This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

D3.3: Intermediate report on inflection and word formation

Author(s): Alexander Fraser, Matthias Huck, Aleš Tamchyna, Ondřej Bojar, Dušan Variš, Anita Ramm, Marion Weller-Di Marco

Dissemination Level: Public

Date: February 1st 2017

Version 1.0

Grant agreement no.	644402	
Project acronym	HimL	
Project full title	Health in my Language	
Funding Scheme	Innovation Action	
Coordinator	Barry Haddow (UEDIN)	
Start date, duration	1 February 2015, 36 months	
Distribution	Public	
Contractual date of delivery	February 1st 2017	
Actual date of delivery	Ferbruary 1st 2017	
Deliverable number	D3.3	
Deliverable title	Intermediate report on inflection and word formation	
Type	Report	
Status and version	1.0	
Number of pages	14	
Contributing partners	CUNI	
WP leader	LMU-MUENCHEN	
Task leader	LMU-MUENCHEN	
Authors	Alexander Fraser, Matthias Huck, Aleš Tamchyna, Ondřej Bojar, Dušan Variš, Anita Ramm, Marion Weller-Di Marco	
EC project officer	Martina Eydner	
The Partners in HimL are:	The University of Edinburgh (UEDIN), United Kingdom	
	Univerzita Karlova V Praze (CUNI), Czech Republic	
	Ludwig-Maximilians-Universitaet Muenchen (LMU-MUENCHEN), Germany	
	Lingea SRO (LINGEA), Czech Republic	
	NHS 24 (Scotland) (NHS24), United Kingdom	
	Cochrane (COCHRANE), United Kingdom	

For copies or reports, updates on project activities and other HimL-related information, contact:

Barry Haddow	bhaddow@staffmail.ed.ac.uk
University of Edinburgh	Phone: +44 (0) 131 651 3173

Contents

1 Executive Summary

This report covers activities in term of modeling inflectional and word formation phenomena. The associated tasks are Task 3.3: Corrective approaches to morphology, Task 3.4: Separating translation from inflection and word formation: German, Task 3.5: Separating translation from inflection and word formation: Czech, and Task 3.6: Separating translation from inflection and word formation, Phase 2, Polish and Romanian.

This report follows Deliverable D3.1, which covered the effort on the same tasks for year 1.

As shown below, Task 3.4 concluded according to the plan, and all other tasks are proceeding according to the plan.

Task	Months	Status
3.3: Corrective approaches to morphology	1–36	AS PLANNED
3.4: Separating translation from inflection and word formation: German	1–24	CONCLUDED AS PLANNED
3.5: Separating translation from inflection and word formation: Czech	7–30	AS PLANNED
3.6: Separating translation from inflection and word formation, Phase 2, Polish and Romanian	13–36	AS PLANNED

2 Task 3.3: Corrective approaches to morphology

In this section, CUNI presents current status of development of MLFix, an automatic post-editing tool for statistical machine translation. In Deliverable 3.1, we described a general architecture of MLFix, the process of preprocessing and postprocessing of the input data, design of a used feature set and the outline of the feature extraction process. In this deliverable we present a more detailed description of the statistical post-editing components and evaluation results on both EN-CS and EN-DE translation outputs.

In Deliverable 3.1, we listed main sources of data available for training and development of the MLFix system. Additionally we included data available for the WMT16 post-editing task[1]. Training data available for this task were taken from the QT21 project (which we already mentioned in the previous deliverable). However, we decided to use the WMT16 test set containing 3,000 sentences translated by various MT systems together with reference sentences as an additional source of training data.

2.1 MLFix: Statistical components

After examining the task at hand, we decided to split the post-editing process into two tasks: identifying words with an erroneous wordform (error detection) and assigning a new wordform to the erroneous words based on a corrected morphology (morphology prediction). Both components work with word tokens as their basic units. The error detection component uses a binary classifier to solve the task while the morphology prediction component fits a multiclass classifier and predicts several different categories (Interset features) at once[2] Both tasks had their own specific issues that we had to solve.

Error detection task was problematic mainly due to a lack of available annotated data. The only information available is the source sentence, MT translation output and post-edited or reference sentence. Our preprocessing provides us with additional morphological and some syntactic information together with a word alignment between all sentences. However, this information is usually not enough to identify the type of error or its span. Thus we designed a heuristic to label tokens which have an incorrect surface form in our training data (in the MT output). The word gets an error label assigned if and only if these three conditions are satisfied:

- *word.form* is not equal to *word.ref.form*
- *word.parent* was already modified OR *word.parent.form* is equal to *word.parent.ref.form*
- *word.grandparent.lemma* is equal to *word.grandparent.ref.lemma*

This heuristic still manages to label some instances incorrectly. We evaluated the heuristic using a perfect (Oracle) classifier (a classifier that sees and uses the correct answers). Since the automatic evaluation metrics are not well suited for this evalution we used a single human annotator to evaluate the performance of the classifier on the HimL test set. The evaluation was performed on pairs of MT output and post-edited output, randomly shuffled to avoid biasing evaluator's decisions. Table 1 contains results of the evaluation. For each sentence pair, we examined if the post-edited sentence was better (+), worse (−) than the MT output or if the result was indecisive (0). Then, we computed precision (1) and the impact (2) of the Oracle's post-edits with following formulas:

[1] http://www.statmt.org/wmt16/ape-task.html

[2] This is due to availability of such classifier implementations within the Scikit-Learn toolkit.

