Automatic Construction of Morphologically Motivated Translation Models for Highly Inflected, Low-Resource Languages

John Hewitt johnhew@seas.upenn.edu
Department of Computer and Information Science, University of Pennsylvania
Philadelphia, PA 19104

Matt Post post@cs.jhu.edu
David Yarowsky yarowsky@jhu.edu
Center for Language and Speech Processing, Johns Hopkins University
Baltimore, MD 21211

Abstract

Statistical Machine Translation (SMT) of highly inflected, low-resource languages suffers from the problem of low bitext availability, which is exacerbated by large inflectional paradigms. When translating into English, rich source inflections have a high chance of being poorly estimated or out-of-vocabulary (OOV). We present a source language-agnostic system for automatically constructing phrase pairs from foreign-language inflections and their morphological analyses using manually constructed datasets, including Wiktionary. We then demonstrate the utility of these phrase tables in improving translation into English from Finnish, Czech, and Turkish in simulated low-resource settings, finding substantial gains in translation quality. We report up to +2.58 BLEU in a simulated low-resource setting and +1.65 BLEU in a moderate-resource setting. We release our morphologically-motivated translation models, with tens of thousands of inflections in each of 8 languages.

1 Introduction

Statistical machine translation systems are typically trained on large bilingual parallel corpora (bitext). Low-resource machine translation focuses on translation of languages for which there exists little bitext, and where translation quality is subsequently often poor. Highly inflected languages—those that exhibit large inflectional paradigms of words with a common dictionary entry—excacerbate the problems of a low-resource setting. Many inflections of words in an inflectional paradigm are complex and rare, and their translations are unlikely to be well-estimated even in a moderately large parallel corpus. For example, Koehn (2005) point to the highly inflected nature of Finnish as a reason for poor translation performance into English even in high-resource settings.

However, even where bitext may be lacking or scarce, there are often many other resources available. One source of rich morphological information is Wiktionary.[1] This paper describes a method for using resources extracted from Wiktionary to automatically map inflections in paradigms of morphologically rich languages to ranked sets of English phrasal translations. This is done by the following procedure:

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

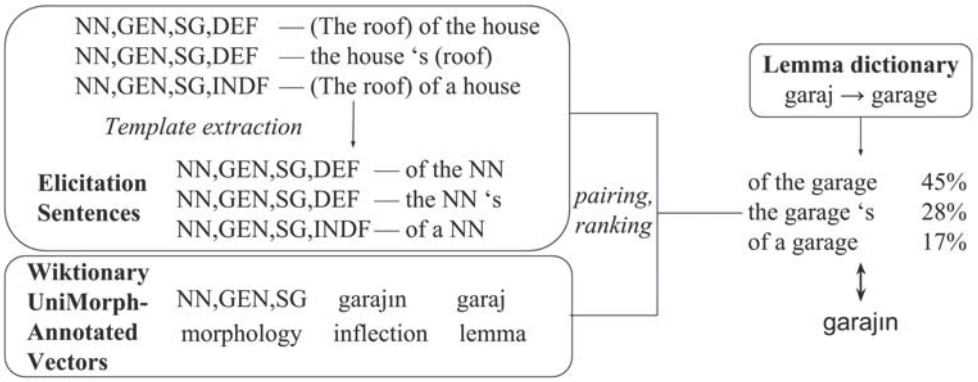

Figure 1: The translation model (TM) construction pipeline. This depicts the process by which we map each morphologically annotated inflection to a ranked set of English phrasal translations.

1. We begin with resources from the UniMorph project (Sylak-Glassman et al., 2015), which produced millions of tuples pairing inflected words forms with their lemmas and a rich morphological tag (which we refer to as a UniMorph tag or vector) that was designed to be a universal representation of morphological features (§3).

2. We next take a small set of pairs of UniMorph vectors and short English sentences that were produced in an *Elicitation Corpus*, designed to collect inflections that in English are expressed phrasally instead of morphologically (§4).

3. We then produce phrasal translation pairs by extracting English phrases from these sentences and pairing them with the foreign language through the UniMorph tag (§5). We investigate different methods for extracting and scoring phrase pairs.

4. Finally, we evaluate the utility of these phrase pairs to improve machine translation in simulated low-resource settings (§6).

A depiction of the full pipeline is in Figure 1.

2 Prior Work

Maximizing the utility of a baseline phrase table has been the focus of a large body of prior work in translating from morphologically rich languages. Habash (2008) work on the OOV problem in Arabic, mapping OOV types to in-vocabulary (INV) types by orthographic and morphological smoothing methods. Mirkin et al. (2009) take inspiration from the Textual Entailment (TE) problem, using WordNet to determine a set of entailed alternatives for English OOV tokens. However, since this OOV-resolution scheme is dependent on the existence of a semantic resource like WordNet in the source language, it is unsuitable in general low-resource settings. Yang and Kirchhoff (2006) implement a backoff model for Finnish and German, stemming and splitting OOV tokens at test time and searching a baseline phrase table for the resulting simplified forms.

