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

D3.1: Initial report on inflection and word formation

Author(s): Ondřej Bojar, Alexander Fraser, Anita Ramm, Dušan Variš, Marion Weller

Dissemination Level: Public

Date: February 1st 2016

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 2016	
Actual date of delivery	Ferbruary 1st 2016	
Deliverable number	D3.1	
Deliverable title	Initial 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	Ondřej Bojar, Alexander Fraser, Anita Ramm, Dušan Variš, Marion Weller	
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.

As shown below, all 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	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	START M13 AS PLANNED

2 Task 3.3: Corrective approaches to morphology

In this section, CUNI presents the progress in the development of the automatic post-editing tool for statistical machine translation called MLFix.

MLFix is based on the very successful Depfix (Rosa, 2014) which uses handcoded rules to correct frequent errors (esp. agreement but also e.g. negation) in MT output. Depfix was developed specifically for English-to-Czech translation and in HimL, we promised to reduce this language dependence and use machine learning techniques to automatically identify rules for many language pairs.

2.1 Available data for learning correction rules

We will use supervised ML techniques in MLFix, so we need data collections where corrections can be learned from.

So far, we came across four different data sources: Khan's school human post-edits of manually translated (EN→CS) subtitles, the Autodesk triparallel data[1], log files of human post-editing done by Lingea for the HimL test set and the data from the QT21 project[2]. In the following subsections, we take a closer look at each of the datasets.

The sources above (each one to the different degree) can be considered a knowledge base for examining the behaviour of a human post editor as well as training data for our system. However, we believe that for training, we can also use any parallel data (*source sentence* + *reference translation*), translate the source sentences with a specific SMT system and acquire data (*source* + *SMT* + *reference*) that can be used to train MLFix for that specific SMT system[3]. This method should help to overcome the data acquisition bottleneck since there is generally much more parallel data then post-edited sentences. For Czech, the natural choice of the parallel corpus would be CzEng 1.0[4] Bojar *et al.* (2012). Even if we do not use this approach for the training of the final MLFix models, the volume of the data might help us with feature selection and preliminary model evaluation.

The basic summary of the available data is shown in Table 1: the number of parallel sentences and the number of tokens in the English source, the MT output ("MT") and the post-edited MT output ("PE"). Only English-Czech data is listed since these datasets for other target languages (where available) are similar in volume. For Khan's school, we only provide estimates. The CzEng dataset is not translated by any SMT at the moment, so the related information is omitted. We do not include information about the QT21 data since we are yet to explore them.

	# Sentences	# Tokens		
		English	Czech (MT)	Czech (PE)
Khan's school	NA	˜93k	˜93k	˜93k
Autodesk	46,916	490,005	456,697	441,645
HimL-Lingea	3892	60,142	51,428	56,485
CzEng 1.0	15M	206M	NA	150M

Table 1: Summary of the available post-editing data.

[1] https://autodesk.app.box.com/Autodesk-PostEditing

[2] http://www.qt21.eu/deliverables/annotations/

[3] Obviously, the MLFix training data have to be different from the parallel data used for the training of the SMT system, so some jackknife sampling should be used with limited training data.

[4] http://ufal.mff.cuni.cz/czeng

2.1.1 Khan's school

The data provided by Khan's school consist of en-cs subtitles, where the Czech part (usually manually translated from English) was manually edited. During the preliminary analysis, we've noticed that most of the time, the corrections were made mostly on a lexical level, so they might not be completely suitable for the task at hand. This is due to Czech subtitles not being created by SMT, but rather created by human translators.

2.1.2 Autodesk

Autodesk data consist of English sentences which were machine translated into a set of target languages (cs, de, pl etc.) complemented with human post-edits of MT output. Despite the different domain (IT vs. our medical), the data can provide a good basis for testing the machine learning methods we are going to implement.

2.1.3 HimL-Lingea logs

The data provided by Lingea were collected when official HimL test sets were created. As documented in D5.1, the original English sentences were first machine-translated to HimL languages and then post-edited by professional translators using Lingea's post-editing tool. The data are probably the most detailed ones since they consist of complete logfiles describing elementary actions taken by human post-editors (such as selecting phrases in a translated sentence, looking up alternative translations in a dictonary etc.).

When we examined the data more closely, we noticed that it is rather difficult to determine which actions are useful for our machine learning process. For the time being, we thus simply extract a triparallel data from these logs (the source sentence + SMT output + result of the human post-editing).

2.1.4 QT21 data

Another EU project, QT21 (and one of its predecessors, QTLaunchpad) produced also datasets relevant for our learning objective. We still need to examine the data in a close detail, but at the first glance, they seem like an interesting alternative to the datasets listed above.

