Marrying Universal Dependencies and Universal Morphology

Arya D. McCarthy[1], **Miikka Silfverberg**[2], **Ryan Cotterell**[1],
Mans Hulden[2], and **David Yarowsky**[1]

[1]Johns Hopkins University
[2]University of Colorado Boulder
{arya,rcotter2,yarowsky}@jhu.edu
{miikka.silfverberg,mans.hulden}@colorado.edu

Abstract

The Universal Dependencies (UD) and Universal Morphology (UniMorph) projects each present schemata for annotating the morphosyntactic details of language. Each project also provides corpora of annotated text in many languages—UD at the token level and UniMorph at the type level. As each corpus is built by different annotators, language-specific decisions hinder the goal of universal schemata. With compatibility of tags, each project's annotations could be used to validate the other's. Additionally, the availability of both type- and token-level resources would be a boon to tasks such as parsing and homograph disambiguation. To ease this interoperability, we present a deterministic mapping from Universal Dependencies v2 features into the UniMorph schema. We validate our approach by lookup in the UniMorph corpora and find a macro-average of 64.13% recall. We also note incompatibilities due to paucity of data on either side. Finally, we present a critical evaluation of the foundations, strengths, and weaknesses of the two annotation projects.

1 Introduction

The two largest standardized, cross-lingual datasets for morphological annotation are provided by the Universal Dependencies (UD; Nivre et al., 2017) and Universal Morphology (UniMorph; Sylak-Glassman et al., 2015; Kirov et al., 2018) projects. Each project's data are annotated according to its own cross-lingual schema, prescribing how features like gender or case should be marked. The schemata capture largely similar information, so one may want to leverage both UD's token-level treebanks and UniMorph's type-level lookup tables and unify the two resources. This would permit a leveraging of both the token-level UD treebanks and the type-level UniMorph tables of paradigms. Unfortunately, neither resource perfectly realizes its schema. On a dataset-by-dataset basis, they incorporate annotator errors, omissions, and human decisions when the schemata are underspecified; one such example is in Figure 1.

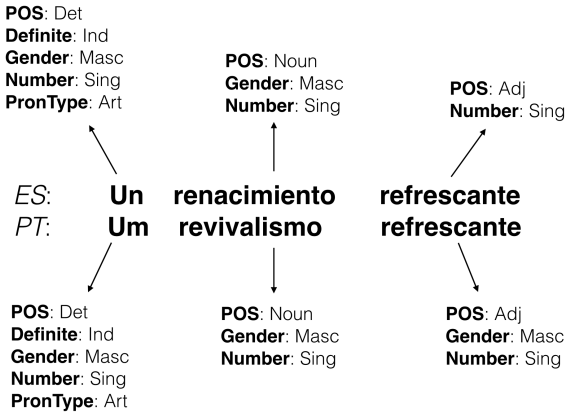

Figure 1: Example of annotation disagreement in UD between two languages on translations of one phrase, reproduced from Malaviya et al. (2018). The final word in each, "*refrescante*", is not inflected for gender: It has the same surface form whether masculine or feminine. Only in Portuguese, it is annotated as masculine to reflect grammatical concord with the noun it modifies.

A dataset-by-dataset problem demands a dataset-by-dataset solution; our task is not to translate a *schema*, but to translate a *resource*. Starting from the idealized schema, we create a rule-based tool for converting UD-schema annotations to UniMorph annotations, incorporating language-specific post-edits that both correct infelicities and also increase harmony between the datasets themselves (rather than the schemata). We apply this conversion to the 31 languages with both UD and UniMorph data, and we report our method's recall, showing an improvement over the strategy which just maps corresponding schematic features to each other. Further, we show similar downstream performance for each annotation scheme in the task of morphological tagging.

Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 91–101
Brussels, Belgium, November 1, 2018.

This tool enables a synergistic use of UniMorph and Universal Dependencies, as well as teasing out the annotation discrepancies within and across projects. When one dataset disobeys its schema or disagrees with a related language, the flaws may not be noticed except by such a methodological dive into the resources. When the maintainers of the resources ameliorate these flaws, the resources move closer to the goal of a universal, cross-lingual inventory of features for morphological annotation.

The contributions of this work are:

- We detail a deterministic mapping from UD morphological annotations to UniMorph. Language-specific edits of the tags in 31 languages increase harmony between converted UD and existing UniMorph data (§5).
- We provide an implementation of this mapping and post-editing, which replaces the UD features in a CoNLL-U file with UniMorph features.[1]
- We demonstrate that downstream performance tagging accuracy on UD treebanks is similar, whichever annotation schema is used (§7).
- We provide a partial inventory of missing attributes or annotation inconsistencies in both UD and UniMorph, a guidepost for strengthening and harmonizing each resource.