Many systems attempt to address the incorrect independence assumptions traditional phrase-based MT systems impose on inflections in the same paradigm. Koehn and Haddow (2012) train a baseline phrase-based translation model, and back off to a factored model that

Inflection	Lemma	Mood	POS	Tense	Gender	Number	Animacy	Person
bude absolvovat	absolvovat	IND	VB	FUT		2		SG
absolvuj	absolvovat	IMP	VB			2		SG
absolvujete	absolvovat	IND	VB	PST	MASC	2	ANIM	SG
absolvoval jste	absolvovat	IND	VB	PST	MASC	2	INAN	SG

Table 1: Czech verb inflections and partial annotations from Wiktionary. Empty cells indicate that the inflection is not marked in that dimension.

decomposes OOV tokens into `lemma+morphology`. Dyer (2007) address the intuition that even INV tokens may have poorly estimated translations. They decode on a confusion network of tranformations of source sentences, imposing penalties for backing off from surface tokens.

Each of these approaches attempts to use pre- or post-processing to make up for poorly estimated, low-coverage phrase tables. We take advantage of an entirely new resource, a very large, massively multilingual, morphologically-annotated dictionary, to directly improve phrase table coverage and provide improved estimations for INV types. Thus, the practical translation gains we see through our methods should be orthogonal to those of prior work. This paper presents a method of constructing English phrases that express inflectional features, and a novel system of mapping these phrases to foreign inflections, with the following qualities:

- We use no bitext, and no language-dependent morphological information.
- We apply our system to the Wiktionary dataset, constructing substantial phrase tables for 8 languages, with the capacity to build a model for each of the 73 languages with more than 10,000 inflections in Wiktionary.
- We demonstrate the utility of these phrase tables, finding substantial gains when augmenting low-resource MT of Czech, Finnish, and Turkish.
- We present insights on the utility of morphological information in translation by conducting an ablation study in two dimensions, analyzing the effects of varying available bitext and available morphological information for each language.

3 UniMorph Inflectional Paradigms

The Universal Morphological Feature Schema (UniMorph) represents fine-grained distinctions in meaning expressed through inflectional morphology cross-lingually (Sylak-Glassman et al., 2015). The schema defines 23 distinct dimensions across which inflections can vary independently. Though English does not express many of these features through morphology, their purposes can still be intuitive:

> of the houses : NN, GEN, PL, DEF (Noun, Genitive, Plural, Definite)
> with a hammer : NN, COM, SG, INDF (Noun, Comitative, Singular, Indefinite)

The dimensions are as follows:

> Aktionsart, Animacy, Aspect, Case, Comparison, Definiteness, Deixis, Evidentiality, Finiteness, Gender+, Information Structure, Interrogativity, Mood, Number, Part of Speech, Person, Polarity, Politeness, Switch-Reference, Tense, Valency, Voice

A total of 212 possible values are distributed across the dimensions. The morphological information of each inflection is encoded in its "UniMorph vector."

Inflection	Lemma	Case	Number
absurdity	absurdita	GEN	SG
absurdit	absurdita	GEN	PL
absurdit	absurdita	DAT	SG
absurditm	absurdita	DAT	PL

Table 2: Czech noun inflections and partial annotations from Wiktionary.

Sylak-Glassman et al. (2015) scraped Wiktionary, extracted inflectional paradigms in hundreds of languages, and transformed the inflectional annotations to UniMorph vectors. The editors of Wiktionary often include all inflections for each lemma, so full inflectional paradigms are scraped. The result, as shown for nouns in Table 2 and verbs in Table 1, is a list of inflections with corresponding lemmata, and inflectional values as a UniMorph vector. Each language has a corresponding table of inflection entries with their corresponding UniMorph vector. The size of the Wiktionary dataset varies from language to language, including 2,990,482 Finnish inflections and 15,132 Swahili inflections.

To a large extent, combining a lemma with the information in the UniMorph vector reconstructs the meaning of the inflection. We do not claim perfect reconstruction, as no inflectional morphological schema can perfectly encode language-independent meaning. However, practical gains in translation quality do not require a perfect schema. We instead focus on the substantial signal provided by UniMorph annotations, and show that they are highly effective at providing cross-linguistic information.