Reference	Evaluated	Changed	+	−	0	Precision	Impact
Post-edits	800	95	61	16	18	79.2%	7.6%

Table 1: Results of the manual evaluation of the ideal system based on the chosen heuristic on the HimL testset.

$$precision = \frac{better}{better + worse} \tag{1}$$

$$impact = \frac{better}{evaluated} \tag{2}$$

We can see that Oracle improved around 7.6% of the sentences and of all performed changes, 79.2% were for better. This gave us an optimistic ground for training our classifiers.

Naturally, a low impact of the Oracle classifier implies that our heuristic produces very unbalanced training data. Only less then 1/10 of the training sentences contain at least one positive prediction instance (a morphology error) which makes training of an error classifier with satisfying recall very difficult (not labeling any word as an error gives us ˜95% accuracy). We decided to counter this problem by including only sentences containing erroneous instances in our training data and weighting the desired target class. This, however, further lowered the overall number of training instances that were used for training of the classifier. This issue can be resolved by collecting more training data[3] and it is going to be one of the aspects of our future work.

Morphology prediction task is supposed to predict new morphological categories for words labeled by the error detection component. Interset contains more than 40 different features, however, many of these (e.g. puncside, style) features are not required for correctly generating a surface form from a lemma. We analysed our data (labeled by our heuristic) to determine which features are changed most frequently during the post-editing process. We discovered that nouns and adjectives are the most edited part-of-speech (POS) classes in our data, usually by modifying either a case, gender or number of a word. For this reason, we decided to focus mainly on predicting the Interset subset containing these three categories.

The data used for training were even smaller than those used for error detection because we used only instances of words marked as erroneous by our heuristic. We did not use the whole dataset since a subset of features used during prediction contains information about the morphology of the MT word. On the other hand, including the correct instances (without any modified Interset features) might help counter situations where the error detection component makes an error. Until now, we did not compare these two aproaches, however, it will be part of our future work.

2.2 System evaluation

We performed an automatic evaluation of the EN-CS pipeline on various datasets listed in Deliverable 3.1. MLFix supports a combination of multiple models which we used to perform an evaluation in a leave-one-out manner. We trained a separate model for each dataset, both for error detection and morphology prediction, and we used different methods for model combination. To combine results of multiple binary classfiers we examined these three approaches:

- **majority vote** - the prediction represented by the majority of classifiers is chosen,
- **at-least-one vote** - we mark the word as erroneous if at least one classifier labels it,
- **average vote** - we label the word if the average confidence (probability of the positive class) exceeds a specific threshold

After performing experiments on the isolated error detection component we chose average vote with threshold 0.5 for the final evaluation. So far we did not investigate the influence of various threshold values on the MLFix performance and it is going to be a subject of future research. Since morphology prediction classifies more than two classes we decided to simply pick a class that received the highest score across all classifiers.

Table 2 shows results of the automatic evaluation. We can see that MLFix was able to improve almost every dataset, however, often by only a small margin. Interesting is the performance during post-editing of the NMT output. Even though it looks promising it requires further investigation in the future.

We also performed a manual evaluation on the HimL testset and the WMT16 testset translated by CU Chimera. The evaluation was performed in the same manner as during the Oracle evaluation. Two independent annotators were used. The results are shown in Table 3. In the end, 126 out of 3,800 sentences were modified by MLFix; these were annotated by the annotators. MLFix was able to improve about four-fifths of these sentences. The inter-annotator agreement reached 62% (87% if we disregar the indefinite changes). Therefore, these results still need to be confirmed by a larger scale evaluation.

[3] The easiest solution is to include data that contain only reference sentences instead of post-edited sentences.

Dataset	System	Base	MLFix	Depfix
Autodesk	NA	47.82	**47.89 (+0.06)**	47.63 (-0.19)
HimL	Moses	20.66	20.69 (+0.02)	**21.02 (+0.35)**
WMT10	CU Bojar	15.66	15.76 (+0.10)	**15.91 (+0.25)**
WMT16	UEDIN NMT	26.31	**26.49 (+0.18)**	26.15 (-0.15)
	CU Chimera	21.72	**21.79 (+0.07)**	21.75 (+0.02)

Table 2: System-wide evaluation of MLFix using Bleu. Values are multiplied by 100 for easier reading. Average voting method was used for error detection. The performance of Depfix is shown for comparison.

	Evaluated	Changed	+	−	0	Precision	Impact
HimL-A	800	26	17	5	4	77.2%	2.1%
WMT16-A	2,999	100	73	21	6	77.6%	2.4%
HimL-B	800	26	5	5	16	50%	0.6%
WMT16-B	2,999	100	71	12	17	85.5%	2.3%
Total	7,598	252	166	43	43	79.4%	2.1%

Table 3: Results of manual evaluation of the best MLFix configuration. Annotators A and B are distinguished by a suffix for each dataset.

