Morphology Without Borders: Clause-Level Morphology

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

Reut Tsarfaty
Bar Ilan University, Israel
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, which arise from the lack of a clear linguistic and operational definition of *what is a word*, and which 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, that encapsulates all functions realized in a saturated clause. We deliver MIGHTYMORPH, 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 allows assessing the morphological knowledge encoded in these models and their usability for morphological tasks. Taken together, this work opens up new horizons in the study of computational morphology, leaving ample space for studying neural morphology 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 (Batsuren et al., 2022), a large multi-lingual dataset that provides morphological analyses of standalone words, organized into inflection tables in over 170 languages. Indeed, 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,

Transactions of the Association for Computational Linguistics, vol. 10, pp. 1455–1472, 2022. https://doi.org/10.1162/tacl_a_00528
Action Editor: Richard Sproat. Submission batch: 4/2022; Revision batch: 8/2022; Published 12/2022.
© 2022 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

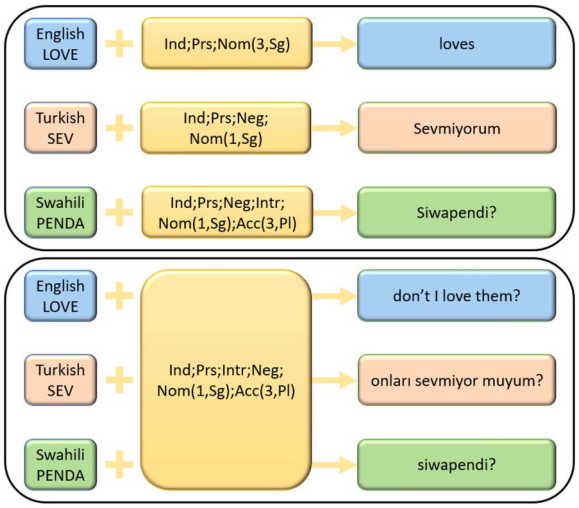


Figure 1: In word-level morphology (top), inflection scope is defined by 'wordhood', and lexemes are inflected to different sets of features in the bundle depending on language-specific word definitions. In our proposed clause-level morphology (bottom) inflection scope is fixed to the same feature bundle in all languages, regardless of white-spaces.

they are in fact built upon language-specific sets 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 side-stepping 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*, which 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 number 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 (see Figure 1).

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 that 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 modeling, leaving ample space to test 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 for all languages, 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* (gender, number, case, etc.) and their possible *values*

lexeme=PENDA	IND		IND;PERF		COND	
PRS;DECL;NOM(2,SG)	POS	NEG	POS	NEG	POS	NEG
ACC(1,SG)	unanipenda	hunipendi	umenipenda	hujanipenda	ungenipenda	usingenipenda
ACC(1,PL)	unatupenda	hutupendi	umetupenda	hujatupenda	ungetupenda	usingetupenda
ACC(2,SG,RFLX)	unajipenda	hujipendi	umejipenda	hujajipenda	ungejipenda	usingejipenda
ACC(2,PL)	unawapendeni	huwapendini	umewapendeni	hujawapendeni	ungewapendeni	usingewapendeni
ACC(3,SG)	unampenda	humpendi	umempenda	hujampenda	ungempenda	usingempenda
ACC(3,PL)	unawapenda	huwapendi	umewapenda	hujawapenda	ungewapenda	usingewapenda

(a) Swahili inflection table

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

(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 number of words needed to realize each cell. In practice, we did not inflect English clauses for number in 2nd person, so we did not use the *y'all* pronoun and it is given here for the illustration.

(e.g., gender:{masculine,feminine,neuter}). A specific *attribute:value* pair defines an *inflectional feature* (henceforth, a *feature*) and a specific combination of features is called an *inflectional feature bundle* (here, a *feature bundle*). An inflection table includes, for a given lexeme l_i, an exhaustive list of m inflected word-forms $\{w^{l_i}_{b_j}\}^m_{j=0}$, corresponding to all available *feature bundles* $\{b_j\}^m_{j=0}$. 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 b_j that specifies the target word-form. The output is the inflected word-form $w^{l_i}_{b_j}$ realizing the feature bundle. Example (1b) is an example in the French verbal paradigm for the lemma *finir*, inflected to an indicative IND future tense FUT with a 1st person singular subject 1;SG.

