# IGT2P: From Interlinear Glossed Texts to Paradigms

**Sarah Moeller  Ling Liu  Changbing Yang  Katharina Kann  Mans Hulden**
University of Colorado
first.last@colorado.edu

## Abstract

An intermediate step in the linguistic analysis of an under-documented language is to find and organize inflected forms that are attested in natural speech. From this data, linguists generate unseen inflected word forms in order to test hypotheses about the language's inflectional patterns and to complete inflectional paradigm tables. To get the data linguists spend many hours manually creating interlinear glossed texts (IGTs). We introduce a new task that speeds this process and automatically generates new morphological resources for natural language processing systems: IGT-to-paradigms (IGT2P). IGT2P generates entire morphological paradigms from IGT input. We show that existing morphological reinflection models can solve the task with 21% to 64% accuracy, depending on the language. We further find that (i) having a language expert spend only a few hours cleaning the noisy IGT data improves performance by as much as 21 percentage points, and (ii) POS tags, which are generally considered a necessary part of NLP morphological reinflection input, have no effect on the accuracy of the models considered here.

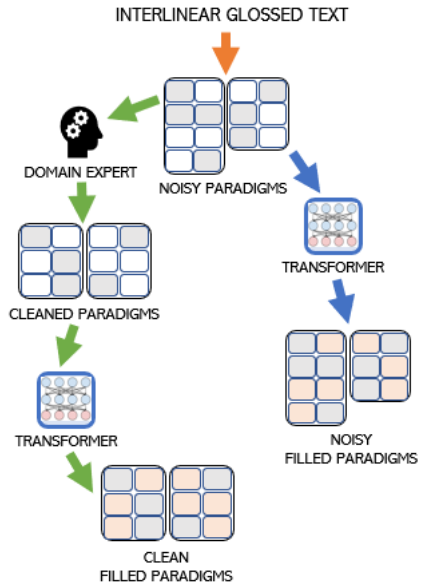


Figure 1: Inflected word forms attested in interlinear glossed texts (IGT) train transformer encoder-decoder to generalize morphological paradigmatic patterns and generate word forms when given known morphosyntatic features of missing paradigm cells. Noisy paradigms are automatically constructed from IGT and a language expert creates "cleaned" paradigms. Both sets are tested on the same missing word forms and the results are compared.

## 1 Introduction

Over the last few years, multiple shared tasks have encouraged the development of systems for learning morphology, including generating inflected forms of the canonical form—the lemma—of a word. NLP systems that account for morphology can reduce data sparsity caused by an abundance of individual word forms in morphologically rich languages (Cotterell et al., 2016, 2017a, 2018; McCarthy et al., 2019; Vylomova et al., 2020) and help mitigate bias in training data for natural language processing (NLP) systems (Zmigrod et al., 2019). However, such systems have often been limited to languages with publicly available structured data, i.e. languages for which tables containing inflectional patterns can be found, for example, in online dictionaries like Wiktionary.[1] This limits the development of NLP systems for morphology to languages for which morphological information can be easily extracted.

Here, we propose to instead make use of a resource which is much more common, especially for low-resource languages: we explore how to leverage interlinear glossed text (IGT)—a common artifact of linguistic field research—to generate unseen forms of inflectional paradigms, as illustrated in Figure 1. This task, which we call **IGT-to-paradigms** (IGT2P), differs from the ex-

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



*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5251–5262,
November 16–20, 2020. 

isting *morphological inflection* (Yarowsky and Wicentowski, 2000; Faruqui et al., 2016) task in three aspects: (1) inflected forms extracted from IGT are noisier than curated training data for morphological generation, (2) since lemmas are not explicitly identified in IGT, systems cannot be trained on typical lemma-to-form mappings and, instead, must be trained on form-to-form mappings, and (3) part-of-speech (POS) tags are often unavailable in IGT. IGT2P can thus be seen as a noisy version of morphological *re*inflection (Cotterell et al., 2016), but without explicit POS information. Our experiments show that morphological reinflection systems following preprocessing are strong baselines for this task.

We further perform two analyses:

(i) Part-of-speech (POS) tags are usually considered necessary inputs for learning morphological generation. However, they are frequently missing from IGT, since they result from a later step in a linguist's pipeline. Thus, we ask: are POS tags necessary for morphological generation? Surprisingly, we find that POS tags are of little use for morphological generation systems.

(ii) How much does manual cleaning of IGT data by a domain expert improve performance? As expected, cleaning the data improves performance across the board with a transformer model: by 1.27% to 16.32%, depending on the language.

We examine which inflection model performs better on noisy and cleaned IGT data and how the performance varies across languages and data quality or size.

## 2 A New Morphological Task: IGT2P

### 2.1 Background: Morphological Generation

An inflectional paradigm is illustrated in tables, such as Table 1. Paradigms can be large; for example, Polish verbs paradigms can have up to 30 cells and other languages may have several more. Here we define the notation related to morphological inflection systems for the remainder of this paper.

