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

D3.4: Final report on inflection and word formation

Author(s):	Alex Fraser, Matthias Huck, David Mareček, Anita Ramm, Dušan Variš
Dissemination Level:	Public
Date:	January, 31st 2018

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	January, 31st 2018	
Actual date of delivery	January, 31st 2018	
Deliverable number	D3.4	
Deliverable title	Final report on inflection and word formation	
Type	Report	
Status and version	1.0	
Number of pages	22	
Contributing partners	LMU, CUNI	
WP leader	LMU	
Task leader	LMU	
Authors	Alex Fraser, Matthias Huck, David Mareček, Anita Ramm, Dušan Variš	
EC project officer	Tünde Turbucz	
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 Introduction 4

2 Target-side Word Segmentation Strategies for Neural Machine Translation 4

2.1 Word Segmentation Strategies . . . 5

2.1.1 Byte Pair Encoding . . . 5

2.1.2 Compound Splitting . . . 5

2.1.3 Suffix Splitting . . . 6

2.1.4 Prefix Splitting . . . 6

2.1.5 Cascaded Application of Segmenters . . . 6

2.1.6 Reversibility . . . 7

2.2 Machine Translation Experiments . . . 7

2.2.1 Experimental Setup . . . 7

2.2.2 Experimental Results . . . 8

2.3 Conclusion . . . 8

3 Modeling target-side inflection in NMT 8

3.1 Two-step procedure for NMT . . . 9

3.1.1 Baseline NMT models and training settings . . . 9

3.1.2 English→Czech . . . 10

3.1.3 English→German . . . 10

3.1.4 Discussion . . . 11

3.2 Further development and experiments of the two-step approach for English→German . . . 12

3.2.1 Replacing parsing with tagging . . . 12

3.2.2 Performance on a large medical domain data set . . . 13

3.3 Further development and experiments of the two-step approach for English → Czech . . . 15

3.4 Further development and experiments of the two-step approach for English → Polish and English → Romanian 15

3.5 Conclusion . . . 16

4 Placement and inflection of verbs in the German NMT outputs 16

4.1 Evaluation data . . . 16

4.2 Position of the German verbs . . . 16

4.3 Inflection of the German verbs . . . 17

4.4 Conclusion . . . 19

5 MLFix 19

5.1 Postediting of the NMT outputs . . . 20

6 Conclusion 20

1 Introduction

The aim of this report is to present the work done on morphology in the third year of the project.

All tasks in Workpackage 3 were completed as planned. The switch from SMT to NMT was not as disruptive for year 3 of this workpackage as it was for other work packages, because we could implement simpler systems addressing the same morphological formula (e.g., consider the NMT tag stems representation versus SMT two-step). The deliverable first discusses work on target-side word segmentation strategies for German. This work provides a superior linguistic segmentation versus byte-pair-encoding, and was used in the final Year 3 system. In the next section, we present work on English to Czech and English to German inflection, which is the work that most closely follows work in Year 1 and Year 2 on two-step for SMT, with the difference being here that stems and tags are generated together using NMT; this was not deployed in Year 3 as it is more complex than linguistic segmentation and does not result in a large performance gain. This tag-stem approach was also tried for Polish and Romanian, where it also did not produce improvements, see Section 4. Following this, in Section 5, we present a negative result for using techniques we developed in earlier years for modeling German verbs; it seems that this is not critical for NMT. Finally Section 6 presents the MLFix work, which seemed promising but did not result in improvements on NMT outputs.

Overall, the work on inflection and word formation led to many new insights in both SMT (earlier in the project) and NMT (later in the project). The final Year 3 systems used linguistic segmentation of German instead of BPE, but no linguistic generalizations for Czech, Polish and Romanian, as the new NMT systems for these languages were better than proposed improved systems, as shown in this deliverable.

2 Target-side Word Segmentation Strategies for Neural Machine Translation

Inflection and nominal composition are morphological processes which exist in many natural languages. Machine translation into an inflected language or into a compounding language must be capable of generating words from a large vocabulary of valid word surface forms, or ideally even be open-vocabulary. In NMT, though, dealing with a very large number of target symbols is expensive in practice.

While, for instance, a standard dictionary of German, a compounding language, may cover 140 000 vocabulary entries,[1] NMT on off-the-shelf GPU hardware is nowadays typically only tractable with target vocabularies below 100 000 symbols.

This issue is made worse by the fact that compound words are not a closed set. More frequently occurring compound words may be covered in a standard dictionary (e.g., “Finanztransaktionssteuer”, English: “financial transaction tax”), but the compounding process allows for words to be freely joined to form new ones (e.g., “Finanztransaktionssteuerzahler”, English: “financial transaction tax payer”), and compounding is highly productive in a language like German.

Furthermore, a dictionary lists canonical word forms, many of which can have many inflected variants, with morphological variation depending on case, number, gender, tense, aspect, mood, and so on. The German language has four cases, three grammatical genders, and two numbers. German exhibits a rich amount of morphological word variations also in the verbal system. A machine translation system should ideally be able to produce any permissible compound word, and all inflections for each canonical form of all words (including compound words).

Previous work has drawn on byte pair encoding to obtain a fixed-sized vocabulary of subword units. We investigate word segmentation strategies for NMT which are linguistically more informed. Specifically, we explore and empirically compare:

- Compound splitting.
- Suffix splitting.
- Prefix splitting.
- Byte pair encoding (BPE).
- Cascaded applications of the above.

Our empirical evaluation focuses on target-language side segmentation, with English→German translation as the application task. Our proposed approaches improve machine translation quality by up to +0.5 Bleu and −0.9 Ter, respectively, compared with using plain BPE.

Advantages of linguistically-informed target word segmentation in NMT are:

1. *Better vocabulary reduction* for practical tractability of NMT, as motivated above.

[1] Duden, 26th ed., 2013, cf. http://www.duden.de/ueber_duden/auflagengeschichte.

suffixes
-e, -em, -en, -end, -enheit, -enlich, -er, -erheit, -erlich, -ern, -es, -est, -heit, -ig, -igend, -igkeit, -igung, -ik, -isch, -keit, -lich, -lichkeit, -s, -se, -sen, -ses, -st, -ung
prefixes
ab-, an-, anti-, auf-, aus-, auseinander-, außer-, be-, bei-, binnen-, bitter-, blut-, brand-, dar-, des-, dis-, durch-, ein-, empor-, endo-, ent-, entgegen-, entlang-, entzwei-, epi-, er-, extra-, fehl-, fern-, fest-, fort-, frei-, für-, ge-, gegen-, gegenüber-, grund-, heim-, her-, hetero-, hin-, hinter-, hinterher-, hoch-, homo-, homöo-, hyper-, hypo-, inter-, intra-, iso-, kreuz-, los-, miss-, mit-, mono-, multi-, nach-, neben-, nieder-, non-, pan-, para-, peri-, poly-, post-, pro-, prä-, pseudo-, quasi-, schein-, semi-, stock-, sub-, super-, supra-, tief-, tod-, trans-, ultra-, um-, un-, unab-, unan-, unauf-, unaus-, unbe-, unbei-, undar-, undis-, undurch-, unein-, unent-, uner-, unfehl-, unfort-, unfrei-, unge-, unher-, unhin-, unhinter-, unhoch-, unmiss-, unmit-, unnach-, unter-, untief-, unum-, ununter-, unver-, unvor-, unweg-, unwider-, unzer-, unzu-, unüber-, ur-, ver-, voll-, vor-, voran-, voraus-, vorüber-, weg-, weiter-, wider-, wieder-, zer-, zu-, zurecht-, zurück-, zusammen-, zuwider-, über-

Table 1: German affixes which our suffix splitter and prefix splitter separate from the word stem.

2. *Reduction of data sparsity*. Learning lexical choice is more difficult for rare words that appear in few training samples (e.g., rare compounds), or when a single form from a source language with little inflection (such as English) has many target-side translation options which are morphological variants. Splitting compounds and separating affixes from stems can ease lexical selection.

3. *Better open vocabulary translation*. With target-side word segmentation, the NMT system can generate sequences of word pieces at test time that have not been seen in this combination in training. It may produce new compounds, or valid morphological variants that were not present in the training corpus, e.g. by piecing together a stem with an inflectional suffix in a new, but linguistically admissible way. Using a linguistically informed segmentation should better allow the system to try to learn the linguistic processes of word formation.