(1) a. $\langle l_i, b_j \rangle \mapsto w^{l_i}_{b_j}$

b. $\langle$ *finir*, IND;FUT;1;SG$\rangle \mapsto$ *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 $w^{l_i}_{b_j}$ from some lexeme l_i is given as input accompanied by its own feature bundle b_j, and the model reinflects it to a different feature bundle b_k, resulting in the word $w^{l_i}_{b_k}$ (2a). In (2b) we illustrate for the same French lemma *finir*, a reinflection from the indicative present tense with a first person singular subject *'finis'* to the subjunctive past and second person singular *'finisses'*.

(2) a. $\langle b_j, w^{l_i}_{b_j} \rangle, \langle b_k, \_\_\_\rangle \mapsto w^{l_i}_{b_k}$

b. $\begin{matrix} \langle \text{ IND;PRS;1;SG, } \textit{finis} \rangle, \\ \langle \text{ SBJV;PST;2;SG, } \_\_\_ \rangle \end{matrix} \mapsto$ *finisses*

[1]The 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).

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) a. $w^{l_i}_{b_j} \mapsto \langle l_i, b_j \rangle$

b. *finirai* $\mapsto \langle$*finir*, IND;FUT;1;SG$\rangle$

2.2 UniMorph

The most significant source of inflection tables for training and evaluating all of the aforementioned tasks is UniMorph[2] (Sylak-Glassman et al., 2015; Batsuren et al., 2022), 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 $\langle l,b,w \rangle$ with *lemma* l, a *feature bundle* b, and a *word-form* w. 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 Wiktionary[4] 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. 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 closely related languages in the same language family, for example, 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. On the other hand, a decades-long debate about a 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 a consistent feature set to inflect lexemes 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 $\{b_j\}$ 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

[2] https://unimorph.github.io.

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

[4] https://www.wiktionary.org.

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 feature types expressed at word level in *some* language. This set essentially defines a *saturated clause*.

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 in most European languages, *negation* is expressed at the word level in many Bantu languages (Wilkes and Nkosi, 2012; Mpiranya, 2014), and *interrogativity*—in, for example, 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:

(4) a. Georgian: გაგიშვებთ
Trans: ''**we** will let **you** go''
IND;FUT;NOM(1,PL);ACC(2,SG)

b. Spanish: *dímelo*
Trans: ''tell **it** to **me**''
IMP;NOM(2,SG);ACC(3,SG,NEUT);DAT(1,SG)

c. Basque: *dakarkiogu*
Trans: ''**we** bring **it** to **him/her**''
IND;PRS;ERG(1,PL);ABS(3,SG);DAT(3,SG)

Following Anderson's (1992) feature layering approach, we propose the annotation of arguments to be done as complex features, that is, features that allow a feature set as their value.[5] So, the Spanish verb form *dímelo* (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 use personal pronouns to realize the relevant feature-bundles, for example, 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 in this work, we also limit 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's (1992, p. 219) 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''. Thus, our suggested feature set is not only diverse and inclusive in practice, it is also theoretically sound.[6]

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 our proposal, We amend the task definitions in Section 2 to refer to forms in general $f^{l_i}_{b_j}$ rather than words $w^{l_i}_{b_j}$:

(5) Clause-Level Morphological Tasks

a. inflection $\langle l_i, b_j \rangle \mapsto f^{l_i}_{b_j}$

b. reinflection $\langle b_j, f^{l_i}_{b_j} \rangle, \langle b_k, \_\_ \rangle \mapsto f^{l_i}_{b_k}$

c. analysis $f^{l_i}_{b_j} \mapsto \langle l_i, b_j \rangle$

See Table 2 for detailed examples of these tasks for all the languages included in this work.

[5] This is reminiscent of feature structures in Unification Grammars (Shieber, 2003) such 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 holds: When adding a new language with new word-level features, these features will be realized for all languages.

	Input		Output
Eng	give	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	I will give him to her
Deu	geben	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	ich werde ihn ihr geben
Tur	vermek	IND;FUT;NOM(1,SG);ACC(3,SG);DAT(3,SG)	onu ona vereceğim
Heb	נתן	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	אתן אותו לה
Heb$_{voc}$	נָתַן	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	אֶתֵּן אוֹתוֹ לָהּ