We denote the paradigm of a lemma $\ell$ as:

$$\pi(\ell) = \left\langle f(\ell, \vec{t}_\gamma) \right\rangle_{\gamma \in \Gamma(\ell)} \quad (1)$$

where $f : \Sigma^* \times \mathcal{T} \rightarrow \Sigma^*$ defines a mapping from a tuple consisting of the lemma and a vector $\vec{t}_\gamma \in \mathcal{T}$

| | present | | past | |
|---|---|---|---|---|
| | **sing.** | **pl.** | **sing.** | **pl.** |
| **1 person** | am | are | was | were |
| **2 person** | are | are | were | were |
| **3 person** | is | are | was | were |

Table 1: The inflectional paradigm of the English verb "to be". This verb has more inflected forms than any other English lemma, but is quite small compared to paradigms in many other languages.

of morphological features to the corresponding inflected form. $\Sigma$ is an alphabet of discrete symbols, i.e., the characters used in the natural language. $\Gamma(\ell)$ is the set of slots in lemma $\ell$'s paradigm. We will abbreviate $f(\ell, \vec{t}_\gamma)$ as $f_\gamma(\ell)$ for simplicity. Using this notation, we now describe the most important generation tasks from the computational morphology literature.

**Morphological inflection.** The task of morphological inflection consists of generating unknown inflected forms, given a lemma $\ell$ and a feature vector $\vec{t}_\gamma$. Thus, it corresponds to learning the mapping $f : \Sigma^* \times \mathcal{T} \rightarrow \Sigma^*$.

**Morphological reinflection.** Morphological *re*inflection is a generalized version of the previous task. Here, instead of having a lemma as input, system are given some *inflected form* $f(\ell, \vec{t}_{\gamma_1})$ – optionally together with $\vec{t}_{\gamma_1}$ – and a target feature vector $\vec{t}_{\gamma_2}$. The goal is then to produce the inflected form $f(\ell, \vec{t}_{\gamma_2})$.

**Paradigm completion.** The task of paradigm completion consists of, given a *partial* paradigm $\pi_P(\ell) = \left\langle f(\ell, \vec{t}_\gamma) \right\rangle_{\gamma \in \Gamma_P(\ell)}$ of a lemma $\ell$, generating all inflected forms for all slots $\gamma \in \Gamma(\ell) - \Gamma_P(\ell)$. Training data for this task consists of entire paradigms.

**Unsupervised morphological paradigm completion.** For the *unsupervised* version of the paradigm completion task, systems are given a corpus $\mathcal{D} = w_1, \ldots, w_{|\mathcal{D}|}$ with a vocabulary $V$ of word types $\{w_i\}$ and a lexicon $\mathcal{L} = \{\ell_j\}$ with $|\mathcal{L}|$ lemmas belonging to the same part of speech. However, no explicit paradigms are observed during training. The task of unsupervised morphological paradigm completion then consists of generating the paradigms $\{\pi(\ell)\}_{\ell \in \mathcal{L}}$ of all lemmas $\ell \in \mathcal{L}$.

### 2.2 IGT-to-Paradigms

The task we propose, IGT-to-paradigms (IGT2P), can be described as the paradigm completion problem above, with an additional step of inference regarding which of the attested forms is associated with which lemma.

Formally, systems are given IGTs consisting of words with – potentially empty – morphological feature vectors: $\mathcal{D} = (w_1, \vec{t}_1) \ldots, (w_{|\mathcal{D}|}, \vec{t}_{|\mathcal{D}|})$ and a list $\mathcal{U} = \{u_j\}$ with $|\mathcal{U}|$ inflected words, $u_j = f(\ell_j, \vec{t}_{\gamma_j})$. The goal of IGT2P is to generate the paradigms $\{\pi(\ell_j)\}_{f(\ell_j, \vec{t}_{\gamma_j}) \in \mathcal{U}}$.

Similar to unsupervised paradigm completion, we do not assume information about the lemma to be explicit. Similar to morphological reinflection, the input includes word forms with features, and a system has to learn to generate inflections from other word forms and morphological feature vectors. IGT2P is further similar to paradigm completion in that we aim at generating *all* inflected forms for each lemma.[2]

### 2.3 Why IGT2P?

Descriptive linguistics aims to objectively analyze primary language data in new languages and publish descriptions of their structure. This work informs our understanding of human language and provides resources for NLP development through academic literature, which informs projects such as UniMorph (Kirov et al., 2016), or through crowdsourced effort such as Wiktionary. Yet with most descriptive work performed manually with very little NLP assistance, language resources for thousands of under-described languages remain limited. This includes languages with millions of speakers, such as Manipuri in India.

However, there exists a type of labeled data that is available in nearly all languages where a linguist has undertaken any scientific endeavor: *interlinear glossed texts* (IGT), illustrated in Table 2. They are the output of early steps in a field linguist's pipeline which consist of recording natural speech, transcribing it, and then identifying minimal meaningful units—the morphemes—and using internally consistent tags to label the morphemes' morphosyntactic features. IGTs serve as vital sources of morphological, syntactic, and higher levels of linguistic information. They are often archived in long-term repositories, and openly accessible for non-commercial purposes, yet they are rarely utilized in NLP.

IGT2P has potential benefits for NLP (by increasing available resources in low-resource languages) but also for linguistic inquiry. First, since machine-assistance has been shown to increase speed and accuracy of manual linguistic annotation with just 60% model accuracy (Felt, 2012), such a model could assist the initial analysis of morphological patterns in IGT. Second, by quickly learning morphological patterns from word forms attested in IGT, IGT2P generates forms that fill empty cells in a lemma's paradigm. Since IGTs are unlikely to contain complete paradigms of lemmas, an accompanying step in fieldwork is that of elicitation of inflectional paradigms for selected lemmas. Presenting candidate words to a native speaker for acceptance or rejection is often easier than asking the speaker to grasp the abstract concept of a paradigm and to generate the missing cells in a table. With the help of IGT2P, linguists could use the machine-generated word forms to support this elicitation process. IGT2P then becomes a tool for the discovery of morphological patterns in under-described and endangered languages.