2.1 Word Segmentation Strategies

2.1.1 Byte Pair Encoding

A technique in the manner of the Byte Pair Encoding (BPE) compression algorithm (Gage, 1994) can be adopted in order to segment words into smaller subword units, as suggested by Sennrich *et al.* (2016b). The BPE word segmenter conceptionally proceeds by first splitting all words in the whole corpus into individual characters. The most frequent adjacent pairs of symbols are then consecutively merged, until a specified limit of merge operations has been reached. Merge operations are not applied across word boundaries. The merge operations learned on a training corpus can be stored and applied to other data, such as test sets.

An advantage of BPE word segmentation is that it allows for a reduction of the amount of distinct symbols to a desired order of magnitude. The technique is purely frequency-based. Frequent sequences of characters will be joined through the merge operations, resulting in common words not being segmented. Words containing rare combinations of characters will not be fully merged from the character splitting all the way back to their original form. They will remain split into two or more subword units in the BPE-segmented data. On the downside, the BPE algorithm has no notion of morphosyntax, narrowing down its capabilities at modeling inflection and compounding. BPE also has no guidelines for splitting words into syllables. This way no phonetic or semantic substructures are taken into account. Therefore BPE splits often appear arbitrary to the human reader, since it appears frequently that subword units ignore syllable boundaries entirely.

Nevertheless, NMT systems incorporating BPE word segmentation have achieved top translation quality in recent shared tasks (Sennrich *et al.*, 2016a; Bojar *et al.*, 2016). We designed our linguistically-informed segmentation techniques by looking at the shortcomings of BPE segmentations.

2.1.2 Compound Splitting

BPE word segmentation operates bottom-up from characters to larger units. Koehn and Knight (2003) have proposed a frequency-based word segmentation method that starts from the other end, top-down inspecting full words and looking into whether they are composed of parts which are proper words themselves. Any composed word is segmented into parts such that the geometric mean of word frequencies of its parts (counted in the original corpus) is maximized. This technique represents a suitable approach for compound splitting in natural language processing applications. It has been successfully applied in numerous statistical machine translation systems, mostly on the source language side, but sometimes also on the target side (Sennrich *et al.*, 2015).

The difference in nature between BPE word segmentation and frequency-based compound splitting (bottom-up and top-down) leads to quite different results. While BPE tends to generate unintuitive splits, compound splitting nearly always comes up with reasonable word splits. On the other hand there are many possible intuitive word splits that compound splitting does not catch.

2.1.3 Suffix Splitting

Morphological variation in natural languages is often realized to a large extent through affixation. In the German language there are several suffixes that unambiguously mark a word as an adjective, noun, or verb. By splitting these telling suffixes, we can automatically include syntactic information. Even though this underlying relationship between suffix and morphological function is sometimes ambiguous—especially for verbs—reasonable guesses about the POS of a word with which we are not familiar are only possible by considering its suffix.

Information retrieval systems take advantage of this observation and reduce search queries to stemmed forms by means of simply removing common suffixes, prefixes, or both. The Porter stemming algorithm is a well-known affix stripping method (Porter, 1980). In such algorithms, some basic linguistic knowledge about the morphology of a particular language is taken into account in order to come up with a few hand-written rules which would detect common affixes and delete them. We can benefit from the same idea for the segmentation of word surface forms.

We have modified the Python implementation of the German Snowball stemming algorithm from NLTK[2] for our purposes. The Snowball stemmer removes German suffixes via some language-specific heuristics. In order to obtain a segmenter, we have altered the code to not drop suffixes, but to write them out separately from the stem. Our Snowball segmenter splits off the German suffixes that are shown in Table 1. Some of them are inflectional, others are used for nominalization or adjectivization. The suffix segmenter also splits sequential appearances of suffixes into multiple parts according to the Snowball algorithm's splitting steps, but always retaining a stem with a minimum length of at least three characters.

2.1.4 Prefix Splitting

Similarly to our Snowball suffix segmenter, we have written a small script to split off prefixes. Here, we specifically target verb and adjective prefixes and thus only segment lowercase words, excluding nouns which are written in uppercase in German text. We consider the prefixes as shown in Table 1. We sort them descending by length, checking for longer prefix matches first. Negational prefixes (beginning with *un-*, but not *unter-*) are additionally segmented after *un-*; e.g., *unab-* becomes *un- ab-*. In case the remaining part starts with either of the two verb infixes *-zu-* or *-ge-*, we also segment after that infix. We require the final stem to be at least three characters long.

2.1.5 Cascaded Application of Segmenters

Affix splitting and compound splitting can be applied in combination, by cascading the segmenters and preprocessing the data first with the suffix splitter, then optionally with the prefix splitter, and then with the compound splitter. In a cascaded application, the compound splitter is applied to word stems only, and the counts for computing the geometric means of word frequencies for compound splitting are collected after affix splitting.

When cascading the compound splitter with affix splitting, we introduce a minor modification. Our standalone compound splitter takes the filler letter "*s*" and "*es*" into account, which often appear in between word parts in German noun compounding. For better consistency of the compound splitting component with affix splitting, we additionally allow for more fillers, namely: suffixes, suffixes followed by "*s*", and "*zu*".

The methods for compound splitting, suffix splitting, and prefix splitting provide linguistically more sound approaches for word segmentation, but they do not arbitrarily reduce the amount of distinct symbols. For a further reduction of the number of target-side symbols, we may want to apply a final BPE segmentation step on top of the other segmenters. BPE will not re-merge words that have been segmented before. It can benefit from the prior segmentation provided to it and come up with a potentially better sequence of merge operations. Affixes will be learned as subwords but not joined with the stem. This improves the quality of resulting BPE splits. BPE no longer combines arbitrary second to last syllables with their suffixes, which makes learning the other—non affix—syllables easier.

[2] http://www.nltk.org/_modules/nltk/stem/snowball.html

BPE	sie alle versch ## icken vorsätzlich irreführende Dokumente an Kleinunternehmen in ganz Europa .
compound + BPE	sie alle verschicken vorsätzlich #L irre @@ führende Dokumente an #U klein @@ unter @@ nehmen in ganz Europa .
suffix + BPE	sie all $$e verschick $$en vorsätz $$lich irreführ $$end $$e Dokument $$e an Kleinunternehm $$en in ganz Europa .
suffix + compound + BPE	sie all $$e verschick $$en vorsätz $$lich #L Irre @@ führ $$end $$e Dokument $$e an #U klein @@ Unternehm $$en in ganz Europa .
suffix + prefix + compound + BPE	sie all $$e ver§§ schick $$en vor§§ sätz $$lich #L Irre @@ führ $$end $$e Dokument $$e an #U klein @@ Unternehm $$en in ganz Europa .
English	they all mail deliberately deceptive documents to small businesses across Europe .

Table 2: Different word segmentation strategies applied to a training sentence. ## is a BPE split-point, ver§§ is prefix *ver*, $$en is the suffix *en*, #U and #L are upper and lower case indicators for compounds, @@ indicates a compound merge-point, @s@ would indicate a compound merged with the letter *s* between the parts, etc.

2.1.6 Reversibility

Target-side word segmentation needs to be reversible in postprocessing. We introduce special markers to enable reversibility of word splits. For suffixes, we attach a marker to the beginning of each suffix token; for prefixes to the end of each split prefix.

Fillers within segmented compounds receive attached markers on either side. When a compound is segmented into parts with no filler between them, we place a separate special marker token in the middle which is not attached to any of the parts. It indicates the segmentation and has two advantages over attaching it to any of the parts: (1.) The tokens of the parts are exactly the same as when they appear as words outside of a compound. The NMT system does not perceive them as different symbols. (2.) There is more flexibility at producing new compounds that have not been seen in the training corpus. The NMT system can decide to place any symbol into a token sequence that would form a compound, even the ones which were never part of a compound in training. The vocabulary is more open in that respect.

We adhere to the same rationale for split markers in BPE word segmentation. A special marker token is placed separately between subword units, with whitespace around it. In our experience, attaching the marker to BPE subword units does not improve translation quality over our practice.

The compound splitter alters the casing of compound parts to the variants that appears most frequently in the corpus. When merging compounds in postprocessing, we need to know whether to lowercase or to uppercase the compound. We let the translation system decide and introduce another special annotation in order to allow for this. When we segment compounds, we always place an indicator symbol before the initial part of the split compound token sequence, which can be either *#L* or *#U*. It specifies the original casing of the compound (lower or upper).

The effect of different segmentation strategies on the word splits in an example sentence is shown in Table 2.

