
\textsuperscript{12}For each language and task we trained a separate model for 50 epochs.
|            | Eng word | Eng clause | Deu word | Deu clause | Heb word | Heb clause | Heb\textsubscript{vocalized} word | Heb\textsubscript{vocalized} clause | Tur word | Tur clause | Average word | Average clause |
|------------|----------|-----------|----------|------------|-----------|-----------|-------------------------------|-------------------------------|----------|-----------|--------------|----------------|
| LSTM       | 86.0     | ±1.8      | 68.5     | ±3.8       | 64.5      | ±4.7      | 47.5                          | ±4.0                          | 90.7     | ±1.6      | 82.5         | ±0.6          |
|            |          |           | 91.7     | ±1.1       | 70.0      | ±1.2      |                               |                               | 90.8     | ±0.9      | 81.6         | ±2.1          |
|            |          |           |          |            |           |           |                               |                               | 84.7     | ±1.1      | 70.0         | ±1.2          |
| DEEPSPIN   | 87.3     | ±2.8      | 78.4     | ±1.5       | 78.2      | ±0.5      | 40.0                          | ±0.5                          | 90.9     | ±0.2      | 86.1         | ±0.7          |
|            |          |           | 93.1     | ±2.0       | 71.7      | ±0.7      |                               |                               | 97.5     | ±2.1      | 82.7         | ±1.6          |
|            |          |           |          |            |           |           |                               |                               | 89.4     | ±0.8      | 86.1         | ±0.5          |
|            |          |           |          |            |           |           |                               |                               | 99.4     | ±0.1      | 97.2         | ±0.5          |
|            |          |           |          |            |           |           |                               |                               | 86.7     | ±0.5      | 78.9         | ±0.4          |
| TRANSUDCE  | 86.8     | ±0.4      | 85.4     | ±1.1       | 76.6      | ±2.5      | 71.5                          | ±1.3                          | 89.4     | ±0.6      | 80.4         | ±0.8          |
|            |          |           | 81.1     | ±0.5      | 70.7      | ±1.1      |                               |                               | 60.0     | ±1.1      | 50.6         | ±1.1          |
|            |          |           |          |            |           |           |                               |                               | 99.4     | ±0.1      | 97.2         | ±0.5          |
|            |          |           |          |            |           |           |                               |                               | 86.7     | ±0.5      | 78.9         | ±0.4          |
| MT5        | NA       | ±0.4      | 70.7     | ±1.7       | NA        | ±3.3      | 57.7                          | ±2.9                          | NA       | ±3.3      | 48.0         | ±1.4          |
|            |          |           |          |            |           |           |                               |                               | NA       | ±1.4      | 34.2         | ±1.8          |
|            |          |           |          |            |           |           |                               |                               | NA       | ±1.7      | 48.7         | ±1.1          |
|            |          |           |          |            |           |           |                               |                               | NA       | ±1.1      | 51.9         | ±1.1          |
|            |          |           |          |            |           |           |                               |                               |          |           |              |                |
| LSTM       | 78.2     | ±6.3      | 62.7     | ±2.5       | 53.5      | ±3.5      | 31.0                          | ±1.7                          | 68.4     | ±1.6      | 30.6         | ±2.9          |
|            |          |           | 80.7     | ±1.9       | 31.4      | ±3.6      |                               |                               | 85.2     | ±2.2      | 71.1         | ±1.2          |
|            |          |           |          |            |           |           |                               |                               | 73.2     | ±1.6      | 45.4         | ±9.4          |
| TRANSUDCE  | 82.7     | ±1.1      | 67.1     | ±0.4       | 81.5      | ±0.5      | 35.5                          | ±0.3                          | 77.2     | ±1.2      | 41.5         | ±2.2          |
|            |          |           | 49.2     | ±1.5      | 22.2      | ±1.5      |                               |                               | 6.1      | ±1.8      | 72.5         | ±2.7          |
|            |          |           |          |            |           |           |                               |                               | 84.7     | ±1.1      | 75.1         | ±0.8          |
|            |          |           |          |            |           |           |                               |                               | 95.1     | ±1.0      | 44.5         | ±0.8          |
| MT5        | NA       | ±3.1      | 73.6     | ±2.0       | NA        | ±4.2      | 54.2                          | ±1.9                          | NA       | ±3.7      | 29.7         | ±1.9          |
|            |          |           |          |            |           |           |                               |                               | NA       | ±7.0      | 37.5         | ±1.8          |
|            |          |           |          |            |           |           |                               |                               | NA       | ±1.8      | 45.2         | ±1.8          |
|            |          |           |          |            |           |           |                               |                               |          |           |              |                |
| LSTM       | 81.6     | ±0.4      | 79.8     | ±2.2       | 34.8      | ±2.2      | 25.7                          | ±0.6                          | 34.6     | ±1.3      | 57.7         | ±1.4          |
|            |          |           | 73.3     | ±3.6      | 74.8      | ±2.2      |                               |                               | 85.6     | ±2.7      | 84.4         | ±4.5          |
|            |          |           |          |            |           |           |                               |                               | 62.0     | ±0.9      | 64.4         | ±1.1          |
| TRANSUDCE  | 69.0     | ±1.2      | NA       | ±2.7       | 45.1      | ±3.3      | NA                            |                               | 34.2     | ±1.4      | 46.0         | ±4.5          |
|            |          |           |          |            |           |           |                               |                               | NA       | ±4.5      | 42.8         | ±1.2          |