2 Background: Morphological Inflection

Morphological **inflection** is the act of altering the base form of a word (the **lemma**, represented in `fixed-width type`) to encode morphosyntactic features. As an example from English, `prove` takes on the **form** "proved" to indicate that the action occurred in the past. (We will represent all surface forms in quotation marks.) The process occurs in the majority of the world's widely-spoken languages, typically through meaningful affixes. The breadth of forms created by inflection creates a challenge of data sparsity for natural language processing: The likelihood of observing a particular word form diminishes.

A classic result in psycholinguistics (Berko, 1958) shows that inflectional morphology is a fully productive process. Indeed, it cannot be that humans simply have the equivalent of a lookup table, where they store the inflected forms for retrieval as the syntactic context requires. Instead, there needs to be a mental process that can generate properly inflected words on demand. Berko (1958) showed this insightfully through the "wug"-test, an experiment where she forced participants to correctly inflect out-of-vocabulary lemmata, such as the novel noun `wug`.

Simple label	Form	PTB tag
Present, 3rd singular	"proves"	VBZ
Present, other	"prove"	VBP
Past	"proved"	VBD
Past participle	"proven"	VBN
Present participle	"proving"	VBG

Table 1: Inflected forms of the English verb `prove`, along with their Penn Treebank tags

Certain features of a word do not vary depending on its context: In German or Spanish where nouns are gendered, the word for `onion` will always be grammatically feminine. Thus, to prepare for later discussion, we divide the morphological features of a word into two categories: the modifiable **inflectional features** and the fixed **lexical features**.

A **part of speech (POS)** is a coarse syntactic category (like "verb") that begets a word's particular menu of lexical and inflectional features. In English, verbs express no gender, and adjectives do not reflect person or number. The part of speech dictates a set of inflectional **slots** to be filled by the surface forms. Completing these slots for a given lemma and part of speech gives a **paradigm**: a mapping from slots to surface forms. Regular English verbs have five slots in their paradigm (Long, 1957), which we illustrate for the verb `prove`, using simple labels for the forms in Table 1.

A morphosyntactic **schema** prescribes how language can be annotated—giving stricter categories than our simple labels for `prove`—and can vary in the level of detail provided. Part of speech tags are an example of a very coarse schema, ignoring details of person, gender, and number. A slightly finer-grained schema for English is the Penn Treebank tagset (Marcus et al., 1993), which includes signals for English morphology. For instance, its VBZ tag pertains to the specially inflected 3rd-person singular, present-tense verb form (e.g. "proves" in Table 1).

If the tag in a schema is detailed enough that it exactly specifies a slot in a paradigm, it is

[1] Available at `https://www.github.com/unimorph/ud-compatibility`.

called a **morphosyntactic description (MSD)**.[2] These descriptions require varying amounts of detail: While the English verbal paradigm is small enough to fit on a page, the verbal paradigm of the Northeast Caucasian language Archi can have over 1,500,000 slots (Kibrik, 1998).

3 Two Schemata, Two Philosophies

Unlike the Penn Treebank tags, the UD and UniMorph schemata are cross-lingual and include a fuller lexicon of attribute-value pairs, such as PERSON: 1. Each was built according to a different set of principles. UD's schema is constructed bottom-up, adapting to include new features when they're identified in languages. UniMorph, conversely, is top-down: A cross-lingual survey of the literature of morphological phenomena guided its design. UniMorph aims to be linguistically complete, containing all known morphosyntactic attributes. Both schemata share one long-term goal: a total inventory for annotating the possible morphosyntactic features of a word.

3.1 Universal Dependencies

The Universal Dependencies morphological schema comprises part of speech and 23 additional attributes (also called features in UD) annotating meaning or syntax, as well as language-specific attributes. In order to ensure consistent annotation, attributes are included into the general UD schema if they occur in several corpora. Language-specific attributes are used when only one corpus annotates for a specific feature.

