Fine-grained Morphosyntactic Analysis and Generation Tools for More Than One Thousand Languages

Garrett Nicolai, Dylan Lewis, Arya D. McCarthy, Aaron Mueller, Winston Wu, David Yarowsky
Center for Language and Speech Processing
Johns Hopkins University
(gnicola2, dlewis77, arya, amueller, wswu, yarowsky)@jhu.edu

Abstract

Exploiting the broad translation of the Bible into the world’s languages, we train and distribute morphosyntactic tools for approximately one thousand languages, vastly outstripping previous distributions of tools devoted to the processing of inflectional morphology. Evaluation of the tools on a subset of available inflectional dictionaries demonstrates strong initial models, supplemented and improved through ensembling and dictionary-based reranking. Likewise, a novel type-to-token based evaluation metric allows us to confirm that models generalize well across rare and common forms alike.

Keywords: morphology, low-resource, tools

1. Introduction

Recently, computational linguistics has observed notable gains on a myriad of tasks, with new benchmarks being established and broken at an unprecedented pace. While the breadth of the success is impressive, its depth remains rather shallow - research is concentrated in a small number of languages. Part of this concentration is practical - despite Ethnologue classifying approximately 7,000 living languages as of 2019 (Lewis et al., 2015), more than 50% of internet traffic is concentrated in just 3 languages (English, Mandarin Chinese, and Spanish), with 76% represented by 1. Furthermore, despite the fact that the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) marks the world’s languages along 192 dimensions, these 10 languages observe only 93 (48.4%) of these categories.

While most of the world’s languages are lacking tools and datasets along many linguistic dimensions, we concentrate on the specific task of inflectional morphology. In languages like English, the syntactic role of words in a sentence is largely dictated through word order and the use of function words. In the sentence “Mary wished she had caught the earlier bus.”, it can be inferred that Mary is the subject of the sentence, and an earlier bus is the direct object of the verb — the former precedes the main verb, while the latter follows it. In a language with nominal case, such as Czech, the word order may be free — Mary may precede or follow the verb, but will appear in a distinct form that indicates that she is the subject; likewise, an earlier bus will mark that it is a direct object. Similarly, the verbs may mark features such as tense — which is marked in English — mood, aspect, person, and many other features. In this example, Spanish would need to mark the verb catch as hypothetical — although Mary wished she had caught the bus, she did not do so. Some languages can realize a single hypothetical — although Mary wished she had caught the bus, she did not do so. Some languages can realize a single hypothetical — although Mary wished she had caught the bus, she did not do so. Some languages can realize a single hypothetical — although Mary wished she had caught the bus, she did not do so. Many languages reali...
encourage and facilitate research in a much wider range of the world’s languages.

This paper progresses as follows: Section 2 gives a brief overview of inflectional morphology, and describes the key operations covered by our inflectional tools. Section 3 establishes the current state of affairs in computational inflection, in both high- and lower-resource settings. Section 4 describes the process of inducing inflectional paradigms from projected morphological hypotheses, as well as the process of training the inflectional tools. Section 5 describes the data used to train our tools, including statistics regarding its inflectional diversity and appropriateness as an inflectional dataset. Section 6 describes distribution details for our tools. Section 7 discusses our evaluation metrics, including our new metric for the approximation of token-based evaluation, and evaluates our tools on a subset of our languages. Section 8 concludes the paper.

2. Morphological Analysis and Generation

Languages with productive inflectional morphology will have a very high type-to-token ratio. Whereas English nouns typically realize three forms (e.g. book, books, book’s), and verbs four to five (e.g. see, sees, saw, seeing, seen), other languages produce much a larger number of inflected forms. For instance, a Polish verb may take on over 100 forms (Sadowska, 2012). Even in a very large corpus, a large majority of these inflections will never be observed. Any algorithm acting at the word level will have a very high proportion of out-of-vocabulary words.