## 3 Related Work

**IGT for NLP.** The AGGREGATION project (Bender, 2014) has used IGT to automatically construct *grammars* for multiple languages. This includes inferring and visualizing systems of morphosyntax (Lepp et al., 2019; Wax, 2014). Much of their data comes from the Online Database of INterlinear Text (Lewis and Xia, 2010, ODIN) which is a collection of IGTs extracted from published linguistic documents on the web. Published IGT excerpts, such as those in ODIN, differ from IGTs produced by field linguists such as those used in our experiments. First, noise is generally removed from the published examples. Second, the amount of glossed information in published IGT snippets can vary widely depending on the phenomenon that is the main focus of the publication.

**Computational morphology.** Our work is further related to and takes inspiration from research on the tasks described in Section 2.1.

Most recent work in the area of computational morphology which was concerned with generation (as opposed to analysis) has focused on morpholog-

[2] We currently approximate this during evaluation, since we do not have gold standard paradigms for the languages. Also, our list $\mathcal{U}$ consists of words in $\mathcal{D}$, which we exclude from the input.

| | | | | | |
|---|---|---|---|---|---|
| **Text** | Vecherom | ya | pobejala | v | magazin. |
| **Segmented** | vecher-om | ya | pobeja-la | v | magazin |
| **Glossed** | evening-INS | 1.SG.NOM | run-PFV.PST.SG.FEM | in | store.ACC |
| **Translation** | 'In the evening I ran to the store.' | | | | |

Table 2: An example of typical interlinear glossed text (IGT) with a transliterated Russian sentence, including translation. This paper leverages the original text and gloss lines.

ical inflection or reinflection. Approaches include Durrett and DeNero (2013); Nicolai et al. (2015); Faruqui et al. (2016); Kann and Schütze (2016); Aharoni and Goldberg (2017). Partially building on these, other research has developed models which are more suitable for low-resource languages and perform well with limited data (Kann et al., 2017b; Sharma et al., 2018; Makarov and Clematide, 2018; Wu and Cotterell, 2019; Kann et al., 2020a; Wu et al., 2020). These are the most relevant approaches for our work, since we expect IGT2P to aid documentation of low-resource languages. Accordingly, we use the systems by Wu and Cotterell (2019) and Wu et al. (2020) in our experiments.

Work on paradigm completion – or the *paradigm cell filling problem* (PCFP; Ackerman et al., 2009) – includes Malouf (2016), who trained recurrent neural networks for it, and applied them successfully to Irish, Maltese, and Khaling, among other languages. Silfverberg and Hulden (2018) also trained neural networks for the task. Kann et al. (2017a) differed from other approaches in that they encoded multiple inflected forms of a lemma to provide complementary information for the generation of unknown forms of the same lemma. Finally, Cotterell et al. (2017b) introduced neural graphical models which completed paradigms based on principal parts. The unsupervised version of the paradigm completion task (Jin et al., 2020) has been the subject of a recent shared task (Kann et al., 2020b), with the conclusion that it is exremely challenging for current state-of-the-art systems. Here, we propose to, instead of generating paradigms from raw text, generate them from IGT, a resource available for many under-studied languages.

## 4 To POS Tag or Not to POS Tag

In addition to the lemma and the morphological features of the target form, part-of-speech (POS) tags are by default a part of the input to neural morphological reinflection systems. POS tags are assumed to carry valuable information, since, for example, morphemes that are otherwise identical (e.g. "seat") may use one set of inflectional morphemes as nouns (e.g. "many seats") and another as verbs ("be seated").

Since POS tags are typically annotated at a later stage than morpheme boundaries and glosses, IGTs often do not contain POS tags for all words. This makes large parts of the IGT unusable for state-of-the-art reinflection systems if POS tags are assumed necessary. However, the assumption that POS tags improve morphological generation performance has never been empirically verified for recent state-of-the-art systems. We hypothesize that, in fact, POS tags might not be necessary, since they might be implicitly defined by either the morphological features or the input word form. Thus, we ask the following research question: *Are POS tags a necessary or beneficial input to a morphological reinflection system?*

### 4.1 Experimental Setup

To answer this question, we train morphological reinflection systems twice on 10 languages that have been released for the CoNLL-SIGMORPHON 2018 shared task (Cotterell et al., 2018), once with and once without POS tags as input. In order to obtain generalizable results, our selected languages belong to different families and are typologically diverse with regards to morphology, as shown in Table 3.[3] We kept the original training/validation/test splits, and experiment on the three training set sizes: 10,000, 1000, and 100 examples for the *high*, *medium*, and *low* setting, respectively.