The UD schema seeks to balance language-specific and cross-lingual concerns. It annotates for both inflectional features such as case and lexical features such as gender. Additionally, the UD schema annotates for features which can be interpreted as derivational in some languages. For example, the Czech UD guidance uses a COLL value for the NUMBER feature to denote mass nouns (for example, "*lidstvo*" "humankind" from the root "*lid*" "people").[3]

UD represents a confederation of datasets (see, e.g., Dirix et al., 2017) annotated with dependency relationships (which are not the focus of this work) and morphosyntactic descriptions. Each dataset is an annotated treebank, making it a resource of **token-level** annotations. The schema is guided by these treebanks, with feature names chosen for relevance to native speakers. (In §3.2, we will contrast this with UniMorph's treatment of morphosyntactic categories.) The UD datasets have been used in the CoNLL shared tasks (Zeman et al., 2017, 2018 to appear).

3.2 UniMorph

In the Universal Morphological Feature Schema (UniMorph schema, Sylak-Glassman, 2016), there are at least 212 values, spread across 23 attributes. It identifies some attributes that UD excludes like information structure and deixis, as well as providing more values for certain attributes, like 23 different noun classes endemic to Bantu languages. As it is a schema for marking morphology, its part of speech attribute does not have POS values for punctuation, symbols, or miscellany (PUNCT, SYM, and X in Universal Dependencies).

Like the UD schema, the decomposition of a word into its lemma and MSD is directly comparable across languages. Its features are informed by a distinction between **universal categories**, which are widespread and psychologically "real" to speakers; and **comparative concepts**, only used by linguistic typologists to compare languages (Haspelmath, 2010). Additionally, it strives for identity of meaning across languages, not simply similarity of terminology. As a prime example, it does not regularly label a dative case for nouns, for reasons explained in depth by Haspelmath (2010).[4]

The UniMorph resources for a language contain complete paradigms extracted from Wiktionary (Kirov et al., 2016, 2018). Word **types** are annotated to form a database, mapping a lemma–tag pair to a surface form. The schema is explained in detail in Sylak-Glassman (2016). It has been used in the SIGMORPHON shared task (Cotterell et al., 2016) and the CoNLL–SIGMORPHON shared tasks (Cotterell et al., 2017, 2018). Several components of the UniMorph schema have been adopted by UD.[5]

[2] Other sources will call this a morphological tag or bundle. We avoid the former because of the analogy to POS tagging; a morphological tag is not atomic.

[3] Note that NUMBER: COLL does not actually figure in the Czech corpus.

[4] "The Russian Dative, the Korean Dative, and the Turkish Dative are similar enough to be called by the same name, but there are numerous differences between them and they cannot be simply equated with each other. Clearly, their nature is not captured satisfactorily by saying that they are instantiations of a crosslinguistic category 'dative'." (Haspelmath, 2010)

[5] http://universaldependencies.org/v2/features.html#comparison-with-unimorph

Schema	Annotation
UD	VERB MOOD=IND|NUMBER=SING|PERSON=3|TENSE=IMP|VERBFORM=FIN
UniMorph	V;IND;PST;1;SG;IPFV
	V;IND;PST;3;SG;IPFV

Table 2: Attested annotations for the Spanish verb form "*mandaba*" "I/he/she/it commanded". Note that UD separates the part of speech from the remainder of the morphosyntactic description. In each schema, order of the values is irrelevant.

3.3 Similarities in the annotation

While the two schemata annotate different features, their annotations often look largely similar. Consider the attested annotation of the Spanish word "*mandaba*" "(I/he/she/it) commanded". Table 2 shows that these annotations share many attributes.

Some conversions are straightforward: VERB to V, MOOD=IND to IND, NUMBER=SING to SG, and PERSON=3 to 3.[6] One might also suggest mapping TENSE=IMP to IPFV, though this crosses semantic categories: IPFV represents the imperfective *aspect*, whereas TENSE=IMP comes from **imperfect**, the English name often given to Spanish's *pasado continuo* form. The imperfect is a verb form which combines both past tense and imperfective aspect. UniMorph chooses to split this into the atoms PST and IPFV, while UD unifies them according to the familiar name of the tense.

4 UD treebanks and UniMorph tables

Prima facie, the alignment task may seem trivial. But we've yet to explore the humans in the loop. This conversion is a hard problem because we're operating on idealized schemata. We're actually annotating human decisions—and human mistakes. If both schemata were perfectly applied, their overlapping attributes could be mapped to each other simply, in a cross-lingual and totally general way. Unfortunately, the resources are imperfect realizations of their schemata. The cross-lingual, cross-resource, and within-resource problems that we'll note mean that we need a tailor-made solution for each language.