Fortunately, inflectional morphology tends to follow semi-regular patterns, known as inflectional paradigms. Just as the past tense in English is regularly marked with an -ed suffix, inflectional categories in other languages are marked in similar ways across words. If we can identify the inflectional category of a small number of words, we can train morphosyntactic tools to extend this knowledge to other, unknown forms. Figure 2 demonstrates four core inflectional operations, as evidenced on a German verb.

Inflection Generation produces surface realizations by applying morphosyntactic features to a base form, often referred to as the lemma. Going in the opposite direction, Morphological Analysis produces a lemma and a set of morphosyntactic features from a surface realization. Morphological analysis subsumes two individual tasks: Lemmatization, which identifies the lemma from the surface form, and Morphological Tagging, which identifies the set of features. Morphological analysis can be used to reduce the type-to-token ratio - every inflection is reduced to its lemmatic class, while generation can be used to produce surface forms during a generation step, restoring fluent text.

3. Related Work

Morphological inflection has been thoroughly studied by the community, and has seen a renewed interest in the past few years, as neural models cannot afford to devote their limited vocabularies to forms that are only observed a handful of times. Character-based models, and algorithmic approximations of morphology, such as BPE (Sennrich et al., 2015) and SentencePiece (Kudo and Richardson, 2018) have helped these models overcome some of the difficulties presented by inflectional sparsity, but may fail to capture the nuances of true morphological information.

Recently, much of the work in the building of inflectional tools has been promoted by the Shared Tasks in Morphological Inflection (Cotterell et al., 2016; Cotterell et al., 2017; Cotterell et al., 2018a; McCarthy et al., 2019). Side-by-side with the UniMorph project (Kirov et al., 2018), which aims to increase the availability of morphological dictionaries, these tasks have encouraged the rapid progression of the state of the art in inflection generation and lemmatization in more than 100 languages. In high-resource settings (i.e., 10,000 annotated training instances), many of the languages covered by these tasks are near-solved, with inflectional generators producing accuracies in the high 90s. When data is scarcer, however, there is still much room for improvement - the best systems average below 60%, with many languages performing much worse. Furthermore, these tasks assume the presence of gold-annotated inflectional dictionaries. These dictionaries are expensive to create, and for many of the world’s languages, finding annotators can be a difficult task.

An alternative to hand-annotated dictionaries is automatic induction via semi-supervised methods. The class of methods introduced by Yarowsky et al. (2001) project information from a high-resource parse and a bilingual dictionary. Fossum and Abney (2005) and Agic et al. (2015) exploit the parallel nature of the Bible to project POS tags to lower-resources languages, using this projection to then train POS-taggers. Buys and Botha (2016) extend this tagging paradigm to morphological tagging, projecting morphological tags onto a low-resource language, and training a tagger on these morphologically-aware tags. In the op-

---

4How the lemma is defined is a matter of some linguistic discussion, and is beyond the scope of this paper. For our purposes, “lemma” is interchangeable with “citation form”.

---

Figure 2: Four inflectional operations: generation, analysis, lemmatization, and tagging.
Figure 3: Projecting induced morphological categories and lemmas across an alignment from English to Finnish. Red arrows indicate a right-to-left syntactic dependency.

We expand upon the work of Nicolai and Yarowsky (2019), who first richly annotate the English side of a bitext before projecting morphological information onto the low-resource text. The paucity of English morphological tagging is augmented through the heuristic interpretation of syntactic and semantic parses, as well as a reverse projection of a number of other high-resource languages. Our main contribution over their work is the expansion of the language set by a factor of 40 (from 26 languages to more than 1000). We also augment the set of inflectional features covered by their methods and incorporate frequency statistics into their learning model. Details follow in Section 4.

4. Morphological Induction

Traditionally, inflection has been a heavily supervised task, assuming the existence of an inflectional dictionary. In this section, we elaborate and extend the recent work of Nicolai and Yarowsky (2019), who lessen the existence assumption to that of a small parallel corpus and an optional bilingual dictionary.