4 English Inflectional Elicitation

Wiktionary provides us with UniMorph vectors for inflections, but provides no information about how to express these vectors in English. These expressions are difficult to generate, as English expresses inflectional information phrasally. For this, we use a corpus of 8,000 UniMorph-annotated English sentences, which we call the *Elicitation Corpus*. The corpus was developed in an attempt to document the full inflectional variation in 49 languages of morphological interest.

The process worked as follows. For each language, we first collected all its important morphological tags. We then hand-built the Cartesian product of all inflectional variation expressed by the language. Thus, if a language inflected nouns exactly for 4 values of number and 3 values of case, we constructed 12 vectors of (Number $\times$ Case). Then, for each UniMorph tag, keeping the specific language and part of speech in mind, we manually wrote an English sentence with a head word on which the same inflectional features are expressed. For example, for the tag

VB, 3, PRS, PRF, PASS (Verb, 3rd person, Present, Perfect, Passive)

we might generate the sentence

[The apple] has been eaten.

These sentences were given to a bilingual native speaker. The goal of each sentence was to elicit, in the foreign language, the inflection encoded by the lemma we used and the UniMorph vector expressed in English. We wanted to avoid sentential translations, aiming instead for individual inflections. To this end, portions of each sentence necessary for syntactic coherence but deemed unnecessary to express inflectional values were enclosed in brackets. Each line in the corpus is a tuple of (English sentence, UniMorph vector).

By construction, the corpus spans a large amount of the variance of inflectional expression of most languages. Further, equivalent UniMorph vectors in multiple languages were not

English Sentence	UniMorph Vector
[They] were eating [when the door broke.]	IND, PST, PROG, 3, PL
[The view was blocked] by the houses	INS, PL
Was [he] not speaking?	PST, PROG, NEG, INT, 3, SG
[He is] sleeping	PRS, PROG, POS, 3, SG

Table 3: Entries from the Elicitation Corpus. Each sentence on the left was constructed to express the UniMorph vector on the right. Note that not all are fully defined, e.g., missing definiteness, potentially due to the original source language not inflecting for that dimension.

de-duplicated, so many common vectors are paired with multiple English sentences. This permits the corpus to store information about the frequency with which varying ways are used to express the same features. This eventually aids us in ranking phrase templates. For example, the genitive in English is expressed equivalently as *the house's roof* and *the roof of the house*. More examples of sentences in the Elicitation Corpus are given in Table 3.

5 Constructing Morphological Phrase Tables

In this section, we detail the pipeline by which we automatically construct a translation model or phrase table for each language represented in the Wiktionary dataset.

1. We extract and rank English *phrase templates* from the 8000-sentence Elicitation Corpus.
2. We use UniMorph vectors to pair our phrase templates with inflections from Wiktionary, and estimate direct and inverse translation probabilities.
3. We complete phrase templates with lemmata to finish the translation hypotheses.

5.1 Phrase Template Extraction

Each sentence in the Elicitation Corpus expresses a UniMorph vector in English. However, the context and specific head word of the sentence constrain the usefulness of the sentence. Taking our earlier example, we wish to generalize

[The apple] has been eaten.

such that the inflectional values are kept, but the resulting "template" is maximally reusable in different contexts, and for different lemmata. Recalling the UniMorph vector associated with this sentence, we wish to extract

has been VBN : {VB, 3, PRS, PRF, PASS}

The context has been removed, the morphological head word replaced with a part of speech "variable". In effect, the goal of this extraction is to retain exactly the information described by the UniMorph vector. Information like the lemma will be provided by each source language inflection. We use two simple methods to extract phrase templates from the Elicitation Corpus.

5.1.1 Naive Template Extraction

From each sentence in our Elicitation Corpus, we extract a "naive template". After part-of-speech tagging the sentence, all parenthesis-denoted context is removed. Then, we replace the rightmost word whose POS matches the UniMorph vector POS with a variable. For this variable, we choose the most descriptive POS, e.g., VBD instead of VB, to preserve the information necessary to conjugate an English lemma as a replacement for the variable.

Table 4 gives examples in which this naive method produces incorrect or incomplete results. To augment template extraction, we also use a slightly more principled heuristic method.

English Sentence	Naive Template	Generated Templates
(They) were eating (when the door broke.)	were VBG	they were VBG, were VBG
(The dog went) from the boy.	from the NN	from the NN
(The view was blocked) by the leaves.	by the NNS	by the NNS
Was (he) not speaking?	was not VBG*	was he not VBG
(He is) sleeping.	VBG*	he is VBG, is VBG

Table 4: Template extractions. * denotes a clearly incorrect result given the UniMorph vector in Table 3.