Showcasing their schemata, the Universal Dependencies and UniMorph projects each present large, annotated datasets. UD's v2.1 release (Nivre et al., 2017) has 102 treebanks in 60 languages. The large resource, constructed by independent parties, evinces problems in the goal of a universal inventory of annotations. Annotators may choose to omit certain values (like the coerced gender of *refrescante* in Figure 1), and they may disagree on how a linguistic concept is encoded. (See, e.g., Haspelmath's (2010) description of the dative case.) Additionally, many of the treebanks "were created by fully- or semi-automatic conversion from treebanks with less comprehensive annotation schemata than UD" (Malaviya et al., 2018). For instance, the Spanish word "*vas*" "you go" is incorrectly labeled **GENDER:** FEM|**NUMBER:** PL because it ends in a character sequence which is common among feminine plural nouns. (Nevertheless, the part of speech field for "*vas*" is correct.)

UniMorph's development is more centralized and pipelined.[7] Inflectional paradigms are scraped from Wiktionary, annotators map positions in the scraped data to MSDs, and the mapping is automatically applied to all of the scraped paradigms. Because annotators handle languages they are familiar with (or related ones), realization of the schema is also done on a language-by-language basis. Further, the scraping process does not capture lexical aspects that are not inflected, like noun gender in many languages. The schema permits inclusion of these details; their absence is an artifact of the data collection process. Finally, UniMorph records only exist for nouns, verbs, and adjectives, though the schema is broader than these categories.

For these reasons, we treat the corpora as imperfect realizations of the schemata. Moreover, we contend that ambiguity in the schemata leave the door open to allow for such imperfections. With no strict guidance, it's natural that annotators would take different paths. Nevertheless, modulo annota-

[6] The curious reader may wonder why there are two rows of UniMorph annotation for "*mandaba*", each with a different recorded person. The word displays **syncretism**, meaning that a single form realizes multiple MSDs. UniMorph chooses to mark these separately for the sake of its decomposable representation. As this ambiguity is systematic and pervasive in the language, one can imagine a unified paradigm slot V;IND;PST;{1/3};SG;IPFV (Baerman et al., 2005).

[7] This centralization explains why UniMorph tables exist for only 49 languages, or 50 when counting the Norwegian Nynorsk and Bokmål writing forms separately.

tegarg	**latme**-ye	bad-i	be	ba:q-e	man	**zad**.
Hail	damage-EZ	bad-INDEF PAR	to	garden-EZ	1.S	beat-PST.
"The hail caused bad damage to my garden." *or* "The hail damaged my garden badly."						

Figure 2: Transliterated Persian with a gloss and translation from Karimi-Doostan (2011), annotated in a Persian-specific schema. The light verb construction "*latme zadan*" ("to damage") has been spread across the sentence. Multiword constructions like this are a challenge for word-level tagging schemata.

tor disagreement, we assume that within a particular corpus, one word form will always be consistently annotated.

Three categories of annotation difficulty are missing values, language-specific attributes, and multiword expressions.

Missing values In both schemata, irrelevant attributes are omitted for words to which they do not pertain. For instance, an English verb is not labeled GENDER=NULL; the GENDER attribute is simply excluded from the annotation, making the human-readable representations compact. Unfortunately, in both resources, even relevant attributes are intentionally omitted. A verb's positiveness, activeness, or finiteness can be taken as implicit, and it will be omitted arbitrarily on a language-by-language basis. For instance, in our example in Table 2 only UD tags Spanish finite verbs: VERBFORM=FIN. Not only UniMorph makes such elisions: we note that *neither* resource marks verb forms as active—an action entirely permitted by the schemata. This is one source of discrepancy, both between the projects and across languages within a project, but it is straightforward to harmonize.

Language-specific attributes UD records a set of features that are kept language-specific, including POSITION in Romanian, DIALECT in Russian, and NUMVALUE in Czech and Arabic.[8] UniMorph has (potentially infinite) language-specific features LGSPEC1, LGSPEC2, ..., which are sparsely used but opaque when encountered. For instance, LGSPEC1 in Spanish distinguishes between the two (semantically identical) forms of the imperfect subjunctive: the "-se" and "-ra" forms (e.g. "*estuviese*" and "*estuviera*" from "*estar*" "to be"). UD does not annotate the forms differently. If a language has multiple language-specific attributes, their order is not prescribed by the UniMorph schema, and separate notes that explain the use of such tags must accompany datasets.