In addition to the typical "machine translation"-"human post-edit" pair, the data also contain further annotation (provided probably by the same human annotator) indicating the type of the error produced by the MT engine and corrected in the post-edit. We hope that our further analysis of the data will give us useful hints for designing the automatic post-editing rules.

2.2 MLFix development progress

As of now, we have a working MLFix pipeline for English-Czech translation. The pipeline is derived from the existing Depfix Rosa (2014) pipeline. For now MLFix modifies only the grammatical gender, case and number of relevant words in the sentence, one at the time, or a combination of these categories at once depending on the model setup. Since the tool was derived from the original Depfix, the syntactic parsing of the machine translation output is required (or at least recommended). This means that in order to apply MLFix on a new language pair, we need a post-editing model, a morphological analyzer, a syntactic parser (fit for MT output) for the target language, and a word alignment between the source language and the SMT output.

We are not satisfied with the current level of language-independence of the tool, so we are currently trying to avoid some of the limitations, mainly:

1. the need of a target language parser (adapted for the often incorrect sentences due to MT errors),
2. the reliance on a fixed feature set for target languages (the Depfix implementation relies on Czech morphological tags, different morphological features and values are applicable for other languages),
3. the use of one specific classifier for predicting one or several morphological categories.

We want to avoid the need for a specific syntactic parser for each new target language (or even a new parser adapted to each SMT system the errors of which should be corrected), therefore we are removing the syntactic analysis of the SMT output from the pipeline. On the other hand, the information about the parent-child relations in the sentence is crucial for some of the post-editing tasks (e.g. agreement) Rosa (2013). We thus need to extract this information in some other way. At the moment, the parent of a target language node is determined in three steps: get the aligned node in the source language, get the source node's parent and then again through the alignment, get the node on the target side that is aligned to the parent on the source side. We need to evaluate the effect of using this heuristic in contrast to the full target-side parsing.

Since we would like to work with a unified feature set across different target languages, we decided to avoid using POS tags directly (which differ across languages) when extracting morphological features. We use the interlingua-based tagset Interset[5] Zeman (2008) instead. Interset gives us a uniform, language-independent interface to represent common morphological categories, as well as a way to encode-decode most common tagsets into this interface (the set of supported tagsets is still growing). The move to Interset is also in line with D3.2 Common morphology interface description.

Finally, instead of using a single classifier which would predict only one morphological category or a group of categories, we decided to train a separate classifier for each morphological category. Of course this raises some issues to tackle:

- Should we apply the classifiers in a specific order? And how should we decide on the order?
- Or should we apply all the classifiers simultaneously (each classifier would work with the morphological categories before they were modified) instead?

Experimental evidence to make any conclusions is yet to be collected.

On a side note, we also made some minor technical changes. MLFix is still part of the Treex processing Mareček *et al.* (2010) platform[6], however for model training and classification, we are going to use the scikit-learn toolkit[7] and use a Perl wrapper to call the Python scripts from MLFix. We decided for the toolkit because it provides implementation of multiple machine-learning methods with unified and easy-to-use API.

2.3 Feature extraction

Currently, we extract the features from the sentence aligned data containing the source language sentences (English), SMT output and reference sentences (correctly translated sentences). We do morphological analysis of each sentence and a syntactic analysis of the source sentences. Then we construct the word alignment between the source sentence and the SMT output (intersection of two directed alignments) and between the SMT output and the reference sentence (monolingual) using GIZA++. After this preprocessing, we extract features for each sentence token.

The set of features has a hierarchical structure. For each node, we extract information from its aligned source node, aligned source node's parent and the node's parent (if available). For training, we also extract aligned reference node information to get the possible predictions for the morphological categories.

From each one of these "main" nodes, we extract the information specific to them and to their neighbourhood (the parent, grandparent, preceding child, following child, preceding sibling, following sibling), mainly the number of preceeding and following children, direction of the edge coming from the parent and finally, its POS tag. As mentioned before, we chose to substitute the POS tag features by the Interset categories (we create one feature for each category). This leads to redundant features in some cases, however we leave the issue to the feature selection.

In the end, we extract about 2,400 initial (mostly categorical) features. The set was created mostly by intuition so we still need to run experiments, to decide, which ones will be really useful for our task. Due to the size of the feature set and the risk of data sparsity, some feature selection will be necessary.

2.4 Further MLFix development

At the moment, we are finishing the modification of the current MLFix pipeline. Most importantly, we still need to fix some bugs in the core MLFix modules. We reduced the number of required NLP tools (especially the ones that need to be modified to work with the SMT output) and we started experimenting with the scikit-learn toolkit.

We've defined a set of features to work with and are ready to start analysing how helpful they are for the post-editing task.

As far as the classification itself is concerned, we have some initial ideas. However, deeper insight into the problems that might arise is still needed. The next step is to prepare a complete experiment to contrast a simple baseline MLfix setup with the Depfix system.