Table 3: Word and clause results for all tasks, models and languages, stated in terms of exact match accuracy in percentage. Over clause tasks, for every language and task the best performing system is in \textbf{bold}, in cases of statistical insignificance all best systems are marked. Results are averaged over 3 runs.

All models were developed for word-level inflection. **TRANSDUCE** is the SOTA for low-resourced morphological inflection (Cotterell et al., 2017), and DEEPSPIN is the SOTA in the general setting (Vylomova et al., 2020; Goldman et al., 2021). LSTM and TRANSUDCE were modified and applied to reinflection as well,\(^\text{13}\) while only LSTM could be used for analysis due to its general design.

In contrast with word-level tasks, the extension of morphological tasks to the clause-level introduces context, and provides an interface to pretrained contextualized LMs. We used a pretrained text-to-text model as an advanced alternative:

- **MT5**: An encoder-decoder transformer-based model, pretrained by Xue et al. (2021)

MT5’s input and output are tokens provided by the model’s own tokenizer; the morphological features were added to the model’s vocabulary as new tokens with randomly initialized embeddings.

As none of the models were designed to deal with hierarchical complex feature structures, the features’ strings were flattened before training and evaluation. For example, the bundle \textsc{ind};\textsc{prs};\textsc{nom}(1,\textsc{sg});\textsc{acc}(2,\textsc{pl}) is replaced with \textsc{ind};\textsc{prs};\textsc{nom1};\textsc{nomsg};\textsc{acc2};\textsc{accpl}.\(^\text{14}\)

\(^{13}\)For technical reasons, DEEPSPIN could not be adapted in the same way.

\(^{14}\)In future work we intend to take advantage of the hierarchical feature structure with more sophisticated models for

5.1 Results and Analysis

Table 3 summarizes the results for all models and tasks, for all languages. For the generation tasks, inflection and reinflection, the word-level models clearly underperform when applied to the clause-level data, confirming that on average the clause-level generation tasks are more challenging. The only deviation is for **TRANSDUCE** on inflection in English and Turkish, where the differences in performance are still apparent, albeit minimal. Among the two tasks, it is clear that clause-level reinflection is more challenging. This is perhaps not surprising as the models are not given an explicit lemma, but must identify it and only then reinflect.

In terms of languages, English and Turkish seem to pattern together in the generation tasks and are easier than Hebrew and German for the word-level models. This may point to a shared typology, as they both are more fusional than the others in our set. For reinflection over vocalized Hebrew, the drop in performance is exceptionally large. We conjecture that this is due to the fact that non-concatenative morphology expressed in varied vocalization patterns poses a significant hurdle, especially for **TRANSDUCE** that relies on the longest common sequence of characters.

\(^{14}\)In future work we intend to take advantage of the hierarchical feature structure with more sophisticated models for compositional generalization, like Prange et al. (2021)’s model for CCG parsing.
Figure 2: Learning curve for inflection over all languages done by TRANSDUCE. Solid lines are for increasing train set sizes while dashed lines — for using more lexemes.

Figure 3: Learning curve for reinflection over all languages done by TRANSDUCE. Solid lines are for increasing train set sizes while dashed lines — for using more lexemes.

Figure 4: Learning curve for analysis over all languages done by LSTM. Solid lines are for increasing train set sizes while dashed lines — for using more lexemes.

Regarding the analysis task, LSTM performs similarly on the word- and clause-levels on average. The deviations are German, where the clause-level task is significantly harder, and unvocalized Hebrew where the performance on clause-level analysis is surprisingly better than the word-level counterpart. In an error analysis of the LSTM’s analysis output for unvocalized Hebrew, we found that clause-level models have a lot more errors in lemma prediction. We conjecture that this is strictly related to the lack of written vowels for this variant of Hebrew, that forces models and speakers to infer some aspects of meaning from context. This corroborates the findings of More et al. (2019) that context aids morphological disambiguation.