Parallel corpora are themselves rare; however, there do exist several small texts, such as the Universal Declaration of Human Rights (UDHR)\footnote{https://www.ohchr.org/EN/UDHR/} that have been translated into a large number of the world’s languages. Of these corpora, the Bible stands out as a uniquely suitable corpus for projection and induction. First, it is several orders of magnitude larger than the UDHR, with the New Testament containing more than 7,000 sentences, and a full translation consisting of nearly 40,000. Second, the Bible has a very specific structure that simplifies its parallelizability: it is divided into short verses that contain roughly the same semantic content across translations. Each verse is numbered according to a defined canon, allowing the Bible to be approximately parallelized across translations with little linguistic knowledge.

Starting from the Bible, we learn a word-level alignment between the English and target translations. To strengthen the alignment signal, we concatenate 27 translations of the English Bible and align them to each target Bible. The English Bible is syntactically and semantically parsed as well as lemmatized. The resulting parse is used to hypothesize inflectional categories on the English Bible, which are then projected along the alignment to the target Bible.

Nicolai and Yarowsky (2019) heuristically derive morphological feature tags on the English from the syntactic and semantic parse. For example, in the phrase “I baptize you with water”, the syntactic parse indicates that water is governed by the preposition “with”, and the semantic parse indicates an instrumental use. Both highly suggest that “water” is in the instrumental case.

Lemmas are projected in a similar fashion, with the help of a bilingual dictionary. The English lemma is first translated via the dictionary — if there are multiple translations each is considered as a candidate target lemma. Candidates are then compared with the target inflected form, and only those that pass a minimum edit-distance threshold are retained.

Bilingual dictionaries are rare and often small; in the absence of a large bilingual dictionary, we use the alignment to induce one. We consider as translations any target form that is frequently aligned to an English citation form. Only those dictionaries with fewer than 20,000 individual entries are expanded in such a manner. The impoverished inflectional morphology of English means that some English forms will match the citation form, despite being inflected in other languages. Consulting the parses, we only keep lemma hypotheses aligned to nominal subjects, their modifying adjectives and non-finite verbs.

The projection process is illustrated in Figure 3. Note that although some words in the source sentence are not aligned, such as “of”, their information is not necessarily lost — “of” is used to mark “men” as the genitive case, which is then projected onto the Finnish translation.

Nicolai and Yarowsky (2019) only consider two parts of speech — nouns and verbs — predicting plurality and case for the former, and temporality for the latter. We extend the verbal inflectional categories to include person and plurality. English verbs do not mark person explicitly, but much of the information can be inferred from the use of
We also extend the inflectional induction to include adjectives, which must agree with their modified noun in many languages. A syntactic parse marks the adjectival modifier of a noun, and can be used to extend nominal information to an adjective. Figure 4 illustrates the process. The syntactic parse indicates that the adjective “wise” modifies the noun “men”. “Men” has already been tagged as a nominative plural, and this information percolates up the dependency relation, so that “wise” is tagged with the same information.

4.1. Training Morphological Tools
The projection process creates a set of morphological tuples for each target language. Each tuple consists of an inflected form, a lemma, a set of morphosyntactic features that identify the paradigmatic category of the inflection, and the frequency that this inflectional relation occurred in the aligned corpora. We use the notation \{Inflection, Lemma, Inflectional Category, Frequency\} to represent one of these tuples; an inflectional category is composed of several inflectional features defined by the UniMorph (Kirov et al., 2018) schema. Many words may have multiple interpretations depending upon context and thus may be represented by several individual tuples. For example, the German form \textit{gesegnet} may be the past participle of a verb: \{\textit{gesegnet}, \textit{segnen}, VB;PTCP.PST, 15\}, or an adjective: \{\textit{gesegnet}, \textit{gesegnet}, JJ;NOM;SG, 8\}. Note that not only do the morphological categories of the two words differ, but the lemmas differ as well.

