Modeling the position and inflection of verbs in English to German machine translation

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Philosophie (Dr. phil.) genehmigte Abhandlung.

Vorgelegt von

Anita Ramm

aus Vukovar, Kroatien

Hauptberichter	Prof. Dr. Alexander Fraser
1. Mitberichterin	PD Dr. Sabine Schulte im Walde
2. Mitberichter	Prof. Dr. Jonas Kuhn

Tag der mündlichen Prüfung: 23.03.2018

Institut für Maschinelle Sprachverarbeitung
der Universität Stuttgart

2018

Abstract

Machine translation (MT) is automatic translation of speech or text from a source language (SL) into a target language (TL). Machine translation is performed without human interaction: a text or a speech signal is used as input to a computer program which automatically generates translation of the input data. There are two main approaches to machine translation: rule-based and corpus-based. Rule-based MT systems rely on manually or semi-automatically acquired rules which describe lexical, as well as syntactic correspondences between SL and TL. Corpus-based methods learn such correspondences automatically using large sets of parallel texts, i.e., texts which are translations of each other. Corpus-based methods such as statistical machine translation (SMT) and neural machine translation (NMT) rely on statistics which express the probability of translating a specific SL translation unit into a specific TL unit (typically, word and/or word sequence). While SMT is a combination of different statistical models, NMT makes use of a neural network which encodes parallel SL and TL contexts in a more sophisticated way.

Many problems have been observed when translating from English into German using SMT. One of the most prominent errors in the German SMT outputs is related to the verbs. SMT often misplaces or does not generate the German verbs at all. Furthermore, the inflected German verb forms are often incorrect. This thesis describes methods for handling the two respective problems. While the positional problems are dealt with in a pre-processing step which can be seen as a preparation of the English data for training and translation, the verbal inflection is handled in a post-processing step and can thus be seen as an automatic post-editing (or correction) step to the translation.

Consider the position of the verbs *have/habe* and *read/gelesen* in the following English-German sentence pair: *I* ***have read*** *that book* ↔ *Ich* ***habe*** *dieses Buch* ***gelesen***. For SMT, the different position of the participles *read/gelesen* is problematic since the translation step needs to jump over many words, in this case over the words *dieses/this* and *Buch/book*, to place the German participle into the correct position. Such positional differences, caused by grammatical constraints in English and German, are given in almost all sentence types and lead to many errors in the German translations. I correct these errors by applying the so-called *preordering* of the English sentences. Preordering transforms (i.e., reorders) English sentences in a way that they encounter German-like word order. The reordered English texts are then used to train English-German SMT models and also to translate English test sentences. Thus, instead of being trained on the sentence pairs such as *I* ***have read*** *that book* ↔ *Ich* ***habe*** *dieses Buch* ***gelesen***, an English-German SMT system is now trained on the following data: *I* ***have*** *that book* ***read*** ↔ *Ich* ***habe*** *dieses Buch* ***gelesen***. Doing this, SMT does not need to perform problematic search for the correct positions of the German verbs since they correspond

to the positions of their English counterparts. I test the improvement potential of pre-ordering for English→German in many different experimental setups in which the effect of domain, size of the training data, as well as the used language models is taken into account. The experiments show that the German translations generated by an SMT model trained on reordered English sentences have more verbs which are more often correctly placed when compared with translations generated by an SMT model trained on the original parallel corpus.

Correct placement of the verbs does not mean that their inflection is correct. The English→German SMT systems have problems generating correct German verb forms and this problem gets even more severe when reordering of the English data is performed. The difficulty of generating the correct German verb forms is due to the difference in the morphological richness of English and German. English differentiates between only a few forms of a single verb lemma, while in German, a single verb lemma has many different inflectional variants. In the context of (S)MT, this means that a single English verb form may be translated into numerous German verb forms, e.g., *had* ↔ {*hatte, hattest, hatten, gehabt*}. Which of these variants is correct, depends on the context in which the verbs occur. In German, for instance, the subjects require a specific form of the finite verbs, e.g., *I* ***work*** ↔ *Ich* ***arbeite*** or *They* ***work*** ↔ *Sie* ***arbeiten***. This linguistic property is called subject-verb agreement. SMT often fails to capture required contextual dependencies between subjects and verbs which leads to German translations in which the subject-verb agreement is violated: those translations are grammatically incorrect. The German verbal morphology includes not only information about agreement (person and number), but also about tense and mood. Generation of the verb forms with tense and mood properties which do not correspond to the source leads to sentences which may be interpreted incorrectly. Furthermore, if the target language constraints on usage of tense and mood are not met, the translations are, as in the case of false subject-verb agreement, grammatically incorrect. In this thesis, both of the inflection-related problems are tackled with a subsequent generation of the German finite verbs according to morphological features derived by considering relevant contextual information.

Subject-verb agreement errors are dealt with by parsing the German SMT outputs. Given a parse tree, first the subject-verb pairs are identified. Subsequently, the person and number features of the subject are transferred to the corresponding finite verb. This approach ensures that the agreement is established between the generated subjects and the agreeing finite verbs in the German translations. The method works well for the used test set, its success, however, largely depends on the parsing accuracy.

As mentioned above, the generation of the German verbs also requires information about tense and mood. These morphological features are gained with a classifier which is trained on many types of different contextual information derived from the English and German sentences. Although the classification accuracy is relatively high when computed on well-formed test sets, it is not sufficient to generally improve tense and mood of the verbs in the German SMT outputs. In some cases, the predicted values indeed correct false German verbs: particularly the German finite verbs generated as translations of the English non-finite verbs profit from the tense and mood prediction step.

The tense/mood classifier used in this thesis is a first attempt to model tense and mood translation from English to German. A deeper analysis of the translation examples, of the parallel data, as well as of the theoretical research on (human) translation in general shows the whole complexity of the problem. The present thesis includes a summary of the most important findings with respect to the translation of tense and mood for the English→German translation direction. Not only the theoretical knowledge about this topic is required when it comes to its automatic modeling. Tense and mood depend on factors which need to be extractable from the data in terms of their automatic annotation. One of the by-products of my research is an open-source tool for the annotation of tense and mood for English and German in the monolingual context. Along with the results and discussions provided in this theses, the tool provides a strong basis for further work in this research area.