2.2 Machine Translation Experiments

2.2.1 Experimental Setup

We conduct an empirical evaluation using encoder-decoder NMT with attention and gated recurrent units as implemented in Nematus (Sennrich *et al.*, 2017b). We train and test on English–German Europarl data (Koehn, 2005). The data is tokenized and frequent-cased using scripts from the Moses toolkit (Koehn *et al.*, 2007). Sentences with length >50 after tokenization are excluded from the training corpus, all other sentences (1.7 M) are considered in training under every word segmentation scheme. We set the amount of merge operations for BPE to 50K. Corpus statistics of the German data after different preprocessings are given in Table 3. On the English source side, we apply BPE separately, also with 50K merge operations.

For comparison, we build a setup denoted as *top 50K voc. (source & target)* where we train on the tokenized corpus without any segmentation, limiting the vocabulary to the 50K most frequent words on each side and replacing rare words by "*UNK*". In a setup denoted as *suffix + prefix + compound, 50K*, we furthermore examine whether BPE can be omitted in a cascaded

Preprocessing	#types	#tokens
tokenized	303 K	39 M
compound	139 K	45 M
suffix	217 K	54 M
suffix + compound	98 K	60 M
suffix + prefix + compound	88 K	63 M
BPE	46 K	42 M
compound + BPE	46 K	46 M
suffix + BPE	45 K	56 M
suffix + compound + BPE	43 K	60 M
suffix + prefix + compound + BPE	43 K	64 M

Table 3: Target-side training corpus statistics for Europarl English→German translation experiments.

System	test2007		test2008	
	Bleu	Ter	Bleu	Ter
top 50K voc. (source & target)	25.5	60.9	25.2	60.9
BPE	25.8	60.7	25.6	60.9
compound + BPE	25.9	60.3	25.5	60.6
suffix + BPE	**26.3**	60.0	**26.0**	**60.1**
suffix + compound + BPE	26.2	**59.8**	25.8	60.2
suffix + prefix + compound + BPE	26.1	**59.8**	25.9	60.6
suffix + prefix + compound, 50K	25.9	59.9	25.5	60.3
phrase-based (Huck *et al.*, 2015)	22.6	–	22.1	–

Table 4: English→German experimental results on Europarl (case-sensitive Bleu and Ter).

application of target word segmenters. Here, we use the top 50K target symbols after suffix, prefix, and compound splitting, but still apply BPE to the English source.

We configure dimensions of 500 for the embeddings and 1024 for the hidden layer. We train with the Adam optimizer (Kingma and Ba, 2015), a learning rate of 0.0001, batch size of 50, and dropout with probability 0.2 applied to the hidden layer. We validate on the *test2006* set after every 10 000 updates and do early stopping when the validation cost has not decreased for ten epochs.

We evaluate case-sensitive with Bleu (Papineni *et al.*, 2002) and Ter (Snover *et al.*, 2006), computed over postprocessed hypotheses against the raw references with `mteval-v13a` and `tercom.7.25`, respectively.

2.2.2 Experimental Results

The translation results are reported in Table 4. Cascading compound splitting and BPE slightly improves translation quality as measured in Ter. Cascading suffix splitting with BPE or with compound splitting plus BPE considerably improves translation quality by up to +0.5 Bleu or −0.9 Ter over pure BPE. Adding in prefix splitting is less effective. We conjecture that prefix splitting does not help because German verb prefixes often radically modify the meaning. When prefixes are split off, the decoder's embeddings layer may therefore become less effective (as the stem may be confusable with a completely different word).

In order to better understand the impact of the different target-side segmentation strategies, we have analyzed and compared the output of our main setups. We omit the details for the sake of brevity of this report, but refer the interested reader to our published research paper (Huck *et al.*, 2017b).

2.3 Conclusion

Linguistically motivated target-side word segmentation improves neural machine translation into an inflected and compounding language. The system can learn linguistic word formation processes from the segmented data. For German, we have shown that cascading of suffix splitting—or suffix splitting and compound splitting—with BPE yields the best results.

Our cascaded *suffix + compound + BPE* target word segmentation strategy was employed for LMU Munich's participation in the WMT17 shared tasks on machine translation of news and of biomedical texts (Huck *et al.*, 2017a). The system was ranked first in the human evaluation of the WMT17 news translation task (Bojar *et al.*, 2017). Our linguistically motivated target-side word segmentation is also used in the deployed English→German Year 3 HimL translation engine (cf. HimL deliverable D4.3/6).

3 Modeling target-side inflection in NMT

In year 3, the successful two-step procedure to morphology generation for SMT has been adapted to the NMT for the English→ Czech and English→German translation directions (Tamchyna *et al.*, 2017). Although the quality of the NMT outputs is considerably better than the quality of the SMT outputs, there are three main problems associated with rich target-side morphology in NMT: (i) NMT systems have no explicit connection between different surface forms of a single target-side lexeme (lemma), leading to data sparsity, (ii) there is no explicit information about morphological features of target-side words (e.g., whether a subject and verb agree), and (iii) NMT systems can-not systematically generate unseen surface forms of known lemmas.

Token	Lemma	Tag + morphology
Existují	existovat	VB-P—3P-AA—
miliony	milion	NNIP1——A—-
druhů	druh	NNIP2——A—-
pizzy	pizza	NNFS2——A—-
.	.	Z:————-

Table 5: Lemma-tag representation of an example Czech sentence.

Token	Lemma	Tag + morphology
Verwendung	verwenden<V>ung<SUFF>	<+NN><Fem><Nom><Sg><NA>
gemäß	gemäß	[APPR-Nom]
Anspruch	Anspruch	<+NN><Masc><Nom><Sg><NA>
5	5	[CARD]
,	,	[$]
wobei	wobei	[PWAV]
die	die<Def>	<+ART><Fem><Nom><Sg><St>
neuronale	neuronal<Pos>	<+ADJ><Fem><Nom><Sg><NA>
Störung	Störung	<+NN><Fem><Nom><Sg><NA>
Alzheimer-Krankheit	{Alzheimer}-<TRUNC>Krankheit	<+NN><Fem><Nom><Sg><NA>
ist	sein	<+V><3><Sg><Pres><Ind>
.	.	[$]

Table 6: Lemma-tag representation of an example German sentence.

We cope with the above mentioned problems by transforming target language texts into an underspecified *lemma-tag* representation which consists of interleaved lemmas and morphological tags providing the full set of relevant inflection features. The NMT models are then trained on this intermediate representation of the target language data. Decoding is followed by a deterministic inflection generation step. We use the predicted pairs of (tag+features, lemma) as input to a morphological generator which outputs the final inflected surface forms.

3.1 Two-step procedure for NMT

The two-step morphology generation includes lemma-tag representation of the target language data. A single inflected word in the target language texts is represented as a sequence of a lemma and a morphological tag. Morphological tags include morphological information (i.e., morphological features) which are overtly expressed in the source language. Other features such as the case are *predicted* after the lemma-tag output is generated by an NMT model. For Czech, MorphoDiTa toolkit (Straková *et al.*, 2014) is used to linguistically annotate the data in order to derive the lemma-tag representation. For German, the data is parsed with a constituency parser BitPar (Schmid, 2004) to obtain morphological analyses in the sentence context and analyzed with the morphological tool SMOR (Schmid *et al.*, 2004) to obtain lemmas of the inflected German surface forms. Morphological generation of Czech words is based on a lexicon of lemmas and their paradigms and it is fully deterministic. For German, SMOR (Schmid *et al.*, 2004) is used to generate inflected words according to a combination of a lemma with specific morphological features. An example of a lemma-tag representation for Czech is given in Table 5, while an example for German is given in Table 6.

3.1.1 Baseline NMT models and training settings

We use the Nematus toolkit for training all NMT systems. Prior to training, we run BPE splitting of both sides of the training corpus. The models are optimized using Adam (Kingma and Ba, 2015). Furthermore, we use the default early stopping criterion in Nematus. Following Nadejde *et al.* (2017), we set the maximum sequence length to 50 for the baseline and to 100 for systems which produce lemma-tag. Our systems are trained according standard Nematus setups with the parameters given in Table 7.

NMT system results can vary significantly due to randomness in initialization and training. Thus, for Czech, we run system training end-to-end for each variant three times and select the best run based on BLEU as measured on the development set and then evaluate it on the final test set. For German, each of the models is trained two times. We give final BLEU results averaged over the two training runs.