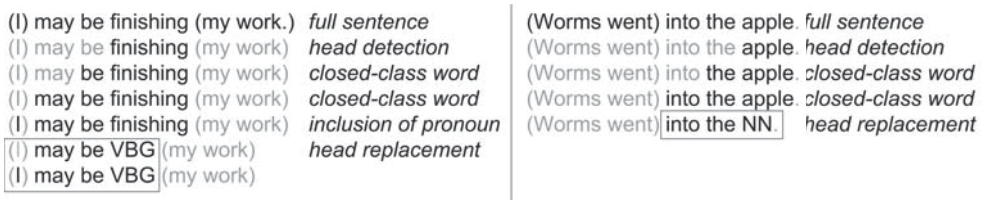

Figure 2: The extraction process, from sentences to templates. Each line shows the next word added to the template. Boxed phrases are final templates. Greyed words have not yet been considered by the algorithm, or have been excluded from the template.

5.1.2 Heuristic Template Extraction

To automatically generate phrase templates, we make the assumption that we're working only with simple sentences, and we assume the presence of context-marking parenthesis. Given these assumptions, we construct an algorithm to extract only the inflectional value-carrying neighbors of the head as part of the phrase table.

1. Determine the head of the sentence by searching for the last word tagged with a POS corresponding to the correct word class. (e.g., VBN and VBP correspond to VB)

2. Walk backwards from the head, prepending every closed-class word to the output template.

3. When an open-class word is seen, stop.

4. Replace the head of the sentence with its part-of-speech tag.

Open-class words such as nouns or verbs are unlikely to encode inflectional values, and are likely to include undesirable specifics for the sentence (such as a verb's subject.) However, there are a few verbs that are necessary for expressing, for example, tense and aspect. As such, we manually compiled a list of these words, and modified our POS tagger to let them pass. Words such as *had, have, going, am, are, did...*, for example, are used to express aspect and mood in English. We used a Brown corpus-trained tagger from the `nltk` python package (Bird et al., 2009).

A few examples of this process are given in Figure 2. The first example demonstrates the multiple potential phrase templates for a single sentence. Because our system is language-independent, we have no information about whether a pronoun in an MT setting will be omitted (as in languages with pro-drop). By extracting phrase templates with and without a pronoun, we expect that the language model will bias towards the with-pronoun phrase in sentences with pro-drop, and towards the without-pronoun phrase in sentences with a marked pronoun.

Inflection	Phrase Template	Lemma Translation	Phrasal Translation
blafovali	they were *VBG*	blafovat, bluff	they were bluffing
filmujme	let 's *VB*	filmovat, film	let 's film
kaupunginvaltuuston	of the *NN*	kaup. . . , city council	of the city council

Table 5: Phrase completion example. Once an inflection has been paired with (a) phrase template(s), we look up its lemma translation, conjugate it, and insert it into the template to complete the phrasal translation.

5.2 Matching Phrase Templates to inflections' UniMorph Vectors

A phrase template t with UniMorph vector $t.v$ is proposed as a candidate translation for inflection i with UniMorph vector $i.v$ if $t.v$ is a superset of the features in $i.v$, where UniMorph vectors are considered unordered sets of inflectional values. We use this superset-match method instead of only constructing phrase pairs with exact morphological matches to account for a large amount of underdefined Wiktionary inflections, each with very few values in its UniMorph vector. The superset matching scheme provides these underdefined inflections with a large number of low-probability phrase templates, reflecting the noise due to the lack of morphological information.

The set of all phrase templates T for vector $i.v$ is

$$\mathrm{T}(i.v) = \{\, t \mid t.v \supseteq i.v \}$$

Each $t \in \mathrm{T}(i.v)$ has a value $\mathrm{freq}(t|\,\mathrm{T}(i.v))$, the count of sentences in the Elicitation Corpus with the phrase template t whose UniMorph vector is a superset of $i.v$. Thus, each vector $i.v$ for which there exists at least 1 template has a total count

$$\mathrm{total}(i.v) = \sum_{t \in \mathrm{T}(i.v)} \mathrm{freq}(t|\,\mathrm{T}(i.v))$$

The direct translation probability, that t is the correct way to express UniMorph vector $i.v$, is thus

$$\mathrm{P}(t|i.v) = \frac{\mathrm{freq}(t|\,\mathrm{T}(i.v))}{\mathrm{total}(i.v)}$$

We also calculate the probability that v is the best morphological analysis of t by calculating the total probability mass of t in all T. Thus, $\mathrm{total}_{\mathrm{P}}(t) = \sum_{i \in I} \mathrm{P}(t|\,\mathrm{T}(i.v))$, where I is the set of all inflections. The inverse translation probability is thus

$$\mathrm{P}(i.v|t) = \frac{\mathrm{P}(t|i.v)}{\mathrm{total}_{\mathrm{P}}(t)}$$

Intuitively, the inverse translation probability discounts highly specified templates (with a rich UniMorph vector) in underspecified settings.