Deutsche Zusammenfassung

Maschinelle Übersetzung (MÜ) befasst sich mit der automatischen Erstellung von Übersetzungen. Für einen Text (oder gesprochene Sprache) in der Quellsprache (QS), wird ein MÜ-System dazu verwendet, automatisch, das heißt ohne Hilfe des Menschen, den äquivalenten Text in der Zielsprache (ZS) zu generieren. Im Jahre 2004 wurde das erste frei verfügbare Programm, genannt *Moses*, zur statistischen maschinellen Übersetzung (SMÜ) veröffentlicht. Dies bedeutete den entscheidenden Durchbruch für den breit gefächerten Einsatz der MÜ. Das Erstellen der SMÜ-Systeme ist denkbar einfach: man benötigt lediglich Texte in QS und ZS, die Übersetzungen voneinander sind. Das SMÜ-Modell lernt aus den Texten, welche Wörter bzw. Wortfolgen in QS und ZS Übersetzungen voneinander sind. Den Übersetzungspaaren werden Übersetzungswahrscheinlichkeiten zugewiesen, die auf der Häufigkeit des Auftretens der ermittelten Übersetzungspaare im gegebenen Textpaar basieren. Für viele Sprachpaare generiert SMÜ gute Übersetzungen, allerdings ist die Qualität der SMÜ-Übersetzungen für Sprachpaare, die sich morphologisch und/oder syntaktisch bedeutend voneinander unterscheiden, immer noch mangelhaft. Zu solchen Sprachpaaren gehören auch Englisch und Deutsch.

Diese Doktorarbeit befasst sich mit den Verben in den deutschen SMÜ-Übersetzungen mit Englisch als Quellsprache. Ausgehend aus dem Englischen, werden deutsche Sätze generiert, die zweierlei Probleme hinsichtlich der Verben aufweisen: (i) Stellung von Verben und (ii) Konjugation von Verben. Aufgrund der großen Unterschiede bezüglich der Stellung von Verben in Deutsch und Englisch sind die Verben in den deutschen Übersetzungen entweder falsch positioniert oder sie werden erst gar nicht generiert. Im Falle eines generierten Verbs wird dieses oft nicht korrekt konjugiert. Das heißt, dass entweder seine Form nicht zum Subjekt passt oder, dass es keinen korrekten Tempus- bzw. Moduswert aufweist. Die genannten Probleme wirken sich negativ auf die Qualität und dadurch auch die Akzeptanz der deutschen Übersetzungen aus. Fehler bezüglich der Verbform erschweren das Verständnis der generierten Übersetzungen oder können sogar zu falschen Interpretationen führen. Auf der anderen Seite sind die Übersetzungen ohne Verb sehr schwer oder gar nicht zu verstehen, was die Motivation für die Behandlung von Verbfehlern in den deutschen Übersetzungen liefert.

Um die korrekte Stellung, sowie Generierung von Verben in den deutschen Übersetzungen sicherzustellen, wird die sog. Umordnungs-Methode angewendet. Dabei werden die Verben in den englischen Sätzen an die Stellen gesetzt, die der Verbstellung im Deutschen entsprechen, z.B. *I* ***have*** *the book* ***read****.* ↔ *Ich* ***habe*** *das Buch* ***gelesen****.* Solche umgeordneten englischen Sätze werden sowohl für das Erstellen von englisch-deutschen SMÜ-Systemen benutzt, als auch als Eingabe im Übersetzungsschritt. Sowohl die automatische als auch die manuelle Auswertung von deutschen Übersetzungen, generiert

basierend auf den umgeordneten englischen Sätzen, zeigen, dass der vorgeschlagene Ansatz sehr wirksam ist, was die Stellung und Generierung von deutschen Verben angeht. Die Übersetzungen beinhalten im Allgemeinen viel mehr Verben im Vergleich zu denen, die von einem SMÜ-System generiert werden, das anhand von ursprünglichen englischen Texten erstellt wurde. Hinzu kommt, dass die Verben nun viel öfter an den korrekten Stellen in den deutschen Sätzen stehen.

Obwohl die Umordnung von englischen Sätzen zum gewünschten Ergebnis führt, was die Stellung von Verben angeht, bringt sie auch gewisse Probleme mit sich, die sich auf die Konjugation von Verben negativ auswirken. Englische Verbformen weisen hohen Synkretismus auf. Das bedeutet, dass eine englische Verbform vielen verschiedenen deutschen Verbformen entsprechen kann, z.B. *had* ↔ {*hatte, hattest, hatten, gehabt*}. Zu den indikativen deutschen Formen im vorangehenden Beispiel kommen zusätzlich Konjunktiv-Formen (z.B. *hätte, hättest, hätten*), die auf der lexikalischen Ebene keine direkte Entsprechung im Englischen haben. Welche dieser vielen möglichen Formen generiert werden muss, hängt vom Kontext ab. SMÜ-Systeme haben bereits Schwierigkeiten, den nötigen Kontext korrekt zu erfassen. In den umgeordneten englischen Sätzen wird das Problem noch eklatanter, da in vielen Fällen die Verben in großer Entfernung zu den Wörtern bzw. Wortfolgen stehen, die die Auswahl der korrekten deutschen Verbformen steuern. Das prominenteste Beispiel dafür ist die Abhängigkeit der Verbform vom Subjekt, z.B. *I* ***work*** ↔ *Ich* ***arbeite*** bzw. *They* ***work*** ↔ *Sie* ***arbeiten***. Da die umgeordneten englischen Sätzen der deutschen Syntax entsprechen, sind die englischen Verben weit entfernt von ihren Subjekten, was in vielen Fällen zur Wahl falscher Verbformen in den deutschen Übersetzungen führt. In dieser Arbeit wird eine Methode vorgestellt, die auf Basis von automatischer Nachbearbeitung von Übersetzungen solche Fehler korrigiert. Dabei werden mithilfe der syntaktischen Analyse von deutschen Übersetzungen die Subjekt-Verb-Paare ermittelt und das Verb wird dem Subjekt entsprechend konjugiert. Die Methode führt zu weniger Fehlern in den deutschen Übersetzungen, allerdings hängt sie sehr davon ab, wie gut die syntaktische Analyse von deutschen Übersetzungen ist.

