Learning Morphosyntactic Analyzers from the Bible via Iterative Annotation Projection across 26 Languages

Garrett Nicolai and **David Yarowsky**
Center for Language and Speech Processing
Johns Hopkins University
gnicola2@jhu.edu yarowsky@jhu.edu

Abstract

A large percentage of computational tools are concentrated in a very small subset of the planet's languages. Compounding the issue, many languages lack the high-quality linguistic annotation necessary for the construction of such tools with current machine learning methods. In this paper, we address both issues simultaneously: leveraging the high accuracy of English taggers and parsers, we project morphological information onto translations of the Bible in 26 varied test languages. Using an iterative discovery, constraint, and training process, we build inflectional lexica in the target languages. Through a combination of iteration, ensembling, and reranking, we see double-digit relative error reductions in lemmatization and morphological analysis over a strong initial system.

1 Introduction

The computational processing of languages such as English and Arabic has undeniably benefited from the construction of annotated datasets such as treebanks and morphological databases. Unfortunately, the construction of even modestly-sized treebanks is very expensive, requiring hundreds of hours of expert annotation. The construction of computational tools is in turn limited by a lack of supervised training data.

One alternative to hand-annotating low-resource languages (LRL) involves using existing tools for a high-resource language (HRL), such as English, and projecting these annotations to the LRL across a parallel corpus. Consider the example in Figure 1: the English sentence is POS-tagged and dependency parsed by tools that have been trained on large amounts of high-quality data. The sentence is word-aligned to its French translation, and the POS tags and dependency relations follow the alignments to annotate the French words. Note that the projection is not lossless: the aligner could not find a French translation of "doubtless", and has thus been unable to project the RB tag or advmod relation into French.

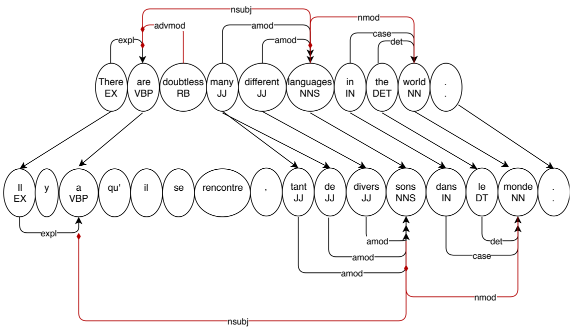

Figure 1: Projection of POS tags and dependency parse from English to French. Black arrows demonstrate left-to-right dependency relations, while red diamonds illustrate right-to-left dependency relations.

Parallel corpora are rare, and even when they do exist, they often only exist between specific pairs of languages. However, the documentation of a language often begins with the creation of several important documents, including a dictionary of key terms, and translations of religious texts. Thus, documents such as the Christian Bible are among the most translated documents in the world (Mayer and Cysouw, 2014). Furthermore, the Bible consists of short, numbered chapters and verses consisting of a small number of sentences. Although not parallel to the standard required in fields such as machine translation, the structure of the Bible means that different Bibles are approximately parallel across verses.

We follow a tradition of projecting POS tags from a high-resource language onto a language with fewer available tools (Yarowsky et al., 2001; Fossum and Abney, 2005; Agić et al., 2015; Buys

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 1765–1774
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

and Botha, 2016). Our contributions, however, lie on the level of morphology and morphosyntax. With no further resources in the target language than a Bible translation and a dictionary, we project English POS tags, dependency relations, and semantic labels across the alignment. Leveraging the alignment and a collaboration of annotations, we are able to hypothesize both a lemma and detailed morphosyntactic features for both inflected nouns and verbs. This information can then be used to inform the construction of morphological analyzers.

We learn to identify morphosyntactic categories including plurality, temporality, and case over nouns and verbs in a test set of 26 diverse languages. By leveraging annotations across a series of alternative Bible translations, we are able to successfully identify lemmas and morphological features, obtaining further improvements from strategies such as ensembling and reranking.

2 Related Work

Automatic morphological induction has had numerous contributions over the years. Here, we list the most relevant to this work, and distinguish this work from what has come before.

The class of methods introduced by Yarowsky et al. (2001) are the most similar to the work described in this paper. Also beginning with aligned Bible data, they recover verbal lemmas by leveraging multi-lingual alignments. However, where they are only interested in recovering the lemma, we simultaneously induce detailed morphological features of the words in the target language, over a wider range of verbal and nominal morphology, and deploy a new set of machine learning techniques to do so. Futhermore, we significantly expand the languages included in our test set, from 3 to 26 typologically diverse languages, substantially increasing the range of morphosyntactic phenomena covered and assessed.