### 4.2 Models

We experiment with two state-of-the-art neural models for morphology learning: the transformer model for character-level transduction (Wu et al., 2020) and the LSTM sequence-to-sequence model with exact hard monotonic attention for character-

[3]The language family and morphological typology for each language is on the UniMorph official website (https://unimorph.github.io).

| Language | POS | transformer model (%) | | | Exact hard mono model (%) | | |
|---|---|---|---|---|---|---|---|
| | | high Δ | medium Δ | low Δ | high Δ | medium Δ | low Δ |
| Adyghe | N, ADJ | 0.0 | -0.3 | 1.7 | 0.2 | -0.3 | -0.5 |
| Arabic | N, V, ADJ | -0.1 | 0.0 | -0.5 | -0.5 | 1.2 | 0.0 |
| Basque | V | -0.2 | 0.0 | -2.8 | -0.3 | 2.1 | -0.4 |
| Finnish | N, V, ADJ | 0.6 | -0.5 | 0.2 | -0.7 | 4.4 | 0.0 |
| German | N, V | 0.6 | -0.6 | -1.6 | -0.1 | 0.0 | -0.7 |
| Persian | V | 0.0 | -1.5 | -0.2 | -0.3 | -0.9 | 1.2 |
| Russian | N, V, ADJ | 0.1 | 1.3 | -0.4 | 0.0 | -0.6 | -0.9 |
| Spanish | N, V | -0.1 | 0.9 | 0.7 | 1.0 | 4.2 | -0.3 |
| Swahili | N, V, ADJ | 0.0 | 0.0 | 0.0 | 0.0 | 3 | 1.0 |
| Turkish | N, V, ADJ | -0.2 | 0.0 | 1.5 | 0.2 | 3.2 | -0.1 |

Table 3: SIGMORPHON languages, their inflected parts of speech used to test the helpfulness of POS tags to neural reinflection tasks, and the difference in accuracy (%) between using and not using POS for the transformer model and the LSTM seq2seq model with exact hard monotonic attention in different training data size settings. Negative scores means that removing POS tags decreased performance.

level transduction (Wu and Cotterell, 2019).[4]

### 4.3 Results

Table 3 illustrates the performance difference when including and not including POS tags for all three training data sizes. The largest difference is a decrease of 4.4 percentage points when POS tags are removed for Finnish at the medium setting using hard monotonic attention. The average difference is about 0.2 percentage points. We therefore conclude that a lack of POS tags does not make a significant difference in the reinflection task.

## 5 IGT2P

### 5.1 Language data

We used IGTs that were primarily transcribed from naturally-occurring oral speech in low-resource and endangered languages. They represent a wide range of projects, which is reflected in the size and quality of the data. The amount of usable data (i.e. glossed words) ranges from approximately 90,000 tokens in Arapaho to about 5,000 in Manipuri. The five languages (see Table 4) are spoken by communities across five continents. They represent different language families and morphological complexity, though all are agglutinating to some degree. Other than the IGT, there is very limited resources for these languages.

[4]It is theoretically possible that the other baselines can outperform these models once we limit our experiments to words with POS information. However, based on our preliminary experiments using POS tags, this seems unlikely.

| Language | ISO | Family | Tokens |
|---|---|---|---|
| Arapaho | arp | Algonquian | 90k |
| Lezgi | lez | Nakh-Daghestanian | 18.7k |
| Manipuri | mni | Tibeto-Burman | 5k |
| Natügu | ntu | Austronesian | 14k |
| Tsez | ddo | Nakh-Daghestanian | 53k |

Table 4: Languages with IGT used in this experiment, their ISO 639-3 identifying codes, and the approximate number of tokens in the database that are interlinearized (i.e. segmented into morphemes and glossed).

### 5.2 Issues specific to IGT

The most notable issue with IGT is the "noise". An inevitable cause is the dynamic nature of ongoing linguistic analysis. As the linguist gains a better understanding of the language's structure by doing interlinearization, early decisions about morpheme shapes and glosses differ from later ones. Another cause is that limited budget and time means IGT are often only partially completed. Another source of noise comes when the project is focused on annotating one particular phenomenon. For example, frequently only one morphosyntactic feature in Manipuri was glossed in each word, meaning different inflected forms looked like they had the same morphosyntactic features. Another source of noise is imprecision introduced by human errors or choices made for convenience to speed tedious annotation. One example of imprecision is glossing different stem morphemes with the same English word. For example, Lezgi has several copula verbs which can

| | | | | | |
|---|---|---|---|---|---|
| вав | SG,AD | вав | SG,AD | вав | SG,AD |
| ваз | SG,DAT | ваз | SG,DAT | ваз | SG,DAT |
| вакай | SG,SBSS,EL | вакай | SG,SBSS,EL | вакай | SG,SBSS,EL |
| | | вун | SG,ABS | вун | SG,ABS |
| бун | SG,ABS | вуна | SG,ERG | вуна | SG,ERG |
| вуна | SG,ERG | ви | SG,GEN | ви | SG,GEN |
| | | | | вавай | SG,AD,EL |
| ви | SG,ABS | | | вавди | SG,AD,DIR |
| | | | | вахъ | SG,POES |
| | | | | вахъай | SG,POES,EL |
| | | | | вахъай | SG,POES,DIR |
| | | | | вак | SG,SBSS |
| | | | | вакди | SG,SBSS,DIR |
| | | | | вал | SG,SPSS |
| | | | | валай | SG,SPSS,EL |
| | | | | валди | SG,SPSS,DIR |
| | | | | ва | SG,INESS |
| | | | | вай | SG,INEL |

Figure 2: Lezgi paradigms were automatically constructed from IGT (left columns) and have typos or incorrect paradigms clusters. Experts filtered or corrected these issues, resulting in "clean" paradigms (middle). These can be compared with the published description (right column) which includes historic forms that are rarely used today.

be narrowly translated as 'be in', 'be at', etc., but most were merely glossed as 'be'. So all copula verbs were initially grouped into one paradigm. A similar situation happened with Arapaho: nuances of meaning were not often distinguished in the glosses; thus, different verb stems are glossed simply as 'give', when, in reality they should be divided into 'hand to someone' in one case, 'give as a present' in another case, and 'give ceremonially, as an honor' in third case.