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

In this section, LMU-MUENCHEN presents the progress on the improved inflectional and word formation prediction proposed in Task 3.4.

[5] http://ufal.mff.cuni.cz/interset
[6] http://ufal.mff.cuni.cz/treex
[7] http://scikit-learn.org/stable/

The LMU-MUENCHEN English to German pipeline which existed before the project began is primarily based on work reported by Fraser *et al.* (2012) and Gojun and Fraser (2012). In the first phase of the work reported on here we have focused on extending this system in three important ways: (1) improving modeling of subcategorization, (2) integrating prediction of verbal morphology into our nominal morphology prediction system, and (3) combining our inflectional systems with rule-based verbal reordering. In addition, we have begun the process of releasing our systems for use in HimL Year 2 systems.

The preliminary experiments reported in Section 3 use standard WMT data sets with standard Moses systems as baselines (except where explicitly indicated). In particular, the systems will be trained on the HimL data sets and tested on the HimL test sets as part of the preparations for the release of the HimL Year 2 systems.

3.1 Improving subcategorization through better modeling of prepositions

One important area of improvement in HimL is improved modeling of subcategorization. Prior to the start of the project we had worked extensively on grammatical case (Weller *et al.*, 2013). Since the beginning of the project, we have focused on extensions of this approach involving the difficult (and often erroneous) choice of prepositions. The work described in this section has already been published in paper form at EAMT Weller *et al.* (2015b) and in abstract form at SSST Weller *et al.* (2015a).

The translation of prepositions is a difficult task for machine translation; a preposition must convey the source-side meaning and also meet target-side constraints. In our approach, we move the selection of prepositions out of the translation system into a post-processing component. During translation, we use an abstract representation of prepositions as a place-holder that serves as a basis for the generation of prepositions in the post-processing step: all subcategorized elements of a verb are considered and allotted to their respective functions – as PPs with an overt preposition or as NPs with an "empty" preposition, e.g. *to call for sth.* → *∅ etw. erfordern*. The language model and the translation rules often fail to correctly model subcategorization in standard SMT systems because verbs and their subcategorized elements are often not adjacent.

input		lemmatized SMT output	prep	morph. feat.	inflected	gloss
∅	⟶	PREP	∅-Acc	–		
what		welch<PWAT>	Acc	Acc.Fem.Sg.Wk	welche	which
role		Rolle<+NN><Fem><Sg>	Acc	Acc.Fem.Sg.Wk	Rolle	role
∅	⟶	PREP	∅-Nom	–		
the		die<+ART><Def>	Nom	Nom.Masc.Sg.St	der	the
giant		riesig<ADJ>	Nom	Nom.Masc.Sg.Wk	riesige	giant
planet		Planet<+NN><Masc><Sg>	Nom	Nom.Masc.Sg.Wk	Planet	planet
has		gespielt<VVPP>	–	–	gespielt	played
played		hat<VAFIN>	–	–	hat	has
in	⟶	PREP	bei-Dat	–	bei	for
the		die<+ART><Def>	Dat	Dat.Fem.Sg.St	der	the
development		Entwicklung<+NN><Fem><Sg>	Dat	Dat.Fem.Sg.Wk	Entwicklung	development
of	⟶	PREP	∅-Gen	–		
the		die<+ART><Def>	Gen	Gen.Neut.Sg.St	des	of-the
solar system		Sonnensystem<+NN><Neut><Sg>	Gen	Gen.Neut.Sg.Wk	Sonnensystems	solar system

Figure 1: Overview of the morphology-aware translation system: prediction of prepositions, morphological features and generation of inflected forms. German cases: Acc-Accusative, Nom-Nominative, Dat-Dative, Gen-Genitive.

We use a morphology-aware SMT system which first translates into a lemmatized representation with a component to generate fully inflected forms in a second step, see Toutanova *et al.* (2008) and Fraser *et al.* (2012). The inflection step requires the modeling of the grammatical *case* of noun phrases, which corresponds to determining the syntactic function. Weller *et al.* (2013) describe modeling *case* in SMT; we extend their setup to cover the prediction of prepositions in both PP and NPs (i.e., the "empty" preposition). The presented work is similar to that of Agirre *et al.* (2009), but is applied to a fully statistical MT system. A detailed presentation of our work including a full literature survey can be found in Weller *et al.* (2015b).

Methodology To build the translation model, we use an abstract target-language representation in which nouns, adjectives and articles are lemmatized, and prepositions are substituted with place-holders. Additionally, "empty" place-holder prepositions are inserted at the beginning of noun phrases. To obtain a symmetric data structure, "empty" place-holders are also added to source-side NPs. When generating surface forms for the translation output, a phrase with a place-holder preposition can be realized as a noun phrase (empty preposition) or as a prepositional phrase by generating the preposition's surface form.