Multiword expressions A final imperfection is how to represent multiword constructions. Both UD and UniMorph are word-level annotations, espousing what has alternately been called the **lexical integrity principle** (Chomsky, 1970; Bresnan and Mchombo, 1995) or **word-based morphology** (Aronoff, 1976, 2007; Spencer, 1991). Unfortunately, not all morphological manifestations occur at the level of individual words. The Farsi (Persian) **light verb construction** illustrates the deficiency (see Karimi-Doostan, 2011). Farsi expresses many actions by pairing a light verb (one with little meaning) with a noun that gives a concrete meaning. The example in Figure 2 uses the light verb construction "*latme zadan*" ("to damage"). The parts of the verb construction are separated in the sentence, seeming to require a morphosyntactic parse. When attempting to annotate these constructs, neither schema provides guidance. In languages where these occur, language-specific decisions are made. It should be noted that multiword expressions are a general challenge to natural language processing, not specifically morphology (Sag et al., 2002).

5 A Deterministic Conversion

In our work, the goal is not simply to translate one schema into the other, but to translate one *resource* (the imperfect manifestation of the schema) to match the other. The differences between the schemata and discrepancies in annotation mean that the transformation of annotations from one schema to the other is not straightforward.

Two naive options for the conversion are a lookup table of MSDs and a lookup table of the individual attribute-value pairs which comprise the MSDs. The former is untenable: the table of all UD feature combinations (including null features, excluding language-specific attributes) would have

[8] The complete list is at http://universaldependencies.org/v2/features.html#inventory-of-features-that-will-stay-language-specific

2.445×10^{17} entries. Of course, most combinations won't exist, but this gives a sense of the table's scale. Also, it doesn't leverage the factorial nature of the annotations: constructing the table would require a massive duplication of effort. On the other hand, attribute-value lookup lacks the flexibility to show how a pair of values interacts. Neither approach would handle language- and annotator-specific tendencies in the corpora.

Our approach to converting UD MSDs to UniMorph MSDs begins with the attribute-value lookup, then amends it on a language-specific basis. Alterations informed by the MSD and the word form, like insertion, substitution, and deletion, increase the number of agreeing annotations. They are critical for work that examines the MSD monolithically instead of feature-by-feature (e.g. Belinkov et al., 2017; Cotterell and Heigold, 2017): Without exact matches, converting the individual tags becomes hollow.

Beginning our process, we relied on documentation of the two schemata to create our initial, language-agnostic mapping of individual values. This mapping has 140 pairs in it. Because the mapping was derived purely from the schemata, it is a useful approximation of how well the schemata match up. We note, however, that the mapping does not handle idiosyncrasies like the many uses of "dative" or features which are represented in UniMorph by argument templates: possession and ergative–absolutive argument marking. The initial step of our conversion is using this mapping to populate a proposed UniMorph MSD.

As shown in §7, the initial proposal is often frustratingly deficient. Thus we introduce the post-edits. To concoct these, we looked into UniMorph corpora for these languages, compared these to the conversion outputs, and then sought to bring the conversion outputs closer to the annotations in the actual UniMorph corpora. When a form and its lemma existed in both corpora, we could directly inspect how the annotations differed. Our process of iteratively refining the conversion implies a table which exactly maps any combination of UD MSD and its related values (lemma, form, etc.) to a UniMorph MSD, though we do not store the table explicitly.

Some conversion rules we've created must be applied before or after others. These sequential dependencies provide conciseness. Our post-editing procedure operates on the initial MSD hypothesis as follows:

1. First, we collect all arguments relating to a possessor or an ergative–absolutive language's argument agreement, because UniMorph represents both categories as a single templatic value.

2. We discard any values that UniMorph doesn't annotate for a particular part of speech, like gender and number in French verb participles, or German noun genders.

3. We make MSD additions when they are unambiguously implied by the resources, like PFV to accompany PST in Spanish "pasado simple", but PST to accompany IPFV in Spanish "pasado continuo".

4. We also incorporate fixes using information outside of the MSD like the LGSPEC1 tag for Spanish's "-ra" forms, as described in §4, and other language-specific corrections, like mapping the various dative cases to the cross-lingually comparable case annotations used in UniMorph.

What we left out We did, however, reject certain changes that would increase harmony between the resources. Usually, this decision was made when the UniMorph syntax or tagset was not obeyed, such as in the case of made-up tags for Basque arguments (instead of the template mentioned above) or the use of idiopathic colons (:) instead of semicolons (;) as separators in Farsi. Other instances were linguistically motivated. UD acknowledges Italian imperatives, but UniMorph does not have any in its table. We could largely alter these to have subjunctive labels, but to ill effect. A third reason to be conservative in our rules was cases of under-specification: If a participle is not marked as past or present in UD, but both exist in UniMorph, we could unilaterally assign all to the majority category and increase recall. This would pollute the data with fallacious features, so we leave these cases under-specified. In other words, we do not add new values that cannot be unequivocally inferred from the existing data.