Finally, we ran several experiments on the EN-DE pipeline. It is similar to the original EN-CS pipeline with some language-specific tools replaced. We use Mate toolkit[4] for German lemmatization and tagging and for parsing of reference sentences. The tagger uses CoNLL2009 style morphological tags. For wordform generation, we first tried a simple Flect model without any success and eventually switched to SMOR[5] generator. Tagset used by SMOR is not identical to the one used by the Mate toolkit. For the time being we were able to use Mate tagset to generate wordforms for nouns and adjectives by SMOR, however, we were unable to do the same for e.g. verbs. It is not an issue at the moment but in future a SMOR-based tagger might be desired.

So far, we performed same steps as in the EN-CS pipeline during model training with little to no adaptation to the German language. We used WMT16 and Autodesk datasets for training. We performed an evaluation on the HimL testest with same voting strategies as in the EN-CS pipeline. We also tried tweaking the threshold in the average voting method. Table 4 shows results of the experiments.

The results so far do not show much promise. We can see that, by increasing the threshold, the resulting score decreases along with the amount of modified sentences. Due to the poor results, we did not perform a manual evaluation. However, we examined a portion of the data post-edited by MLFix and confirmed that there were errors made by both components, the morphology errors usually manifesting due to changes in the morphology of correct words. A more detailed evaluation of the separate components is thus required in the future. As next step we also want to examine whether increasing the amount of training data or simply switching to a different machine learning method can help improve the performance. If unsuccessful, we also consider revision of the training process together with the chosen heuristic to better suit the EN-DE post-editing task.

We suspect that the poor MLFix performance on the EN-DE language pair might be due to large differences between Czech and German morphology. We will explore this by performing evaluations on the EN-PL and EN-RO pipelines. Depending on the results, we will decide on a next direction in which the MLFix research will continue.

2.3 Dockerization

As an intermediate phase of the deployment of automatic error-correction systems, we worked on deploying Depfix by creating a docker image. Depfix and MLfix both rely on the same underlying software platform Treex, so the techniques needed for the deployment of both systems are very similar.

The Depfix docker image can be easily modified in the future to use MLFix instead.

3 Task 3.4: Separating translation from inflection and word formation: German

LMU-MUENCHEN carried out work on adapting pre-processing and post-processing approaches to dealing with morphology for translating to German in the consumer health domain in Task 3.4. There is also related work which is still ongoing (involving further domain adaptation and creation of the Y3 systems).

[4] http://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/matetools.en.html

[5] http://www.cis.uni-muenchen.de/~schmid/tools/SMOR/

	BLEU	Changed
Baseline	30.95	–
Majority	29.93 (-1.02)	230
Average-0.3	30.07 (-0.87)	199
Average-0.5	30,80 (-0.15)	48
Average-0.7	30.90 (-0.05)	16
Average-0.8	30.95 (0.00)	0

Table 4: Results of the EN-DE MLFix automatic evaluation using BLEU on the HimL testset. Values are multiplied by 100 for easier reading. We also list the number of sentences changed by MLFix.

As already documented in Deliverable 3.1, we began with our successful approach to dealing with nominal morphology and added support for prediction of German verbal information into this pipeline, additionally combining this with our work on rule-based reordering for English→German translation, all of which is included in the HimL Y2 system which was evaluated. We also carried out significant work improving the prediction of arguments, focusing on grammatical case and on the translation of prepositions, which are often subcategorized for in the verbal frame (which was documented in Deliverable 3.1). For details see Weller *et al.* (2015b,a).

In Y2, we looked at synthesizing new phrase table entries to deal with the sparsity in longer phrases caused by prepositions, which will be discussed further below, see also Weller-Di Marco *et al.* (2016). We extended our previous work on verbal complexes by combining our nominal inflection system with verbal reordering and modification of verbal inflection, and in particular correcting subject-verb agreement using a dependency parser (similar to CUNI's approach in DepFix) and trying to model verbal inflection, which is also described further below, see also Ramm and Fraser (2016).

Synthesizing new phrase table entries SMT output is often incomprehensible because it confuses complement types (noun phrases/NPs vs. prepositional phrases/PPs) by generating a wrong grammatical case, by choosing an incorrect preposition, or by arranging the complements in a meaningless way. However, the choice of complement types in a translation represents important information at the syntax-semantics interface: The case of an NP determines its syntactic function and its semantic role; similarly, the choice of preposition in a PP distinguishes between argument PPs and adjunct PPs, and also sets the semantic role of the prepositional phrase.

While the lexical content of a target-language phrase is defined by the source sentence, the exact choice of preposition and case strongly depends on the target context, and most specifically on the target verb. For example, the English verb phrase *to call for sth.* can be translated into German by *etw. erfordern* (subcategorizing a direct-object NP but no preposition) or by *(nach) etw. verlangen* (subcategorizing either a direct-object NP or a PP headed by the preposition *nach*). Differences in grammatical case and syntactic functions between source and target side include phenomena like subject-object shifting: *[I]*$_{\text{SUBJ}}$ *like [the book]*$_{\text{OBJ}}$ vs. *[das Buch]*$_{\text{SUBJ}}$ *gefällt [mir]*$_{\text{OBJ}}$. Here, the English object corresponds to a German subject, whereas the English subject corresponds to the indirect object in the German sentence.