	German	Czech
vocabulary	30k	50k
dropout	yes	no
dimension	500	500
dropout_emb	0.2	
dim	1024	1024
dropout_hid	0.2	–
lrate	0.0001	0.0001
dropout_src	0.1	–
opt	adam	adam
dropout_trg	0.1	–
maxlen	50 (100)	50 (100)

Table 7: NMT training settings for the English→Czech and English→German NMT experiments.

Corpus	Sentences	SRC tokens	TRG tokens
train	114k	2,309k	1,908k
test2012	1,385	25,150	20,682
test2013	1,327	28,454	24,107

Table 8: Sizes of English-Czech corpora.

3.1.2 English→Czech

We use the IWSLT training and test sets in English→Czech experiments.[3] The training set consists of transcribed TED talks as collected in the WIT3 corpus (Cettolo *et al.*, 2012). We use IWSLT test set 2012 as the held-out set and the 2013 test set for evaluation. Table 8 summarizes the basic data statistics.

In addition to the baseline and the lemma-tag experiments, we also evaluate a third setting where we train the system to output sequences of morphological tags interleaved with the surface. Table 9 shows the obtained results. In our main experiment, our two-step system achieves a substantial improvement of roughly 1.7 BLEU points, showing that two-step in the neural context works for English to Czech translation for this data size. In the serialization experiment, we see that, surprisingly, the serialization system does not outperform the baseline setup. It is possible that with larger training data, serialization might still outperform the baseline, but our main result has shown that morphological generalization on this data size is beneficial.

We run a targeted experiment with larger sizes of parallel training data to determine whether the improvements hold. We always use the main training set described above but additionally, we add a random sample from the CzEng 1.0 parallel corpus[4] to achieve training data sizes of 250 thousand up to 2 million parallel sentences (total). Table 10 shows the results. We observe the highest difference in the 500k setting (over 3 BLEU points absolute) and while the improvement decreases slightly as we add more data, the difference is still around 2.3 BLEU points even in the largest evaluated setting, which is an encouraging result.

3.1.3 English→German

Similarly to Czech, the first set of experiments for German is run on IWSLT training and test data. The training data consists of 184,879 parallel sentences after filtering out sentences shorter than 5 or longer than 50 words, as well as sentences that could not be parsed. The system is optimized on the 2012 dev-set (1165 sentences), and tested on the 2013 test-set (1363 sentences) and the 2014 test-set (1305 sentences). In addition to the lemma-tag experiment, we also run an experiment which includes simple compound splitting and merging (lemma-tag-split) in order to assess the benefit of compound handling in the context of the NMT for English→German translation direction. The evaluation results of the different NMT models for English→German are

[3] http://workshop2016.iwslt.org/

[4] https://ufal.mff.cuni.cz/legacy/czeng/czeng10/

System	BLEU (dev)	BLEU (test)
Baseline	12.60	12.89
Lemma-tag	14.05	14.57
Serialization	11.49	12.07

Table 9: English-Czech: BLEU scores of NMT system variants.

	Baseline	**Lemma-tag**	Δ
IWSLT	12.89	14.57	1.68
250k	14.87	17.51	2.64
500k	16.96	20.05	3.09
1M	18.07	20.95	2.88
2M	20.04	22.31	2.27

Table 10: English-Czech: BLEU scores of systems with larger parallel training data.

TED'13	**run-1**	**run-2**	**avg.**
baseline	19.87	20.15	20.01
lemma-tag	20.73	20.98	20.86
lemma-tag-split	20.88	21.18	21.03
TED'14	**run-1**	**run-2**	**avg.**
baseline	19.02	18.68	18.85
lemma-tag	20.01	19.93	19.97
lemma-tag-split	20.07	20.76	20.42

Table 11: Lowercased BLEU scores for two English→German test sests (1363 and 1305 sentences).

given in Table 11. On both test sets, the system generating inflected forms based on stems and features is better than the baseline. The addition of compound splitting leads to a minor further improvement. We consider this a promising result, indicating that segmentation using the rich information provided by SMOR can be helpful; we plan to explore this further in future work.

To assess the influence of domain and corpus size, we also evaluate the approach to model German morphology in a larger news corpus setting. To obtain a training corpus that is diverse, but still restricted in size, we combined randomly selected sentences (between 5-50 words) from the 4 parallel corpora provided for English↔German translation at the WMT'17 shared task[5] (selected in equal parts from Europarl, CommonCrawl, News-Commentary and RapidCorpus), resulting in a set of 250k and 500k sentences. The model is optimized on newstest'15 and evaluated on newstest'16. Table 12 shows the results for the surface form baseline and the morphological generation systems with and without compound handling. As for the TED data set, the morphological generation systems outperform the systems trained on surface data, but the improvement for the system trained on 500k sentences is slightly lower than for the system trained on 250k sentences. The systems with additional compound splitting obtained the same result as the basis morphological generation system (250k) or were slightly better (500k).

3.1.4 Discussion

English→Czech We take a deeper look into the Czech translations to see whether there are cases where the generator failed to produce the surface form. We found only a handful of them. These mostly involve unknown proper names (Braper, Hvanda). In just four cases, the tag proposed by the network is not compatible with the lemma (i.e., the network made an error). In order to determine where the improvement comes from, we analyze the number of novel surface forms produced by the system. We find that indeed, unseen word forms are generated by the system but not nearly as many as we expected: only 125 novel tokens are found in the test set (114 word types). Out of these, 14 forms are confirmed by the reference sentences (note that the unconfirmed words may still be correct within the system output). While morphological generalization does indeed occur, it is not the source of most of the observed improvement. When we use surface forms together with the annotations (in our serialization experiment), we see no improvement.

In addition to automatic evaluation in terms of BLEU, we also carried out a blind manual annotation contrasting outputs of baseline and lemma-tag systems. For each instance, the annotator had access to the reference translation and both outputs. The task was to rank which translation is better or to mark both as equal quality. The annotator analyzed 200 sentences. In 130 cases,

[5] http://www.statmt.org/wmt17/translation-task.html

	baseline	**lemma-tag**	**lemma-tag-split**
250k	18.75	20.55	20.51
500k	21.39	22.79	23.00

Table 12: English→German: lowercased BLEU for newstest'16 (2169 sentences) trained on 250k and 500k sentences news-mix data.

freq	part	freq	part	freq	part
2469	ten	1257	sten	1077	ern
2157	te	1214	es	1077	-
1738	en	1169	ter	1058	den
1607	er	1148	gen	1040	s
1474	ung	1 078	ischen	1015	ungen

Table 13: The most frequent fragments on word ends after BPE from the German surface data.

	vocabulary	**vocabulary**
	size	**size w/ BPE**
DE surface data	121,892	22,712
DE lemma-tag	97,587	21,663
DE lemma-tag-split	68,533	21,892

Table 14: Overview of vocabulary size in the German TED data (BPE: Byte Pair Encoding).

the translations are judged as equal. Out of the remaining 70 sentences, the lemma-tag system is marked as better in 48 cases and the baseline won in 22 cases.

English→German A closer look at the German translation outputs reveals that, as also observed for Czech, there are indeed new word forms generated by the lemma-tag system. For the TED'13 set, for example, the lemma-tag system output a total of 261 words that are not in the training data or the English input sentence. Of these, 112 are names or nonsense words produced by concatenating BPE segments.[6] The other 149 words are morphologically well-formed, though not necessarily semantically sound (e.g. *Schokoladenredakteur (chocolate editor)* as proposed translation for *smart-ass editor*), or appropriate in the translation context. Thus, we compared the novel words with the reference translations: 23 words (21 nouns, 2 adjectives) were found in the reference of the respective sentence.

Despite SMOR's complicated structure, the resulting stems are generally well-formed; for uninflectable stems (mostly made-up words such as *Parunelogramm<+NN><Neut><Gen><Sg>*), the markup is simply removed.

One of the main objectives of the two-step approach is to reduce the target-side vocabulary size. Table 13 shows the most frequent fragments on the end of words obtained through BPE splitting on the German surface data – most tend to be inflectional suffixes. Table 14 shows the reduction of vocabulary in the lemma-tag representation: replacing inflected forms with their stems leads to a considerable reduction of the vocabulary size. Further reduction is achieved when compound splitting is performed.

3.2 Further development and experiments of the two-step approach for English→German

3.2.1 Replacing parsing with tagging

In the first set of experiments carried out for the translation direction English→German outlined in Section 3.1.3, we used parsed German data to acquire morphological information needed to create lemma-tag representation of the German part of the training data. Since parsing is a time-consuming process, we examined the possibility of using morphological tagging instead of parsing.