Figure 1 illustrates the process: for the English input with the extra null-prepositions (column 1), the SMT system outputs a lemmatized representation with place-holder prepositions (column 2). In a first step, prepositions and *case* for the SMT output are predicted (column 3). Then, the three remaining inflection-relevant morphological features *number*, *gender* and *strong/weak* are predicted on "regular" sentences without place-holders, given the prepositions from the previous step (column 4). In the last step, fully inflected forms are produced based on features and lemmas (column 5).

Abstract Representation and Prediction Features Initial experiments showed that replacing prepositions by simple place-holders decreases the translation quality. As an extension to the basic approach with plain place-holders, we thus experiment with

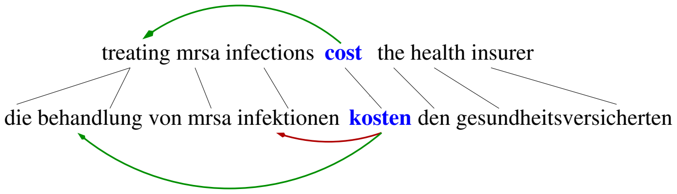
treating mrsa infections **cost** the health insurer
die behandlung von mrsa infektionen **kosten** den gesundheitsversicherten

Figure 2: An example clause pair with erroneous subject–verb agreement in the German translation.

enriching the place-holders such that they contain more relevant information and represent the content of a preposition while still being in an abstract form. For example, the representation can be enriched by annotating the place-holder with the grammatical case of the preposition it represents: for overt prepositions, case is often an indicator of the content (e.g. direction/location), whereas for NPs, case indicates the syntactic function. Other variants contain information of the governing verb/noun, and whether the represented preposition is functional.

For the prediction of prepositions, we combine the following feature types into a linear-chain CRF: *target-side context* (lemmas, POS-tags), *source-side context* (the aligned phrase), *projected source-side information* (relevant target-side words obtained based on source-side parses) and *target-side subcategorizational preferences* (distributional subcategorization information). These features address both functional and content-bearing prepositions, but do not require an explicit distinction between the two categories.

Experiments and Discussion We compare the approach of generating prepositions on the target-side with a morphology-aware SMT system with no special treatment for prepositions. When using "plain" place-holders, there is a considerable drop in BLEU (16.81) in comparison to the baseline (17.38). The annotation of *case* on the place-holders, the best of the abstract representation variants, leads to an improvement (17.23), but still does not surpass the baseline. Additionally, we assess the translation accuracy of prepositions. To allow for an automatic evaluation, we restrict the evaluation to cases where the relevant parts, namely the governing verb and the noun governed by the preposition, are the same in reference and MT output. While there is a minor improvement over the baseline, the difference is very small.

Our approach aims at assigning subcategorized elements to their respective functions and to inflect them accordingly which allows to handle structural differences in source and target language. While the systems fail to improve over the baseline, our experiments show that a meaningful representation of place-holders during translation is a key factor. In particular, the annotation of *case* helps, which can be considered as a "light" semantic annotation. This observation is also in line with English-to-Czech two-step experiments of Bojar and Kos (2010) where prepositions were annotated for case.

Later in the project, we will continue to investigate such semantic annotation, as well as investigating an approach to predicting prepositions that is integrated into the decoder.

3.2 Handling verbal inflection and integrating rule-based reordering

Translating English to German is difficult. English has poor morphology and German has complicated nominal declension and verbal conjugation. German verbs have to match their subject in person and number, and moreover, tense and mood are strongly reflected in German verbal morphology. In many cases, German SMT output cannot be understood correctly due to problems with verbal inflection. In addition, there is a considerable difference in the position of verbs in English and German which leads to German translations in which verbs are often missing or placed incorrectly.

We focus on getting the verbs right when translating from English to German. The key challenge is to correctly predict verbal morphology using contextual information to establish subject–verb agreement and, even more interesting, to model tense/mood translation from English to German.

3.2.1 Verbs in English–German SMT

Position and inflection problems Consider the clause pair given in Figure 2. The subject of the verb *costs* is *treating* which is a 3rd person singular. The German translation *kosten*, however, does not match the subject *behandlung*, but instead the object *infektionen* which is a 3rd person plural. The translation therefore has an erroneous subject–verb agreement which makes correct interpretation difficult.

Consider further the clause pair given in Figure 3. The English verb *confused* is in past tense, while the German translation *verwirrt* is in present tense. Such tense mismatches often lead to false interpretation of the translation. Moreover, *verwirrt* is wrongly placed (it should be placed at the clause end) which is a typical error.

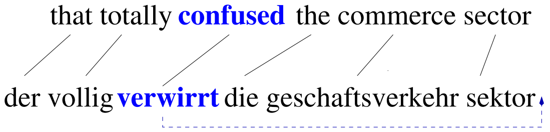
that totally confused the commerce sector
der vollig verwirrt die geschaftsverkehr sektor

Figure 3: An example clause pair with erroneous position and inflection of the finite verb in the German translation.