Selecting the wrong complement type or an incorrect preposition obviously has a major effect on the fluency of SMT output, and also has a strong impact on the perception of semantic roles. Consider the sentence *John looks for his book.* When the preposition *for* is translated literally by the preposition *für*, the meaning of the translated sentence *John sucht für sein Buch* shifts, such that *the book* is no longer the object that is searched, but rather a recipient of the search. To preserve the source meaning, the prepositional phrase headed by *for* must be translated as a direct object of the verb *suchen*, or as a PP headed by the preposition *nach*.

Since prepositions tend to be highly ambiguous, the choice of a preposition depends on various factors. Often, there is a predominant translation, such as *for* → *für*, which is appropriate in many contexts, but does not lead to valid translations in other contexts. Such translation options are often difficult to override, even when there are clues that the translation is wrong. Furthermore, even though prepositions are highly frequent words, there can be coverage problems if a preposition is not aligned with the specific preposition required by the context, due to structural mismatches between the source sentence and the hypothesized target translation.

We worked on two novel approaches to improve the modeling of complement types in an English-to-German SMT system. A simple approach introduces an abstract representation of "placeholder prepositions" at the beginning of noun phrases on the source and target sides. The insertion of these placeholder prepositions leads to a more symmetric structure and consequently to a better coverage of prepositions, as all NPs are effectively transformed into PPs, and prepositions in one language without a direct equivalent in the other language can be aligned. Furthermore, the placeholder prepositions function as explicit phrase boundaries and are annotated with grammatical case, so they provide flat structural information about the syntactic function of the phrase. The placeholder representation leads to a significant improvement over a baseline system without prepositional placeholders.

Our second approach enhances the abstract placeholder representation, and integrates source-side context into the phrase table of the SMT system to model different complement types. This is done by generating synthetic phrase-table entries containing contextually predicted prepositions. With this process, we aim to (i) improve the preposition choice conditioned on the source sentence, and to (ii) manipulate the scores in the generated entries to favour context-appropriate translations. Generating phrase-table entries allows to create prepositions in contexts not observed in the parallel training data. The resulting phrase-table entries are unique for each context and provide the best selection of translation options in terms of complement realization on token-level. Our work showed that this variant significantly outperforms the baseline (the baseline had a BLEU score of 19.76, while the improved system had a BLEU score of 19.86, the difference is significant using bootstrap resampling but very small). Additionally, it is slightly better than the system with inserted placeholder prepositions.

In this work, we carried out experiments showing that explicit information about different complement types (insertion of empty placeholders) leads to improved SMT quality. The results of the synthetic-phrases system are slightly better than those of the preposition-informed system, with two variants being significantly better. As the differences are rather small and apply only to some system pairs (see the paper for details), it is difficult to draw a clear conclusion concerning the effectiveness of the synthetic-phrases method. Our analysis showed, however, that newly generated phrases are indeed used within the systems and help to improve translation quality. We consider this a confirmation that the generation of synthetic phrases for handling subcategorization is a sound approach.

Combining nominal and verbal inflection prediction with pre-translation reordering Statistical machine translation of English into German faces two main problems involving verbs: (i) correct placement of the verbs, and (ii) generation of the appropriate inflection for the verb.

The position of verbs in German and English differs greatly and often large-range reorderings are needed to place the German verbs in the correct positions. Gojun and Fraser (2012) showed that the *preordering* approach applied on English–to–German SMT overcomes large problems with both missing and misplaced verbs.

Fraser *et al.* (2012) proposed an approach for handling inflectional problems in English to German SMT, focusing on the problems of sparsity caused by nominal inflection. However, they do not handle the verbs, ensuring neither that verbs appear in the correct position (which is a problem due to the highly divergent word order of English and German), nor that verbs are correctly inflected (problematic due to the richer system of verbal inflection in German). In many cases, verbs do not match their subjects (in person and number) which makes understanding of translations difficult. In addition to person and number, the German verbal inflection also includes information about tense and mood. If these are wrong (i.e. do not correspond to the tense/mood in the source), very important information, such as point of time and modality of an action/state expressed by the verb, is incorrect. This can lead to false understanding of the overall sentence.

In our initial work on modeling verbal inflection we used the nominal inflection modeling for translation to German presented by Fraser *et al.* (2012) and combined it with the reordering of the source data Gojun and Fraser (2012). We presented a method for correction of the agreement errors, and an approach for modeling the translation of tense and mood from English into German. A summary of the results is in table 5.

	BLEU$_{ci}$
Surface	21.59
Baseline	22.00
Verbal inflection	22.05
Agreement	22.08
Tense/mood	21.95

Table 5: BLEU scores of MT outputs with corrected verbal inflection.