We experiment with *MarMoT* (Mueller *et al.*, 2013), a CRF-based morphological tagger for German. For tagging, we use the pretrained model available on the Internet.[7] MarMoT output is in the ConLL format: Table 15 shows the annotation output for an example German sentence. Note that the table only includes the relevant columns. Derivation of the lemma-tag representation of German follows the same steps like those developed for the German parse trees. The only difference is that the context-dependent morphological information is now not gained from the parses but from the tagged sentences.

In the first set of experiments, we test the comparability of the German NMT translations generated by two different NMT models:

- a model trained on the lemma-tag representation derived from parses (see Section 3.1.3),
- a model trained the lemma-tag representation derived from the MarMoT output.

[6] Into this category, we also count non-wellformed generations by SMOR caused by incorrect transitional elements in compounds, e.g. *Oszillationengenerator* vs. *Oszillationsgenerator*.

[7] http://cistern.cis.lmu.de/marmot/models/CURRENT/

Token ID	Token	POS	Morphology
1	Verwendung	NN	case=nom|number=sg|gender=fem
2	gemäß	APPR	_
3	Anspruch	NN	case=nom|number=sg|gender=masc
4	5	CARD	_
5	,	$,	_
6	wobei	PWAW	_
7	die	ART	case=nom|number=sg|gender=fem
8	neuronale	ADJA	case=nom|number=sg|gender=fem|degree=pos
9	Störung	NN	case=nom|number=sg|gender=fem
10	Alzheimer-Krankheit	NN	case=nom|number=sg|gender=fem
11	ist	VAFIN	number=sg|person=3|tense=pres|mood=ind
12	.	$.	_

Table 15: Relevant output of the Marmot tagger for an example German sentence.

	Baseline	Parsing-based lemma-tag	Tagging-based lemma-tag
BLEU	18.75	20.55	20.30

Table 16: BLEU scores for the NMT systems trained on lemma-tag representation gained from different sentences analyses. The systems are trained on a set of 250,000 parallel sentences an tested on the news test set from the WMT 2016.

The NMT training settings for both systems are equal and follow the training specifications given in Section 3.1.1. The results in terms of BLEU are shown in Table 16. There is a small difference in the systems, however, compared to the baseline, both parsing- as well as tagging-based derivation of the lemma-tag representation leads to the improvement of the German NMT translations. The experiments discussed in the following Section are thus based on the lemma-tag representation derived from the tagged German texts.

3.2.2 Performance on a large medical domain data set

Experiments described in Section 3.1.3 include testing of the lemma-tag approach on three different data sets in terms of the size, as well as of the domain. In this section, we outline experiments carried out on the medical part of the UFAL corpus (see Deliverable D1.1). By doing this, we aim at answering the following two questions:

- Does the lemma-tag approach improves NMT even if a larger training data set is available as observed for English→Czech?
- What performance is gained if a NMT model is trained solely on a relatively small set of the in-domain (i.e., medical) data compared to the HimL models trained on large corpora with in- and out-of domain texts?

The UFAL corpus is a collection of texts from different domains. Each of the sentence pairs in the corpus is enriched with the information about the domain that the pair belongs to and the source of the data as shown in Table 17. Via a simple grep-based search for the word *medical* in the third column of the corpus, we pull out all sentence pairs belonging to the medical domain. From a total set of more than 37 million sentence pairs, we create a medical subcorpus consisting of 3,036,581 sentences.

Experimental setups We compare the lemma-tag approach in two different experimental setups: unconstrained and length-filtered. As mentioned in Section 3.1, the length of the lemma-tag sentences is doubled compared to the original data. Prior to training, the training data is split using the BPE method (Sennrich *et al.*, 2016b) which additionally changes the length of the

German	English	Domain	Source
Doch, es ist dringend.	- It's urgent.	general_corpus	OpenSubtitles
Verwendung gemäß Anspruch 5, wobei die neuronale Störung Alzheimer-Krankheit ist.	The use according to claim 5, wherein the neuronal disorder is an Alzheimer's disease.	medical_corpus	PatTR_Medical

Table 17: Example sentence pairs extracted from the UFAL corpus.

	No filtering						Length filtering					
	NHS24		Cochrane		Himl dev		NHS24		Cochrane		Himl dev	
	BL	LT	BL	LT	BL	LT	BL	LT	BL	LT	BL	LT
BLEU$_{ci}$	24.70	25.63	33.43	34.67	29.78	41.81	24.41	26.12	33.61	33.27	29.64	41.17
	+0.93		+1.04		–		+1.71		-0.34		–	
BLEU$_{cs}$	24.26	20.76	32.91	32.87	–	–	24.00	20.73	32.88	31.60	–	–
	-3.5		-0.09		–		-3.27		-1.28		–	

Table 18: Evaluation of lemma-tag (LT) and the baseline (BL) NMT systems. The validation scores on the HimL dev set are also given (note that the validation scores for LT are computed based on the lemma-tag output-reference comparison.

Input	If **you** do not wish to use it, please ∅ leave it untranslated and ∅ do not publish it.
NMT output	wenn [KOUS] **sie [PPER]** es [PPER] nicht [PTKNEG] verwenden <+V><Inf> mögen <+V><1><Pl><Past><Subj> , [$] verlassen <+V><1><Pl><Pres><Ind> **sie [PPER]** es [PPER] untranslatiert [ADV] und [KON] veröffentlichen <+V><1><Pl><Pres><Ind> **sie [PPER]** es [PPER] nicht [PTKNEG] . [$]
NMT inflected	Wenn **sie** es nicht verwenden möchten, verlassen **sie** es untranslatiert und veröffentlichen **sie** es nicht.

Table 19: Example of an imperative English test sentence and the corresponding German outputs. In the inflected, cased, detokenized output, all occurrences of the pronoun *sie* are lowercased which is in this context false. In fact, all of them should be capitalized since they are used in the polite form of address.

training data in terms of words, i.e. BPE segments. For the training of an NMT model, we limit the length of the surface training sentences to 50 words, while the length of the lemma-tag sentences is limited to 100 words. The length limit is applied on the segmented training data which may lead to a situation in which not exactly the same set of parallel sentences is used for building the surface baseline and for the lemma-tag model.

To ensure that both the surface baseline, as well as the lemma-tag models are trained on exactly the same set of parallel sentences, in the second experimental setup, we perform a length-filtering of the training data after the BPE-splitting step. The filtering function is defined as follows: for each English-German sentence pair e$_i$-d$_i$, if the number of segments in e$_i$ is less or equal to 50 and the number of segments in d$_i$ is less or equal to 100, then e$_i$-d$_i$ is added to the training data set.

The original medical training data consists of 3,036,581 while the length-filtered set contains 2,355,920 parallel sentence pairs. The filtering thus removes 680,661 sentences. We train NMT models using the two data setups. All systems have exactly the same training settings summarized in Table 7. The systems are evaluated on the 1st version of the HimL test set and the results are summarized in Table 18.

General conclusions Lemma-tag leads to better results compared to the baseline in the unconstrained experimental setup. Performance on both NHS24, as well as Cochrane gets better when the lemma-tag approach is applied. Unfortunately, the results for the length-filtered experimental setup are not that clear. While the lemma-tag outperforms the surface baseline on the NHS24 data, it leads to slightly worse translations of the Cochrane test set.

Case-sensitive evaluation Case-sensitive BLEU scores are generally lower than the corresponding case-insensitive scores. While the performance drop for the surface baselines is around 0.4 BLEU points, the drop for the lemma-tag is considerably higher. For the NHS24 test set, for instance, the drop is -0.4 for the baseline and -5.39 BLEU for the lemma-tag, respectively. For the Cochrane data, the drop is a bit lower: -1.4 and -1.67 BLEU points, respectively. Manual inspection of the translations revealed that there is a problem of stemming (and subsequently generating) the German pronoun *sie*. There is namely the difference between the lowercased and capitalized pronoun *sie*: the lowercased version is used to refer to the 3rd person singular or to the 3rd person plural. The 3rd person plural variant is however also used for the polite from of address. In that case, it has to be capitalized. In the used lemma-tag representation, the stem markup did not include information about the capitalization. Hence, in the inflection generation step, all forms of *sie* are lowercased. This poses a problem for our medical data set since many sentences include imperatives in which in the German sentences, *Sie* must be used. An example of erroneous *sie* pronouns in the German NMT translations is shown in Table 19.