5.3 Phrase Template Completion

So far, we have described a system for mapping morphological feature vectors to sets of English phrase templates. The Wiktionary dataset provides a map from foreign inflections to their corresponding feature vectors. Composing the two, we map foreign inflections to phrase templates. The final step in the process is to complete, or compile out the phrase templates, replacing the part-of-speech variable with an inflected English word whose corresponding lemma translates to the foreign inflection's lemma. The Wiktionary dataset provides a mapping from foreign

inflection to corresponding foreign lemma. We use a lemma dictionary built from Wiktionary and Google Translate to map between foreign lemma and English lemma. (Note that for all languages in Wiktionary, including those missing from Google Translate, a lemma dictionary is extractable with the inflections used.) Finally, we inflect the English lemma using an English pattern library (De Smedt and Daelemans, 2012). The input and output of phrase template completion are shown in Table 5.

5.4 Phrase Table Construction

Running our system on the Wiktionary inflections for Finnish, Czech, Russian, Korean, Georgian, Swahili, Turkish, and Urdu, we construct a phrase table for each language, containing the top-5 phrasal translations for each inflection, as well as their computed direct and inverse translation probabilities. We release all constructed models for use in morphological analysis as well as end-to-end SMT.[2]

6 Experimental Design

We evaluate our system by examining its effectiveness in improving the quality of end-to-end MT from Czech, Finnish, Russian and Turkish into English. We use the Joshua decoder (Post et al., 2015) with Hiero grammars (Chiang, 2007). For many of our target languages, the only bitext available is the Bible. We thus simulate a low-resource setting by training on the Bible. We simulate moderately higher-resource settings by appending differing numbers of lines of modern bitext (described below) to the bible. We test on newswire from the Workshop on Statistical Machine Translation (Bojar et al., 2015a).

For all translation models, we use a gigaword-trained 5-gram language model (LM). We anticipated that the English phrases in our tables might be discounted by the language model due to higher-order (3- and 4-gram) misses in the gigaword corpus. In preliminary experiments, we trained language models with `lmplz` (Heafield et al., 2013) on the training data with our phrases appended. However, we saw best performance when using the gigaword LM, and by using it across all TMs, the BLEU scores are kept comparable.

Table 6 presents statistics on how many tokens and types in the test data are found in our Wiktionary inflections, anchoring the potential benefit of the system. We note that for Finnish, Czech, and Turkish, our system covers a large number of both OOV and in-vocabulary (INV) tokens in each of the resource settings. This points to potential translation quality gains through coverage of previously OOV wordforms as well as improved translation of poorly estimated wordforms.

6.1 Morphological Information Ablation Study

Along with testing the validity of our particular generation of morphologically-motivated phrasal translations, we present an ablation study, highlighting the effects of using varying portions of the morphological information provided to us, in 4 cases.

1. We consider the use of no morphological information. This completely unmodified Joshua system is our baseline.

2. We test the inclusion of a lemma dictionary with the bitext, including no morphological information. As a small dictionary largely comes for free for even low-resource languages, we see this as a stronger baseline.

3. We test the inclusion of an *inflection dictionary*. This is made possible through the Wiktionary dataset. This augmentation method pairs each inflection with a corresponding bare

[2] `https://github.com/john-hewitt/morph16`

	Bible			Bible +20k			Bible+20k+Wiktionary			Abs %
	Covered	Total	%	Covered	Total	%	Covered	Total	%	Added
Finnish Types	1,999	8,497	23.5	4,558	8,497	53.6	5,896	8,497	69.3	**15.75**
Finnish Tokens	7,456	16,447	45.3	11,613	16,447	70.6	13,188	16,447	80.1	**9.58**
Czech Types	3,445	16,581	20.7	9,616	16,581	57.9	10,110	16,581	60.9	**2.98**
Czech Tokens	26,377	51,373	51.3	41,972	51,373	81.7	42,618	51,373	82.9	**1.26**
Turkish Types	1,658	3,677	45.0	2,698	3,677	73.3	2,824	3,677	76.8	**3.43**
Turkish Tokens	4,268	7,337	58.1	6,036	7,337	82.2	6,205	7,337	84.5	**2.30**
Russian Types	4,112	15,346	26.8	9,949	15,346	64.8	9,950	15,346	64.8	0.01
Russian Tokens	25,882	45,390	57.0	38,407	45,390	84.6	38,408	45,390	84.6	0.00

Table 6: The first column reports the number of tokens and types in the newswire test set that are in-vocabulary in a model trained on the Bible. It also reports the total number of tokens and types in the test set, and the percent of coverage. The second column reports the same statistics for a model trained on the Bible and 20,000 sentences of modern text. The third column reports the same statistics for a model that is trained on the Bible plus 20,000 sentences of modern text, and includes our inflections. The column *Abs % Added* reports how many percentage points of type and token coverage are gained when adding the inflections.