Similarly, Fossum and Abney (2005) and Agić et al. (2015) exploit the parallel nature of the Bible to project POS tags and train taggers in the target languages, leveraging the signal from multiple languages to improve the tagger accuracy. We focus, instead, on the induction of detailed morphological categories.

Soricut and Och (2015) induce morphological transformation rules in an unsupervised manner. While this is analogous to lemmatization, part of our motivation is to also produce detailed morphological features that might be useful to train low-resource taggers, or to more richly annotate morphologically sparse languages such as English.

Buys and Botha (2016) train morphological taggers in morphologically rich languages from an English projection. However, their method is dependent upon an English corpus tagged with more morphologically aware tags than are typically produced by an off-the-shelf English POS tagger. We instead argue that much of this information is recoverable from syntactic and semantic parses, allowing us to use massively-parallel corpora such as the Bible.

Kirov et al. (2017) notes the morphological sparsity of English, and reverses our setup, projecting morphologically rich tags from Czech into English. Rather than add another potentially noisy projection step (i.e., Czech to English to LRL), we instead leverage dependency and semantic parses to more richly tag English.

In the area of contraint-based discovery, our methodology most closely resembles the constrained discovery systems of Lin et al. (2016) and particularly Upadhyay et al. (2018). Starting from a high-quality seed, a learning algorithm generalizes observed patterns, iteratively increasing the seed data with confident examples, while discarding examples that fail to pass certain heuristics. However, unlike previous work, we assume no gold seed annotations for our system - our seed is extracted exclusively from a noisy bitext word alignment.

3 Methods

In this section, we describe our methods for inducing lemmas and morphological features pertaining to plurality, temporality, and case from aligned English-target Bibles. Our process is outlined in Figure 2. After annotating English Bibles for POS, dependency relations, and semantic roles, these observations are projected across an alignment to a target language. Candidate analyses are first *discovered* from the projection. These analyses are then *constrained* with a number of noise-reduction heuristics. Finally, inflection tools are trained on the candidates, and used to *generate* new hypotheses, and the process is repeated.

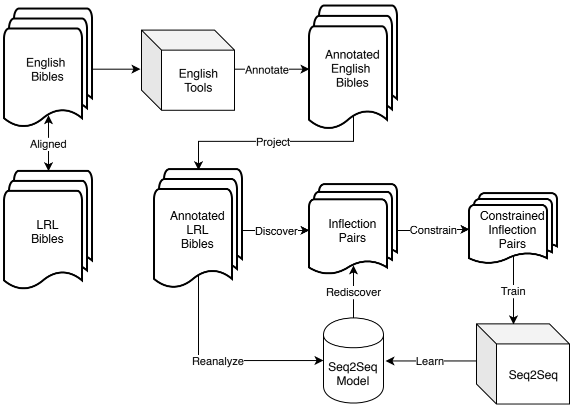

Figure 2: The discovery, constraint, and generation process. Beginning in the top-left, our method proceeds towards the lower-right corner, which forms an iterative cycle that can be repeated until convergence.

3.1 Tagging and Projection

We begin with a series of 27 English Bible translations, each verse-aligned to at least one Bible in a target language. Many of these Bibles are based on translations that are hundreds of years old, and preserve archaic conventions for literary reasons. Unfortunately, modern NLP tools are usually trained on modern text data, and the presence of archaic linguistic forms can seriously degrade the quality of the annotation.

Fortunately, many archaicisms in the Bible are older verbal inflections that follow a small set of consistent patterns: 2nd person verbs end in "-est" instead of a null affix, and 3rd person verbs end in "-eth" instead of "s" (i.e., "seest" and "believeth"). Before tagging and parsing, we normalize these forms, as well as other common archaic forms, such as "thou", to their modern equivalents.[1]

The English Bibles are then lemmatized, POS-tagged, and syntactically and semantically parsed. POS tags are directly projected between aligned words in the source and target: if a word in English aligns with multiple target words, its annotations are projected to all of them. Conversely, if many English words align to a single target word, all of the annotations are projected onto the target word (for induction, each of these tags is given equal, reduced weight).

Parses are similarly projected across the alignment, however unlike tags, parses are tuples containing a head, a relation, and a modifier (or a predicate and its arguments, for a semantic parse). Semantic parses behave similar to POS tags, and can be projected directly onto the target words. For syntactic parses, we project the relation onto the modifier, with a back pointer to the head.

When working with a noisy alignment, such as are common in low-resource situations, it is possible that either the head or the modifier will not have an aligned translation in the target. If the modifier is not aligned, then the dependency relation is lost, such as is the case with "doubtless" in Figure 1. However, if the head is not aligned, the relation will still be projected onto the modifier. For our purposes, it is far more informative to know that a particular noun is a nominal subject, without knowing the verb, than to know that a verb has a subject, but not knowing what the subject is.