Die Generierung von deutschen Verben hängt nicht nur von der Person und des Numerus des Subjekts ab, sondern auch von Tempus und Modus. Um diese beiden Werte zu ermitteln, wird zunächst eine datengetriebene Analyse von Tempus und Modus in englisch-deutschen Übersetzungen präsentiert. Basierend auf dem linguistischen Wissen sowie den Schlüssen, die die betrachteten Texte nahe legen, wird ein Klassifikator entwickelt, der die beiden Werte für jedes deutsche Verb vorhersagt. Der Klassifikator erhält Zugriff zu unterschiedlichen Informationen im gegebenen englischen Satz und lernt automatisch, welcher Tempus bzw. Modus in der deutschen Übersetzung zu generieren ist. Die Auswertung des Klassifikators und der Verben, die entsprechend der Vorhersage konjugiert werden, zeigt, dass der Klassifikator in vielen Fällen ungenaue Vorhersagen macht und somit zu den Fehlern in den deutschen Übersetzungen führt. Allerdings ist der Klassifikator in der Lage, richtige Werte für die Sätze vorherzusagen, die im Englischen gar keine Zeit/Modus-Werte aufweisen, nämlich Infinitiv- bzw. gerundive Sätze. Infinite englische Sätze werden oft als finite deutsche Sätze, also mit Zeit/Modus-Werten, übersetzt (z.B. *Not* ***knowing*** *that...* ↔ *Da wir nicht* ***wussten****, dass...*). Das MÜ-System

macht an dieser Stelle oft Fehler, die der Klassifikator zu korrigieren imstande ist.

Das in dieser Arbeit beschriebene, eher einfache, Modell zur Vorhersage von Tempus und Modus für die Verben in den deutschen Übersetzungen ist der erste Versuch, sich an dieses Thema heranzutasten. Die Analyse von Klassifikationsfehlern macht deutlich, dass die Tempus/Modus-Vorhersage ein sehr komplexes Problem ist, das oft von Faktoren abhängt, die im Text (d.h. in den Wörtern und Wortsequenzen) nicht explizit herauszulesen sind. Die Nichtverfügbarkeit solcher Informationen hat zur Folge, dass oft fehlerhafte Vorhersagen gemacht werden, was wiederum zu Fehlern in den deutschen Übersetzungen führt. Die zusammenfassende Beschreibung des Problems aus der monolingualen Perspektive, sowie im Kontext der (maschinellen) Übersetzung in Kombination mit der Analyse von Tempus/Modus-Übersetzung in einem regelbasierten MÜ-System liefert Grundlage zu weiterführenden Arbeiten im Bereich der automatischen Modellierung der Übersetzung von Tempus und Modus.

List of Related Publications

Parts of the research described in this thesis have been published in:

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

- Cap, F., Weller, M., Ramm, A., and Fraser, A. (2014). CimS - The CIS and IMS joint submission to WMT 2014: translating from English into German. In *Proceedings of the 9th Workshop on the Statistical Machine Translation (WMT)*, Baltimore, Maryland, USA

- Cap, F., Weller, M., Ramm, A., and Fraser, A. (2015). CimS - The CIS and IMS Joint Submission to WMT 2015 addressing morphological and syntactic differences in English to German SMT. In *Proceedings of the 10th Workshop on the Statistical Machine Translation (WMT)*, Lisbon, Portugal

- Ramm, A. and Fraser, A. M. (2016). Modeling verbal inflection for English to German SMT. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers (WMT)*, Berlin, Germany

- Ramm, A., Loáiciga, S., Friedrich, A., and Fraser, A. (2017a). Annotating tense, mood and voice for English, French and German. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (ACL), system demonstrations*, Vancouver, Canada

- Ramm, A., Superbo, R., Shterionov, D., O'Dowd, T., and Fraser, A. (2017b). Integration of a Multilingual Preordering Component into a Commercial SMT Platform. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, Prague, Czech Republic

Contents

List of Abbreviations

BLEU	Bilingual evaluation understudy
CRF	Conditional random field
DE	German
EC	European Commission
EN	English
LM	Language model
LRM	Lexicalized reordering model
LSK	Linke Satzklammer
maxent	Maximum entropy
MF	Mittelfeld
MT	Machine translation
MÜ	Maschinelle Übersetzung
NF	Nachfeld
NMT	Neural machine translation
NP	Noun phrase
PBSMT	Phrase-based statistical machine translation
POS	Part-of-speech
PP	Prepositional phrase
QS	Quellsprache
RSK	Rechte Satzklammer
SL	Source language
SMT	Statistical machine translation
SMÜ	Statistische maschinelle Übersetzung
SOV	Subject-object-verb
SVO	Subject-verb-object
SVOV	Subject-verb-object-verb clause

SVM	Support vectors machines
TL	Target language
TM	Translation model
VC	Verbal complex
VE	Verb-end clauses
VF	Vorfeld
VFIN	Finite verb
VP	Verb phrase
V1	Verb-first clauses
V2	Verb-second clauses
ZS	Zielsprache

List of Figures

List of Tables

1. Introduction

Machine translation is a process of automatically translating speech or text from a source language into a target language. Automatically means that no humans are involved in the translation process. Instead, computers which make use of computational models are used to translate between languages. Since the first fairly simple ideas about how to perform machine translation were proposed and implemented in the 1940s, the quality of the automatically generated translations has continuously grown reaching the level which is meanwhile comparable with the translations produced by humans.