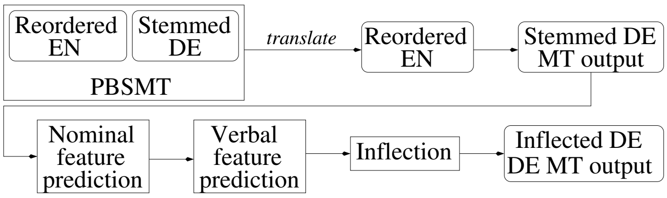

Figure 4: Processing pipeline.

Proposed problem solution We propose a method for generation of correctly inflected verbs which is realized as a post-processing step to phrase-based decoding. The processing pipeline of our system is sketched in Figure 4. To ensure that the MT output contains as many correctly placed German finite verbs as possible, we reimplemented the approach proposed by Gojun and Fraser (2012) which reorders English prior to training and translation. We combine reordering with nominal inflection of German proposed by Fraser *et al.* (2012) and add a novel component for verbal inflection. We follow Fraser *et al.* (2012) and train an SMT system on the reordered English and *stemmed* German text and apply it on the reordered English test set. The MT output consists of stemmed German words which are subsequently inflected.

Reordering and inflection We compare our approach to a baseline in which many non-finite verbs, which are the semantic heads of the verbal complexes, are missing. The reordering leads to the generation of considerably more non-finite verbs which is a major improvement (see Table 2). However, the inflection of the finite verbs is often erroneous. Subject–verb agreement suffers from a large distance between subjects and verbs, and tense and mood are often not properly translated. We improve on these problems in the next section.

3.2.2 Verbal morphology as classification

We train a classifier which predicts verbal morphological features for *finite* verbs in stemmed German MT output: person (1, 2, 3), number (singular, plural), tense (present, past) and mood (indicative, subjunctive).

Both the baseline system, as well as the reordering-nominal inflection system only have access to the information in the phrase pairs used, combined with the language model, to establish the subject–verb agreement. If the verb is far from the subject, it is quite likely that the subject does not match the verb. While person and number depend only on the subject in the given clause, the choice of tense and mood depends on many factors which we describe in more detail in Section 3.2.2.

Classifier training The classifier training samples are extracted from the English parse trees Charniak and Johnson (2005) and stemmed German sentences. The parallel sentences are processed clause-wise. We use automatically computed word alignments Och and Ney (2003) to find verb pairs, i.e. parallel clauses. The training data is tagged with `RFTagger` Schmid and Laws (2008) whose tags for the finite verbs (e.g. *1.Sg.Pres.Ind*: 1st person singular present indicative) are used as classification labels.

Features used for classification The feature set for prediction of person and number is simple while the features for tense and mood are more complex and thus grouped and presented by feature type. The features are given in Table 3.

	finite V	non-finite V	sum V
`Ref`	5165	2843	8008
`B`	4818	2589	7407
`R`	4766	3324	8090
`RN`	4591	3099	7690

Table 2: Number of verbs in MT and the reference. `B`: baseline, (`R`): reordered English, `RN`: `R` and inflection of stemmed German noun phrases

	Person/number	**English**	**German**
	subject person	+	+
	subject number	+	+
	clause type	+	–
	finite verb	+	+
	POS of the finite verb	+	+
	Tense/mood		
lexical	finite verb	+	+
	main verb	+	+
	previous sentence main verb	+	+
	participle	–	+
	infinitive	–	+
	conjunction	+	+
	negation	–	+
	2 words left/right to the finite verb	+	–
syntax	finite verb POS	+	+
	participle POS	–	+
	infinitive POS	–	+
	clause type	+	–
discourse	clause tense	+	–
	sentence tense	+	+
	sentence tense of the previous sentence	+	–

Table 3: Features for person/number and tense/mood prediction.

Number and person We identify the potential subject in the German clause and extract information about its person and number. This is done by searching for a nominative (pro)noun (case information comes from the nominal prediction step, cf. Figure 4). Since the subject may be missing or the nominal feature annotation may be erroneous, we perform the subject search also in the English clause and use its person and number values as additional features. We further consider the verbs in both languages, as well as their POS and the type of the given English clause.

Tense and mood Finite verbs in German can be in present or past. The tense of the German verbal complex depends on the predicted tense for the finite verb, as well as on the existence of other non-finite verbs. Suppose we had a German clause with the finite verb *haben/have* and the participle *arbeiten/work*; if the finite verb is in present, then the tense of the verbal complex is in present perfect (*(ich/I) habe/have gearbeitet/worked*), if the finite verb is in past, then we get the German tense past perfect ((ich/I) *hatte/had gearbeitet/worked*).