As in most of experimental setups, the vanilla Moses system is worse than the baseline, our stemmed system. The new work shows that modeling of tense/mood translation is highly problematic, both in terms of modeling but also in terms of knowing the correct answer for a particular context (i.e., what is required for evaluation). For an extensive study with many examples see Ramm and Fraser (2016). The subject-verb agreement problems were dealt with successfully, by using a dependency parser-driven approach similar to CUNI's DepFix. The final result is an improvement from 21.59 (surface) to 22.05 BLEU.

Concluding the English-to-German work This task is concluded with this deliverable. We carried out extensive experimentation to determine the best combination of the work we have reported on here and in Deliverable 3.1.

In our experiments, we studied all three linguistic levels we examined separately previously, by combining our approaches which were previously studied only independently. We explored system variants that combine target-side morphological modeling, structural adaptation between source and target side and a discriminative lexicon enriched with features relevant for support verb

system	basic	VW-1 pos/lem	VW-2 pos/lem/dep
Surface	19.45	19.81*	19.90*
Surface V-Reordered	19.71*	20.24*	20.27*
MorphSys	19.81*	19.80*	19.93*
MorphSys V-Reordered	20.08*	20.51*	20.50*

Table 6: Morpho-syntactic and lexical strategies.
***: significantly better than Surface-basic (19.45)**

constructions and verbal inflection. We showed that the components targeting the different linguistic levels are complementary, but also that applying only verbal pre-ordering can introduce problems on the morpho-lexical level; our experiments indicated that the discriminative classifier developed as part of our work in HimL (Tamchyna *et al.*, 2016) can overcome these problems.

Individual strategies aiming at one linguistic level are established and usually improve translation, but prior to our experimentation it was not clear (i) whether individual gains add up when combining approaches and (ii) how individually targeting one linguistic level impacts other levels. We addressed these questions for the combined strategies of *source-side reordering* (preprocessing), *discriminative classifier* (at decoding time) and *target-side generation* of nominal inflection (post-processing). For (ii), we focused on source-side reordering and investigate whether introducing German clause ordering in the English data entails new problems: while in "regular" English verbs and their arguments are close to each other, they can be separated by large distances in the German-structured English.

Reordering improves translation quality, but separating the verb from its arguments has also negative consequences. First, the agreement in number between verbs and subjects is impaired because subjects and verbs are separated Ramm and Fraser (2016). Second, there can be a negative effect on the lexical level, for example when translating multi-word expressions. Consider the phrase *to cut interest rates*: if the parts occur close to each other, there is enough context to translate *cut* into *senken* ('to decrease'). However, with too large a gap between *cut* and *interest rates*, it becomes difficult to disambiguate *cut*, leading to the wrong translation *schneiden*: ('to cut with a knife').

The column "basic" in table 6 shows the results for combining strategies at the morphological and the syntactic level: "Surface" refers to a baseline system trained on surface forms; "MorphSys" denotes the inflection prediction system. "V-Reordered" refers to systems built on reordered source-side data. Combining the two strategies adds up to a statistically significant gain of 0.63 between the basic system (19.45) and the system with morphological modeling and source-side reordering (20.08).

4 Task 3.5: Separating translation from inflection and word formation: Czech

LMU-MUENCHEN and CUNI have conducted research towards better modeling of Czech morphology in English→Czech phrase-based machine translation, with a particular focus on covering morphological variants in Czech that are not present in the parallel training data, and can therefore not be learned.

Translation into morphologically rich languages such as Czech poses problems to statistical machine translation systems because many of the valid morphological variants are not observed in training, even when large amounts of parallel training data are available. Typical state-of-the-art phrase-based translation systems are able to generate an inflected word form on the target side if the respective morphological variant has been seen in combination with the given input word in the training corpus. However, if this is not the case, the system has no means of producing the correct inflection.

CUNI has developed an experimental pipeline for morphology prediction. We address the task using a linear discriminative classifier which uses features both from the source and the target side, along with all of their combinations. The classifier is integrated in the Moses toolkit which allows for efficient queries. So far, we have experimented with English→Czech translation in a single step, using the classifier to judge the morphological adequacy of the translation. One of the main engineers of Task 3.5, Aleš Tamchyna of CUNI, visited LMU-MUENCHEN in 2016, allowing LMU-MUENCHEN and CUNI to work jointly. The result was a new discriminative lexicon which is able to utilize target-side features in search, resulting in an ACL paper (Tamchyna *et al.*, 2016); the evaluation showed that the approach leads to BLEU score gains in all four HimL languages.