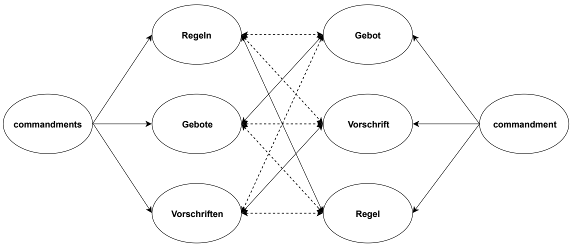

Figure 3: Projecting lemmas across alignments. Dashed lines can be eliminated with an edit-distance threshold.

3.2 Lemma Discovery

Although it is straightforward to project tags across an alignment, lemmas provide a more significant challenge. In this section, we describe our method of discovering lemmas that can later be used to train lemmatizers and morphological analyzers.

Our lemma-induction approach is similar to that proposed by Yarowsky et al. (2001). Each English word forms a set with the target words with which it is aligned. Likewise, each English *lemma* forms a set with a group of target words. In the best case, the lemma set contains translations obtained from a bilingual dictionary, but if a dictionary is sparse, the set can be supplemented with the words aligned with the English lemma. These sets are then used to create a complete bipartite graph such that each edge corresponds with a candidate plural-singular word pair. Pairs that fail to meet an edit-distance threshold can be discarded. An example is shown in Figure 3.

In this example, "commandments" has been

[1] Although Bibles in other languages can also be written in older forms of the language, we leave target normalization to future work.

aligned to three German words. Similarly, its lemma “commandment” has been aligned to three words. Completing the graph, we establish 9 candidate plural-singular pairs. However, some of these pairs, such as *Regeln–Gebot* are obviously false, and can be eliminated by an edit-distance threshold. Three pairs: *Regeln–Regel*, *Gebote–Gebot*, and *Vorschriften–Vorschrift*, remain.

3.3 Discovery of Morphological Features

Lemmatization is itself an important application, as it can reduce data sparsity in inflectionally-rich languages. However, lemmas are only one of many available English annotations that may be able to benefit LRLs. In this section, we describe our methods for leveraging English syntactic and semantic parses to discover morphological features in our target languages. We consider three types of morphological information: nominal plurality, case, and temporality.

Our first task is to identify, for a given noun, whether it is singular or plural. This information is readily available from the English POS, and we can thus create an inflection triple for each word tagged as a noun. This triple contains the inflected form, the hypothesized lemma, and a morphological tag identifying whether the noun is singular or plural. For example, “women” would produce the triple `{women, woman, PL}`.

Although English does not, for the most part, decline its nouns, some case information has been translated into syntax: direct objects of verbs are in the accusative case, indirect objects are in the dative case, and nouns in prepositional phrases headed by “of” are in the genitive case. We approximate case by using a set of heuristics to translate a syntactic and semantic parse into a nominal case. With these heuristics, we are able to construct 12 nominative cases. Details concerning the rules used to construct the cases can be found in the Appendix.[2]

Finally, we extract verbal temporality. Namely, we extract whether a verb describes an event in the past, the present, or the future. While many languages further distinguish between other temporal actions such as completion or habituality, we restrict our work here to a tripartite extraction, as temporality features are ready available from an English POS tagger and a syntactic parse.

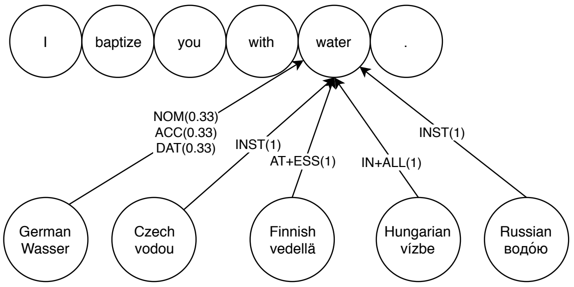

Figure 4: Forming a consensus from morphologically-informed languages.

For every verb in our English Bible, we label it as either `past`, `present`, or `future`, and project the label onto the target language. Present and simple past verbs can be determined directly from the POS tags, while the perfect and future tenses are informed by the syntactic parse. Past participles (i.e., VBN), governed by a form of “have” is marked as past tense. Similarly, any past participle or infinitive governed by an auxiliary form of “will” or “shall” is marked as future tense.

Rule-based systems, however, can be brittle, so we also investigate a secondary case signal: other target languages. The Bible is not only bilingually parallel – each translation is approximately parallel with *every* other language. Other languages than English may be better-suited to annotating the case of a target language.