Another issue is that IGT annotators do not usually differentiate between different types of morphemes. Thus, we do not always distinguish between them. Derivational and inflectional morphemes were only differentiated where we were able to easily identify and eliminate derivational glosses. For example, in Arapaho we were able to group derived stems into separate paradigms because they were glossed distinctly. Also, clitics are often not distinguished from affixes. This means that the morphological patterns that the models learn are not always, strictly speaking, inflectional paradigms, but it does mean that the models learn all attested forms related one lemma.

### 5.3 Approach

As a first step, partial inflectional paradigms were automatically extracted from the IGT. Words were organized into paradigms based on the gloss of the stem morpheme. Then, these stem glosses were removed, leaving only the affix glosses which serve as morphosyntactic feature tags.

**Step 1: Preprocessing paradigms.** The automatically extracted paradigms were preprocessed in two ways. The resulting data is publicly available.[5] In the first preprocessing method, a language domain expert was asked to "clean" the automatically extracted paradigms. Example results are in shown Figure 2. Experts reorganized words into correct inflectional paradigms, for example, by regrouping Lezgi copula verbs. They also completed missing morphosyntatic information; for example, adding PL (plural) or SG (singular) where the nouns were otherwise glossed identically. Finally, they removed any words that are not inflected in the language. This usually included words that are morphologically derived from another part of speech but not inflected. For example, an affix might derive an adverb from a noun root, and if the adverbializing affix was glossed, then the word form would have been extracted automatically, resulting in more noise since it displays derivational morphology and no inflectional morphology. Experts were asked to spend no more than six hours on the cleaning task.

For the second preprocessing method, the automatically extracted paradigms were surveyed by a non-expert. Since non-experts could not be expected to identify and correct most issues, they simply removed obvious mistakes such as glosses of stem morphemes that were misidentified as affix glosses and word forms with obviously incomplete glosses or ambiguous glosses (due to identi-

[5] https://github.com/LINGuistLIU/IGT

| Language | paradigms | single-entry | total words | train | validation | test | unannotated |
|---|---|---|---|---|---|---|---|
| arp clean | 16,857 | 10,857 | 56,644 | 283,714 | 14,151 | 14,150 | 6,877 |
| arp noisy | 14,389 | 8,855 | 56,922 | 435,430 | | | |
| ddo clean | 982 | 330 | 7,221 | 35,773 | 2,173 | 2,172 | 9,408 |
| ddo noisy | 945 | 295 | 7,315 | 36,875 | | | |
| lez clean | 301 | 202 | 543 | 539 | 88 | 88 | 3,054 |
| lez noisy | 298 | 188 | 588 | 1,254 | | | |
| mni clean | 479 | 126 | 2,860 | 9,917 | 853 | 852 | 2,593 |
| mni noisy | 428 | 165 | 2,192 | 15,958 | | | |
| ntu clean | 316 | 123 | 1,654 | 5,774 | 473 | 472 | 1,661 |
| ntu noisy | 365 | 167 | 1,646 | 7,886 | | | |

Table 5: Data sizes for noisy extracted paradigms and paradigms cleaned by experts. The columns show the total number of inflectional paradigms extracted from the IGT, the number of paradigms with only a single word entry, the number of three-tuples (source, target, features) in the train/validation/test sets before adding unannotated forms and finally the number of additional unannotated/uninflected word forms.

cal glosses on one or more word forms). For some languages, this cleaning-by-removal made these paradigms smaller than the "cleaned" dataset.

**Step 2: Preparing reinflection data.** The typical morphological reinflection data is in tuple format of `(source form, target form, target features)`. We convert the paradigm data into this format in preparation for reinflection. Table 5 presents the data sizes.

For each language, we prepare the validation and test sets by using the the expert-cleaned data language in the following way: If the paradigm has more than one form, pick a random form as the source form and select the remaining forms in the paradigm with a probability of 0.3 to be "unknown", i.e. to be predicted from the first form. Half of the "unknown" data transformed in this way is used for validation and the other half for testing. The validation and test sets for each language is shared across all the experiments we conduct for that language.

To prepare the training data from the noisy and clean paradigms, we first map each form in the data to itself and add them to the training data. Paradigms with a single entry have only self-to-self mapping. If a paradigm has more than one form, all possible pairs of forms in a paradigm are generated and added to the training data, excluding those that are part of testing or validation set, i.e. "unknown".

**Step 3: Reinflection models and experimental setup.** We experiment with two state-of-the-art models for morphological reinflection, the transformer model for character-level transduction (Wu et al., 2020) and the LSTM sequence-to-sequence model with exact hard monotonic attention for character-level transduction (Wu and Cotterell, 2019). For all the models, we used the implementation of the SIGMORPHON 2020 shared task 0 baseline (Vylomova et al., 2020),[6] and our hyperparameters are the same as the shared task baseline.

After paradigms are extracted and preprocessed, we conduct two experiments to generate "unknown" inflected forms. We then expand those experiments by two data augmentation techniques. First, we add all unannotated/uninflected words from the IGT data to the training data. When tokens that were either unannotated or uninflected are added, they are self-mapped as the source and target forms (as we do with single-entry paradigms), and their morphosyntactic features are annotated with a special tag: `XXXX`. Second, we augment the training data by generating 10,000 artificial instances with the implementation in the SIGMORPHON 2020 shared task 0 baseline of the data hallucination method proposed by (Anastasopoulos and Neubig, 2019). Finally, we combine both additions. These augmentations are intended to overcome data scarcity.