Finally, we turn to the performance of mT5. While this strong pretrained LM proved itself useful in many downstream tasks, it was not proven to be a silver bullet in these morphological tasks. Its clause-level performance is on par with the word-level models in reinflection, but it significantly underperforms in inflection and analysis. Its performance also seems to be biased towards the Western languages, English and German. However, its performance is the most invariable across tasks, hinting at its general applicability. Moreover, its good performance on reinflection is not surprising given that this is the only task that relates to full clauses in both input and output, suggesting that more compatible modeling may make this model better also for the other tasks.

Data Sufficiency How much labeled data should suffice for training clause-morphology models? To answer this, let us first note that the nature of morphology provides (at least) two ways to increase the amount of information available for the model. One is to increase the absolute number of sampled examples to larger training sets, while using the same amount of inflection tables; alternatively, the number of inflection tables can be increased, for a fixed size of the training set. The former is especially easy in languages with larger inflection tables, where each table can provide hundreds or thousands of inflected forms per lexeme, but the lack of verity in lexemes may lead to overfitting. To examine which dimension is more important for the overall success in the tasks, we tested both.
The resulting curves are provided in Figures 2, 3, and 4, for inflection, reinflection and analysis, respectively. In each Figure, the solid lines are for the results as the absolute train set size is increased, and the dashed lines are for increasing the number of lexemes in the train set while keeping its size fixed. The balance between the options is different for each task. For inflection (Figure 2), increasing the size and the lexeme-variance of the training set produce similar trends, indicating that one dimension can compensate for the other. The curves for reinflection (Figure 3) show that for this task the number of lexemes used is more important than the size of the training set, as the former produces steeper curves and reaches better performance with relatively little amount of lexemes added. On the other hand, the trend for analysis (Figure 4) is the other way around, with increased train set size being more critical than increased lexeme-variance.

6 Related and Future Work

Wordhood in Linguistic Theory The quagmire surrounding words and their demarcation is longstanding in theoretical linguistics. In fact, no coherent definition of word has been provided by the linguistic literature as of yet, despite many attempts. For example, Zwicky and Pullum (1983) enumerate 6 different, sometimes contradictory, ways to discern between words, clitics and morphemes. Haspelmath (2011) also names 10 criteria for wordhood before concluding that no cross-linguistic definition of this notion can currently be found.

Moreover, words may be defined differently in different areas of theoretical linguistics. For example, the prosodic word (Hall, 1999) is defined in phonology and phonetics independently of the morphological word (Bresnan and Mchombo, 1995). And in general, many different notions of a word can be defined (e.g. Packard, 2000, for Chinese).

However, the definition of morpho-syntactic words is inherently needed for the contemporary division of labour in theoretical linguistics, as it defines the boundary between morphology, the grammatical module in charge of word construction, and syntax, that deals with word combination (Dixon and Aikhenvald, 2002). Alternative theories do exist, including ones that incorporate morphology into the syntactic constituency trees (Halle and Marantz, 1993), and others that expand morphology to periphrastic constructions (Ackerman and Stump, 2004) or to phrases in general (Anderson, 1992). In this work we follow that latter theoretical thread and expand morphological annotation up to the level of full clauses.

The definition of words is also relevant to historical linguistics, where the common view considers items on a spectrum between words and affixes. Diachronically, items move mostly towards the affix end of the scale in a process known as grammaticalization (Hopper and Traugott, 2003) while occasional opposite movement is also possible (Norde et al., 2009). However, here as well it is difficult to find precise criteria for determining when exactly an item moved to another category on the scale, despite some extensive descriptions of the process (e.g. Joseph, 2003, for Greek future construction).

The vast work striving for cross-linguistically consistent definition of morpho-syntactic words seems to be extremely Western-biased, as it aspires to find a definition for words that will roughly coincide with those elements of text separated by white-spaces in writing of Western languages, rendering the endeavour particularly problematic for languages with orthographies that do not use white-spaces at all, like Chinese whose grammatical tradition contains very little reference to words up until the 20th century (Duanmu, 1998).

In this work we wish to bypass this theoretical discussion as it seems to lead to no workable word definition, and we therefore define morphology without the need of word demarcation.