Consider the example in Figure 4. A dependency parser might inform us that “water” is a nominal modifier of “with”, but “with” is an ambiguous preposition, corresponding to both instrumental uses such as “He caught fish with a net”, and comitative uses “He sat down with his apostles”. We can observe which words in morphologically-rich languages have aligned to “water” in this verse. The case of these words can then be identified via a morphological dictionary.

Morphological dictionaries are expensive to construct, but exist for a small number of languages; a consensus of high-resource languages can be used to inform the annotation in a low-resource one.[3] In Figure 4, water is identified as clearly being used in the *instrumental* case in both Czech and Russian, and as in the *essive* and *allative* case in Finnish and Hungarian, respectively. German has a weaker signal, with an identical re-

[2] These rules are by no means complete. They merely serve as an approximation to find some examples of the desired inflections.

[3] If the relevant word form is not present in any of the dictionaries, we back-off to the rule-based method.

alization in three different cases. A simple voting scheme can annotate this use of “water” with the *instrumental* case. This annotation is then simply another piece of information to be projected across the alignment onto the target language.

3.4 Constraint

To filter out noisy candidate pairs, we implement a series of sequential heuristics. These heuristics leverage the projected annotation to remove false positives while preserving as many of the true pairs as possible.

We note that in the English translations of the Bible, if a word is present in its plural form, it is also often present as a singular. Furthermore, the singular form is regularly more frequent. Our first heuristic discards any pairs for which a proposed singular form occurs less frequently in the corpus than the plural.

Secondly, we ensure that both inflected and lemma candidates have been regularly tagged as such. Polysemy, syncretism, and alignment errors mean that each word may have had many tags projected upon it. For example, a past tense verb may occasionally incorrectly receive a present tag – we do not want this infrequent mistake to identify false morphological phenomena. We compromise between a desire to remove noise, while preserving true candidates. For each word, we calculate the average frequency across all of its tags. A pair is kept if the desired tag occurs more frequently than average.

Next, we discard any pair that demonstrates an unlikely character transformation. These transformations are discovered through the use of an unsupervised character aligner. The inflected forms are aligned with their discovered analysis. A pair is discarded if its normalized alignment likelihood does not fall within 2 standard deviations of the average likelihood. Consider the triple `praised,praise,TAG`. This inflection and lemma will pass an edit-distance threshold, but is much more likely to be a verbal inflection than a nominal one. The pair will be discarded if the task is plurality detection, as `d`→`PL` is an unlikely sequence. However, `d`→`PST` is very common, and thus the pair would be retained for temporality detection.

Our preliminary nominal lemma detection is based solely on a singular/plural distinction, with no regards to case. It is possible that the hypothesized lemma is a singular form other than the citation form. To limit the singular forms in the discovered set to citation forms, we use the dependency parse and a target dictionary to restrict lemmas to nominal subjects that occur in the dictionary.

3.5 Generation

After denoising our initial lexicon, we train models that learn to transform an inflected form into a citation form.[4] After training, we attempt to analyze all verbs and nouns in the corpus.

We then limit the hypotheses to high-confidence analyses, and pairs for which the predicted lemma appeared in the original target Bible. This restricted hypothesis list is then constrained via the heuristics in Section 3.4, and new models are learned. By augmenting the training data with hypotheses generated by the original models, we can exploit words that were in the original Bibles, but that our original induction methods missed, due to a missing alignment, a poor parse, or other noise. Development experiments demonstrated that one iteration of supplementing the training data was beneficial across our languages; subsequent iterations led to little further gain.

4 Experiments

In this section, we describe the data and tools that we use to label our English Bibles and generate our morphological analyses. We also outline our evaluation metrics and describe our experimental results.

Our Bible data is obtained from the corpus of Mayer and Cysouw (2014), which consists of verse-parallel Bible data across 591 languages, including 27 English Bibles. The English and target Bibles are aligned using the Berkeley aligner (Liang et al., 2006), and POS tagged and syntactically parsed using the Stanford NLP toolkit (Manning et al., 2014). We semantically parse the Bibles using the Deep Semantic Role Labeler (He et al., 2017). The alignment filter is implemented using M2M aligner (Jiampojamarn et al., 2007), and our dictionaries come from PanLex (Kamholz et al., 2014); statistics concerning dictionary and training sizes are contained in the appendix.

[4] For languages such as Arabic and Hebrew, where the citation form is not an attested word, we use the unmarked nominative singular form, instead.