The use of mood and tense in German is rather flexible: there are no fixed rules for using specific tense (combinations) within a sentence. Nevertheless, there are some regularities and usage preferences which are reflected on the syntactical level. Since we want to reproduce English content in German, we also use clues from English to predict tense and mood in German.

Lexical features Following the theory of a *tense as anaphor* (cf. e.g. Lee (2011)), we add verbs from both English and German to the feature set. Lexical forms of the verbs are also helpful for the prediction of mood. For example, the verbs of speaking often take a subjunctive verb in the subordinate clause expressing indirect speech (cf. e.g. Fabricius-Hansen and Sæbø (2004)). To model the dependency between clauses, we use information about the main verb of the previous clause to predict the tense of the current clause. We extract conjunction from the current German clause since specific conjunctions such as *als (ob)* (meaning as (if)) give clues about mood (e.g. 'Als/*as-if* könnte$_{Subj}$/*could* er/*he* das/*that* wissen/*know*'). Temporal adverbials are also used as a feature. Finally, we use the context of the English finite verb.

Syntactic features German verbal complexes can be simple (containing only one verb) or composed (consisting of more than one verb). The structure of a composed verbal complex may have restrictions regarding the tense of the finite verb. We thus use the knowledge about the composition of the given German verbal complex in terms of POS tags. The clause type also plays a role for choosing a specific tense and mood in German. This information is derived from the English parse tree.

	Agreement		Tense/Mood		All	
B	0.86	0.31	0.83	0.30	0.69	0.25
R	0.85	0.50	0.81	0.47	0.66	0.39
RN	0.85	0.47	0.82	0.46	0.68	0.38
RNV	0.86	0.48	0.80	0.45	0.66	0.37

Table 4: Precision and recall of verbal features (alignment used, no lemma match required). RNV: RN with verb handling.

Classifier	Agreement	Tense/Mood	All
$\mathtt{C}_{me}$	87%	82%	69%
$\mathtt{C}_{seq}$	86%	82%	69%

Table 5: Accuracy of the verb features in clean data experiment (WMT newstest 2014). *seq*: sequential, *me* maximum entropy classifier.

Discourse features The tense in an English sentence and its German translation does not necessarily have to be the same. However, a past action in the source sentence should usually be translated as a past action in the target sentence. We thus use the tense information in English as a feature. The tense is derived for each clause, but also for the entire sentence. The *sentence tense* is the tense of the main clause in the English sentence. In order to account for the tendency of using the same tense (or tense group) within a text part (cf. Weinrich (2001)), we also use the information about the sentence tense of the previous English sentence.

3.2.3 Performance of the classifiers

The comparison of the verbal feature values of different MT outputs with the reference (see Table 4) showed that the reordering reduces agreement and tense/mood precision, however it considerably increases the recall. A more detailed evaluation of the predicted verbal features (cf. RNV, Table 4) showed that the prediction of person and mood seems to be more reliable than prediction of tense and number. Most of the subjects in our training and testing data are 3rd person, so it is not surprising that the classifier performs well on person prediction. It however has difficulties to decide whether the number should be singular or plural. Similarly to person and number, the decision on mood seems to be easier than on tense: most verbs in the training and testing data are indicative. The prediction of tense is least accurate showing the difficulty of defining features for better tense prediction.

We assume that the verbal features have influence on each other. Furthermore, we investigate whether the verb feature prediction should be viewed as a sequential prediction problem. In preliminary experiments, adding English extra-sentential context to the feature set led to improvements of the classifier performance. Moreover, the access to the tense and mood sequence within the German sentence seems to help the decision on these two verbal features. We trained a sequential and a maximum entropy classifier on the same feature set and the same training data. Both classifiers perform equally well on clean data, see Table 5. The agreement (person/number) accuracy is 86-87%, while the accuracy of tense/mood predictions is 82%. Overall, for about 69% of the verbs all predicted morphological feature values are correct.

According to the evaluations in Tables 4 and 5, the noisy MT output seems to hurt the prediction of tense and mood, while the prediction of person and number works equally well for both noisy MT output and well-formed clean data.

3.2.4 Verbal morphology prediction for MT output

We trained a standard English to German PBSMT system using a subset of the WMT 2014 data and tested it on the WMT 2014 test set. The stemmed MT output was post-processed as shown in Figure 4. The resulting fully-inflected German translations are evaluated with BLEU in Table 6.

The reordering combined with nominal inflection clearly leads to an improved translation quality. However, the verbal inflection

	B	RN	$\mathtt{RNV}_{seq}$	$\mathtt{RNV}_{me}$
BLEU_{ci}	16.24	16.98	16.86	16.87

Table 6: Case-insensitive BLEU scores of MT outputs.

prediction does not seem to have an impact on translation quality. Both RN and RNV have errors concerning verbal inflection. Agreement errors are usually made when the subject is a coordination or the sentence structure is complex. True tense and mood errors can be difficult to identify, as in some cases a different tense and/or mood can be used without having a negative impact on the translation.