English lemma. This encodes *some* morphological information, as it groups all inflections of a paradigm together by their common lemma.

4. We test our full system, mapping inflections to English phrasal translations motivated by the morphological features of the inflection.

6.2 How much does morphology help when I have X much data?

Our methods are particularly exciting as they come largely for free from the Wiktionary corpus, and do not depend on the amount of bitext available for a language. However, it is interesting to examine the utility of our work as we vary the definition of low-resource. We evaluate our work by training models on Bible bitext with varying amounts of bitext and analyzing the benefit of augmenting each model with our system:

1. We train a model only on translated portions of the Bible.

2. We consider a slightly higher-resource setting, appending 1000 sentences of Europarl (Koehn, 2005), SETIMES[3] (Tyers and Alperen, 2010), extracted from OPUS (Tiedemann, 2009), or Common Crawl (Bojar et al., 2015b) (depending on the language, described below) to the Biblical training set.

3. We train a model with 20,000 lines of modern data and the biblical training set.

4. For Finnish and Czech only, we consider the highest of our low-resource settings, appending 50,000 lines of modern data to the biblical training set.

6.3 Augmentation Method

The inclusion of outside data as augmentation for an existing translation model is non-trivial, as the probabilities in the outside data are unrelated to those of the table. The well-estimated translations of each must be given substantial probability mass in the resulting combined table without adversely promoting poorly estimated phrase pairs. We test a *dual grammar* method,

[3] http://nlp.ffzg.hr/resources/corpora/setimes/

Language	# of Inflections
Finnish	2,990,482
Russian	326,361
Turkish	275,342
Czech	145,230
Georgian	121,625
Korean	76,739
Swahili	15,132
Urdu	13,682

Table 7: The number of annotated inflections provided by the Wiktionary dataset for each of the eight languages for which we build phrase tables. Note that these numbers come from the latest release of UniMorph, and may differ modestly from the number of inflections used in our experiments.

wherein our artificial phrases and their translation probabilities are constructed as a Joshua phrase table, and assigned weights tuned by Joshua in tandem with its bitext-derived phrase table. We also test a *bitext augmentation* method, wherein the artificial phrases are appended to the bitext. In particular, we concatenate the bitext to itself 10X, and allocate a quota of 10 lines total for all translation candidates of each inflection in our artificial data. Finally, we test the two methods in combination.

We found no consistent best method out of the three. Thus, we present either the best of the three, or just the dual grammar method, as it simplified manually identifying when artificially constructed phrases were used by the decoder.

6.4 Training, Tuning, Testing Sets

For the Finnish-English, Czech-English, Russian-English, and Turkish-English language pairs, we train a Hiero model on 29,000 sentences of Biblical data. Separately, for Finnish and Czech, we train models on 29,000 sentences of Biblical data with 1,000, 20,000, and 50,000 sentences of Europarl. For Russian and Turkish, we train models on 29,000 sentences of biblical data with 1,000 and 20,000 sentences of CommonCrawl and SETIMES data, respectively.

Our tuning and test sets are consistent, per language, through all experiments. We tuned Finnish-English on the Workshop on Machine Translation (WMT) 2015 dev set (1810 sentences), and tested on the WMT 2015 test set (1724 sentences). We tuned Czech-English on the WMT 2013 test set (3000 sentences) and tested on the WMT 2014 test set (3287 sentences). We tuned Turkish-English on half of the WMT 2016 test set (506 sentences) and tested on the other half (505 sentences). We tuned Russian-English on the WMT 2013 test set (3000 sentences) and tested on the WMT 2014 test set (3308 sentences).

7 Results and Analysis

Figure 3 illustrates that our system is able to correctly generate the English expression of complex inflectional information, both intelligently filling in OOVs and providing better translations for poorly-estimated inflections. Table 8 shows the results of our end-to-end tests. Our study of end-to-end MT proceeds in two dimensions, varying both the amount of training data and the amount of morphological information. We focus our analysis on Finnish, Czech, and Turkish, for which we see substantial gains in BLEU. We present a negative result for Russian, documented in Table 8. Post-hoc examination of the tokens and types of test corpora in Table 6 explains this: only a few, low-frequency types in the Russian test corpus are covered in the