(a) Inflection examples

	Input		Output
Eng	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	I will give him to her	
	IND;PRS;NOM(1,PL);ACC(2);DAT(3,PL);NEG		we don't give you to them
Deu	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	ich werde ihn ihr geben	
	IND;PRS;NOM(1,PL);ACC(2,SG);DAT(3,PL);NEG		wir geben dich ihnen nicht
Tur	IND;FUT;NOM(1,SG);ACC(3,SG);DAT(3,SG)	onu ona vereceğim	
	IND;PRS;PROG;NOM(1,PL);ACC(2,SG);DAT(3,PL);NEG		seni onlara vermiyoruz
Heb	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	אתן אותו לה	
	IND;PRS;NOM(1,PL,MASC);ACC(2,SG,MASC);DAT(3,PL,FEM);NEG		אנחנו לא נותנים אותך להן
Heb$_{voc}$	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)	אֶתֵּן אוֹתוֹ לָהּ	
	IND;PRS;NOM(1,PL,MASC);ACC(2,SG,MASC);DAT(3,PL,FEM);NEG		אֲנַחְנוּ לֹא נוֹתְנִים אוֹתָךְ לָהֶן

(b) Reinflection examples

	Input	Output	
Eng	I will give him to her	give	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)
Deu	ich werde ihn ihr geben	geben	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)
Tur	onu ona vereceğim	vermek	IND;FUT;NOM(1,SG);ACC(3,SG);DAT(3,SG)
Heb	אתן אותו לה	נתן	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)
Heb$_{voc}$	אֶתֵּן אוֹתוֹ לָהּ	נָתַן	IND;FUT;NOM(1,SG);ACC(3,SG,MASC);DAT(3,SG,FEM)

(c) Analysis examples

Table 2: Examples for the data format used for the inflection, reinflection, and analysis tasks.

4 The MIGHTYMORPH Benchmark

We present MIGHTYMORPH, the first multilingual clause-level morphological dataset. Like UniMorph, MIGHTYMORPH 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 four languages from three language families: English, German, Turkish, and Hebrew.[7] 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 at least 500 inflection tables.

Our data is currently limited to clauses constructed from verbal lemmas, as these are typical clause heads. Reserved for future work is the expansion of the process described below to nominal and adjectival clauses.

4.1 Data Creation

The data creation process, for any language, can be characterized by three conceptual components: (i) Lexeme Sampling, (ii) Periphrastic Construction, and (iii) Argument Marking. We describe each of the different phases in turn. As a running example, we will use the English verb *receive*.

Lexeme Sampling. To create MIGHTYMORPH, we first sampled frequently used verbs from UniMorph. We assessed the verb usage by the position of the lemma in the frequency-ordered vocabulary of the FastText word vectors (Grave et al., 2018).[8] We excluded auxiliaries and any lemmas frequent due to homonymy with non-verbal lexemes.

[7] For Hebrew, we annotated a vocalized version in addition to the commonly used unvocalized forms.

[8] https://fasttext.cc/docs/en/crawl-vectors.html.

Periphrastic Constructions We expanded each verb's word-level inflection table to include all periphrastic constructions using a language-specific rule-based grammar we wrote and the inflection tables of any relevant auxiliaries. That is, we constructed forms for all possible TAM combinations expressible in the language, regardless of the number of words used to express this combination of features. For example, when constructing the future perfect form with a 3rd person singular subject for the lexeme *receive*, equivalent to IND;FUT;PRF;NOM(3,SG), we used the past participle from the UniMorph inflection table *received* and the auxiliaries *will* and *have* to construct *will have received*.

Argument Marking At first, we added the pronouns that the verb agrees with, unless a pro-drop applies. For all languages in our selection, the verb agrees only with its subject. A place-holder was then added to mark the linear position of the rest of the arguments. So the form of our example is now *he will have received ARGS*.

In order to obtain a fully saturated clause, but also not to over-generate redundant arguments—for example, a transitive clause for an intransitive verb—an exhaustive list of *frames* for each verb is needed. The frames are lists of cased arguments that the verb takes. For example, the English verb *receive* has 2 frames, {NOM, ACC} and {NOM, ACC, ABL}, where an accusative argument indicates theme and the an ablative argument marks the source. When associating verbs with their arguments we did not restrict ourselves to the distinction between intransitive, transitive, and ditransitive verbs, we allow arguments of any case. We treated all argument types equally and annotated them with a case feature, whether expressed with an affix, an adposition, or a coverb. Thus, English *from you*, Turkish *senden*, and Swahili *kutoka kwako* are all tagged with an ablative case feature ABL(2,SG).