The first statistical models for machine translation were presented in the early 1990s (Brown et al., 1990, 1993). While the first SMT models were word-based models which supported word-to-word translation, the *phrase-based* SMT models (PBSMT) developed in the middle of the first decade of the 2000s (Och and Ney, 2004; Koehn et al., 2003; Koehn, 2004) allowed translation of word sequences rather than single words. SMT models are automatically trained on *parallel texts*, i.e., texts which are translations of each other. Given a source-target sentence pair, PBSMT automatically extracts *translation phrases*, i.e., sequences of the source and target language words which correspond to each other. The translation phrases are assigned frequency-based probabilities which indicate the likelihood of a target language phrase being the translation of a source language phrase.

The release of *Moses* – the first open-source tool for building the SMT systems – combined with the availability of large amounts of parallel data for different language pairs (e.g., (Koehn, 2005)) had a great impact on the further research activities in the field of machine translation in general. SMT achieved great success due to its simplicity and effectiveness: relatively simple statistical models automatically induced from parallel text collections suddenly allowed to translate great amounts of source language texts in a short time providing translations of sufficient quality.

The potential of machine translation has initially been recognized by institutions of the public sector. Since World War II, large amounts of important information was encoded in many languages which motivated the public sector to invest into machine translation

research. Meanwhile, machine translation has found its way into globally operating enterprises and multi-national institutions such as the European Commission which need to provide large amount of information in and acquire from many different languages. Here, machine translation is often used in the context of post-editing: raw translations are gained by means of the automatic translation and are subsequently post-edited (i.e., corrected or adapted to the in-house translation quality requirements). Although in this scenario, the translation process is not fully automatic, it considerably speeds up the translation process and thus lowers the translation costs (Plitt and Masselot, 2010).

Not only big companies use machine translation: in the era of the *World Wide Web*, people all over the world use machine translation to translate foreign-language contents found on the Internet. One of the most famous online translators *Google translate*[1] translates more than 100 billion words per day.[2] Currently the most widely used social network platform *Facebook*[3] automatically produces 2 billion translations a day.[4] These overwhelming numbers indicate very nicely the importance, as well as the acceptance of machine translation in the age of worldwide digitization and globalization.

1.1. Motivation

Acceptance and usage of machine translation depend greatly on the quality of the automatically generated translations. Since 2003, for more than 10 years, PBSMT has been the state-of-the-art approach to machine translation. Although the development of SMT was a breakthrough in the research on machine translation, the assumptions made by SMT[5] models cause errors in the translations related to different linguistic phenomena. For instance, SMT relies on the translation of relatively short word sequences (phrases) which is a powerful device for automatic modeling of the translation process. However, the phrase-based approach has difficulties to capture specific syntactic or morphological dependencies between the words across the phrase boundaries. These *long-distance dependencies* often have a negative impact on, for instance, generation of the target language words with the correct inflection. Not only the generation of the correct target language word forms is problematic, but also the placement of the generated words.

[1] `https://translate.google.com/`

[2] `http://www.k-international.com/blog/google-translate-facts/` retrieved on January 2nd, 2018.

[3] `https://www.facebook.com/`

[4] `https://techcrunch.com/2016/05/23/facebook-translation/` retrieved on November 9th, 2017.

[5] Henceforth, we use the acronym SMT to refer to the PBSMT which is the main topic of this thesis.

SMT needs to perform many reorderings during the translation step in order to generate translations with target-language like syntax. The bigger the difference regarding the position of the source words and their target language equivalents (which often require *long-range reorderings*), the higher the probability that the translations expose the erroneous word order.

Long-distance dependencies as well as long-distance reorderings are particularly problematic when translating between languages with great differences regarding the syntax and morphology. One such problematic language pair is English-German. Due to divergent syntactic and morphological properties, many different errors are observed, particularly in the German SMT outputs. One of the most prominent issues is generation and placement of the verbs in the German translations. Since the positions of the verbs in English and German differ in many types of clauses, the verbs are often missing in the German translations or they are placed incorrectly. Especially the problem of the missing verbs is critical since it hinders the correct interpretation of the German translations: verbs are one of the most informative words in a sentence and in cases where they are not present, it is almost impossible to derive the meaning of a sentence. In the translations in which the verbs are available, they are usually placed incorrectly. The presence of the verbs allows us to understand the translation, however, for instance in the commercial usage of such translations, the verb placement errors need to manually be corrected. Manual correction is not only required to correct the placement of the verbs in the German outputs, but also to correct their inflected forms. German has rich verbal morphology: the verb forms match the subjects in terms of person and number and they bear tense and mood information. SMT often has problems choosing the correct German verb forms which results in grammatically incorrect sentences, as well as sentences which may be misinterpreted.

Positional, as well as inflectional problems regarding the German verbs may have a negative impact on the willingness to use computer programs to automatically translate English texts (or speech) into German using the statistical approach to machine translation. This is a strong motivation to explore possibilities for reducing the respective errors in the German translations. The main topic of this thesis is analysis, development and implementation of methods which aim at reducing errors related to the verbs in the German SMT outputs. Regarding the positional problems, we explore the effectiveness of the *preordering approach* which relies on the reduction of the syntactic differences between English and German. The simple idea of placing the words in the source language into the target-language specific positions prior to training and translation has

been proven to work for many different language pairs. The reason is fairly simple: by putting the words in the source sentences into the positions in which their target language counterparts are expected, SMT does not need to perform problematic reorderings which often include jumps over a big number of the words (i.e., long-range reorderings). Instead, we allow the SMT to translate in a monotonic fashion where the target language words have the same position as the words they are translations of. Besides preordering which accounts for the problematic syntactic differences between English and German, we additionally explicitly model inflection of the German verbs. While preordering is a pre-processing step to the training and translation, the verbal inflection modeling is implemented as a post-processing step: it is applied after the German translations have been generated and aims at *automatically correcting* the inflection of the German finite verbs. The correction step is based on the prediction of the morphological features for the German finite verbs given the information about the context in which the verbs occur. The features that are required to generate the German verbs are *person, number, tense* and *mood*. While the agreement features, person and number, are determined by the morphological properties of the corresponding subjects, tense and mood often depend on factors which are not overtly expressed in the contexts of the respective verbs.