Large mixed corpus vs. domain in-domain corpus The HimL medical NMT models evaluated in D5.6 are trained on large training data sets consisting not only of the texts from the medical domain, but also from other domains such as news,

vocabulary	90k
dim_word	500
dim	1024
lrate	0.0001
opt	adam
maxlen	50 (100)
batch_size	80
enc_recurrence_transition_depth	4
dec_base_recurrence_transition_depth	8

Table 20: Nematus parameters for English-to-Czech experiments on medical domain.

	Cochrane Y3	NHS24 Y3	HimL Y2	HimL dev
form→form	25.88	**20.86**	**26.94**	**28.10**
stc→stc	**26.34**	20.30	25.34	26.85
stc→lemma-tag	23.08	20.11	22.87	23.54

Table 21: English-Czech experiments on the HimL data. "HimL Y2" and "Him dev" denote concatenations of respective Cochrane and NHS24 parts.

political discussions, etc. In addition to the existing parallel texts, the training data also includes parallel texts gained via automatic translation of the monolingual target language data (i.e., backtranslated data). The BLEU score of the Y3 system for English→German is 36.52: our models trained on a rather small in-domain training data set reach BLEU scores which are about 6 points lower. This leads to the conclusion that additional, out-of-domain data seems to be quite beneficial. Although our experiments are carried out only for the English→ German translation direction, we assume that the same results may also be expected for other language pairs.

3.3 Further development and experiments of the two-step approach for English → Czech

We applied the two-step approach on the medical domain English-Czech dataset. We used a deep version of Nematus with the parameters specified in Table 20.

We performed three experiments, all of them were trained on the UMC (UFAL Medical Corpus) data.

- **form→form** - The source and the target texts were tokenized by Morphodita tokenizer with additional processing after that. The tokens were split by BPE with dictionary size of 90,000.
- **stc→stc** - The source and the target texts were tokenized the same way as the previous ones and then truecased. Then split by BPE with dictionary size of 90,000.
- **stc→lemma-tag** - The two-step approach: The source side is the same as in previous case, the target side is additionally morphologically analyzed by Morphodita. Instead of the word tokens, there are lemma-tag (produced by Morphodita) pairs separated by space as described in Section 3.1.

The results of these experiments are shown in Table 21. Unfortunately, they do not show any improvements over the baseline. Using the truecaser improved the results only on one dataset. Using the two-step approach performed much worse than the baseline. The reason, why the results are not good as reported in previous sections may be the different domain but also a newer version of Nematus, which uses the deeper reccurence transitions.

3.4 Further development and experiments of the two-step approach for English → Polish and English → Romanian

The experiments using the two-step approach for Polish and Romanian performed very badly. They ended up with BLEU scores only around 15. To compare, the baseline BLEU scores were about 20. We looked on the translated text and propose the following findings:

- It seems that there is no problem with the BPE splitting. Even though the POS tags for Polish and Romanian are very long (often from 30 to 50 characters), because they are concatenations of feature-value pairs of individual morphological features, they were not split by BPE.

Range	0-9	10-20	20-29	30-39	40-49	50-59	60-69	70+
Total sentences	10	10	10	10	10	10	10	4
Total VCs	11	25	28	45	52	59	49	30
Avg. VC/sentence	1.1	2.5	2.8	4.5	5.2	5.9	4.9	7.5

Table 22: Statistics about the test set used to examine the performance of NMT regarding the German verbs.

- The translations seem correct with respect to alternation of lemmas and tags.
- The surface form generation was performed using a vocabulary containing triplets of wordform-lemma-tag. These triplets were extracted from the training data using morphological tools provided by Lingea. Using this vocabulary, we performed a lookup for each lemma-tag pair generated by the MT system to generate the surface form. We returned the lemma in case of the unsuccessful lookup. During extraction, we counted how many times was each triplet encountered and in case of an ambiguity, the most frequent surface form was returned.

  However, from the development data translation, it seems that such a generation process is not enough successful. A wrong word form is generated quite often and this probably causes the low BLEU scores.

3.5 Conclusion

The initially promising results for the tag-lemma representation (which were on smaller data) did not turn out to scale well for German and Czech. For Polish and Romanian, there may have been additional problems with the deterministic morphological generation step. As a result, we have decided not to further pursue the tag-lemma representation in HimL.

4 Placement and inflection of verbs in the German NMT outputs

Studies about the quality of the German NMT outputs carried out by Bentivogli *et al.* (2016) and Popović (2017) have shown that NMT makes almost no errors regarding both the placement, as well as the inflection of the German verbs. Bentivogli *et al.* (2016), for instance, observed that NMT contains 70% less verb position related errors than PBSMT. However, the German NMT outputs are not completely error-free with respect to the verbs. While shorter sentences indeed have almost no verb-related errors, longer German translations still tend to contain errors with respect to the verbs. In order to get a clearer picture about the verbs (which are the most important word category for understanding the meaning of the translations) in the German NMT outputs, we manually examined randomly selected German NMT outputs.

4.1 Evaluation data

For evaluation, we selected a random set of sentences of the different length (i.e. number of tokens) from the WMT'17 news test set[8]. The translations are obtained with the English→German NMT system developed by Sennrich *et al.* (2017a). The sentences are grouped by the sentence length ranges: 0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+ tokens. For each of the ranges, with an exception of the range 70+, 10 sentences are considered (for the range 70+, only 4 sentences were available). Since we expected to find errors regarding the translation of the verbal complexes (VCs)[9], the evaluation is focused on the VCs found in the chosen test set. The statistics about the test set are given in Table 22.

4.2 Position of the German verbs

The evaluation includes is a binary judgment whether the given translation includes at least one verb order related error. As such, we consider the wrongly placed verbs, as well as the non-generated verbs. The results are presented in Table 23. From a total of 74 sentences, 13 of them (i.e., 17%) have at least one verb order related error. Their distribution with respect to the sentence length ranges is however highly imbalanced. Sentences up to the length of 50 words have a very few errors and can be considered not to be problematic for NMT. Errors start to occur in sentences longer than 50 words, i.e., sentences consisting of more than 5 subclauses (i.e., 5 VCs).

The numbers in Table 23 indicate that 50% of the sentence from the length range 50-59 have at least one verb order related error. Let us put the number of the verb order errors into the relation to the total number of the VCs in the respective sentences. The VC-based counts are shown in Table 24. The results show that only 20% of the VCs in the respective English sentences are

[8] http://www.statmt.org/wmt17/translation-task.html

[9] A verbal complex is composed of the verbs, verb particles, the negation and infinitival particle placed within a single clause.

Range	0-9	10-20	20-29	30-39	40-49	50-59	60-69	70+
Total sentences	10	10	10	10	10	10	10	4
OK	10	10	9	9	9	5	6	3
Error	0	0	1	1	1	5	4	1

Table 23: Number of the German NMT outputs with at least one verb order error.

Range	Order	Missing	Total VCs
50-59	3	3	29
60-69	0	3	10
70+	0	1	10

Table 24: Number of the erroneously translated English VCs in sentences with token number greater than 50 words having at least one verb order related error.

translated incorrectly into German. If we further relate the given error counts to the total number of the VCs in all test sentences, the VC translation errors rate decreases to a total of 7% (10 out of 138 VCs).

The existence of (a few) position-related errors raised the question whether preordering, which we successfully used to improve the German SMT translation with respect to the position of the verbs, may help to eliminate those errors in the German NMT outputs. We examined this by training an NMT model on reordered English data. The evaluation of the German translations is given in Table 25.

In terms of BLEU, preordering leads to the worsening of the German translations. Although the manual evaluation of the translation is slightly in favor of the model trained on preordered English data, preordering indeed leads to lower NMT quality. This was observed also for other language pairs. Details on the respective experiments can be found in Ramm *et al.* (2017) and Du and Way (2017).

4.3 Inflection of the German verbs

Similarly to the conduction of verb placement errors, we also evaluated the German NMT output with respect to the inflection of the finite verbs. Finite verb inflection in German includes morphological features person/number (agreement), tense and mood. The evaluation in Table 26 shows evaluation results for our test set regarding the verbal morphological features. Total amount of errors is very small. While there are no errors regarding agreement, there are a few errors with respect to tense and mood. Table 27 shows these example sentences which we discuss in the following in more detail.