For each frame we exhaustively generated all suitably cased pronouns without regarding the semantic plausibility of the resulted clause. So the clause *he will have received you from it* is in the inflection table since it is grammatical—even though it sounds odd. In contrast, *he will have received* is not in the inflection table, as it is strictly ungrammatical, missing (at least) one obligatory argument.

Notably, we excluded adjuncts from the possible frames, defined here as argument-like elements that can be added to all verbs without regards to their semantics, like beneficiary and location.

We manually annotated 500 verbs in each language with a list of frames, each listing 0 or more arguments. This is the only part of the annotation process that required manual treatment of individual verbs.[9] It was done by the authors, with the help of a native speaker or a monolingual dictionary.[10]

We built an annotation framework that delineates the different stages of the process. It includes an infrastructure for grammar description and an interactive frame annotation component. Given a grammar description, the system handles the sampling procedure and constructs all relevant periphrastic constructions while leaving an additional-arguments place-holder. After receiving the frame-specific arguments from the user, the system completes the sentence by replacing the place holder with all prespecified pronouns for the frame. The framework can be used to speed up the process of adding more languages to MIGHTYMORPH.[11] Using this framework, we have been able to annotate 500 verb frames in about 10 hours per language on average.

4.2 The Annotation Schema

Just as our data creation builds on the word-level inflection tables of UniMorph and expands them, so our annotation schema is built upon UniMorph's.

In practice, due to the fact that some languages do use a single word for a fully saturated clause, we could simply apply the UniMorph annotation guidelines (Sylak-Glassman, 2016) both as an inventory of features and as general guidelines for the features' usage. Adhering to these guidelines ensures that our approach is able to cover essentially all languages covered by UniMorph. In addition, we extended the schema with the layering mechanism described in Section 3 and by

[9] Excluding the manual work that may have been put in constructing the UniMorph inflection tables to begin with.

[10] For German we used Duden dictionary, and for Turkish we used the Türk Dil Kurumu dictionary.

[11] The data and annotation scripts are available at https://github.com/omagolda/mighty_morph_tagging_tool.

	Table size		Feat set size		Feats per form		Form length	
	UM	MM	UM	MM	UM	MM	UM	MM
Eng	5	450	6	32	2.8	12.75	6.84	29.63
Deu	29	512	12	43	4.62	12.67	9.18	31.28
Heb	29	132	13	25	4.46	13.55	5.20	20.47
Heb$_{voc}$	29	132	13	25	4.46	13.55	9.80	32.02
Tur	703	702	25	30	7.87	11.95	17.81	28.71

Table 3: Comparison of statistics over the 4 languages common to UniMorph (UM) and MIGHTYMORPH (MM). In all cases, the values for MIGHTYMORPH are more uniform across languages.

Guriel et al. (2022), and officially adopted as part of the UniMorph schema by Batsuren et al. (2022).

See Table 4 for a detailed list of features used.

4.3 Data Analysis

The MIGHTYMORPH benchmark represents inflectional morphology in four typologically diverse languages, yet the data is both more uniform across languages and more diverse in the features realized for each language, compared with the de facto standard word-level morphological annotations.

Table 3 compares aggregated values between UniMorph and MIGHTYMORPH across languages: the inflection table size,[12] the number of unique features used, the average number of features per form, and the average form-length in characters.

We see that MIGHTYMORPH is more cross-lingually consistent than UniMorph on all four comparisons: The size of the tables is less varied, so English no longer has extraordinarily small tables; the sets of features that were used per language are very similar, due to the fact that they all come from a fixed inventory; and finally, forms in all languages are of similar character length and are now described by feature bundles whose feature length are also highly similar. The residual variation in all of these values arises only from true linguistic variation. For example, Hebrew does not use features for aspects as Hebrew does not express verbal aspect at all. This is a strong empirical indication that applying morphological annotation to clauses reintroduces universality into morphological data.

In addition, the bigger inflection tables in MIGHTYMORPH include phenomena more diverse, 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 more universal than the previously studied word-level morphology.

[12] Since the table size is dependent on the transitivity of the verb, the clause level is compared to an intransitive table.

5 Experiments

Goal We set out to assess the challenges and opportunities presented to contemporary models by 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 the clause-level.

Splits For each task we sampled from 500 inflection tables 10,000 examples (pairs of examples in the case of reinflection). We used 80% of the examples for training and the rest was divided between the validation and test sets. We sampled the same number of examples from each table and, following Goldman et al. (2022), we split the data such that the lexemes in the different sets are disjoint. So, 400 lexemes are used in the train set, and 50 are for each of the validation and test sets.