Both reordering, as well as prediction of the verbal features include interesting research questions:

- Can the main syntactic differences between English and German regarding the verbs be formally described?
- What representation of the English sentences in needed to have access to information which is required in order to transform English into a German-like form?
- Is the deterministic preordering of English sufficient to improve translation quality given a relatively flexible word order in German?
- What is the optimal method for establishing the agreement between subjects and finite verbs in the German translations?
- What kind of knowledge is needed to predict tense and mood of the verbs and verbal complexes in the German translations?
- Is there a general description of how the (human or machine) translation of tense and mood is carried out?

1.2. Contributions

This thesis focuses on handling problems in the German translations related to the verbs. The problems are twofold:

(i) due to the positional differences of the verbs in English and German, the verbs in the German translations are often misplaced or even omitted;

(ii) due to the morphological richness of German, the finite verbs in the German SMT outputs are often incorrectly inflected.

The positional problems are handled with the *preordering* approach which reduces the syntactic differences between English and German. The inflectional problems are tackled with a post-processing method which includes prediction of the morphological features *person, number, tense* and *mood* for every single finite verb in the German translations and the subsequent generation of the inflected forms for the respective verbs.

Preordering Encouraged by the success of the preordering approach for SMT, I adapt preordering to English→German translation direction (Gojun and Fraser, 2012). I identify clause boundaries, clause and verbal complex types as a crucial contextual information needed to transform English into German-like sentence structure which motivates the use of the constituency parse trees as an underlying representation of the English sentences. The **syntactic differences** regarding the position of the verbs in English and German are first described in a **formal way**. The formal description is then used to manually formulate the **rules which describe movements** of the specific subtrees of a given parse tree in such a way that the enclosed English words are moved to the positions which are typical for German. It needs to be noted that preordering presented in this thesis cannot be seen as a simple reversement of the preordering for German→English SMT described in Collins et al. (2005). The translation from English to German is more challenging since the positions of the verbs in German differ depending on the clause type, as well as on the type of the given verbal complex. Thus, there are **more contexts that need to be considered** than when translating into the opposite translation direction. Additionally, the parts of a single German verbal complex may be placed in different positions. In many cases, this requires **splitting the English verbal complexes** into parts that carry enough contextual information in order to allow SMT to generate correct verbs in German.

In a small multilingual study on applying the preordering approach in the commercial setting, we develop a **language-independent component for the deterministic**

preordering for three different language pairs (Ramm et al., 2017b). We examine the **performance of preordering in terms of speed and choice of a parser**. The experiments confirm the **benefit of preordering** for English→German SMT regardless the domain, amount of training data and the underlying method to compute word alignment needed to train SMT models.

Verbal inflection Modeling of the verbal inflection extends the framework of modeling the nominal inflection for English→German SMT originally proposed by Fraser et al. (2012) and further improved by Weller et al. (2013). Similarly to the approach for modeling nominal inflection, the implemented method for handling verbal inflection relies on the **prediction** of the morphological features of the German finite verbs (Ramm and Fraser, 2016). Regarding the agreement features, I show that the predicted values are often incorrect due to **syntactic differences** between English and German, as well as **translation-related discrepancies** between constituents in the source and target language. In order to overcome these problems, I apply the **parsing-based method to correct the agreement** of the finite verbs in the German SMT outputs. This method has been successfully tested on the English→Czech translation direction in the past (Rosa et al., 2012). Generation of the German finite verbs also requires the morphological features *tense* and *mood*. I use a pre-trained classifier to predict these features. Despite relatively **high prediction accuracy**, the predicted labels do not always lead to the improved translations. Similar findings were previously reported by Gispert and Mariño (2008) for English→Spanish SMT. In a minor study in the context of the factored SMT, I show that the direct integration of tense and mood information into SMT leads to higher quality of the German SMT outputs. Hereby, the provision of the target **side tense/mood information**, in our case German, proves to be more appropriate than making the corresponding monolingual information explicitly available.

Translation of tense and mood Despite a simple assumption that the tense given in a source sentence needs to be transferred to a target sentence, it is difficult to model the translation of tense and mood. In the bilingual context, we not only need to consider **bilingual correspondence of the tenses** in the source and target language, but also the **usage of tense and mood in a target language**, i.e., in the monolingual context. My tense/mood classifier may be seen as a prototype, as a first attempt to tackle this complex issue for the English→German translation direction. In order to provide hints for the future work on this topic, I carry out a corpus-based analysis of the tense/mood

correspondences in the English-German parallel texts. Moreover, I study the usage of tense and mood in German which reveals that the usage often underlies factors or criteria which are not accessible from the representations of the (meaning of the) English and German sentences that were used in this work. To those belong **genre and register specifics**, as well as **translator's and author's preferences**. A detailed analysis of theses aspects described in the thesis will serve as a basis for the future work on this topic.

Automatic annotation of tense, mood and voice The corpus study, as well as training of the tense/mood classification models require a parallel English-German corpus annotated with syntactic tense and mood information. Morphological analyzers available for the two languages annotate the *morphological tense* of the finite verbs, however, there are **no tools which annotate *syntactic* tense, mood and voice** for the two languages. Therefore, I implement a **tool for the automatic annotation of the syntactic tense, mood and voice** for English, German and French (Ramm et al., 2017a). The annotation is based on the dependency trees of the input sentences and a set of **morpho-syntactic language-specific annotation rules**. Depending on the language, the rules include information such as lemma, morphological analysis and part-of-speech (POS) tag. The sequence of the POS tags within a given verbal complex plays the central role for the annotation rules. The thesis includes an exhaustive list of the English and German verbal complexes in terms of their POS sequences and morphological properties needed to distinguish between the different syntactic tense, mood and voice forms.