Nicolai and Yarowsky (2019) use these sets of tuples to train morphological analyzers, but they disregard the frequency of the inflectional processes. Alignment and projection are noisy operations. Without considering frequency, rare processes are given the same weight as frequent ones. This noise then unfairly biases the learning algorithms. We mitigate this issue by incorporating frequency statistics into training. The training data is then used to train a sequence-to-sequence character transducer. Morphological analyzers are trained by presenting the inflected form as input, producing a lemma and inflectional category; the data is reversed to train morphological generators (see Figure 2).

4.2. Reranking
Due to the inherent levels of noise in operations such as alignment and projection, the inflectional tools may produce the correct output, but not as the top prediction for a given input. One simple method to promote desirable forms is dictionary reranking. Under this paradigm, a gazetteer of known valid forms is combined with an initial list of hypotheses, with observed forms being promoted to the top of the list. For analysis, a list of lemmas can promote more likely analyses; in the opposite direction, we can promote a list of observed forms from a monolingual corpus. Figure 5 demonstrates this process. Both generators and analyzers can produce an \(n\)-best list that contains the predicted output, its rank, and a confidence score. Any output form that is observed in the supplied gazetteer has its score increased by the score of the previous best hypothesis. In this example, “sah” is originally ranked third in a list of hypotheses, but neither of the predictions above it are observed forms. By increasing the confidence for forms listed in the gazetteer, we promote attested words to the top of the list while preserving the order of the hypotheses.

5. Data
All of our generators and analyzers are trained on inflectional paradigms extracted from the Bible corpus of McCarthy et al. (under review), which contains 4032 parallel translations in 1108 languages. Among these are 27 English translations, whose archaic forms have been replaced with their modern equivalents (i.e., “believeth” is replaced with “believes”). The English and target Bibles have been aligned using the Berkeley aligner (Liang et al., 2018).

---

6The 2nd person plural and singular pronouns are identical in modern English. 2nd person verbs are thus marked for number only. We leave the 2nd person plurality distinction to future work.
POS-tagged and syntactically-parsed using the Stanford NLP toolkit (Manning et al., 2014), and semantically-parsed using the Deep Semantic Role Labeler (He et al., 2017). If a target Bible has multiple translations, they are all used to extract inflectional paradigms.

The Bible corpus contains languages representing a wide spectrum of morphological phenomena, including languages that do not inflect at all. Many Polynesian languages, such as Indonesian, Tonga, and Balinese, for example, have very little inflectional morphology on nouns, verbs, and adjectives. Likewise, the Sino-Tibetan language family, containing languages like Mandarin Chinese, observes very little inflectional morphology (cf. Tibetan).

The URIEL typological database (Littel et al., 2016) categorizes almost 8000 languages and dialects along 290 dimensions. Unlike other typological databases, such as WALS (Dryer and Haspelmath, 2013), or PHOIBLE (Moran and McCloy, 2019), which require expert annotations of typological features, URIEL is an automatically-derived database that extrapolates from other typological sources, allowing for much wider coverage. It is the only database that covers all of the languages in the Bible corpus.

Of the dimensions marked by URIEL, 12 are exclusively concerned with morphology. Table 1 gives a breakdown of how many languages in the Bible corpus observe each of the appropriate features, as well as the percentage of the entire corpus such coverage represents.