Models As baselines, we applied contemporary models designed for word-level morphological tasks (henceforth: word-level models). The application of word-level models will allow us to assess the difficulty of the clause-level tasks comparing to their word-level counterparts. These models generally handle characters as input and output, and we applied them to clause-level tasks straightforwardly by treating white-space as yet another character rather than a special delimiter. For each language and task we trained a separate model for 50 epochs. The word-level models we trained are:

- LSTM: An 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 using sparsemax instead of softmax, by Peters and Martins (2020).

Attribute		Value
Tense		PST(past),PRS(present),FUT(future)
Mood		IND(indicative) IMP(imperative) SBJV(subjunctive) INFR†(inferential) NEC†(necessitative) COND(conditional) QUOT(quotative)
Aspect		HAB(habitual) PROG(progressive) PRF(perfect) PRSP(prospective)
Non-locative Cases		NOM(nominative) ACC(accusative) DAT(dative) GEN(genitive) COM(comitative) BEN(benefactive)
Locative Cases		LOC†(general locative) ABL(ablative) ALL(allative) ESS(essive) APUD(apudessive) PERL†(perlative) CIRC(near) ANTE(in front) CONTR†(against) AT(at, general vicinity) ON(on) IN(in) VON†(about)
Sentence Features		NEG(negative) Q(interrogative)
Argument Features	Person	1(1st person) 2(2nd person) 3(3rd person)
	Number	SG(singular) PL(plural)
	Gender	MASC(masculine) FEM(feminine) NEUT(neuter)
	Misc.	FORM(formal) RFLX†(reflexive)

Table 4: A list of all features used in constructing the data for the 4 languages in MIGHTYMORPH. Upon addition of new languages the list would expand. Features not taken from Sylak-Glassman (2016) are marked with †.

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 (Goldman et al., 2022). We modified TRANSDUCE to apply to reinflection, while only the generally designed LSTM could be used for all tasks.

In contrast with word-level tasks, the extension of morphological tasks to the clause-level introduces *context* of a complete sentence, which provides an opportunity to explore the benefits of pre-trained *contextualized* LMs. Success of such models on many NLP tasks calls for investigating their performance in our setting. We thus used the following pretrained text-to-text model as an advanced modeling alternative for our clause-level tasks:

- 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 used as a prompt and were added to the model's vocabulary as new tokens with randomly initialized embeddings.[13]

[13]As none of the models were designed to deal with hierarchical feature structures, the features' strings were flattened before training and evaluation. For example, the bundle IND;PRS;NOM(1,SG);ACC(2,PL) is replaced with IND;PRS;NOM1;NOMSG;ACC2;ACCPL.

5.1 Results and Analysis

Table 5 summarizes the results for all models and all tasks, for all languages. When averaged across languages, the results for the inflection task show a drop in performance for the word-inflection models (LSTM, DEEPSPIN, and TRANSDUCE) on clause-level tasks, indicating that the clause-level task variants are indeed more challenging. This pattern is even more pronounced in the results for the reinflection task which seems to be the most challenging clause-level task, presumably due to the need to identify the lemma in the sequence, in addition to inflecting it. In the analysis task, the only word-level model, LSTM, actually performs better on the clause level than on the word level, but this seems to be the effect one outlier language, namely, unvocalized Hebrew, where analysis models suffer from the lack of diacritization and extreme ambiguity.

Moving from words to clauses introduces context, and we hypothesized that this would enable contextualized pretrained LMs to shine. However, on all tasks MT5 did not prove itself to be a silver bullet. That said, the strong pretrained model performed on par with the other models on the challenging reinflection task—the only task involving complete sentences on both input and output—in accordance with the model's pretraining.

In terms of languages, the performance of the word-level models seems correlated across languages, with notable under-performance over all tasks in German. In contrast, MT5 seems to be