Verbs in the German NMT outputs The main topic of my research are verb-related problems in the German SMT outputs. However, since 2015, there is a new promising approach to MT, namely neural machine translation (NMT). To explore the importance of handling the verbs in the German NMT outputs, I carry out an analysis of the different German NMT outputs. I combine preordering with NMT which shows that **preordering hurts NMT quality** (Ramm et al., 2017b). While the German NMT outputs indeed have almost no positional errors, there are a few **contexts in which the verbs are erroneously inflected**. The identification and discussion of those contexts will serve as a basis for the future research with respect to the inflection of the verbs in German NMT outputs.

1.3. Road map

Machine translation Chapter 2 includes a brief introduction to phrase-based SMT. The introduction is focused on the problems which we aim at solving with methods described in this thesis, namely long-range reorderings, as well as inflection of the verbs. The Chapter presents the SMT submodels and indicates their properties which lead to the verb-related errors in English→German SMT.

Linguistic background The methods for handling the problems regarding the verbs in the German SMT outputs require a deep understanding of the linguistic properties of the verbs in English and German. These are presented and discussed in Chapter 3. First, the linguistic terms used throughout this work are introduced. Subsequently, the analysis of the linguistic phenomena relevant for this work is given. The analysis includes the description of the verb-related positional differences in the two languages (syntax), as well as inflectional properties of the verbs (morphology). A special attention is given to the data-driven bilingual analysis of tenses in English and German.

Reordering In Chapter 4, the method for dealing with the positional problems of the verbs in the German SMT outputs is described. First, the related work which describes different variants of preordering for SMT implemented for numerous language pairs is presented. Subsequently, a detailed description of the reordering method used in this work is given. Hereby, a thorough discussion of the developed reordering rules as well as of the crucial details regarding the implementation is presented.

SMT experiments with reordering Chapter 5 presents the evaluation of the preordering method described in Chapter 4. The method is evaluated in many different experimental setups in order to estimate its performance for different size of the training data and language models, with respect to different domains and approaches to compute automatic word alignment. Furthermore, we analyze the adequacy of preordering in a combination with different parsers which provide the underlying syntactic representation of the source language sentences used by the implemented preordering approach.

Inflection Chapter 6 is dedicated to the modeling of the verbal inflection in the English→German SMT. First, the handling of the verbal inflection is motivated. Afterwards, the related work is presented, whereby the relevant previous findings are grouped by the verbal morphological features (i.e., *agreement* and *tense/mood*). We then present

the classification-based method developed and implemented within this work which aims at correcting the inflection of the finite verbs in the German translations. In addition to the classification-based approach, we also present a parsing-based method for handling agreement errors.

SMT experiments with inflection The methods for modeling verbal inflection are evaluated in Chapter 7. We first investigate the potential improvement which may be gained by correcting the inflection of the finite verbs in the used test set, i.e., its German SMT output. Subsequently, we apply our automatic post-editing method to correct verbs in the German baseline translations and evaluate their corrected variants. In addition to the experiments with automatic post-editing of the German phrase-based SMT outputs, we also present experiments with factored SMT, particularly to investigate whether explicit tense/mood information in form of factors may help SMT to generate more appropriate German translations.

Verbs in English→German NMT The methods presented in Chapters 4 and 6 may also be combined with NMT. In contrast with SMT, NMT produces considerably better translations, also with respect to the German verbs. In Chapter 8, we present a thorough analysis of the German NMT outputs with respect to the verbs. We also present results for combining preordering with NMT. Regarding the inflection, we point to a few specific cases in which NMT has problems generating correct German tense forms.

Revisiting tense/mood in (machine) translation In Chapter 9, we give an analysis of tense and mood both in the monolingual, as well as the bilingual context. The analysis points to a number of different aspects which need to be taken into account when dealing with this complex problem. We specify contextual features which can be used to account for the respective aspects and give an overview of tools which automatically annotate texts with tense/mood related properties. The theoretical analysis, as well as the discussion of the availability of the annotations presented in Chapter 9 represent a solid basis for further research in this area.

Conclusion Chapter 10 includes a summary of the main findings of the work described in this thesis, as well as proposed future research directions.

2. Machine translation

Chapter 2 provides a short description of SMT whereby the focus lies on the presentation of the properties of SMT which may lead to problems which are handled in this thesis. In Section, 2.1, a general description of SMT is given. In Section 2.2, SMT is discussed with respect to the modeling of the word order. In Section 2.3, we then analyze SMT regarding the choice of the correct inflected forms. We automatically evaluate our systems in terms of the BLEU score which is introduced in Section 2.4. Finally, 2.5 summarizes the most important facts about SMT.

2.1. Statistical machine translation

The aim of the methods described in this thesis is to improve German translations generated by a standard phrase-based SMT system. SMT models are a log-linear combination of different submodels each of them used to model different aspects of the linguistic phenomena important for the process of the (automatic) translation.

As the name already suggests, the phrase-based SMT relies on the translation units consisting of word sequences, i.e, *phrases*. Translation phrase pairs are automatically extracted from a set of word-aligned parallel sentences, the so-called *training corpus*, as shown in Figure 2.1. Translation phrase pairs are not necessarily linguistically motivated. In fact, they rather capture sequences of the source and target words which are connected to each other by means of the automatically computed word alignment. Translation pairs are assigned with translation probabilities φ as shown in Table 2.1. The automatically computed translation scores reflect how often the given phrase pair has been seen in the training corpus. These scores are used in the translation step to choose between the different translation options of a single source phrase. The submodel which contains the phrase pairs along with their translation probabilities is called the *translation model* (TM).

The translation model solely cannot provide translations of sufficient quality. The languages differ in many aspects, one of them being the word order. For example,

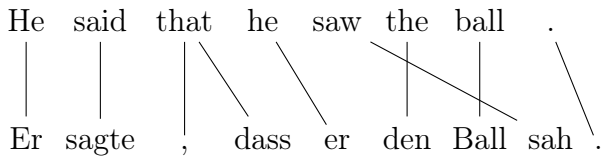


Figure 2.1.: Example of an English-German word-aligned sentence pair.