All models and techniques were tested on the same held-out set chosen randomly from multi-

[6] https://github.com/shijie-wu/neural-transducer/tree/f1c89f490293f6a89380090bf4d6573f4bfca76f

entry paradigms in each language.

### 5.4 Results

We compared results when training on the noisy paradigms and on the expertly cleaned paradigms and found that the limited involvement of experts always improved results. We also found the transformer outperformed the LSTM with hard monotonic attention on cleaned data in all instances and on noisy data overall. When comparing results from augmenting the data by artificial and uninflected/unannotated tokens, we find varied results. The results are displayed in Table 6.

There is no clear correlation between accuracy and the total number of annotated tokens or training paradigms (see Tables 4 and 5). Tsez and Arapaho [arp] achieved over 60% accuracy and these languages do have more training data (35K and 283K triples, respectively) than the others (less than 10K). However, even though Arapaho has considerably more training data, its accuracy is lower than Tsez. A slight correlation between accuracy and amount of multi-entry paradigms does exist. Languages with a higher proportion of multi-entry paradigms tend to have better results. Fewer single-entry paradigms may indicate more complete paradigm information.

Any correlation between results and linguistic factors such as language family or morphological type is uncertain because of the limited number of testing languages. Tsez [ddo] gave best results overall. This could be due to its limited allomorphy and very regular inflection which may explain why its relative Lezgi [lez] perform better than languages with more data. Arapaho's poorer performance could be due to its polysynthetic morphology (Cowell and Moss, 2008) which is more complex than the fairly straightforward agglutination in Tsez (Job, 1994) and Lezgi (Haspelmath, 1993). The models do seem less sensitive in recognizing the word structure in Arapaho. When the front part of a stem is incidentally the same as a common inflection affix, the stem is often generated incorrectly.

The factor that seems most clearly correlated with accuracy is the consistency and thoroughness of IGT annotations. The Arapaho, Tsez, and Natügu [ntu] corpora were noticeably more complete (i.e. most morphemes were glossed) and polished. This probably explains why Tsez not only had the best results but also showed the smallest improvements after cleaning. Interestingly, augmentation techniques also helped these languages the least (only artificial data augmentation helped Tsez slightly). It seems, therefore, that results are highest and data augmentation is most helpful when original manual annotations are least consistent or complete.

As might be expected with limited data, errors were most common with irregular or rare forms. For example, the best performing model incorrectly inflected many Lezgi pronouns which have an inflection pattern identical to nouns except for a unpredictable change in the stem vowel. Perhaps related to this, the model also misidentified some epenthetic vowels in several Lezgi nouns. Another interesting pattern involved unique Nakh-Daghestanian (Tsez and Lezgi) case-stacking, where nominal affixes concatenate, rather than substitute each other, to form several peripheral cases such as SUPERELATIVE or POSTDIRECTIVE. The more common affixes in the concatenation string were often generated correctly but the less common concatenated affixes were not. Allomorphy also causes difficulty. Models struggle generating the right form when multiple forms are possible. For example, in Arapaho the third person singular inflection has variations (e.g. -oo, -o, or -'). On the other hand, models learned regular inflectional patterns well enough to correctly inflect forms even where the expert had left misspellings of that form in the clean data.

Finally, we clearly see expert cleaning improved performance across the board (with two negligible exceptions for Tsez and Lezgi on the hard monotonic attention model). Experts were asked to spend no more than six hours and actually spent up to seven but as little as two hours on each language. This indicates that expert labor is well worth its "cost".

## 6 Conclusion

We proposed a new morphological generation task called IGT2P, which aims to learn inflectional paradigmatic patterns from interlinear gloss texts (IGT) produced in linguistics fieldwork. We experimented with neural models that have been used for morphological reinflection and new preprocessing steps as baselines for the task. Our experiments show that IGT2P is a promising method for creating new morphological resources in a wide range of low-resource languages.

| | T | +aug | +uninfl | +both | mono | +aug | +uninfl | +both |
|---|---|---|---|---|---|---|---|---|
| arp clean | **62.08** | 61.39 | 61.58 | 60.78 | 15.93 | 15.75 | 15.58 | 15.94 |
| arp noisy | 57.77 | 57.64 | *58.04* | 57.51 | 14.51 | 14.64 | 14.52 | 14.69 |
| ddo clean | 65.38 | **66.53** | 65.19 | 65.42 | 59.9 | 60.87 | 59.53 | 60.64 |
| ddo noisy | 63.54 | 63.95 | 62.89 | *64.04* | 59.12 | 58.66 | 57.87 | 57.97 |
| lez clean | 46.59 | 32.95 | 46.59 | **48.86** | 32.95 | 35.23 | 31.82 | 31.82 |
| lez noisy | *35.23* | 29.55 | 32.95 | 27.27 | 30.68 | 28.41 | 20.45 | 31.82 |
| mni clean | 30.63 | 30.87 | 31.81 | **32.04** | 23.24 | 25.7 | 21.95 | 24.77 |
| mni noisy | 21.48 | *22.3* | 21.60 | 21.83 | 18.78 | 18.31 | 19.37 | 20.31 |
| ntu clean | **53.18** | 46.82 | 49.15 | 48.52 | 29.66 | 33.9 | 28.18 | 33.05 |
| ntu noisy | 36.86 | 45.55 | 45.34 | *45.76* | 31.99 | 33.69 | 31.78 | 30.93 |

Table 6: Accuracy percentages of reinflection task for transformer model (T) and the LSTM seq2seq model with exact hard monotonic attention (mono) with/out artificial data augmention (+aug), unannotated/uninflected word forms (+uninfl) and both together. Boldface indicates best result; italics indicate best result on noisy paradigms.