		Average		Eng		Deu		Heb		$\text{Heb}_{vocalized}$		Tur	
		word	clause	word	clause	word	clause	word	clause	word	clause	word	clause
inflec.	LSTM	84.7 ±1.1	70.0 ±1.2	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
	DEEPSPIN	89.4 ±0.8	71.8 ±0.5	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
	TRANSDUCE	86.7 ±0.5	**78.9 ±0.4**	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	60.0 ±1.1	99.4 ±0.1	**97.2 ±0.5**
	MT5	NA	51.9 ±1.1	NA	70.7 ±1.7	NA	57.7 ±3.3	NA	48.0 ±3.3	NA	34.2 ±1.4	NA	48.7 ±1.7
reinflec.	LSTM	73.2 ±1.6	**45.4 ±9.4**	78.2 ±6.3	62.7 ±2.5	53.5 ±3.5	31.0 ±1.7	68.4 ±1.6	30.6 ±29.8	80.7 ±1.9	**31.4 ±36.4**	85.2 ±2.2	**71.1 ±1.2**
	TRANSDUCE	75.1 ±0.5	**44.5 ±0.8**	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	6.1 ±1.8	84.7 ±1.1	**72.5 ±2.7**
	MT5	NA	**45.2 ±1.8**	NA	**73.6 ±3.1**	NA	**54.2 ±2.0**	NA	30.8 ±4.2	NA	**29.7 ±1.9**	NA	37.5 ±7.0
analysis	LSTM	62.0 ±0.9	**64.4 ±1.1**	81.6 ±0.4	**79.5 ±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 ±0.7	**84.4 ±4.5**
	MT5	NA	42.8 ±1.2	NA	69.0 ±1.2	NA	**45.1 ±2.7**	NA	48.0 ±3.3	NA	34.2 ±1.4	NA	46.0 ±4.5

Table 5: 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 **bold**, in cases that are too close to call, in terms of standard deviations, all best systems are marked. Results are averaged over 3 runs with different initializations and training data order.

somewhat biased towards the western languages, English and German, especially in the generation tasks, inflection, and reinflection.

Data Sufficiency To illustrate how much labeled data should *suffice* for training clause-morphology models, 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 number of inflection tables; alternatively, the number of inflection tables can be increased for a fixed size of the training set, increasing not the size but the variation in the 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 variety in lexemes may lead to overfitting. To examine which dimension is more important for the overall success in the tasks, we tested both.

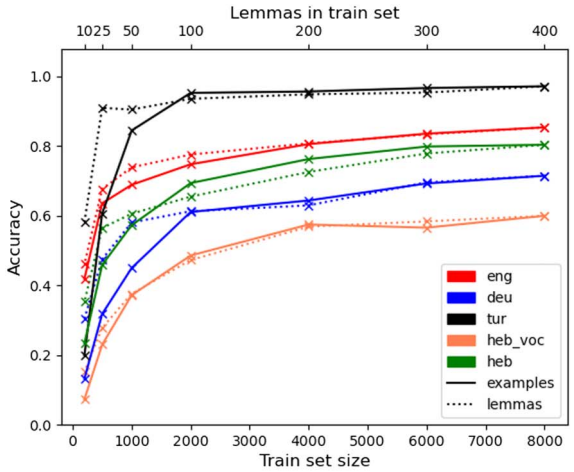
The resulting curves are provided in Figure 2. In each sub-figure, the solid lines are for the results as the absolute train set size is increased, and the dashed lines are for increasing the number of lexemes in the train set while keeping the absolute size of the train set fixed.

The resulting curves show that the balance between the options is different for each task. For inflection (Figure 2a), 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 2b) 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 number of lexemes added. On the other hand, the trend for analysis (Figure 2c) is the other way around, with increased train set size being more critical than increased lexeme-variance.