Errors with respect to the verbal inflection can be related to the following linguistic issues:

- Non-finite (tenseless) English VC translating into finite German VCs
- Ambiguous English VCs with respect to tense
- Indicative English VCs translating into subjunctive German VCs
- Translation into the German subjunctive mood

Non-finite English VCs English often uses non-finite VCs in subordinate clauses which are often translated into the German finite constructions:
the police patrolled the area closing parts of the street and preventing visitors from accessing house →
die Polizei patrouillierte den Bereich, sperrte Teile der Straße und hielt Besucher davon ab, das Haus zu betreten. In this example, *closing* and *preventing* are translated into finite German VCs *sperrte* and *hielt ab* which both have specific tense/mood values (tense = past, mood = indicative). In the English source VCs, these features are not overtly given which, at least in this example, leads to false tense (i.e., present tense) of the verbs in the German NMT output.

	Baseline		Preordering	
	BLEU	Human	BLEU	Human
EN→DE	38.26	49.2	36.74	50.08

Table 25: Evaluation results for the preordering combined with the English→German NMT.

Range	0-9	10-20	20-29	30-39	40-49	50-59	60-69	70+
Total sentences	10	10	10	10	10	10	10	4
OK	10	9	8	10	9	7	9	10
Tense error	0	1	1	0	0	0	2	0
Mood error	0	0	1	0	1	2	2	0
Agreement error	0	0	0	0	0	0	0	0

Table 26: Number of the German NMT outputs with at least one verb inflection error.

Input	NMT output	Reference	Error
" Tourists that way , tomatoes back here , " as a report in Vedemosti newspaper phrased it .	"Touristen auf diese Weise, Tomaten wieder hier", wie ein Bericht in Vedemosti Zeitung formuliert.	Ein Bericht in Vedemosti formulierte es folgendermaßen: "Touristen dort lang, Tomaten hier lang."	phrased → formuliert
Heavy rain and widespread flooding in Louisiana lead the governor to declare a state of emergency on Friday ...	Starker Regen und weit verbreitete Überschwemmungen in Louisiana führen den Gouverneur dazu, am Freitag ...	Heftige Regenfälle und großflächige Überschwemmungen in Louisiana zwangen den Gouverneur dazu, am Freitag ...	lead → führen
Angela Rochelle Liner , Stefanie R. Ellis and Ellis ' daughter Maleah were shot on June 12 while the car was parked in a driveway , authorities said .	Angela Rochelle Liner, Stefanie R. Ellis und Ellis 'Tochter Maleah seien am 12. Juni erschossen worden, während das Auto in einer Einfahrt abgestellt war, teilten die Behörden mit.	Angela Rochelle Liner, Stefanie R. Ellis and Ellis Tochter Maleah wurden am 12. Juni erschossen, während das Auto in einer Einfahrt geparkt war, sagten die Behörden.	were shot → seien erschossen worden
Donald Trump wouldn 't really mind if he lost the US presidential election in November ...	Donald Trump würde es nicht wirklich ahnen, wenn er die US-Präsidentschaftswahl im November verloren hat...	Verliert Donald Trump die US-Präsidentschaftswahlen im November, wäre ihm das relativ egal...	lost → verloren hat
After the rooms on the ground floor , the Kadterschmiede , were cleared on 22 June by 300 police officers , the police patrolled the area for three weeks , according to official statements , with 30 to 40 officers and three vans , closing parts of the street and preventing visitors from accessing house 94 .	Nachdem die Räume im Erdgeschoss, die Kadterschmiede, am 22. Juni von 300 Polizisten geräumt wurden, patrouillierte die Polizei den Bereich nach offiziellen Angaben mit 30 bis 40 Beamten und drei Kleintransportern, die Teile der Straße schließen und die Besucher vom Zugang zu Haus 94 hindern.	Nachdem am 22. Juni die Räume im Erdgeschoss, die Kadterschmiede, mit 300 Polizeibeamten geräumt worden waren, patrouillierte die Polizei drei Wochen lang offiziellen Aussagen zufolge mit 30 bis 40 Beamten und drei Mannschaftswagen vor Ort, sperrte Teile der Straße und hielt Besucher davon ab, das Haus mit der Nummer 94 zu betreten.	closing → schließen; preventing → hindern
After plot , parcelling and accessibility questions had been answered , and applications for measurements had been made , there was nothing standing in the way of the sale of the plots in the residential areas of " Straßlweg " and " Schönau-West " (near Binderstraße - development section 1) to interested parties .	Nach Grundstück, Parcelling und Barrierefreiheit seien Fragen beantwortet worden, und Anträge für Messungen seien gestellt worden, es stehe dem Verkauf der Grundstücke in den Wohngebieten "Straßlweg" und "Schönau-West" (nahe Binderstraße - Entwicklung Abschnitt 1) an Interessenten nichts im Wege.	Nachdem die Grundstücks-, Parzellierungs- und Erschließungsfragen geklärt und die Anträge auf Vermessung gestellt werden konnten, steht dem Verkauf der Grundstücke in den Wohnbaugebieten "Straßlweg" und "Schönau-West" (Bereich Binderstraße - Ausbauabschnitt 1) an Interessenten nichts mehr im Wege.	had been made → seien gestellt worden; was standing → stehe

Table 27: Tense and mood errors in the German NMT outputs.

Ambiguous English verbs English verbs are highly ambiguous with respect to their morphological features. For instance, in one of the examined translations, the English verb *lead* was wrongly translated into *führen* in the present tense. False translations is however obvious only by looking at the context in which *lead* occurs: *Heavy rain and widespread flooding in Louisiana lead the governor to declare a state of emergency on Friday...* The occurrence of *lead* in a combination with the event of raining and with a temporal PP *on Friday* allows to interpret the verb form as a form in the as tense. NMT failed to capture these contextual dependencies and generated the German verb in the present tense.

Indicative → subjunctive An interesting case of mood-related errors has been found in the context of a conditional sentence: *Donald Trump wouldn't really mind if he lost the US presidential election in November...* The German NMT translation contains the indicative VC *hat verloren* as a translation of *lost* which is in the conditional context wrong. The correct translation would be *würde verlieren* (*Konjunktiv II*) which indicates a conditional in the present/future tense.

Translation into German subjunctive mood In addition to the conditional contexts in which, in German, the subjunctive mood *Konjunktiv II* has to be used, subjunctive is also used to indicate indirect speech. However, not only subjunctive moos is allowed, but also indicative. The choice is often described as a matter of author's preference and genre/domain specifics. In our test set, we found two sentences in which NMT seems to be confused about the German subjunctive mood. In the first example, namely, *Angela Rochelle Liner,Stefanie R. Ellis and Ellis' daughter Maleah were shot ..., authorities said* ., NMT translated *were shot* as *seinen erschossen worden* while the reference translation suggest the indicative translation in the past tense *wurden erschossen*. Although we count the NMT translation as false because it does not match the reference, both of these possibilities are correct. There is however a small pragmatic difference between them: subjunctive mood indicates the non-assertion of the author regarding the proposition of the utterance which is typical in the news domain when reporting about what other people said. In spoken German, the indicative form is however more often used.

In the second example, the generation of the German subjunctive forms is indeed not appropriate: *After plot, parcelling and accessibility questions had been answered, and applications for measurements had been made, there was nothing standing in the way of the sale of the plots in the residential areas of "Straßlweg" and "Schönau-West" (near Binderstraße - development section 1) to interested parties* . Here, *had been made* is translated into *seien gestellt worden*, while *was standing* is translated as *stehe*. Both German VCs are subjunctive. Without access to the preceding context of the given English source sentence, *Konjunktiv I* is rather inappropriate in this context. However, in a context in which in the preceding sentence(s) somebody is writing about findings reported by someone else, the use of *Konjunktiv I* would be correct. Since in the reference translation, indicative is used, we count the two subjunctive translations as wrong.

We observed this kind of overgeneration of *Konjunktiv I* in the German translations several times. NMT correctly captures the dependency of using *Konjunktiv I* in a combination with the so-called reporting verbs such as *(to) answer* or with quotation marks. However, it is sometimes not able to distinguish between context in which these contextual clues should not lead to the subjunctive German VCs.

4.4 Conclusion

The simpler errors related to verbal inflection that we considered in SMT seem largely solved in NMT, while exactly how to solve more complex errors is difficult for even humans to agree on. We learned a great deal about verbal inflection in translation in HimL, and we may continue this work outside of HimL, perhaps focusing on linguistic analysis first.