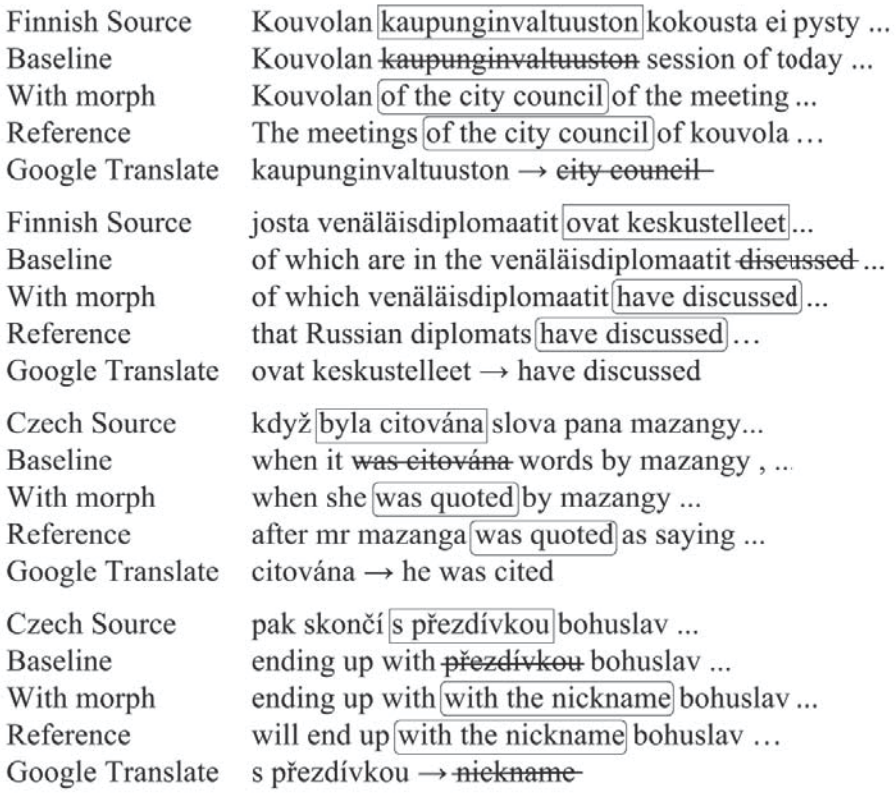

Figure 3: Newswire sentences from the test set where a source inflection is either OOV or poorly estimated in a low-resource setting, and a precise translation is generated by our system. The baseline system is trained on the Bible + 20,000 sentences of Europarl. Our system's translation is denoted by "With morph". Targeted inflections are boxed, and their translations from our system are in rounded boxes along with the reference translations. Also provided, for reference, is Google Translate's translation of the inflection. We note that Google's Finnish and Czech models were not constrained in the amount of training data, but still fail to capture the genitive case of *city council* in Finnish and the instrumental case of *nickname* in Czech. Manual analysis showed that for the Czech *citovna*, the baseline *was* was not generated in the system translation; instead, the Czech inflection was expressed fully as *was quoted*.

	Unmodified	Lemmata	Inflections	Full Morph
Finnish Bible	4.70	5.29†	**5.85	**7.28
Bible + 1k Europarl	5.39	6.07†	**6.59	**7.73
Bible + 20k Europarl	8.58	9.07†	**9.51	**10.37
Bible + 50k Europarl	9.76	10.43	10.61†	****11.41**
Czech Bible	5.29	5.69†	**5.91	**6.17
Bible + 1k Europarl	6.98	7.33†	**7.64	** 8.01
Bible + 20k Europarl	13.98	14.04	14.22	*14.33
Bible + 50k Europarl	16.23	16.20	*16.02	****16.69**
Turkish Bible	3.58	4.09†	4.02	4.23
Bible +1k SETIMES	4.78	5.00	4.96	*5.28
Bible +20k SETIMES	7.85	8.05	8.19	**8.26**
Russian Bible	1.18	1.32†	1.26	1.20
Bible + 1k Common	6.24	6.26	6.25	6.26
Bible + 20k Common	11.20	**11.22**	11.17	11.21

Table 8: All experiment results. * and ** denote significantly better results than the lemma-lemma model of on the same bitext, at $p < 0.05$ and 0.01, respectively. To motivate our use of the lemma-lemma model as a sronger baseline for significance testing than the unaugmented MT system, we denote with a † where the lemma-lemma model is significantly better than the unaugmented system. We use the bootstrap resampling method of Koehn (2004) to estimate significance tests.