Table 1: Number and percentage of the dataset that observes specific morphological phenomena. Note that several languages observe multiple phenomena, so the percentages will add up to more than 100%.

| Feature                  | Number | % of dataset |
|--------------------------|--------|--------------|
| Plural Prefix            | 111    | 10.0         |
| Plural Suffix            | 635    | 57.5         |
| Plural (Any)             | 746    | 67.5         |
| Case Prefixing           | 54     | 4.9          |
| Case Suffixing           | 290    | 26.2         |
| Case Proclitic           | 14     | 1.3          |
| Case Enclitic            | 57     | 5.2          |
| Case Other               | 112    | 10.5         |
| Case (Any)               | 526    | 47.6         |
| Adjective Agreement      | 590    | 53.4         |
| Past Vs. Present         | 599    | 54.2         |
| Future Affix             | 584    | 52.9         |
| 3-way distinction        | 389    | 35.2         |
| Tense/Aspect/Mood Prefix | 294    | 26.6         |
| Tense/Aspect/Mood Suffix | 586    | 53.0         |
| Tense/Aspect/Mood (Any)  | 880    | 79.6         |
| None                     | 79     | 7.1          |

These statistics demonstrate that the Bible corpus is a morphologically diverse dataset and furthermore that it follows typical linguistic trends observed in the literature. Namely, suffixing is much more frequent than prefixing (Dryer, 2013), and that past-tense (Osten Dahl and Velupillai, 2013) and case marking (Iggesen, 2013) are roughly evenly distributed between languages that observe it and those that do not. We also note that verbal marking is much more prevalent than nominal marking, which in turn occurs more frequently than adjective agreement. Furthermore, languages that inflect nouns are much more likely to mark plurality than case.

Table 2: The average number of times that any verbal lemma will appear as a given inflection in a single translation of the Bible.

We also collect statistics on the inflectional coverage appearing in the English Bibles. Verbal statistics are collected in Table 2. For each inflectional category, we calculate both the number of individual lemmas observed, as well as the average frequency across all English Bibles. There are 3642 verbal lemmas inflected across the 27 English Bibles, including 1009 hapaxes. While the past tense and the 3rd person singular form a majority of finite verb forms, even the rarer categories are observed with some regularity. For example, the 1st person plural future form is observed on average 8.6 times in each English Bible; that is, 66 different verbs are inflected to this category on average 0.13 times in each Bible. When projecting across to a target language Bible, each category demonstrates both moderate frequency and wide lemmatic coverage.

Table 3 demonstrates similar observations regarding nouns. The 27 translations of the English Bible contain 10,118 nominal lemmas, including 3,292 hapaxes. As would be expected, the four Germanic cases (nominative, accusative, dative, and genitive) are highly represented, with several hundred individual instances occurring in each Bible. However, other cases are also represented, with a wide distribution of lemmas across Bibles, and a small number of individual instances within a single Bible.

6. Distribution Details

We train two separate sequence transduction models: one neural, and one non-neural, and construct an ensemble that combines their hypotheses. We make these models available to the community, but understanding that portability

---

7Words that occur only once.
We provide a script that enables the use of the generators and analyzers described herein. It is written in Python 3, and requires that the appropriate training tools be installed locally. It requires a list of words to analyze, the 3-character ISO 639-3 code for the language (with the option -l), a configuration file (-c) and a location to write the output (-o). If no other options are provided, it will assume that the user wishes to perform morphological analysis. The user can specify inflection generation with the (~g) option. The configuration file is a tab-separated file indicating the locations of the trained models for each language. Each row in the file must include the ISO code of the language, followed by 3 database locations: the location of a inflectional dictionary, the location of the DTL trained model, and the the location of the neural model. If such a model is non-existent, the file may specify N/A.

The wrapper will first look for the input word in the provided inflectional dictionary; if the word is found, the analysis from the dictionary is provided; if not, then the provided models will be run, and ensembled, if possible. The user can specify the length of the n-best list of the models (~n). 1 is the default.

If a lemma-dictionary or type-dictionary is available, it can be provided with the (~d) option, which will enable dictionary reranking. The dictionary should consist of a single column of acceptable word forms.