3.2.5 Discussion

Predicting verbal morphological features for SMT is a challenging task. Especially modeling of tense and mood is known to be difficult (e.g. Gispert and Mariño (2008). For languages which have strict rules for using tense, it seems to be easier to model tense as shown by Meyer *et al.* (2013) in their work on translating English into French. In German, there are no strict rules for using a specific tense or mood. In the linguistic literature, the usage is often described as register-dependent and highly dependent on the speaker or text author. The use of mood, particularly of the subjunctive (indirect speech, *unreal world* actions), is even more complicated since it requires access to discourse knowledge.

We predict tense and mood for finite verbs assuming that non-finite verbs are already part of the MT output (and correct). In the future, we will possibly predict the tense of the entire clause, i.e. one of the six German tenses, instead of predicting *present* or *past* for the finite verbs. This approach has already been successfully applied for Chinese to English SMT Gong *et al.* (2012) where a target language *tense model* interacts with the decoding process. Applying this idea to German will lead us to an interesting new problem of abstracting over the often discontiguous German verbal complex.

3.3 Progress towards releasing LMU-MUENCHEN pipeline

There are four important components in the LMU-MUENCHEN pipeline relevant to a release. The first is the preparation of training data for the Moses system, which conceptually translates from English words to German stems (with some additional markup). This will be carried out in-house at LMU-MUENCHEN for the Year 2 system.

The second component is preparation of the linear-chain CRF-based inflection prediction system, and in particular the training of the CRF system. This will also be carried out in-house at LMU-MUENCHEN.

Given translation output of unseen (e.g., test) English from Moses (consisting of German stems and some additional markup), the next component must predict the full representation (e.g., nominal and verbal morphology). This component will be released, and consists of trained models for the Wapiti linear-chain CRF classifier and various input and output scripts.

Finally, given a sequence of stems and together with morphologically rich POS representations, the final component produces the final surface form. This component depends on SMOR, a finite state morphological analyzer. SMOR will be applied to all stems observed in the parallel traning data used to train Moses, and a flat file will be created containing all possible surface forms generable from these stems. This flat file and associated scripts will be part of the surface form generation component, which will be released.

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

Task 3.5 is an interesting area of collaboration between CUNI and LMU-MUENCHEN. The goal here is to achieve the same sort of generalization in inflection prediction as in the LMU-MUENCHEN system, but in an appropriate way for Czech, which has a much less predictable word order than German.

We previously implemented a discriminative model of phrasal translation with a rich set of features. The Vowpal Wabbit (VW) classifier[8] is fully integrated in the Moses decoder, allowing for efficient training and decoding.

In the HimL project, an advanced goal for this framework is to use this model for predicting target-side morphological features, for instance in a two-step scenario. In order to fully solve this problem we will need to utilize target-side context information. But at this early stage in the task (Task 3.5 started in M7), we have initially experimented with a simpler setting: using source-context information to disambiguate surface phrasal translations. After we discuss this work, we will briefly present our plans for the currently ongoing work which utilizes target-side features.

4.1 Source Side Context: Experiments

We run our experiments primarily between English and Czech. As parallel training data, we use (subsets of) the CzEng 1.0 corpus (Bojar *et al.*, 2012). For tuning, we use the WMT13 test set (Bojar *et al.*, 2013) and we evaluate the systems on the WMT14 test set (Bojar *et al.*, 2014). Our baseline system is a standard phrase-based Moses setup. The phrase table in both

[8] http://hunch.net/~vw/

data size (k sents)	100	300	500	700	900	1000	2000	3000	4000	5000
baseline	11.5	13.3	14.0	14.6	14.8	14.8	15.4	**15.8**	16.0	16.1
+VW	**11.7**	**13.6**	**14.4**	**15.0**	**15.3**	**15.0**	**16.1**	15.5	**16.2**	**16.7**

Table 7: BLEU scores of Moses with and without the discriminative classifier.

cases is factored and outputs also lemmas and morphological tags. The system uses a 5-gram language model trained on the full CzEng corpus.

We train both VW and the standard phrase table on the same dataset. We use leave-one-out in VW training to avoid overfitting (VW might otherwise learn to simply trust very long phrase pairs which were extracted from the same training sentence). Our source-side features use surface forms, lemmas, morphological attributes and syntactic information from dependency trees. On the target side, we use surface forms, lemmas and POS tags within the target phrase (wider target-side context is not used at the moment). Initial experiments were not promising but through careful feature engineering and setting of the classifier (hyper)parameters, we obtain a small but significant improvement of the BLEU score which scales well with the training data size.