5 MLFix

In this section, CUNI summarizes current status of development of an automatic post-editing tool, MLFix. The tool development was focused mainly on correcting morphological agreement in the outputs of statistical machine translation (SMT) systems, which was reported in Deliverable 3.3. In this deliverable, we present our findings on MLFix performance when combined with neural machine translation (NMT) systems. Based on these finding we draw a conclusion on possibilities of further MLFix development.

In Deliverables 3.1 and 3.3, we presented the MLFix architecture, the SMT output preprocessing, the process of output correction, data used for training of the statistical components and the training procedure. There were no major changes prior to writing this section.

5.1 Postediting of the NMT outputs

While MLFix development was focusing mainly on the SMT output correction, current rapid development in the research of the NMT systems made the neural-based models new state-of-the-art. The current design of the MLFix tool allows us postediting outputs regardless of the type of the input system, therefore, we decided to evaluate MLFix performance when postediting state-of-the-art NMT.

As our baseline NMT system, we chose deep recurrent neural network model implemented in Nematus[10] (Sennrich *et al.*, 2017b). The network consists of standard encoder-decoder architecture with attention (Bahdanau *et al.*, 2014), using deep transitions (Zhou *et al.*, 2016) and layer normalisation (Ba *et al.*, 2016). The datasets that we used for MLFix analysis was then translated using an ensemble of the last 4 models that were saved during training. The resulting outputs were postedited by MLFix using models presented during evaluation in the Deliverable 3.3.

The results of the automatic evaluation indicated that postedits made by MLFix slightly worsen the NMT output. Also, when compared to the postediting of SMT, even lower percentage of the translated sentences (around 1% of the translated sentences) was affected by the postediting system. When we analysed the translation results, we noticed that the outputs produced by the NMT system already have high fluency resulting in a very small number of errors that can be corrected by MLFix.

Recently, there were several attempts at deeper analysis of the current NMT systems (Sennrich, 2017; Burlot and Yvon, 2017), assessing morphological and grammatical errors made by these translation systems. The results of these studies show that the state-of-the-art NMT models are able to produce target side translations with high fluency, often modeling agreement over a very long distance. Therefore, automatic correction of morphology in the outputs might not be feasible in the future. On the other hand, there are still issues NMT that needs to be resolved. For example, sometimes a slight change in the input sentence can have a strong impact on the adequacy of the resulting output translation. Additionally, linguistic phenomena spanning outside of scope of single sentence (e.g. coreference, overall text cohesion) are still very problematic for the current state-of-the-art systems. The future research in automatic postediting should therefore focus on these kind of errors. The current state of the MLFix framework might still be useful for the following research, although, some design changes might be necessary in the future.

6 Conclusion

All tasks in Workpackage 3 were completed as planned.

This deliverable documents the work completed in Year 3.

We discussed the following work:

- Target-side word segmentation for German
- Two-step (tag-stem) representation for Czech, German, Polish and Romanian
- Application of German verbal modeling work (carried out in Year 2) to NMT
- MLFix: automatic post-editing of NMT outputs

Overall, the work on inflection and word formation led to many new insights in both SMT (earlier in the project) and NMT (later in the project). The final Year 3 systems used linguistic segmentation of German instead of BPE, but no linguistic generalizations were used in the Czech, Polish and Romanian Year 3 sytems because the new NMT systems for these languages were better than proposed improved systems, as was shown in this deliverable.

References

Ba, Lei Jimmy, Ryan Kiros, and Geoffrey E. Hinton. 2016. “Layer normalization.” *CoRR*, abs/1607.06450.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. “Neural machine translation by jointly learning to align and translate.” *CoRR*, abs/1409.0473.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. “Neural versus Phrase-Based Machine Translation Quality: a Case Study.” *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, Texas, USA.

[10] https://github.com/EdinburghNLP/nematus

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. "Findings of the 2017 Conference on Machine Translation (WMT17)." *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, 169–214. Copenhagen, Denmark.

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.

Burlot, Franck and François Yvon. 2017. "Evaluating the morphological competence of machine translation systems." *Proceedings of the Second Conference on Machine Translation, Volume 1: Research Papers*, 43–55. Copenhagen, Denmark.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. "Wit^3: Web inventory of transcribed and translated talks." *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*. Trento, Italy.

Du, Jinhua and Andy Way. 2017. "Pre-Reordering for Neural Machine Translation: Helpful or Harmful?" *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*. Prague, Czech Republic.

Gage, Philip. 1994. "A New Algorithm for Data Compression." *C Users J.*, 12(2):23–38.

Huck, Matthias, Alexandra Birch, and Barry Haddow. 2015. "Mixed-Domain vs. Multi-Domain Statistical Machine Translation." *Proc. of MT Summit XV, vol.1: MT Researchers' Track*, 240–255. Miami, FL, USA.

Huck, Matthias, Fabienne Braune, and Alexander Fraser. 2017a. "LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts." *Proceedings of the Second Conference on Machine Translation*, 315–322. Copenhagen, Denmark.

Huck, Matthias, Simon Riess, and Alexander Fraser. 2017b. "Target-side Word Segmentation Strategies for Neural Machine Translation." *Proceedings of the Second Conference on Machine Translation*, 56–67. Copenhagen, Denmark.

Kingma, Diederik P. and Jimmy Ba. 2015. "Adam: A method for stochastic optimization." *CoRR*, abs/1412.6980.

Koehn, Philipp. 2005. "Europarl: A Parallel Corpus for Statistical Machine Translation." *Proceedings of the MT Summit X*. Phuket, Thailand.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. "Moses: Open Source Toolkit for Statistical Machine Translation." *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, 177–180. Prague, Czech Republic.

Koehn, Philipp and Kevin Knight. 2003. "Empirical Methods for Compound Splitting." *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, 187–194. Budapest, Hungary.

Mueller, Thomas, Helmut Schmid, and Hinrich Schütze. 2013. "Efficient higher-order CRFs for morphological tagging." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, Washington, USA.

Nadejde, Maria, Siva Reddy, Rico Sennrich, Tomasz Dwojak, Marcin Junczys-Dowmunt, Philipp Koehn, and Alexandra Birch. 2017. "Syntax-aware neural machine translation using ccg."

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. "Bleu: a Method for Automatic Evaluation of Machine Translation." *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA.

Popović, Maja. 2017. "Comparing Language Related Issues for NMT and PBMT between German and English." *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*. Prague, Czech Republic.

Porter, Martin. 1980. "An algorithm for suffix stripping." *Program: electronic library and information systems*, 14(3):130–137.

Ramm, Anita, Riccardo Superbo, Dimitar Shterionov, Tony O'Dowd, and Alexander Fraser. 2017. "Integration of a Multilingual Preordering Component into a Commercial SMT Platform." *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*. Prague, Czech Republic.

Schmid, Helmut. 2004. “Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors.” *20th International Conference on Computational Linguistics (COLING)*. Geneva, Switzerland.

Schmid, Helmut, Arne Fitschen, and Ulrich Heid. 2004. “SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection.” *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal.

Sennrich, Rico. 2017. “How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs.” *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, 376–382.

Sennrich, Rico, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017a. “The University of Edinburgh’s Neural MT Systems for WMT17.” *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*.

Sennrich, Rico, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. 2017b. “Nematus: a toolkit for neural machine translation.” *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 65–68. Valencia, Spain.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016a. “Edinburgh Neural Machine Translation Systems for WMT 16.” *Proceedings of the First Conference on Machine Translation*, 371–376. Berlin, Germany.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016b. “Neural Machine Translation of Rare Words with Subword Units.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–1725. Berlin, Germany.

Sennrich, Rico, Philip Williams, and Matthias Huck. 2015. “A tree does not make a well-formed sentence: Improving syntactic string-to-tree statistical machine translation with more linguistic knowledge.” *Computer Speech & Language*, 32(1):27–45.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. “A Study of Translation Edit Rate with Targeted Human Annotation.” *Proceedings of the Conference of the Association for Machine Translation in the Americas (AMTA)*, 223–231. Cambridge, MA, USA.

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*. Baltimore, Maryland.

Tamchyna, Aleš, Marion Weller-Di Marco, and Alexander Fraser. 2017. “Modeling target-side inflection in neural machine translation.” *Proceedings of the 2nd Conference on Machine Translation, WMT 2017*. Copenhagen, Denmark.

Zhou, Jie, Ying Cao, Xuguang Wang, Peng Li, and Wei Xu. 2016. “Deep recurrent models with fast-forward connections for neural machine translation.” *TACL*, 4:371–383.