### 7. Evaluation

We construct both inflection generators and morphological analyzers for more than 1000 languages. In this section, we evaluate the quality of the tools that we present to the community. It is not possible to collect evaluation sets for each and every language represented in this dataset — the evaluation of morphological tools requires annotated inflectional dictionaries, which do not exist for a majority of languages. We instead evaluate on a subset of the languages for which we do have inflectional dictionaries. We turn to UniMorph ([Sylak-Glassman et al., 2015]; [Kirov et al., 2018]), a collection of morphological dictionaries that spans more than 100 languages. Of these languages, 50 overlap with our languages, and can be used as a test set. For each language, we extract a validation set of 500 randomly-sampled tuples of the form \{LEMMATIZATION\}, \{INFLECTED\}, \{MORPH\}, and a test set of 1000 instances. For example, an English instance of the word “played” would appear as \{play, played, PST\}. The validation set is used to tune hyper-parameters, and for early stopping of the neural models. The languages are listed in Table 3.

From these tuples, we construct test sets for both generation and morphological analysis. Generators must correctly produce an inflected form when given a lemma and inflectional category as input. If a given input pair has multiple correct outputs, such as \{dive, dived/dove, PST\}, a generator is evaluated as correct if it produces any of the correct forms; however, each type is only evaluated once.

In the opposite direction, we evaluate two separate productions. Recall from Figure 2 that morphological analysis is a composite of two sub-tasks — we report the accuracy of our tools on both complete analysis, as well as the sub-task of lemmatization. Analysis accuracy requires both the correct lemma and the correct inflectional category for a given inflected form. Lemmatization accuracy evaluates the same hypothesis, but only requires that the correct lemma be returned. For example, the German type “gehend” - going can either be a present participle with the lemma gehen, or an adjective with the lemma gehend. Thus, while the analysis \{gehend, VB;PST,PTCP\} has the correct lemma,
Table 4: Distribution of Evaluation Languages

| Family         | Number of Languages |
|----------------|---------------------|
| Albanian       | 1                   |
| Armenian       | 1                   |
| Balto-Slavic   | 11                  |
| Bantu          | 2                   |
| Celtic         | 1                   |
| Dravidian      | 2                   |
| Germanic       | 7                   |
| Hellenic       | 1                   |
| Indo-Aryan     | 2                   |
| Iranian        | 1                   |
| Italic         | 1                   |
| Kartvelian     | 1                   |
| Quechuan       | 1                   |
| Romance        | 6                   |
| Semitic        | 3                   |
| Turkic         | 5                   |
| Uralic         | 4                   |

Table 5: Generation, Lemmatization, and Analysis accuracy at the type level.

| System            | Generation | Lemma | Analysis |
|-------------------|------------|-------|----------|
| DTL@1             | 21.1       | 23.6  | 10.5     |
| M&C@1             | 25.0       | 24.1  | 9.7      |
| Ensemble@1        | 24.1       | 25.2  | 11.6     |
| Ensemble+RR@1     | 24.1       | 44.1  | 16.9     |
|                  |            |       |          |
| DTL@5             | 35.6       | 39.1  | 20.2     |
| M&C@5             | 43.9       | 46.6  | 19.9     |
| Ensemble@5        | 43.0       | 47.3  | 23.7     |
| Ensemble+RR@5     | 43.0       | 60.3  | 31.0     |
|                  |            |       |          |
| DTL@50            | 48.8       | 52.1  | 31.0     |
| M&C@50            | 60.7       | 66.9  | 32.9     |
| Ensemble@50       | 57.7       | 68.0  | 37.0     |
| Ensemble+RR@50    | 57.7       | 69.3  | 39.5     |

7.2. Token Accuracy

While type accuracy can demonstrate how well a given system can generalize to new lemmas, it is in some ways artificial: while a Polish verb can be inflected in hundreds of different ways, a large number of the inflectional slots will only rarely be used in standard text and speech. A system that correctly predicts the 80% least frequent types with 100% accuracy is very different than one that correctly analyzes types with 80% accuracy, regardless of type frequency. Ideally, a token-based evaluation metric would consider a corpus annotated for morphological inflection, and weight each type by its frequency. This is infeasible for a number of reasons. Even if such a corpus existed in more than a handful of languages, morphological fecundity prevents all but a very small number of inflections from appearing in any text - most types would receive a weight of 0.