To evaluate the quality of the lexica that are produced, we extract gold validation and heldout sets from UniMorph (Kirov et al., 2018). Using the URIEL typological database (Littel et al., 2016), we limit the languages to those that include affixing verbal and nominal inflection, and that distinctly mark plurality and temporality.[5] Our evaluation set consists of 26 languages belonging to several language families such as Semitic, Germanic, Italic, Slavic, Uralic, and Bantu. For each of these languages, we randomly select a validation set of 5000 instances, and 1000 heldout instances.[6] For our declension experiments, we approximate case from a majority of higher-resource morphological dictionaries, as described in Section 3.3. For these experiments, the majority is obtained from the 10 largest nominal databases in our language set. Further information is included in the appendix.

4.1 Data

We consider two learning algorithms for the generation phase of lexicon creation. The first is the bidirectional, hard-attentional RNN (RNN) over edit actions of Makarov and Clematide (2018). We use 100 hidden units on the input layer, and 200 on the encoder and decoder. We train the system using the ADADELTA optimizer for a maximum of 60 epochs, with 50% dropout. The second is DirecTL+ (DTL; Jiampojamarn et al., 2010), a semi-Markov model that learns transduction actions over sequences of characters; an n-gram size of 9 is used, with a joint n-gram size of 3. We further ensemble the two models by adding the normalized confidence scores produced by each model (Ensemble). We also consider a simple reranking (RR) scheme where any analysis with a lemma appearing in a dictionary has its confidence score incremented by the score of the best original hypothesis. In this way, forms that appear in the dictionary appear at the top of the list, in the same order as they were generated by the original model.

We evaluate against two simple baselines that provide estimates of the difficulty of the task. The first baseline simply produces the inflected form as the lemma (Identity). The second baseline compares an inflected form with every citation form in a dictionary, and identifies the lemma as the citation form with the lowest edit distance from the inflected form (DictED). For morphological analysis, both baselines return the most common inflectional class from the training data. All systems are evaluated on accuracy@1, accuracy@5, and accuracy@50. Accuracy@n rewards a system if it returns one of the correct solutions in its first n predictions. While we focus our analysis on the accuracy@1, containing the correct solution in an n-best list can also be desirable when recall is valued more highly than precision.

4.2 Singularization

Morphological analysis produces a lemma and bundle of inflectional features, given an inflected wordform. In our first set of experiments, we investigate a special case of analysis: singularization. By focusing on singularization, we can establish which of our filtering heuristics are effective in a task where we can be relatively certain that the lemma exists somewhere in the text. In these experiments, we sequentially accumulate the heuristic filters described in Section 3.4, beginning with the plural-singular pairs hypothesized by our dictionary-independent lemma extraction. The average singularization accuracy over all 26 languages is detailed in Table 1.

We see that DirecTL+ and the RNN behave very differently when the training data is filtered. DirecTL+ improves marginally for each successive filter. Contrarily, the morphological filter, in particular, leads to a decrease in accuracy for the RNN, while all of the filters sharply limit the number of correct candidates that appear lower in the list. Some of this decrease can be attributed to smaller training data, and most of the loss is recovered via a second iteration, which increases the size of the training data. However, we hypothesize that the morphological filter, in particular, is too aggressive. It removes instances that contain infrequent transformations that allow the RNN to produce correct candidates further down the list.

Our systems are trained exclusively from Bible data, but are able to generalize well to modern terms with a number of different pluralizing strategies. For example, in German, even the projection baseline can correctly generalize affix deletion and umlaut: “Ämter”→“Amt” (department), as well

[5] Of our languages, six do not contain declension information in UniMorph. For these languages, the declension models will be identical to the plurality ones.

[6] Several of the UniMorph corpora contain fewer than 6000 suitable inflection-lemma pairs; in these cases, the size of the validation set is adjusted accordingly.

System	Projection	+Lemma	+Morph	+Align	+Dep	+Dict	I2	+RR
Identity	9.1	9.1	9.1	9.1	9.1	9.1	9.1	9.1
DictED	N/A	N/A	N/A	N/A	N/A	31.0	31.0	31.0
DTL@1	15.2	16.8	16.9	16.7	17.8	21.3	33.0	43.1
RNN@1	17.7	17.8	16.5	17.4	18.7	22.8	30.6	36.9
Ensemble@1	17.5	19.3	18.9	18.9	20.6	25.5	36.4	**43.5**
DTL@5	31.6	31.9	33.1	33.0	33.8	37.1	47.3	53.3
RNN@5	40.3	33.7	30.9	32.2	31.7	37.1	46.3	52.0
Ensemble@5	43.4	40.2	39.8	40.0	40.3	45.6	57.9	**61.5**
DTL@50	44.9	49.7	50.4	50.8	50.8	50.6	57.8	57.8
RNN@50	63.2	52.0	49.7	51.0	50.8	54.3	60.9	60.9
Ensemble@50	63.5	58.5	57.6	58.3	58.4	60.8	70.2	**71.4**