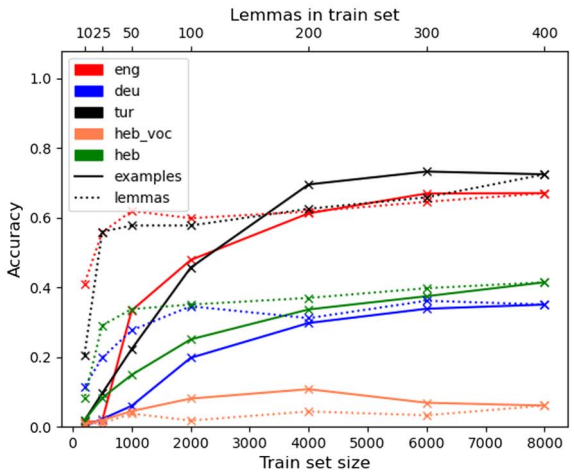
6 Related Work

6.1 Wordhood in Linguistic Theory

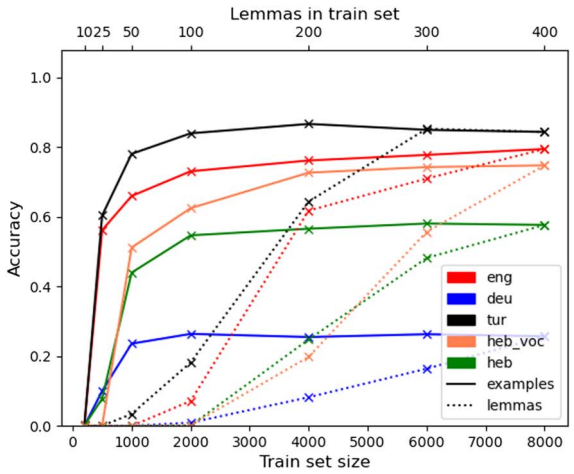
The quagmire surrounding words and their demarcation is long-standing in theoretical linguistics. In fact, no coherent word definition has been provided by the linguistic literature despite many attempts. For example, Zwicky and Pullum (1983) enumerate 6 different, sometimes contradictory, ways to discern between words, clitics, and morphemes. Haspelmath (2011) names 10 criteria for wordhood before concluding that no



(a) Task: inflection; model: TRANSDUCE



(b) Task: reinflection; model: TRANSDUCE



(c) Task: analysis; model: LSTM

Figure 2: Learning curves for the best performing model on each task. Solid lines are for increasing train set sizes while dashed lines are for using more lexemes.

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. This approach is theoretically leaner and requires less decisions that may be controversial, for example, regarding morpheme boundaries, empty morphemes, and the like.

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.

6.2 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., Richens and Booth, 1952; 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 connect 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 (Bārzdiņš 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, and this work can provide a stepping-stone for such a solution.

7 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. Here we discuss some of the limitations and opportunities of this work for the future of morpho(syntactic) models in NLP.

The Derivation-inflection Divide. Our definition or clause-level morphology does not solve the long-debated demarcation of boundary between inflectional and derivational morphology (e.g., Scalise, 1988). Specifically, we only refered here to inflectional features, and, like UniMorph, did not provide a clear definition of what counts as inflectional vs. derivational. However, we suggest here that the lack of a clear boundary between inflectional and derivational morphology is highly similar to the lack of definition for words that 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 to ours, that will define a single framework for the entire inflectional–derivational morphology continuum without positing a boundary between them.

Overabundance. Our shift to clause-level morphology does not solve the problem of overabundance, where several forms are 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, a further extension of the inflection table is conceivable, to accommodate sets of forms in every cell, rather than a single one.

Implications to Syntax. Our solution for annotating morphology at the clause level blurs the boundary between morphology and syntax as it is often presupposed in NLP, and thus has implications also for *syntactic* tasks. Some previous studies 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 syntax for inter-clausal operations. 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. These are all fascinating research directions that extend the present contribution, and we reserve them for future work.

Polysynthetic Languages. As a final note, we wish to make the observation that a unified morpho-syntactic system, whose desiderata are laid out in the previous paragraph, is essential for providing a straightforward treatment of some highly polysynthetic languages, specifically those that employ noun incorporation to regularly express some multi-lexemed clauses as a single word.

For example, consider the Yupik clause *Mangteghangllaghyugtukut* translated *We want to make a house*[14] containing 3 lexemes. Its treatment with the current syntactic tools is either non-helpful, as syntax only characterizes inter-word relations, or requires ad hoc morpheme segmentation not used in other types of languages. Conversely, resorting to morphological tools will also provide no solution, due to the lexeme–inflection table paradigm that assumes single-lexemed words. With a single morpho-syntactic framework, we could annotate the example above by incorporating the lemmas into their respective positions on the nested feature structure we used in this work, ending up with something similar to *yug*;IND;ERG(1;PL);COMP(*ngllagh*;ABS(*-mangtegha*;INDEF)). Thus, an annotation of this kind can expose the predicate-argument structure of the sentence while also being naturally applicable to other languages.

Equipped with these extensions, our approach could elegantly deal with polysynthetic languages and unlock a morpho-syntactic modeling ability that is most needed for low-resourced languages.

[14]Example adopted from Yupik UD (Park et al., 2021).

8 Conclusions

In this work we expose the fundamental inconsistencies in contemporary computational morphology, namely, the inconsistency of *wordhood* across languages. To remedy this, we deliver MIGHTYMORPH, the first labeled dataset for clause-level morphology. We derive training and evaluation data for the 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 universality into computational morphology. Moreover, we showed that standard (re)inflection models struggle on the clause-level compared to their performance on word-level tasks, and that the challenge is not trivially solved, even by contextualized pretrained LMs such as MT5. In the future we intend to further expand our framework for more languages, and to explore more sophisticated models that take advantage of the hierarchical structure or better utilize pretrained LMs. Moreover, future work is planned to expand the proposal and benchmark to the inclusion of derivational morphology, and to a unified morpho-syntactic framework.