Wiktionary dataset. We also note that the potential benefit of our system is constrained by the quality and size of the Wiktionary dataset for a given language. Table 7 gives the size the dataset for each of the eight languages for which we release a phrase table.[4]

7.1 Full Morphological Translation Model

Augmenting an SMT system with a translation model built by our full morphological analysis and generation improves translation quality significantly across Finnish, Czech, and Turkish, even at higher levels of resources. We expected that the potential benefit of adding morphological information would decrease as the training set size of the baseline model increased. At increasing sizes of training corpora, the gains from morphology decrease, but remain significant for Finnish and Czech. This may suggest that morphological information aids in the translation of poorly estimated inflections even in settings of moderate resources. It is worth noting that adding even 20k sentences of Europarl to the training data improved over a baseline Bible system more than adding the entirety of our morphological information. However, when adding the morphology *on top of* the added bitext, there are substantial gains.

7.2 Inflection-Lemma Model

Augmenting an SMT system with a translation model that naively pairs all Wiktionary inflections with their English lemma equivalents also improves translation quality, often to a substantial percentage of the full model's gains. These gains may be due to OOV coverage, even with poor translations. It is encouraging to see that this model does not perform as well as the full morphological model. This points to the utility of the UniMorph vectors and our phrases in providing the capacity for *good* translation estimates, not just OOV coverage. For Finnish and Czech, the percentage of a full morphological system's gains recovered by an inflection-lemma

[4] http://unimorph.org

model seems to be independent of the resource setting. This model recovered an average of 50% of the gains of the full Finnish system, and 50.2% for Czech.

7.3 Lemma-Lemma Model

Augmenting an SMT system with a translation model that pairs entries in a dictionary also improves translation at most resource levels. However, it is an impoverished system that lacks the OOV coverage of the inflection-lemma model. Except for Turkish at the 0k and 1k levels, the lemma-lemma model underperforms the inflection-lemma model, as expected. Regardless, it is a stronger baseline than an unmodified MT system, significantly outperforming the unmodified system in half of the experimental settings.

8 Summary and Future Work

Translation of highly inflected languages presents compounding data scarcity problems for SMT systems with little bitext. The information encoded in the inflections of highly inflected languages is formalized in UniMorph, and a large, multilingual, freely available repository of UniMorph-annotated inflections exists in the Wiktionary dataset. Using a small UniMorph-annotated English corpus, we generalize English inflectional phrase templates to express a wide range of UniMorph vectors. We then use the inflectional information to assign English phrase templates as translation candidates to inflections from Wiktionary. Building translation models from these pairs, we substantially improve the quality of MT in a range of low-resource settings.

Analyzing a range of resource settings and levels of morphological information, we find that a full morphological system outperforms inflection-lemma mappings and lemma-lemma mappings. We also find that morphological information is less useful in higher-resource settings, but can still provide substantial gains. We believe this approach holds promise for constructing translation systems for language pairs that do not have much in the way of bitext. In service of this goal, we release translation models constructed for Finnish, Czech, Russian, Korean, Georgian, Swahili, Turkish, and Urdu.

For future work, we believe it would be useful to investigate how well these phrase-table augmentation techniques work when combined with other approaches to low-resource machine translation into English, such as lemmatized backoff models.

References

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python*. " O'Reilly Media, Inc.".

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015a). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015b). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Chiang, D. (2007). Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

De Smedt, T. and Daelemans, W. (2012). Pattern for python. *J. Mach. Learn. Res.*, 13:2063–2067.

Dyer, C. J. (2007). The'noisier channel': translation from morphologically complex languages. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 207–211. Association for Computational Linguistics.

Habash, N. (2008). Four techniques for online handling of out-of-vocabulary words in arabic english statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, pages 57–60. Association for Computational Linguistics.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing prague czech-english dependency treebank 2.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, İstanbul, Turkey. ELRA, European Language Resources Association.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable modified kneser-ney language model estimation. In *ACL (2)*, pages 690–696.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. Barcelona, Spain.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P. and Haddow, B. (2012). Interpolated backoff for factored translation models. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Mirkin, S., Specia, L., Cancedda, N., Dagan, I., Dymetman, M., and Szpektor, I. (2009). Source-language entailment modeling for translating unknown terms. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 791–799. Association for Computational Linguistics.

Post, M., Cao, Y., and Kumar, G. (2015). Joshua 6: A phrase-based and hierarchical statistical machine translation system. *The Prague Bulletin of Mathematical 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. Association for Computational Linguistics.

Tiedemann, J. (2009). News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In Nicolov, N., Bontcheva, K., Angelova, G., and Mitkov, R., editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria.

Tyers, F. M. and Alperen, M. S. (2010). South-east european times: A parallel corpus of balkan languages. In *Proceedings of the LREC Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 49–53.

Yang, M. and Kirchhoff, K. (2006). Phrase-based backoff models for machine translation of highly inflected languages. In *EACL*, pages 3–7.