Table 1: Accumulative lemmatic recall in the top-1, top-5, and top-50 hypotheses. Projection does not filter training candidates, other than by edit distance. Lemma implements the lemma heuristic, Morph the morphological one, Align the alignment one, Dep the dependency parses, Dict the dictionary, I2 applies a second iteration, and RR reranks the target hypotheses.

as null inflection: “Kochlöffel”→“Kochlöffel” (cooking spoon).

Limiting the target candidates by case has a marked impact upon the systems. By removing false lemmas like the German genitive “Geistes” (of the spirit), the Hungarian inessive “temploban” (in the temple), and Danish definite forms l ike “skidet” (the boat), the systems are more likely to produce the citation form: German “*Ingenieurs”→“Ingenieur” (engineer); Hungarian “*gõzhajóban”→“gõzhajó” (steamboat); Danish “*rygradet”→“rygrad” (backbone). By removing these noisy forms, we see large gains; the lemmas returned by the Finnish, Hungarian, and Turkish system without noise reduction are correct less than 10% of the time, while filtering the data increases the accuracy to approximately 26%, 56%, and 70%, respectively.

Supplementing the system with a second iteration strengthens the signal of correct inflection patterns, relatively weakening the effect of noise. For example, German nouns ending in “-ung“ are very likely to pluralize with an “-en” suffix, but the projection baseline discovers no correct “-ung” pairs. However, “-en” is a common plural suffix in German, and the systems systematically strip the “-en” from “-ungen” nouns, although often lower in the hypothesis list. These correct pairs become training examples in the second iteration, outnumbering noisy examples, and improving system accuracy.

If we have access to a dictionary, simply choosing the singular form closest to the inflection provides a surprisingly strong baseline – indeed, our systems do not surpass this simple heuristic until we implement a second iteration. Noting that the dictionary and iteration process contribute significantly more than any of the filtering heuristics, we investigate moving the dictionary earlier in the pipeline. Instead of creating a lemma list from the words aligned with the English lemma, as in Section 3.2 we use a list of translations of the English lemma.

By moving the dictionary to the “front-of-the-line” in such a matter, we see astounding gains, with the @1 recall of the reranked ensemble improving to 58.5%. In our further experiments, we thus adopt the dictionary in the lemma extraction method.

4.3 Lemmatization

Singularization is a simplified version of lemmatization, as it assumes that all input forms are in the plural. In our next experiments, we train models that take as input an inflected word form, and produce a morphological tuple containing a lemma and morphological features. We train separate models to annotate plurality, temporality, and case. In this section, we evaluate the quality of the lemmas produced by these systems, before evaluating the quality of the complete analyses in Section 4.4.

Table 2 shows the accuracy of our nominal and verbal lemmatizers. In particular, verbal lemmatization appears to be a more difficult task than its nominal equivalent. Both baselines struggle to

System	Nouns			Verbs		
	I1	I2	+RR	I1	I2	+RR
Identity	9.6	9.6	9.6	2.8	2.8	2.8
DictED	34.8	34.8	34.8	18.6	18.6	18.6
DTL@1	44.8	48.1	**59.3**	46.3	49.4	50.6
RNN@1	45.7	47.6	55.7	47.6	51.5	49.5
Ens@1	51.0	51.0	57.5	51.1	52.8	**53.7**
DTL@5	61.0	63.8	68.9	59.6	60.9	63.6
RNN@5	55.0	55.1	61.1	58.4	60.5	62.0
Ens@5	66.6	64.8	**71.1**	65.0	65.9	**68.7**
DTL@50	71.1	74.7	74.7	68.2	70.0	70.0
RNN@50	71.2	68.9	68.8	70.6	69.2	69.2
Ens@50	**78.7**	77.2	78.4	74.8	75.2	**75.9**

Table 2: Average Lemmatization accuracy on nouns and verbs. I1 uses the dictionary-based lemma extraction, I2 implements a second iteration, RR adds a reranker to I2.

produce the correct lemma – nouns are about 4 times as likely to observe null-inflection as verbs, and even plural nouns tend to drift significantly from their lemmas, to the point that another citation form has a smaller edit-distance. However, we note little difference between nouns and verbs for any of our systems - in fact, our verbal system prior to reranking is slightly better than the nominal system. Ensembling neural and traditional systems augments performance,

The ensemble makes use of complementary information to improve over either the RNN or DTL, even when neither system correctly predicts the lemma as its top candidate. For example, DTL predicts the lemma of the Estonian "lõpetagem" as "*lõpemama", while the RNN predicts "*lõpetamine. Both predict the correct "lõpetama" (to finish) in 2nd place, which is exploited by the ensemble system.