Output The Universal Dependencies data are presented in the CoNLL-U format.[9] Each sentence

[9] http://universaldependencies.org/format.html

is represented in tabular form to organize annotations like lemmas, parts of speech, and dependencies of each word token. The MSDs are held in a column called `FEATS`. Our MSD conversion tool produces a CoNLL-U file whose `FEATS` column now contains a UniMorph-style MSD. For more straightforward interface with UniMorph, the feature bundle includes the part of speech tag. As the `POS` column of the CONLL-U file is preserved, this can easily be stripped from the `FEATS` column, depending on use case.

Why not a learned mapping? One can imagine learning the UniMorph MSD corresponding to a UD dataset's MSD by a set-to-set translation model like IBM Model 1 (Brown et al., 1993). Unfortunately, statistical (and especially neural) machine translation generalizes in unreliable ways. Our goal is a straightforward, easily manipulable and extensible conversion that prioritizes correctness over coverage.

6 Experiments

We evaluate our tool on two tasks:

Intrinsic assessment: Once we convert UD MSDs to UniMorph MSDs, how many of the converted ones are attested in UniMorph's paradigm tables.

Extrinsic assessment: Whether performance on a downstream task is comparable when using pre- and post-conversion MSDs.

To be clear, our scope is limited to the schema conversion. Future work will explore NLP tasks that exploit both the created token-level UniMorph data and the existing type-level UniMorph data.

Data We draw our input data from the UD v2.1 treebanks (Nivre et al., 2017). When multiple treebanks exist for a language, we select the one with a basic name, e.g. "Spanish" instead of "Spanish-AnCora". We leave the construction of additional converters to future work, and we invite the community to participate in designing the mappings for all UD treebanks. UniMorph modifies its language packs individually instead of offering versioned releases. Our UniMorph lookup tables are the latest versions at the time of writing.[10] There are 31 languages which possess both a UD and a UniMorph corpus.

[10] As of 19 June 2018, the latest modification to a UniMorph language resource was to Finnish on 3 August 2017.

6.1 Intrinsic evaluation

We transform all UD data to the UniMorph. We compare the simple lookup-based transformation to the one with linguistically informed post-edits on all languages with both UD and UniMorph data. We then evaluate the recall of MSDs without partial credit.

Calculating recall Because the UniMorph tables only possess annotations for verbs, nouns, adjectives, or some combination, we can only examine performance for these parts of speech. We consider two words to be a match if their form and lemma are present in both resources. Syncretism allows a single surface form to realize multiple MSDs (Spanish "*mandaba*" can be first- or third-person), so we define success as the computed MSD matching *any* of the word's UniMorph MSDs. This gives rise to an equation for recall: of the word–lemma pairs found in both resources, how many of their UniMorph-converted MSDs are present in the UniMorph tables?

Why no held-out test set? Our problem here is not a learning problem, so the question is ill-posed. There is no *training* set, and the two resources for a given language make up a test set. The quality of our model—the conversion tool—comes from how well we encode prior knowledge about the relationship between the UD and UniMorph corpora.

6.2 Extrinsic evaluation

If the UniMorph-converted treebanks perform differently on downstream tasks, then they convey different information. This signals a failure of the conversion process. As a downstream task, we choose morphological tagging, a critical step to leveraging morphological information on new text.

We evaluate taggers trained on the transformed UD data, choosing eight languages randomly from the intersection of UD and UniMorph resources. We report the macro-averaged F1 score of attribute-value pairs on a held-out test set, with official train/validation/test splits provided in the UD treebanks. As a reference point, we also report tagging accuracy on those languages' untransformed data.

We use the state-of-the-art morphological tagger of Malaviya et al. (2018). It is a factored conditional random field with potentials for each attribute, attribute pair, and attribute transition. The potentials are computed by neural networks, predicting the values of each attribute jointly but not

monolithically. Inference with the potentials is performed approximately by loopy belief propagation. We use the authors' hyperparameters.

We note a minor implementation detail for the sake of reproducibility. The tagger exploits explicit guidance about the attribute each value pertains to. The UniMorph schema's values are globally unique, but their attributes are not explicit. For example, the UniMorph MASC denotes a masculine gender. We amend the code of Malaviya et al. to incorporate attribute identifiers for each UniMorph value.