$\boldsymbol{e}$		$\boldsymbol{f}$	$\varphi(f\|e)$	$\varphi(e\|f)$
he saw the ball	$\leftrightarrow$	er den Ball sah	0.21	0.11
he	$\leftrightarrow$	er	0.53	8.02
saw	$\leftrightarrow$	sah	0.82	0.04
the ball	$\leftrightarrow$	den Ball	0.44	0.95

Table 2.1.: Excerpt of the translation pairs with example translation probabilities.

the English verbs are usually placed in the 2nd position in a sentence, while in many cases, in German, the verbs are at the clause end. When translating from English to German, this means that the position of the target language side of a given phrase pair does not correspond to the position of the source side of it. In other words, it is required to rearrange the target language phrases to achieve grammatical correctness of the generated translation. To cope with positional differences between source and target languages, SMT uses the so-called *lexicalized reordering model* (LRM) which describes by means of frequency distributions which types of phrase movements are required to generate correct target language sentences.

There are many different possibilities to split the source side sentences into phrases. Furthermore, typically, there are also many different possibilities to translate a single source side phrase. The decision how to segment the source side data and which of the translation variants to choose for the given source phrase is supported by the so-called target *language model* (LM). Language models are trained on the target language data and indicate correctness of the target language word sequences. In other words, the language models consist of n-grams of the target language words assigned with probabilities which express how probable it is that a specific n-gram is a valid word sequence in the target language.

As stated at the beginning of this section, SMT is a combination of different submodels, namely the translation model TM, the reordering model LRM and the language model LM. Mathematically, SMT is defined as shown in Equation (2.1). The log-linear combination of these models aims at generating the translation with the highest

probability e given the source sentence f.

$$\begin{aligned} p(e,a|f) = argmax_{e,a}\ & \lambda_{TM} log(P_{TM}(f|e)) \\ & + \lambda_{LRM} log(P_{LRM}(a)) \\ & + \lambda_{LM} log(P_{LM}(e)) \end{aligned} \tag{2.1}$$

Each of the models contributes to the estimation of how good the translation f matches the source e. In order to model the trustfulness of each of the models, the models are weighted with λ. The weights are automatically learned from the training data which is called *parameter optimization* or *tuning*. There are several methods for tuning the model weights. One of the most popular ones which is also used in this work, is *minimum error rate training* (Och, 2003). The idea is to find model parameters which lead to the least translation errors. The model weights are adjusted in a way that they maximize the translation probability of a small bilingual *tuning set*. Each time, a specific set of weights is assumed, the source side of the tuning set is translated and the translation quality of the generated translation in terms of BLEU[1] is computed. Weight adjustments proceeds until the translation quality of the tuning set cannot be improved further.

2.2. Word order within SMT

Each of the submodels within SMT is used to cope for a different mono-/bilingual language phenomenon. In the following, we describe how the order of the words in the source and target language sentences is reflected in each of the different SMT components.

2.2.1. Word alignment

The first step to train a SMT model is to compute alignment between source and target language words. Back in the 1990's, Brown et al. (1993) proposed 5 models, so-called IBM models, for automatic computation of alignment between words in a bilingual parallel text. Brown et al. (1993) already realized the importance of explicit consideration of the positional differences of the words in a bilingual text. Their IBM Model 2 thus incorporates a model which predicts the source word positions conditioned on the gener-

[1]The BLEU metric is explained in Section 2.4.

ated target word positions. The model does not incorporate any lexical information, but rather models position-based probability distribution derived from the parallel text. For computation of the model parameters, not only the word positions are considered, but also the respective sentence length is taken into account. Instead of the position-based alignment probability defined in Model 2, the IBM Model 3 uses the so-called *distortion probability distribution* which predicts target word positions based on the source side positions. IBM Model 4 improves the distortion model implemented in Model 3 by defining the *relative distortion model*. The underlying assumption is that the position of a generated target word particularly depends on the position of the previously translated source word.

In this work, we use two different tools for the automatic computation of the word alignment, namely *GIZA++* (Och and Ney, 2003) and *FastAlign* (Dyer et al., 2013). While GIZA++ includes the IBM Models 1 and 4, FastAlign is based on a modification of IBM Model 2. The different quality of the SMT models trained on the output of the two tools shows the importance of handling the word order at the very early stage of training an SMT model.[2] The simplicity of FastAlign has a very big impact on the speed, however, its performance is lower compared with the performance of Giza++. This indicates that explicit (and time-consuming) modeling of the word order differences between the languages is needed to achieve accurate word alignment and thus to increase the quality of the SMT models.

2.2.2. Translation model

Phrase-based SMT is based on translation phrase pairs composed of a sequence of arbitrary source and target language words. In other words, SMT is able to capture positional differences of the words within the extracted phrase pairs.

Assume we have a small parallel corpus and a set of extracted phrase pairs given in Table 2.2. We are particularly interested in the phrases containing the verbs, i.e., phrases including the English verb *saw* and the German verb *sah*. In the context of a subordinate clause, the German finite verbs are always placed at the end of a clause – in the example sentence (a), *after* the object noun phrase *'den Ball' (the ball)*. In English, the verb *saw* is placed *before* the object *'the ball'*. The translation phrase pair (1) contains both verbs and, even more importantly, it already captures the positional differences of the two verbs in the given context. Thus in the decoding process, if this

[2]This hypothesis is supported by the experiments which are presented and discussed in Chapter 5.

Training corpus		
	English	**German**
(a)	He said, that he saw the ball.	Er sagte, dass er den Ball sah.
(b)	He said, that he saw the ball under the black desk.	Er sagte, dass er den Ball unter dem schwarzen Tisch sah.
Translation phrase pairs extracted from (a)		
(1)	he saw the ball	er den Ball sah
(2a)	he	er
(2b)	saw	sah
(2c)	the ball	den Ball

Table 2.2.: Example of English-German phrase pairs derived from the given pair of sentences.

phrase pair is used, it is able to generate the verb *sah* in the German MT output in the correct position.