Acknowledgments

We would like to thank the TACL anonymous reviewers and the action editor for their insightful suggestions and remarks. This work was supported funded by an ERC-StG grant from the European Research Council, grant number 677352 (NLPRO), and by an innovation grant by the Ministry of Science and Technology (MOST) 0002214, for which we are grateful.

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. https://doi.org/10.1007/978-94-011-3474-3_10

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. https://doi.org/10.1093/acprof:oso/9780199547548.003.0003

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. https://doi.org/10.1017/CBO9780511586262

Guntis Bārzdiņš, 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. https://doi.org/10.18653/v1/2021.eacl-main.117

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina J. Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. Unimorph 4.0: Universal morphology. In *Proceedings of the 13th Language Resources and Evaluation Conference*.

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. https://doi.org/10.18653/v1/2020.findings-emnlp.414

Joan Bresnan, Ash Asudeh, Ida Toivonen, and Stephen Wechsler. 2015. *Lexical-Functional Syntax*. John Wiley & Sons. https://doi.org/10.1002/9781119105664

Joan Bresnan and Sam A. Mchombo. 1995. The lexical integrity principle: Evidence from Bantu. *Natural Language & Linguistic Theory*, 13(2):181–254. https://doi.org/10.1007/BF00992782

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. https://doi.org/10.1075/tsl.9

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. https://doi.org/10.18653/v1/K17-2001

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. https://doi.org/10.18653/v1/W16-2002

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308. https://doi.org/10.1162/coli_a_00402

Robert M. W. Dixon and Alexandra Y. Aikhenvald. 2002. Word: a typological framework. *Word: A Cross-Linguistic Typology*, pages 1–41. https://doi.org/10.1017/CBO9780511486241.002

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. https://doi.org/10.2307/2185122

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*, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.96

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David Guriel, Omer Goldman, and Reut Tsarfaty. 2022. Morphological reinflection with multiple arguments: An extended annotation schema and a georgian case study. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.21

T. Alan Hall. 1999. The phonological word: a review. *Studies on the Phonological Word*, 174(1):22. https://doi.org/10.1075/cilt.174.02hal

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. https://doi.org/10.1515/flin.2011.002

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. https://doi.org/10.18653/v1/2021.acl-long.279

Paul J. Hopper and Elizabeth Closs Traugott. 2003. *Grammaticalization*. Cambridge University Press. https://doi.org/10.1017/CBO9781139165525

Brian D. Joseph. 2003. Morphologization from syntax. *Handbook of Historical Linguistics*, pages 472–492. https://doi.org/10.1002/9780470756393.ch13

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. https://doi.org/10.18653/v1/P18-1007

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. https://doi.org/10.18653/v1/2020.acl-main.703

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.

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.

Fidèle Mpiranya. 2014. *Swahili Grammar and Workbook*. Routledge. https://doi.org/10.4324/9781315750699

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. https://doi.org/10.3115/1220355.1220422

Muriel Norde et al. 2009. *Degrammaticalization*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199207923.001.0001

Jerome L. Packard. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press. https://doi.org/10.1017/CBO9780511486821

Hyunji Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.14

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. https://doi.org/10.18653/v1/2020.sigmorphon-1.4

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. https://doi.org/10.18653/v1/2021.sigmorphon-1.25

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

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.

Richard H. Richens and A. Donald Booth. 1952. Some methods of mechanized translation. In *Proceedings of the Conference on Mechanical Translation*.

Sergio Scalise. 1988. Inflection and derivation. *Linguistics*, 26(4):561–582. https://doi.org/10.1515/ling.1988.26.4.561

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. https://doi.org/10.1109/ICASSP.2012.6289079

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. https://doi.org/10.18653/v1/P16-1162

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. https://doi.org/10.18653/v1/D18-1315

John Sylak-Glassman. 2016. The composition and use of the universal morphological feature schema (unimorph schema). Technical report, Department of Computer Science, Johns Hopkins University.

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 and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-2111

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, Project MAC, MIT.

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: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Arnold M. Zwicky and Geoffrey K. Pullum. 1983. Cliticization vs. inflection: English n't. *Language*, 59(3):502–513. https://doi.org/10.2307/413900