With sufficient IGT annotations, IGT2P obtains reasonable performance from noisy data. We investigated the effect of manual cleaning on model performance and showed that even very limited cleaning effort (2-7 hours) drastically improves results. The inherent noisiness in IGT and other linguistic field data can be overcome with limited input from domain experts. This is a significant contribution considering the extensive effort—on the order of months and years—to produce the curated structured data normally used to train NLP models. In languages with the noisiest data performance is improved even further by data augmentation techniques. Finally, since field data does not often include POS annotation, we investigated the usefulness of POS tags for morphological reinflection and find that, surprisingly and in contrast to common assumptions, they are not beneficial to recent state-of-the-art systems. This is a useful discovery for researchers who wish to optimize their inflection systems.

There is room for future improvement. Better techniques for further cleaning might be useful since accuracy seems to have close related to data quality. However, at some point more cleaning will return less improvement. Upper bounds could be established by comparing results on languages with gold standard inflection tables, although polysynthetic languages like Arapaho would make this difficult since their tables do not always include noun incorporation. Better use of experts' time might involve identification of lemmata that could be used to train a lemma-to-form model, rather than the form-to-form mapping used here. Another approach would be to compare improvements between manual-only cleaning and cleaning done by a linguist working with someone who can write scripts to automatically correct repeated patterns of noise.

IGT2P also has implications for the documentation of endangered languages and addressing digital inequity of speakers of marginalized languages. It could be integrated into linguists' workflow in order to improve the study of inflection and increase IGT data. For example, the generated inflected forms could be used for automated glossing of raw text. IGT2P could speed the discovery and description of a language's entire morphological structure. An elicitation step with native speakers could be added to strategically augment data. This would integrate well with linguists' workflow. IGT2P results could serve as to prompt speakers for forms that are rare in natural speech. It might also be integrated into linguistic software such as FLEx.

## Acknowledgments

We are obliged to Drs. Brenda Boerger, Shobhana Chelliah, Bernard Comrie, and Andy Cowell, as well as Chuck Donet, and Andrew Brumleve for generously sharing their field data and to Mary Burke, Drs. Boerger and Cowell, and Andrew Brumleve who did expert cleaning of the data. Also, to anonymous reviewers for their insightful comments.

## References


Farrell Ackerman, James P. Blevins, and Robert Malouf. 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In James P. Blevins and Juliette Blevins, editors, *Analogy in Grammar: Form and Acquisition*, pages 54–82. Oxford University Press.

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

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Emily M. Bender. 2014. Language CoLLAGE: Grammatical Description with the LinGO Grammar Matrix. In *Proceedings of the Ninth International Conference of Language Resources and Evaluation (LREC-2014)*, pages 2447–2451.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017a. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017b. Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 759–765, Valencia, Spain. Association for Computational Linguistics.

Andrew Cowell and Alonzo Moss. 2008. *The Arapaho Language*. University Press of Colorado.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

Paul Felt. 2012. *Improving the Effectiveness of Machine-Assisted Annotation*. Thesis, Brigham Young University.

Martin Haspelmath. 1993. *A grammar of Lezgian*. Mouton de Gruyter, Berlin; New York.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya D. McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. *arXiv*.

Michael Job, editor. 1994. *The indigenous languages of the Caucasus. Volume 3: The North East Caucasian languages. Part 1*, volume 3. Caravan Books, Delmar, N.Y.

Katharina Kann, Samuel R Bowman, and Kyunghyun Cho. 2020a. Learning to learn morphological inflection for resource-poor languages. *arXiv preprint arXiv:2004.13304*.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017a. Neural multi-source morphological reinflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 514–524, Valencia, Spain. Association for Computational Linguistics.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017b. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1993–2003, Vancouver, Canada. Association for Computational Linguistics.

Katharina Kann, Arya McCarthy, Garrett Nicolai, and Mans Hulden. 2020b. The sigmorphon 2020 shared task on unsupervised morphological paradigm completion. *arXiv preprint arXiv:2005.13756*.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Geraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, and Arya D. McCarthy. 2016. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA)*, pages 1868–1873.

Haley Lepp, Olga Zamaraeva, and Emily M. Bender. 2019. Visualizing inferred morphotactic systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 127–131, Minneapolis, Minnesota. Association for Computational Linguistics.

William D. Lewis and Fei Xia. 2010. Developing ODIN: A multilingual repository of annotated language data for hundreds of the world's languages. *Literary and Linguistic Computing*, 25(3):303–319.

Peter Makarov and Simon Clematide. 2018. Imitation learning for neural morphological string transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2877–2882, Brussels, Belgium. Association for Computational Linguistics.

Robert Malouf. 2016. Generating morphological paradigms with a recurrent neural network. *San Diego Linguistic Papers*, 6:122–129.