7 Results

We present the intrinsic task's recall scores in Table 3. Bear in mind that due to annotation errors in the original corpora (like the "*vas*" example from §4), the optimal score is not always 100%. Some shortcomings of recall come from irremediable annotation discrepancies. Largely, we are hamstrung by differences in choice of attributes to annotate. When one resource marks gender and the other marks case, we can't infer the gender of the word purely from its surface form. The resources themselves would need updating to encode the relevant morphosyntactic information. Some languages had a very low number of overlapping forms,[11] and no tag matches or near-matches between them: Arabic, Hindi, Lithuanian, Persian, and Russian. A full list of observed, irremediable discrepancies is presented alongside the codebase.

There are three other transformations for which we note no improvement here. Because of the problem in Basque argument encoding in the UniMorph dataset—which only contains verbs—we note no improvement in recall on Basque. Irish also does not improve: UD marks gender on nouns, while UniMorph marks case. Adjectives in UD are also underspecified. The verbs, though, are already correct with the simple mapping. Finally, with Dutch, the UD annotations are impoverished compared to the UniMorph annotations, and missing attributes cannot be inferred without external knowledge.

For the extrinsic task, the performance is reasonably similar whether UniMorph or UD; see Table 4. A large fluctuation would suggest that the two annotations encode distinct information. On the contrary, the similarities suggest that the UniMorph-mapped MSDs have similar content. We recognize

[11] Fewer than 250 overlapping form–lemma pairs. The other languages had overlaps in the thousands.

Language	CSV	Post-editing
Ar	0.00	-
Bg	34.61	87.88
Ca	23.23	99.78
Cs	0.48	81.71
Da	1.55	4.70
De	17.20	60.81
En	42.17	90.10
Es	17.20	97.86
Eu	0.00	0.00
Fa	0.00	-
Fi	59.19	92.81
Fr	18.61	99.20
Ga	0.41	0.41
He	4.08	46.61
Hi	0.00	-
Hu	15.46	24.94
It	22.32	94.89
La	11.73	64.25
Lt	0.00	-
Lv	0.17	90.58
Nb	2.11	38.88
Nl	12.12	12.12
Nn	2.40	40.21
Pl	7.70	88.17
Pt	20.11	99.34
Ro	0.00	25.16
Ru	0.00	-
Sl	37.57	90.27
Sv	13.20	83.44
Tr	0.00	65.14
Uk	4.06	96.45
Ur	0.00	55.72

Table 3: Token-level recall when converting Universal Dependencies tags to UniMorph tags. CSV refers to the lookup-based system. Post-editing refers to the proposed method.

Language	UD F1	UniMorph F1
Da	90.58	92.59
Es	78.31	96.44
Fi	93.78	94.98
Lv	84.20	86.94
Pt	95.57	95.77
Ru	89.89	89.95
Bg	95.54	95.79
Sv	92.39	93.83

Table 4: Tagging F1 using UD sentences annotated with either original UD MSDs or UniMorph-converted MSDs

that in every case, tagging F1 increased—albeit by amounts as small as 0.16 points. This is in part due to the information that is lost in the conversion. UniMorph's schema does not indicate the type of pronoun (demonstrative, interrogative, etc.), and when lexical information is not recorded in UniMorph, we delete it from the MSD during transformation. On the other hand, UniMorph's atomic tags have more parts to guess, but they are often related. (E.g. IPFV always entails PST in Spanish.) Altogether, these forces seem to have little impact on tagging performance.

8 Related Work

The goal of a tagset-to-tagset mapping of morphological annotations is shared by the Interset project (Zeman, 2008). Interset decodes features in the source corpus to a *tag interlingua*, then encodes that into target corpus features. (The idea of an interlingua is drawn from machine translation, where a prevailing early mindset was to convert to a universal representation, then encode that representation's semantics in the target language. Our approach, by contrast, is a direct flight from the source to the target.) Because UniMorph corpora are noisy, the encoding from the interlingua would have to be rewritten for each target. Further, decoding the UD MSD into the interlingua cannot leverage external information like the lemma and form.

The creators of HamleDT sought to harmonize dependency annotations among treebanks, similar to our goal of harmonizing across resources (Zeman et al., 2014). The treebanks they sought to harmonize used multiple diverse annotation schemes, which the authors unified under a single scheme.

Petrov et al. (2012) present mappings into a coarse, "universal" part of speech for 22 languages. Working with POS tags rather than morphological tags (which have far more dimensions), their space of options to harmonize is much smaller than ours.