Later in Y2, LMU-MUENCHEN and CUNI have investigated an approach based on synthesized morphological variants. A morphological generation tool is utilized to synthesize all valid morphological forms from Czech word lemmas. The phrase table of the baseline system is then augmented with additional entries. Phrase translation probabilities and lexical translation probabilities cannot be calculated in the usual manner for artificial entries. The valid Czech inflection also largely depends on the context. LMU-MUENCHEN and CUNI have explored features that go beyond only relying on large *n*-gram language models for scoring synthesized morphological forms. Particularly, we take context on both source and target side into account (as in our

Case	Surface Form	50K	500K	5M	50M
1	čéšky	●	●	●	●
2	čéšek	–	●	●	●
3	čéškám	–	–	●	●
4	čéšky	○	○	●	●
5	čéšky	○	○	○	○
6	čéškách	–	●	●	●
7	čéškami	–	–	–	●

Table 7: Morphological variants of the Czech lemma "čéška". For differently sized corpora (50K/500K/5M/50M), "•" indicates that the variant is present, and "○" that the same surface form realization occurs, but in a different syntactic case.

previous work, but now combined with new phrase table entries), along with the use of an array of feature functions to control when synthesized new forms are used. We provide more details on the latter approach in the following subsection (Section 4.1), along with experimental results on the HimL test set.

Our new approach differs from previously investigated "two-step" techniques while retaining many of their benefits. By devising dedicated components that strictly separate translation from inflection and word formation in terms of both modeling and inference, previous "two-step" paradigms counteract sparsity and allow for the generation of morphological word forms that are unobserved in the training data. The same capabilities are provided with our new method, however, inference is not spread across two consecutive search steps. A morphology classifier is integrated directly into the weighted model combination used by the machine translation decoder, allowing the decoder to make use of information that is only available in a second inference step in "two-step" paradigms. Furthermore, we solve the problem of the generation of unseen morphological variants by means of creating all of them at training time from lemmatized forms and providing them to the decoder, rather than producing them on-the-fly in some sort of postprocessing, as done by the "two-step" systems. Our approach owes inspiration to the "two-step" techniques but can be interpreted as a more *soft* separation of translation from inflection and word formation.

4.1 Producing unseen morphological variants using synthesized phrase table entries

4.1.1 Motivation

Morphologically rich languages exhibit a large amount of inflected word surface forms for most lemmas, which poses difficulties to current statistical machine translation technology. SMT systems, such as phrase-based translation engines (Koehn *et al.*, 2003), are trained on parallel corpora and can learn the vocabulary that is observed in the data. After training, the decoder can output words which have been seen on the target side of the corpus, but no unseen words.

Sparsity of morphological variants leads to many linguistically valid morphological word forms remaining unseen in practical scenarios. This is a substantial issue under low-resource conditions, but the problem persists even with larger amounts of parallel training data. When translating into the morphologically rich language, the system fails at producing the unseen morphological variants, leading to major translation errors.

Consider the Czech example in Table 7. A small parallel corpus of 50K English-Czech sentences contains only a single variant of the morphological forms of the Czech lemma "čéška" (plural of English: "kneecap"), out of seven syntactically valid cases. The situation improves as we add in more training data (500K/5M/50M), but we can generally not expect the SMT system to learn all variants of each known lemma. In Czech, the number of possible variants is even larger for other word categories such as verbs or adjectives.

LMU-MUENCHEN and CUNI propose an extension to phrase-based SMT that allows the decoder to produce *any* morphological variant of all known lemmas. We design techniques for integrating and scoring unseen morphological variants right in phrase-based search, with the decoder being able to choose freely amongst all possible morphological variants.

4.1.2 Generating unseen morphological variants

We investigate an approach based on synthesized morphological variants, in a manner comparable to (Chahuneau *et al.*, 2013). A morphological generation tool is utilized to synthesize all valid morphological forms from target-side lemmas. The phrase table is then augmented with additional entries to provide complete coverage.

We process single target-word entries from the baseline phrase table and feed the lemmatized target word into the morphological generation tool. If its output contains morphological forms that are not known as translations of the source side of the phrase, we add these morphological variants as new translation options. We consider two settings: (1.) **word**, where morphological word

Feature Type	Configurations
source indicator	l, t
source internal	l, l+r, l+p, t, r+p
source context	l (-3,3), t (-5,5)
target context	l (2), t (2)
target indicator	l, t
target internal	l, t

Table 8: Feature templates for the discriminative classifier: l (lemma), t (morphosyntactic tag), r (syntactic role), p (lemma of dependency parent). Numbers in parentheses indicate context size.

forms are generated from phrase table entries of length 1 on both source and target side, and (2.) **mtu** (for "minimal translation unit"), where the phrase source side can have arbitrary length.

Morphological generation for Czech can be performed with the MorphoDiTa toolkit (Straková *et al.*, 2014), which we use in our experiments. MorphoDiTa knows a dictionary of most Czech lemmas and can generate all their morphological variants (Hajič, 2004).

When not restricted, the morphological generator also produces forms which do not match in number, tense, degree of comparison, or even negation. This may be undesirable and we therefore define a *tag template*. The template only allows freedom in the following morphological categories: gender, case, person, possessor's number and gender. All other attributes must match the original word form. We mark this configuration with an asterisk (★) in our experiments.

4.1.3 Scoring unseen morphological variants

Assigning dependable model scores to synthesized morphological forms is a primary challenge. During decoding, the artificially added phrase table entries compete with baseline phrases that had been directly extracted from the parallel training data. The correct choice has to be determined in search based on model scores.