Re-incorporating the dictionary back in as a reranking step also provides gains, particularly to nominal lemmatization. This is even true with very small dictionaries: although the Northern-Sami and Zulu dictionaries both contain fewer than 5000 entries, North-Sami nominal lemmatization accuracy increase from 40 to 44 %, and Zulu from 38 to 40%.

4.4 Morphological Analysis

In our next series of experiments, we consider not only the accuracy of the lemmas produced by our systems, but of the complete morphological analyses. The task of morphological analysis subsumes lemmatization: a correct analysis must find not only the correct lemma, but also the correct set of morphological features that transformed the lemma into the inflected form. Analyzing the same systems as in Section 3.2, we report the accuracy of complete analyses in Table 3.

We note that with the exception of temporality, arriving at a consensus for the morphological tag is superior to deriving it from a simple heuristic. While the English signal is strong enough to recover some morphological information, perhaps unsurprisingly, the signal from languages that have maintained their nominal declension is stronger. Given enough languages, the signal is strong enough to overcome idiosyncratic properties of the languages individually.

The heuristics that extract case from English can be confused by complex clauses. In the sentence "He ordered his soldiers to remove him from his midst" the soldiers are the nominative subject of the verb "remove", but the dative object of the verb "order". Relying on the dependency parse alone allows dative plurals such as the Polish "żołnierzom" (soldiers) to enter the training data erroneously tagged as a nominative plural. The model then incorrectly tags other words ending in "-om", a distinctly dative suffix, as nominatives. Achieving a consensus from other languages correctly identifies the form as a dative, even though it is used as a subject.

4.5 Further Analysis

In the previous sections, we averaged our results over 26 languages exhibiting various morphological phenomena. In this section, we provide a more nuanced investigation of the types of languages suited to our methods.

We claim that the Bible is a suitable resource for learning the morphology of low-resource languages, but due to the necessity of gold morphological dictionaries, many of our evaluation languages cannot be considered low-resource. However, the only available resources we assume to exist are a translated Bible and a bilingual dictionary. By grouping languages by the size of their dictionaries, we can determine the impact that the size of the dictionary has on our methods, and extrapolate how they might work in a true low-resource scenario. Table 4 demonstrates how the dictionary size influences two steps in our method: lemma extraction, and reranking.

We see that although the dictionary has some impact on the accuracy of trained lemmatizers, it is not the only contributing factor. The lan-

System	Plurality				Temporality				Case			
	RB	Maj	I2	RR	RB	Maj	I2	RR	RB	Maj	I2	RR
Identity	8.9	8.9	8.9	8.9	2.0	2.0	2.0	2.0	8.7	8.7	8.7	8.7
DictED	20.9	20.9	20.9	20.9	8.9	8.9	8.9	8.9	10.1	10.1	10.1	10.1
DTL@1	32.1	37.5	39.2	47.0	37.2	36.4	38.8	38.7	18.9	21.9	23.6	**27.9**
RNN@1	34.1	36.8	37.7	42.9	37.0	38.7	40.2	38.2	17.0	16.3	17.7	19.6
Ensemble@1	36.6	43.4	41.2	**47.8**	41.4	40.4	41.3	**41.5**	21.1	24.1	24.6	27.4
DTL@5	52.6	56.1	57.3	65.1	53.4	50.3	50.8	55.7	33.3	38.6	39.6	46.7
RNN@5	59.0	62.0	62.9	67.9	56.0	56.3	58.5	60.0	35.4	36.2	40.3	44.4
Ensemble@5	64.7	69.1	67.3	**73.1**	62.1	59.9	61.1	**63.8**	40.8	46.0	46.6	**51.1**
DTL@50	68.6	68.2	71.7	71.7	64.5	61.3	63.9	63.9	47.9	53.1	55.7	55.7
RNN@50	71.9	76.9	75.1	74.9	68.3	69.3	67.4	67.1	54.4	59.2	58.0	58.0
Ensemble@50	76.8	**81.0**	78.8	80.0	**73.0**	71.8	71.4	72.2	58.0	64.2	62.8	**64.5**

Table 3: Average Accuracy of morphological analysis for plurality detection, temporality detection, and case identification. RB denotes a system where case is hypothesized through rules, Maj denotes a majority consensus of other languages, I2 is a second iteration built on top of Maj, and RR applies a reranker to RR.