We propose instead a simple approximation to convert type-based accuracy into token-based accuracy. We only assume the presence of an annotated English corpus, and a raw monolingual corpus in the target language. All reported results are on the best reported system from Figure 5 (ie, the Ensemble with reranking).

We can weight input types by the likelihood that they appear in a monolingual corpus. Due to the inflectional sparsity problem, we use a smoothed Witten–Bell unigram language model to arrive at a normalized likelihood for each inflection in our evaluation set. We do not know the correct distribution of inflectional categories, so instead distribute them uniformly. This is a naive assumption, but it allows us to achieve an initial approximation, which will be modified in subsequent steps. We refer to this metric as our “Surface-based” metric, and use the target translation of the Bible as our monolingual corpus. Our generators and analyzers are evaluated using this metric in Table 6.

We typically see that the surface-based accuracy surpasses
the type-based accuracy by several percentage points, suggesting that our tools are consistent across frequency levels - if our tools were exploiting rare types, it would perform much worse on the token level. It is not overly surprising that the tools can handle common types - these are likely to be discovered by the induction algorithm. If one of these types is irregular, the model should learn its morphology in a highly specialized context.

There are two disincentives to using a metric that establishes token frequency based on the frequency of target types. Smoothing helps mitigate the under-representation of rare types, but the domain of the corpus also heavily biases certain types. The test set of Unimorph is largely modern terminology, which is heavily disjoint with Bible text. While it might be possible for some languages to find a large enough monolingual text in the correct domain, we instead propose a new metric that approximates token frequency from an English corpus and a bilingual dictionary. We first lemmatize our English corpus, and establish the smoothed probability of each lemma as \( P\{\text{Lemma}\} \). We can then determine the likelihood of a target lemma, \( P\{\text{Target}\} \) through the use of a translation model. Ideally, this likelihood would come from a high-quality translation, but lacking such a model, we use a bilingual dictionary, with equal probability assigned to each translation. If a target lemma is not in our dictionary, we assign it the smoothed likelihood of an unseen lemma in the English corpus. Thus, the probability of a target lemma, \( P\{\text{Target}\} \) is equal to Equation 1. From our annotated English corpus, we can obtain the likelihood of each inflectional category in a corpus, denoted \( P\{\text{Morph}\} \). Then, the probability of an inflected form, \( P\{\text{Inflected}\} \) is equal to Equation 2.

\[
P\{\text{Target}\} = P\{\text{Lemma}\} \times P\{\text{Translation}\} \tag{1}
\]

\[
P\{\text{Inflected}\} = P\{\text{Target}\} \times P\{\text{Morph}\} \tag{2}
\]

For example, consider the French type \{finirons, finir, 1;PL;FUT\}. Our dictionary tells us that finir translates as to finish and to complete. Finish appears in the English corpus with a likelihood of 0.00047, and complete appears with a likelihood of 0.00018. The only back-translation of to complete in the dictionary is finir, while to finish back-translates to both finir and terminer. \( P\{\text{finir}\} \) is then equal to 0.5 * 0.00047 + 1.0 * 0.00018 (ie, to finish only has a 50% probability of translating to finir. Finally, the 1;PL;FUT feature set appears with a likelihood of 0.009% in the English Bible, so the final likelihood of \{finirons, finir, 1;PL;FUT\} is 0.00009 * 0.000415 = 3.735 * 10^{-6}.