Table 7 shows the obtained results. Note that scores for sizes lower than 1 million are the average of five independent runs of MERT (tuning of translation system weights), both for the baseline and the improved system. The rest of the scores were obtained by running MERT only once so they are less reliable.

4.2 Target-Side Context

One of our goals is to use also target-side information in the classifier. While the full source sentence contains a lot of information useful for both lexical and morphological disambiguation, having access to the partial translation should allow the classifier to promote more coherent translation outputs.

Similarly to a language model, VW will have access to several preceding target-side tokens. This presents an engineering challenge since the use of target-side information makes the decoding process much more computationally intensive. We will therefore restrict our initial experiments to an n-best re-ranking setting.

5 Conclusion

This initial report on inflection and word formation documents that all work in HimL WP3 is proceeding according to the plan. HimL provides an interesting environment for innovation of the research approaches studied, and their application to the translation of consumer health information.

References

Agirre, Eneko, Aitziber Atutxa, Gorka Labaka, Mikel Lersundi, Aingeru Mayor, and Kepa Sarasola. 2009. “Use of Rich Linguistic Information to Translate Prepositions and Grammatical Cases to Basque.” *EAMT*.

Bojar, Ondřej, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. 2014. “Findings of the 2014 workshop on statistical machine translation.” *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 12–58. Baltimore, MD, USA.

Bojar, Ondřej and Kamil Kos. 2010. “2010 Failures in English-Czech Phrase-Based MT.” *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, 60–66. Uppsala, Sweden.

Bojar, Ondřej, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel, and Aleš Tamchyna. 2012. “The Joy of Parallelism with CzEng 1.0.” *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC’12)*, 3921–3928. Istanbul, Turkey.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. “Findings of the 2013 Workshop on Statistical Machine Translation.” *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 1–44. Sofia, Bulgaria.

Charniak, Eugene and Mark Johnson. 2005. “Coarse-to-fine n-best parsing and MaxEnt discriminative reranking.” *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*. Ann Arbor, Michigan.

Fabricius-Hansen, Cathrine and Kjell Johan Sæbø. 2004. "In a mediative mood: the semantics of the german reportive subjunctive." *Natural Language Semantics*, 12:213–257.

Fraser, Alexander, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. "Modeling Inflection and Word-Formation in SMT." *EACL*.

Gispert, Adrià de and Jose B. Mariño. 2008. "On the impact of morphology in English to Spanish statistical MT." *Speech Communication*, 50(11-12):1034–1046.

Gojun, Anita and Alexander Fraser. 2012. "Determining the placement of German verbs in English-to-German SMT." *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 726–735. Avignon, France.

Gong, Zhengxian, Min Zhang, Chewlim Tan, and Guodong Zhou. 2012. "N-gram-based tense models for statistical machine translation." *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 276–285. Jeju Island, Korea.

Lee, John. 2011. "Verb tense generation." *Procedia. Social and Behavioral Sciences*. Gombak, Malaysia.

Mareček, David, Martin Popel, and Zdeněk Žabokrtský. 2010. "Maximum Entropy Translation Model in Dependency-Based MT Framework." *Proc. of WMT and MetricsMATR*, 207–212. Uppsala, Sweden.

Meyer, Thomas, Cristina Grisot, and Andrei Popescu-Belis. 2013. "Detecting narrativity to improve English to French translation of simple past verbs." *Proceedings of the 1st DiscoMT Workshop at 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Och, Franz Josef and Hermann Ney. 2003. "A systematic comparison of various statistical alignment models." *Computational Linguistics*, 29(1).

Rosa, Rudolf. 2013. *Automatic post-editing of phrase-based machine translation outputs*. Master's thesis, Charles University in Prague, Faculty of Mathematics and Physics, Praha, Czechia.

Rosa, Rudolf. 2014. "Depfix, a tool for automatic rule-based post-editing of SMT." *The Prague Bulletin of Mathematical Linguistics*, 102:47–56.

Schmid, Helmut and Florian Laws. 2008. "Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging." *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*. Machester, UK.

Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp. 2008. "Applying Morphology Generation Models to Machine Translation." *ACL*.

Weinrich, Harald. 2001. *Tempus. Besprochene und erzählte Welt*. C.H.Beck, 6 ed.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2013. "Using Subcategorization Knowledge to Improve Case Prediction for Translation to German." *ACL*.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2015a. "Predicting prepositions for SMT." *Proceedings of the Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 55–56. Denver, Colorado, USA.

Weller, Marion, Alexander Fraser, and Sabine Schulte im Walde. 2015b. "Target-side generation of prepositions for SMT." *Proceedings of EAMT*, 177–184. Antalya, Turkey.

Zeman, Daniel. 2008. "Reusable tagset conversion using tagset drivers." *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, 213–218. Marrakech, Morocco.