#Entries	Nouns	Nouns +RR	Verbs	Verbs +RR
<5K	48.7	52.1	24.1	24.4
5K-20K	38.0	41.2	35.9	38.1
20K-50K	52.5	63.4	62.3	63.0
>50K	57.4	64.5	62.3	63.0

Table 4: Average Lemmatization accuracy@1 on nouns and verbs of the ensemble system for varying dictionary sizes.

Family	NN@1	NN@50	VB@1	VB@50
Armenian	63.0	85.1	37.7	72.1
Bantu	40.4	73.5	1.3	21.4
Hellenic	53.7	77.1	31.7	46.8
Turkic	36.9	62.8	40.5	81.3
Italic	44.1	57.1	33.0	56.3
Semitic	16.7	32.2	10.9	22.1
Uralic	58.1	80.9	51.5	78.6
Balto-Slavic	64.8	84.6	66.1	89.4
Germanic	71.6	93.6	78.1	94.0

Table 5: Average Lemmatization accuracy on nouns and verbs of the ensemble system for varying language Families.

guages with the smallest dictionaries perform approximately as well as larger groups on nominal lemmatization, only starting to degrade after dictionary reranking, which is to be expected. Verbal lemmatization, on the other hand, degrades much faster as the size of the available dictionary is reduced. However, we observe that the reranker – which is entirely dependent on the dictionary – has far less influence on verbs than nouns, even with a large dictionary. The size of the dictionary may be less of a factor than the types of morphology exhibited in the lower-resource languages.

We next observe which languages are most suitable to our methods, by separating our results by linguistic family. Table 5 reports both the accuracy@1 and accuracy@50 for the reranked ensemble. Although our system can accurately lemmatize Bantu nouns, Bantu verbs prove much more difficult. The low accuracy on Bantu verbs appears to be at least partially responsible for the low verbal performance of LRL in Table 4.

Secondly, we note that while our system struggles with Semitic and Bantu language families, our methods of projection and constraint are successful on other language families, even when their morphology differs significantly from English. We correctly lemmatize Uralic and Balto-Slavic languages – languages with large case inventories – with high accuracy. Similarly, the verbal signal is strong enough to train accurate lemmatizers in languages with much more complex inflectional systems than English, such as the agglutinative Turkic and Uralic families.

5 Conclusion

We have presented a method for learning morphosyntactic feature analyzers and lemmatizers from iterative annotation projection. Using no target-language training data, we successfully transferred multiple fine-grained annotations on 27 different English Bible editions to 26 diverse target languages. Using iterative discovery and robust ensembling of multiple high-performance morphological learning algorithms to yield standalone target language systems, we achieve double-digit relative error reductions in both lemmatization and morphosyntactic feature analysis over a strong initial system, evaluated on modern test vocabulary in all 26 languages.

References

Željko Agić, Dirk Hovy, and Anders Søgaard. 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.

Jan Buys and Jan A. Botha. 2016. Cross-lingual morphological tagging for low-resource languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, Berlin, Germany. Association for Computational Linguistics.

Victoria Fossum and Steven Abney. 2005. Automatically inducing a part-of-speech tagger by projecting from multiple source languages across aligned corpora. In *International Conference on Natural Language Processing*, pages 862–873. Springer.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Sittichai Jiampojamarn, Colin Cherry, and Grzegorz Kondrak. 2010. Integrating joint n-gram features into a discriminative training network. In *NAACL-HLT*, pages 697–700, Los Angeles, CA. Association for Computational Linguistics.

Sittichai Jiampojamarn, Grzegorz Kondrak, and Tarek Sherif. 2007. Applying many-to-many alignments and Hidden Markov Models to letter-to-phoneme conversion. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 372–379. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan M Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *LREC*, pages 3145–3150.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.

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.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 104–111. Association for Computational Linguistics.

Ying Lin, Xiaoman Pan, Aliya Deri, Heng Ji, and Kevin Knight. 2016. Leveraging entity linking and related language projection to improve name transliteration. In *Proceedings of the Sixth Named Entity Workshop*, pages 1–10.

Patrick Littel, David R Mortensen, and Lori Levin. 2016. URIEL typological database. *Pittsburgh: CMU*.

Peter Makarov and Simon Clematide. 2018. Neural transition-based string transduction for limited-resource setting in morphology. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 83–93.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. *Oceania*, 135(273):40.

Radu Soricut and Franz Och. 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.

Shyam Upadhyay, Jordan Kodner, and Dan Roth. 2018. Bootstrapping transliteration with constrained discovery for low-resource languages. In *EMNLP*.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 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.