To investigate the variance of this metric, we determine the English lemma likelihood using two differently-sized corpora of different genres. The Brown corpus (Kucera and Francis, 1979) consists of approximately 50K sentences across numerous genres and domains. The LORELEI project (Onyshkevych, 2014) provides parallel disaster snippets for low-resource languages, and consists of approximately 1500 unique sentences. Table 5 summarizes the token-level accuracies of our generators and analyzers.

| System      | Generation | Lemma | Analysis |
|-------------|------------|-------|----------|
| Types@1     | 24.1       | 44.1  | 16.9     |
| Surface@1   | 36.6       | 45.5  | 19.5     |
| Brown@1     | 36.9       | 52.2  | 26.4     |
| LORELEI@1   | 38.3       | 51.9  | 26.1     |
| Types@5     | 43.0       | 60.3  | 31.0     |
| Surface@5   | 59.8       | 66.6  | 40.2     |
| Brown@5     | 57.6       | 71.4  | 46.5     |
| LORELEI@5   | 58.5       | 70.0  | 45.8     |
| Types@50    | 57.7       | 69.3  | 39.5     |
| Surface@50  | 71.1       | 72.9  | 45.2     |
| Brown@50    | 70.2       | 77.1  | 57.1     |
| LORELEI@50  | 71.2       | 77.5  | 56.9     |

Table 6: Generation, Lemmatization, and Analysis accuracy at the type level.

We first note that for generation, our new metric is very close to the original surface-based metric. This suggests that our combination of lemma, translation, and morphology probabilities is a modest approximation of actual type distribution. For lemmatization, we see that the token-based metrics are moderately higher than the surface-based one. While some of this improvement may be attributed to noise, we instead suggest that lemmatization, being an easier task than generation, is simply benefiting from the domain-shift to a modern corpus.

We further observe that although the two corpora are very different in size and genre, they do not vary significantly in their reported numbers, suggesting that they are both reasonably representing common lemmas, with the smoothed probability of less common lemmas also approximating the true distribution.

8. Conclusion

We have learned inflectional generators and analyzers on a scale never before attempted. Leveraging a corpus of parallel Bibles, we project inflectional categories from automatically-annotated English, creating tools that learn nominal, verbal, and adjectival morphology in more than 1000 languages. Intrinsic evaluation demonstrates that the tools are able to generalize beyond their initial Bible lexicons to modern terminology, and a novel type-to-token conversion metric further demonstrates their applicability over a wide range of inflectional types. We freely distribute these tools to the community, in the hopes that they will encourage research in languages that have previously been underserved.

Bibliographical References

Agić, Ž., Hovy, D., and Søgaard, A. (2015). If all you have is a bit of the Bible: Learning POS taggers for truly low-resource languages. 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), volume 2, pages 268–272.

Buys, J. and Botha, J. A. (2016). Cross-lingual morphological tagging for low-resource languages. In Proceedings
M., Mielke, S. J., Heinz, J., et al. (2019). The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. *arXiv preprint arXiv:1910.11493*.

McCarthy, A. D., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (2020). The johns hopkins university bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

McCarthy, A., Wicks, R., Lewis, D., Mueller, A., Wu, W., Adams, O., Nicolai, G., Post, M., and Yarowsky, D. (under review). The Johns Hopkins University Bible Corpus: 1600+ tongues for typological exploration. In *Submitted to LREC 2020*.

Steven Moran et al., editors. (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Nicolai, G. and Yarowsky, D. (2019). Learning morphosyntactic analyzers from the Bible via iterative annotation projection across 26 languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1765–1774, Florence, Italy, July. Association for Computational Linguistics.

Onyshkevych, B. (2014). Low resource languages for emergent incidents (loirelei). *DARPA Broad Agency Announcement (DARPA-BAA-15-04)*.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Sadowska, I. (2012). *Polish: A comprehensive grammar*. Routledge.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Soricut, R. and Och, F. (2015). Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado, May–June. Association for Computational Linguistics.

Sylak-Glassman, J., Kirov, C., Yarowsky, D., and Que, R. (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, July. Association for Computational Linguistics.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.

Östen Dahl and Velupillai, V. (2013). The past tense. In Matthew S. Dryer et al., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.