A phrase-based model with linguistically motivated*factors* (Koehn and Hoang, 2007) enables us to achieve better generalization capabilities when translating into a morphologically rich language. In our baseline systems, we already draw on lemmas and morphosyntactic tags as factors on the target side, in addition to word surface forms. We incorporate *n*-gram LMs over lemmas and morphosyntactic tags, and an operation sequence model (OSM) (Durrani *et al.*, 2013) with lemmas on the target side. These models counteract sparsity, and where models over surface forms fail for unseen variants, they still assign scores which are based on reliable probability estimates.

When enhancing a system with synthesized phrase table entries, we add further features. Since the usual phrase translation and lexical translation log-probabilities over surface forms cannot be estimated for unseen morphological variants, but all new variants are generated from existing lemmas, we utilize the corresponding log-probabilities over target lemmas. A binary indicator distinguishes baseline phrases from synthesized phrases.

The final key to our approach is using a discriminative classifier (**morph-vw**, *Vowpal Wabbit for Morphology*) which can take context from both the source side and the target side into account, as in (Tamchyna *et al.*, 2016). We design feature templates for the classifier that generalize to unseen morphological variants, as listed in Table 8. "Indicator" features are concatenations of words inside the phrase, "internal" features represent each word in the phrase separately. Context features on the source side capture a fixed-sized window around the phrase. Target-side context is only to the left of the current phrase. The features require lemmatization and tagging on both sides and a dependency parse of the source side.

4.1.4 Empirical evaluation

For an empirical evaluation of our technique, we build baseline phrase-based SMT engines using `Moses` (Koehn *et al.*, 2007). We then enrich these baselines with linguistically motivated morphological variants that are unseen in the parallel training data, and we augment the model with the discriminative classifier to guide morphological selection during decoding. Different flavors of synthetic morphological variants are compared, each either combined with the discriminative classifier or standalone.

Experimental setup. We train a phrase-based translation system with three factors on the target side of the translation model (but no separate generation model). The target factors are the word surface form, lemma, and a morphosyntactic tag. We use the Czech positional tagset (Hajič and Vidová-Hladká, 1998) which fully describes the word's morphological attributes. On the source side we use only surface forms, except for the discriminative classifier, which includes the features as shown in Table 8.

	En→Cs HimL Test			
training corpus size	**50K**	**500K**	**5M**	**50M**
setup	Bleu	Bleu	Bleu	Bleu
baseline	14.6	18.4	20.8	23.6
+ morph-vw	14.7	19.6	21.7	23.9
+ synthetic (word)	14.9	18.5	20.9	23.3
+ morph-vw	15.1	19.5	**21.9**	23.9
+ synthetic (word⋆)	15.1	18.3	20.8	23.4
+ morph-vw	15.4	19.5	21.7	24.0
+ synthetic (mtu)	15.1	18.6	20.7	23.7
+ morph-vw	15.2	**19.7**	21.8	**24.1**
+ synthetic (mtu⋆)	15.3	18.6	20.8	23.3
+ morph-vw	**15.6**	19.6	21.7	**24.1**

Table 9: Producing unseen morphological variants using synthesized phrase table entries: English→Czech experimental results on the HimL test set.

We employ corpora that have been provided for the English→Czech News translation shared task at WMT16 (Bojar *et al.*, 2016). Word alignments are created using `fast_align` (Dyer *et al.*, 2013) and symmetrized. The phrase table is pre-pruned by applying a minimum score threshold of 0.0001 on the source-to-target phrase translation probability, and the decoder loads a maximum of 100 best translation options per distinct source side. Pop limit and stack limit for cube pruning are set to 1000 for tuning and to 5000 for testing. Weights are tuned on newstest2013 with *k*-best MIRA (Cherry and Foster, 2012). Translation quality is measured in case-sensitive Bleu (Papineni *et al.*, 2002) on the HimL test set.[6]

Our training data amounts to around 50 million bilingual sentences overall, but we conduct sets of experiments with systems trained using different fractions of this data (**50K**, **500K**, **5M**, **50M**). Whereas English→Czech has good coverage in terms of training corpora, we simulate low- and medium-resource conditions for the purpose of drawing more general conclusions. Irrespective of this, we utilize the same large LMs in all setups, assuming that proper amounts of target language monolingual data can often be gathered, even when parallel data is scarce.

Experimental results. Translation results on the HimL test set are reported in Table 9. Our method is effective at improving Bleu under all the different training resource conditions, showing a gain over the baseline of +1.0 Bleu in the low-resource scenario (50K) and of +0.5 Bleu in the large-scale setting (50M). The Bleu scores on the HimL test set seem to indicate that synthesized phrase table entries are effective mostly when very few parallel data is at hand (50K). The *morph-vw* discriminative classifier boosts translation quality under all conditions, though, and in particular with medium amounts of parallel training data (500K, 5M). We did not observe the exact same tendencies when we evaluated all our setups on WMT *newstest* sets, which may be attributed to out-of-domain effects between the HimL test sets and the training and tuning corpora.[7] However, all our results support the major finding that our method is indeed effective, and that the best translation quality can be achieved with combinations of both the discriminative classifier for morphology *and* synthesized morphological forms in the phrase table.