Wordhood in Language Technology The concept of words has been central to NLP from the very establishment of the field, as most models assume tokenized input (e.g. Winograd, 1971). However, the lack of a word/token delimiting symbol in some languages prompted the development of more sophisticated tokenization methods, supervised (Xue, 2003; Nakagawa, 2004) or statistical (Schuster and Nakajima, 2012), mostly for east Asian languages. Statistical tokenization methods also found their way to NLP of word-delimiting languages, albeit for different reasons like dealing with unattested words and unconventional spelling (Sennrich et al., 2016; Kudo, 2018). Yet, tokens produced by these methods are sometimes assumed to correspond to linguistically defined units, mostly morphemes (Bostrom and Durrett, 2020; Hofmann et al., 2021).

In addition, the usage of words as an organizing notion in theoretical linguistics, separating morphology from syntax, led to the alignment of NLP
research according to the same subfields, with resources and models aimed either at syntactic or morphological tasks. For example, syntactic models usually take their training data from Universal Dependencies (UD; de Marneffe et al., 2021), where syntactic dependency arcs are connecting between words as nodes while morphological features characterize the words themselves, although some works have experimented with dependency parsing of nodes other than words, be it chunks (Abney, 1991; Buchholz et al., 1999) or nuclei (Bārzdinš et al., 2007; Basirat and Nivre, 2021). However, in these works as well, the predicate-argument structure is still opaque in agglutinative languages where the entire structure is expressed in a single word.

Here we argue that questions regarding the correct granularity of input for NLP models will continue to haunt the research. At least until a thorough reference is made to the predicament surrounding these questions in theoretical linguistics. We proposed that given the theoretic state of affairs, a technologically viable word-free solution for computational morpho-syntax is desired.

Limitations and Extensions of Clause-Level Morphology

Our revised definition of morphology to disregard word boundaries does not (and is not intended to) solve all existing problems with morphological annotations in NLP of course.

For example, our definition does not solve the long-debated demarcation of boundary between inflectional and derivational morphology (e.g., Scalise, 1988). The lack of clear boundary between inflectional and derivational morphology is highly similar to the lack of definition for words which operate as the boundary between morphology and syntax. Indeed, in the theoretical linguistics literature, some advocate a view that posits no boundary between inflectional and derivational morphology (Bybee, 1985). Although this question is out of scope for this work, we conjecture that this similar problem may require a similar solution, that blurs that intra-morphological boundary as well and potentially defines a single framework for the entire inflectional–derivational morphology continuum.

Also, our shift to clause-level morphology does not solve the problem of overabundance, where several forms are conceivable for occupying the same cell in the paradigm (for example, non-mandatory pro-drop in Hebrew). As the problem exists also in word-level morphology, we followed the same approach and constructed only one canonical form for each cell. However, for a greater empirical reach of our proposal, an even further extension of the inflection table is conceivable, to accommodate sets of forms in every cell, rather than a single form.

Finally, our solution for annotating morphology at clause level blurs the boundary between morphology and syntax, as often presupposed in NLP, and thus has implications also for syntactic tasks. Some previous works indeed emphasized the cross-lingual inconsistency in word definition from the syntactic perspective (Basirat and Nivre, 2021). Our work points to a holistic approach for morpho-syntactic annotation in which clauses are consistently tagged in a morphology-style annotation, leaving syntactic operations only for relations between clauses. Thus, we suggest that an extension of the approach taken here is desired in order to realize a single morpho-syntactic framework. Specifically, our approach should be extended to include: morphological annotation for clauses with multiple lexemes; realization of morphological features of more clause-level characteristics, e.g. types of subordination and conjunction; and annotation of clauses in recursive structures.

7 Conclusions

In this work we exposed the fundamental inconsistencies in contemporary computational morphology and, to remedy this, delivered MIGHTY-MORPH, the first dataset for clause-level morphology. We derived training data and evaluation for clause-level inflection, reinflection and analysis tasks. Our data analysis shows that the complexity of these tasks is more comparable across languages than their word-level counterparts. This reinforces our assumption that redefinition of morphology to the clause-level reintroduces cross-lingual universality into computation morphology. Moreover, we showed that standard (re)inflection models struggle on clause-level compared to their performance in word-level tasks, and that clause-level tasks present a bigger challenge that is not trivially solved, even by contextualized pre-trained LMs such as M5T5. In the future we intend to further expand our framework for more languages, and to explore more sophisticated models, e.g. by taking advantage of the hierarchical structure or by better utilizing pre-trained LMs. Moreover, we plan to expand the proposal and benchmark to the inclusion of derivational morphology, and to a unified morpho-syntactic framework.
References

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of Modern Standard Arabic. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 587–591, Portland, Oregon, USA. Association for Computational Linguistics.