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

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931, Denver, Colorado. Association for Computational Linguistics.

Abhishek Sharma, Ganesh Katrapati, and Dipti Misra Sharma. 2018. IIT(BHU)–IIITH at CoNLL–SIGMORPHON 2018 shared task on universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 105–111, Brussels. Association for Computational Linguistics.

Miikka Silfverberg and Mans Hulden. 2018. An encoder-decoder approach to the paradigm cell filling problem. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2883–2889. Association for Computational Linguistics.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Ponti, Rowan Hall Maudslay, Ran Zmigrod, Joseph Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrej Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. The SIGMORPHON 2020 Shared Task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

David Allen Wax. 2014. *Automated Grammar Engineering for Verbal Morphology*. Thesis, University of Washington.

Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2020. Applying the transformer to character-level transduction. *arXiv:2005.10213 [cs.CL]*.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 207–216, Hong Kong. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

## Appendix

| | transformer model (%) | | | | | | Exact hard mono model (%) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | High | | Medium | | Low | | High | | Medium | | Low | |
| **Language** | **+pos** | **-pos** | **+pos** | **-pos** | **+pos** | **-pos** | **+pos** | **-pos** | **+pos** | **-pos** | **+pos** | **-pos** |
| Adyghe | 99.9 | 99.9 | 93.1 | 93.4 | 43.2 | 41.5 | 99.9 | 99.7 | 91.4 | 91.7 | 32.5 | 33 |
| Arabic | 95 | 95.1 | 79.5 | 79.5 | 2 | 2.5 | 92.5 | 93 | 66.8 | 65.6 | 0 | 0 |
| Basque | 99 | 99.2 | 93.7 | 93.7 | 24.1 | 26.9 | 98.5 | 98.8 | 73.7 | 71.6 | 0.1 | 0.5 |
| Finnish | 95.7 | 95.1 | 78.9 | 79.4 | 0.3 | 0.1 | 93.1 | 93.8 | 58.6 | 54.2 | 0 | 0 |
| German | 91.1 | 90.5 | 73.3 | 73.9 | 3.8 | 5.4 | 90 | 90.1 | 71.2 | 71.2 | 2.9 | 3.6 |
| Persian | 100 | 100 | 93.2 | 94.7 | 12.1 | 12.3 | 99.7 | 100 | 87.4 | 88.3 | 2.8 | 1.6 |
| Russian | 93.3 | 93.2 | 80.9 | 79.6 | 2.2 | 2.6 | 92 | 92 | 68.7 | 69.3 | 0 | 0.9 |
| Spanish | 97.8 | 97.9 | 90.3 | 89.4 | 8 | 7.3 | 97.5 | 96.5 | 77.8 | 73.6 | 6.2 | 6.5 |
| Swahili | 100 | 100 | 94 | 94 | 35 | 35 | 100 | 100 | 88 | 85 | 3 | 2 |
| Turkish | 98.4 | 98.6 | 88.7 | 88.7 | 6.7 | 5.2 | 97.3 | 97.1 | 74.7 | 71.5 | 0 | 0.1 |

Table 7: **Detailed Results for POS experiments.** Morphological inflection accuracy (%) for languages using and not using POS for the transformer model and the LSTM seq2seq model with exact hard monotonic attention in different training data size settings. *+pos* is including POS in the feature descriptions and *-pos* is excluding POS in the feature descriptions.

| | **T** | **+aug** | **+uninfl** | **+both** | **mono** | **+aug** | **+uninfl** | **+both** |
|---|---|---|---|---|---|---|---|---|
| arp clean | 10:55:55 | 11:46:45 | 14:55:17 | 9:51:25 | 2:02:02 | 2:15:51 | 3:00:02 | 2:14:14 |
| arp noisy | 6:36:37 | 6:18:37 | 10:16:38 | 6:42:19 | 2:42:41 | 2:46:29 | 4:03:22 | 3:14:27 |
| ddo clean | 1:54:09 | 1:57:28 | 3:57:43 | 3:58:00 | 0:09:56 | 0:10:42 | 0:18:54 | 0:15:04 |
| ddo noisy | 1:51:07 | 1:56:24 | 3:23:37 | 3:47:12 | 0:08:34 | 0:10:59 | 0:20:54 | 0:19:41 |
| lez clean | 0:29:05 | 0:37:26 | 1:03:58 | 1:02:38 | 0:00:20 | 0:01:53 | 0:02:02 | 0:04:21 |
| lez noisy | 0:32:02 | 0:37:22 | 0:56:55 | 0:59:00 | 0:00:29 | 0:01:40 | 0:01:52 | 0:02:27 |
| mni clean | 1:15:06 | 1:16:19 | 2:12:52 | 2:05:02 | 0:03:56 | 0:04:42 | 0:08:17 | 0:10:11 |
| mni noisy | 1:16:59 | 1:18:55 | 2:13:06 | 2:14:21 | 0:04:32 | 0:08:41 | 0:07:20 | 0:08:09 |
| ntu clean | 1:09:01 | 0:58:37 | 1:28:45 | 1:29:39 | 0:02:19 | 0:03:34 | 0:02:40 | 0:05:53 |
| ntu noisy | 1:00:25 | 1:01:40 | 1:36:53 | 1:38:05 | 0:02:22 | 0:03:59 | 0:03:08 | 0:05:09 |

Table 8: **Details on Computing.** Training time of our models. All models have been trained on an NVIDIA GP102 [TITAN Xp] GPU.