5 Task 3.6: Separating translation from inflection and word formation: Polish and Romanian

This task began in Year 2. CUNI and LMU-MUENCHEN are coordinating on applying the language neutral techniques developed in T3.4 and T3.5 to translation from English to Polish and Romanian. In our work on the discriminative lexicon which utilizes both source and target-context in the online decoder (Tamchyna *et al.*, 2016), we showed that the approach works for not only Czech and German, but also for Polish and Romanian, using standard well-studied benchmarks studied in the WMT (ACL conference on machine translation) community.

Morphology tools prepared at Lingea to implement the common morphology interface as described in Deliverable 3.2 are about to become available for research at CUNI. Now Lingea is preparing license conditions for their proprietary parts to both protect commercial interests and allow their maximal research use. This will enable use of similar techniques that were employed on Czech and German also on Polish and Romanian.

[6] We evaluate case-sensitive with `mteval-v13a.pl -c`, comparing post-processed hypotheses against the raw reference.

[7] On WMT *newstest* sets, we see larger improvements in the 50K and 500K settings when synthetic phrases are employed standalone without the support by the discriminative classifier; whereas on the other hand, the *morph-vw* appears to be somewhat less effective overall when we evaluate on WMT *newstest* sets.

In the ongoing work in this work package, we are applying these models to HimL systems, and also investigating making them domain-aware. This work will be integrated into the HimL Year 3 systems, and documented in the final morphology deliverable.

6 Conclusion

This deliverable on morphology has described the wide variety of work on morphology in HimL in Year 2. We have reported significant advances in corrective approaches to morphology, concluded the work on separate translation from inflection and word formation in German. The work on Czech has been the key to making our approach language neutral and has shown very satisfactory results in terms of handling surface forms of known lemmas which were not seen in the training data, a critical problem also for Romanian and Polish. In the last year we will continue this work and implement it for Romanian and Polish systems.

References

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. "Findings of the 2016 Conference on Machine Translation." *Proceedings of the First Conference on Machine Translation*, 131–198. Berlin, Germany.

Chahuneau, Victor, Eva Schlinger, Noah A. Smith, and Chris Dyer. 2013. "Translating into Morphologically Rich Languages with Synthetic Phrases." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1677–1687. Seattle, WA, USA.

Cherry, Colin and George Foster. 2012. "Batch Tuning Strategies for Statistical Machine Translation." *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, 427–436. Montréal, Canada.

Durrani, Nadir, Alexander Fraser, and Helmut Schmid. 2013. "Model With Minimal Translation Units, But Decode With Phrases." *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, 1–11. Atlanta, GA, USA.

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. "A Simple, Fast, and Effective Reparameterization of IBM Model 2." *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 644–648. Atlanta, GA, USA.

Fraser, Alexander, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. "Modeling inflection and word-formation in SMT." 664–674. Avignon, France.

Gojun, Anita and Alexander Fraser. 2012. "Determining the placement of German verbs in English-to-German SMT." 726–735. Avignon, France.

Hajič, Jan. 2004. *Disambiguation of Rich Inflection: Computational Morphology of Czech.* Karolinum Press.

Hajič, Jan and Barbora Vidová-Hladká. 1998. "Tagging inflective languages: Prediction of morphological categories for a rich, structured tagset." *Proceedings of the COLING - ACL Conference*, 483–490.

Koehn, Philipp and Hieu Hoang. 2007. "Factored Translation Models." *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 868–876. Prague, Czech Republic.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. "Moses: Open Source Toolkit for Statistical Machine Translation." *Proc. of the ACL 2007 Demo and Poster Sessions*, 177–180. Prague, Czech Republic.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. "Statistical Phrase-Based Translation." *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, 127–133. Edmonton, Canada.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: a Method for Automatic Evaluation of Machine Translation." *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, 311–318. Philadelphia, PA, USA.

Ramm, Anita and Alexander Fraser. 2016. “Modeling verbal inflection for english to german smt.” *Proceedings of First Conference on Machine Translation (WMT2016)*. Berlin, Germany. Peer-reviewed.

Straková, Jana, Milan Straka, and Jan Hajič. 2014. “Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition.” *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. Baltimore, Maryland.

Tamchyna, Aleš, Alexander Fraser, Ondřej Bojar, and Marcin Junczys-Dowmunt. 2016. “Target-Side Context for Discriminative Models in Statistical Machine Translation.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1704–1714. Berlin, Germany.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2015a. “Predicting prepositions for smt (extended abstract).” *Proceedings of the 9th Workshop on Syntax, Semantics and Structure in Statistical Translation at NAACL*. Denver, Colorado. Peer-reviewed.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2015b. “Target-side generation of prepositions for smt.” *Proceedings of EAMT 2015*. Antalya, Turkey. Peer-reviewed.

Weller-Di Marco, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2016. “Modeling complement types in phrase-based smt.” *Proceedings of the First Conference of Machine Translation (WMT2016)*. Berlin, Germany. Peer-reviewed.