Steven P Abney. 1991. Parsing by chunks. In Principle-based parsing, pages 257–278. Springer.

Farrell Ackerman, James P Blevins, and Robert Malouf. 2009. Parts and wholes: Patterns of relatedness in complex morphological systems and why they matter. Analogy in grammar: Form and acquisition, 54:82.

Farrell Ackerman and Gregory Stump. 2004. Paradigms and periphrastic expression: A study in realization-based lexicalism. Projecting morphology, pages 111–157.

Adam Amram, Anat Ben David, and Reut Tsarfaty. 2018. Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from Modern Hebrew. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2242–2252, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Stephen R Anderson. 1992. A-morphous morphology. 62. Cambridge University Press.

Guntis Bārzdinš, Normunds Grūzītis, Gunta Nešpore, and Baiba Saulīte. 2007. Dependency-based hybrid model of syntactic analysis for the languages with a rather free word order. In Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007), pages 13–20, Tartu, Estonia. University of Tartu, Estonia.

Ali Basirat and Joakim Nivre. 2021. Syntactic nuclei in dependency parsing – a multilingual exploration. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1376–1387, Online. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 4617–4624, Online. Association for Computational Linguistics.

Joan Bresnan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. Lexical-functional syntax. John Wiley & Sons.

Joan Bresnan and Sam A Mchombo. 1995. The lexical integrity principle: Evidence from bantu. Natural Language & Linguistic Theory, 13(2):181–254.

Sabine Buchholz, Jorn Veenstra, and Walter Daelemans. 1999. Cascaded grammatical relation assignment. In 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.

Joan L Bybee. 1985. Morphology: A study of the relation between meaning and form, volume 9. John Benjamins Publishing.

Victor Chahuneau, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. Translating into morphologically rich languages with synthetic phrases. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1677–1687, Seattle, Washington, USA. 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.
Marie-Catherine de Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Robert MW Dixon and Alexandra Y Aikhenvald. 2002. Word: a typological framework. *Word: A cross-linguistic typology*, pages 1–41.

San Duanmu. 1998. Wordhood in chinese. In *New approaches to Chinese word formation*, pages 135–196. De Gruyter Mouton.

Gerald Gazdar, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1989. *Generalized phrase structure grammar*. *Philosophical Review*, 98(4):556–566.

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

T Alan Hall. 1999. The phonological word: a review. *Studies on the phonological word*, 174(1):22.

Morris Halle and Alec Marantz. 1993. Distributed morphology and the pieces of inflection. *The view from building 20*, pages 111–176.

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

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre is not superb: Derivational morphology improves BERT’s interpretation of complex words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608, Online. Association for Computational Linguistics.

Paul J Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge University Press.

Brian D Joseph. 2003. Morphologization from syntax. *Handbook of historical linguistics*, pages 472–492.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. 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.

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 Krizhanovskiy, 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.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

Amir More, Amit Seker, Victoria Basmova, and Reut Tsarfaty. 2019. Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from Modern
Hebrew. Transactions of the Association for Computational Linguistics, 7:33–48.

Fidèle Mpiranya. 2014. Swahili grammar and workbook. Routledge.

Tetsuji Nakagawa. 2004. Chinese and Japanese word segmentation using word-level and character-level information. In COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, pages 466–472, Geneva, Switzerland. COLING.

Muriel Norde et al. 2009. Degrammaticalization. Oxford University Press.

Jerome L Packard. 2000. The morphology of Chinese: A linguistic and cognitive approach. Cambridge University Press.

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.

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, Katrina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vanina, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardjianto, 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.

Carl Pollard and Ivan A Sag. 1994. Head-driven phrase structure grammar. University of Chicago Press.

Jakob Prange, Nathan Schneider, and Vivek Srikumar. 2021. Supertagging the long tail with tree-structured decoding of complex categories. Transactions of the Association for Computational Linguistics, 9:243–260.

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

Sergio Scalise. 1988. Inflection and derivation.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5149–5152. IEEE.

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.

Stuart M Shieber. 2003. An introduction to unification-based approaches to grammar. Microtome Publishing.

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.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics.
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.

Donald Humphry Webster. 1968. Let’s learn Eskimo. Summer Institute of Linguistics.

Arnett Wilkes and Nikolias Nkosi. 2012. Complete Zulu Beginner to Intermediate Book and Audio Course: Learn to read, write, speak and understand a new language with Teach Yourself. Hachette UK.

Terry Winograd. 1971. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC.

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.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003:

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