Our extrinsic evaluation is most in line with the paradigm of Wisniewski and Lacroix (2017) (and similar work therein), who compare syntactic parser performance on UD treebanks annotated with two styles of dependency representation. Our problem differs, though, in that the dependency representations express different relationships, while our two schemata vastly overlap. As our conversion is lossy, we do not appraise the learnability of representations as they did.

In addition to using the number of extra rules as a proxy for harmony between resources, one could perform cross-lingual projection of morphological tags (Drábek and Yarowsky, 2005; Kirov et al., 2017). Our approach succeeds even without parallel corpora.

9 Conclusion and Future Work

We created a tool for annotating Universal Dependencies CoNLL-U files with UniMorph annotations. Our tool is ready to use off-the-shelf today, requires no training, and is deterministic. While under-specification necessitates a lossy and imperfect conversion, ours is interpretable. Patterns of mistakes can be identified and ameliorated.

The tool allows a bridge between resources annotated in the Universal Dependencies and Universal Morphology (UniMorph) schemata. As the Universal Dependencies project provides a set of treebanks with token-level annotation, while the UniMorph project releases type-level annotated tables, the newfound compatibility opens up new experiments. A prime example of exploiting token- and type-level data is Täckström et al. (2013). That work presents a part-of-speech (POS) dictionary built from Wiktionary, where the POS tagger is also constrained to options available in their type-level POS dictionary, improving performance. Our transformation means that datasets are prepared for similar experiments with morphological tagging. It would also be reasonable to incorporate this tool as a subroutine to UDPipe (Straka and Straková, 2017) and Udapi (Popel et al., 2017). We leave open the task of converting in the opposite direction, turning UniMorph MSDs into Universal Dependencies MSDs.

Because our conversion rules are interpretable, we identify shortcomings in both resources, using each as validation for the other. We were able to find specific instances of incorrectly applied UniMorph annotation, as well as specific instances of cross-lingual inconsistency in both resources. These findings will harden both resources and better align them with their goal of universal, cross-lingual annotation.

Acknowledgments

We thank Hajime Senuma and John Sylak-Glassman for early comments in devising the starting language-independent mapping from Universal Dependencies to UniMorph.

References

Mark Aronoff. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass.*, 1:1–134.

Mark Aronoff. 2007. In the beginning was the word. *Language*, 83(4):803–830.

Matthew Baerman, Dunstan Brown, Greville G Corbett, et al. 2005. *The syntax-morphology interface: A study of syncretism*, volume 109. Cambridge University Press.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 861–872.

Jean Berko. 1958. The child's learning of English morphology. *Word*, 14(2-3):150–177.

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

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Noam Chomsky. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum, editors, *Reading in English Transformational Grammar*, pages 184–221. Ginn and Co., Waltham.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sebastian Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, Brussels, Belgium. 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, et al. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. *Proceedings of the CoNLL–SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30.

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.

Peter Dirix, Liesbeth Augustinus, Daniel van Niekerk, and Frank Van Eynde. 2017. Universal dependencies for Afrikaans. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May, Gothenburg Sweden*, 135, pages 38–47. Linköping University Electronic Press.

Elliott Franco Drábek and David Yarowsky. 2005. Induction of fine-grained part-of-speech taggers via classifier combination and crosslingual projection. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 49–56. Association for Computational Linguistics.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Gholamhossein Karimi-Doostan. 2011. Separability of light verb constructions in Persian. *Studia Linguistica*, 65(1):70–95.

Aleksandr E. Kibrik. 1998. *Archi (Caucasian–Daghestanian)*. Wiley Online Library.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, and Matt Post. 2017. A rich morphological tagger for English: Exploring the cross-linguistic tradeoff between morphology and syntax. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 112–117.

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large scale parsing and normalization of Wiktionary morphological paradigms. In *LREC*.

Ralph B. Long. 1957. Paradigms for English verbs. *Publications of the Modern Language Association of America*, pages 359–372.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural factor graph models for cross-lingual morphological tagging. In *Proceedings of the 56th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 2652–2662. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. 2017. Universal dependencies 2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for universal dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15. Springer.

Andrew Spencer. 1991. *Morphological theory: An introduction to word structure in generative grammar*, volume 2. Basil Blackwell Oxford.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

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.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. 2013. Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Guillaume Wisniewski and Ophélie Lacroix. 2017. A systematic comparison of syntactic representations of dependency parsing. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 146–152.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *LREC*, volume 2008, pages 28–30.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtskỳ, and Jan Hajič. 2014. Hamledt: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. CoNLL 2017 shared task: multilingual parsing from raw text to universal dependencies. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19.