The capability of modeling positional differences with translation phrases is however limited. On the one hand, erroneous word alignment may lead to the extraction of the less appropriate translations pairs. On the other hand, the length of the phrases is normally limited to a certain number of the words (usually to 5 words). Positional differences including more words than the maximum length of the phrases cannot be captured within the phrases. More concretely, given the sentence pair (b) in Table 2.2 and the maximum phrase length of 5 words, it is not possible to extract a phrase pair which reflects the subject-object-verb order in the German subordinate clause. Even if the phrase length is set to a larger number, the data sparsity would ultimately lead to missing translation phrases since the training corpora are very unlikely to contain all of the possible word combinations for the given language pair.

There are many different translation phrases which can be derived from a parallel corpus (see translation pair set (2*) in Table 2.2). In the translation process, the model needs to find the optimal combination of the translation phrases to form the best target language sentence. In the case of SMT for English→German, this often fails, leading to translations with omitted or misplaced German verbs.

2.2.3. Language model

Language models (LMs) are built of word sequences (n-grams) extracted from the target language texts. For each n-gram, its probability is computed which indicates the appropriateness of the word sequence for the target language. The target language LMs

n-gram	probability
er sah den Ball	0.35
er den Ball sah	0.52

Table 2.3.: Examples of the German n-grams.

help the translation model to choose between the different translation possibilities.

LMs help to eliminate translation options which include non-valid or less appropriate target language word sequences. Given, for example, the German n-grams in Table 2.3: the SMT can use the probabilities to score the translation hypotheses. The sequence *'den Ball sah'* is very frequent in the data, so it is quite probable that the SMT would output the correct translation of the English input clause *'that he saw the ball'*. There is, however, a piece of very important information missing in the example n-grams: the highly probable word sequence can only be used in subordinate clauses. Thus, the dependency that the language model should model is the one between the conjunction *that* and the verb *sah* which is in this context to be placed at the clause end. Such dependencies can be even larger and can again not be captured by the language model which is typically restricted to the maximum length of 5 words. Furthermore, similarly to a TM, a LM cannot contain all possible n-grams of the target language words. Consequently, it cannot contain all possible combinations of the words in the target language.

2.2.4. Linear distortion cost model

A very first attempt to model movements of the translation phrases within the target language is based on the *distortion cost*. In general, the distortion cost penalizes all phrase movements regardless of their appropriateness. For languages with similar syntax, i.e., similar word order, one would want to penalize reorderings during the decoding. However, for syntactically different language pairs such as English and German, reorderings are required. For such languages, the distortion cost model is less appropriate and can cause severe word order problems in the MT output.

In addition to the distortion cost, usually, also the *distortion range* is limited, typically to 5 words. The distortion limit indicates the number of the words in the source language sentence which may be skipped when picking the next phrase which is to be translated. This kind of limitation has a negative impact on the long-range reorderings which would impose jumps larger than 5 words (see example (b) in Table 2.2 where *saw* should be translated at the very end). The distortion limit can theoretically be set to a larger number or even set to the unlimited number of the skipped words. Unfortunately, large

e		*f*	**mono**	**swap**	**disc-left**	**disc-right**
, he saw	→	sah er	0.52	0.16	0.32	0.18

Table 2.4.: Example lexical reordering table entry.

distortion limits lead to a drop in the translation quality (Koehn et al., 2007).

2.2.5. Lexicalized reordering model

A LRM learns the *orientation* of the phrases from the bilingual training data (Koehn et al., 2005). In contrast to the distortion cost, which generally penalizes any kind of phrase reordering, LRMs reward the reordering of the specific phrases supported by the training data (Koehn, 2010).

The model automatically learns whether the target side of the current phrase pair is to be placed in the original position with respect to the previously translated phrase, whether it is to be swapped or whether it is discontinuous. Each of the orientation types is assigned a frequency-based probability score derived from the training data. Consider Table 2.4 for an example lexicalized reordering table entry. The example shows the phrase pair *', he saw → sah er'* and its scores for different orientation types. For example, the model assigned the probability of 0.5 for the phrase being monotonically translated right after translating the phrase that the immediately preceding source word belongs to. The probability of swapping *', he saw → sah er'* before the previous phrase is 0.16. Assumed, we had a partial source sentence *'yesterday, he saw'* where *yesterday* was translated into *gestern*, then the probability of placing *'sah er'* after *gestern* following the monotonic orientation probability is much higher than producing the sequence *'sah er gestern'* according to the swap orientation probability.

The lexicalized reordering model consists of the translation phrases which are enriched with information about their orientation type. Due to the limitation of the phrase length, the phrases may contain insufficient information for a certain type of reordering.

2.3. Verb inflection within SMT

Inflectional variants are only indirectly captured within SMT.[3] Typically, the SMT sub-models are trained on the fully inflected parallel data which means that particularly the

[3]We discuss here verbal inflectional variants, however the discussion applies to all inflectional word categories: nouns, adjectives, articles pronouns and verbs.

Training corpus		
	English	**German**
(a)	He said, that he saw the ball.	Er sagte, dass er den Ball sah.
(b)	He said, that we saw the ball under the black desk.	Er sagte, dass wir den Ball unter dem schwarzen Tisch sahen.
Translation phrase pairs extracted from (a)		
(1)	he saw the ball	er den Ball sah
(2)	that we	dass wir
(3)	he	er
(4)	saw	sah
(5)	saw	sahen
(6)	the ball	den Ball
(7)	under the black desk	unter dem schwarzen Tisch

Table 2.5.: Example of English-German phrase pairs derived from the given pair of sentences.

translation model is composed of the translation phrase pairs which are built on the sequences of the inflected words. For example, we might have a translation pair *'he works - er arbeitet'* in which the verbs *'(to) work'* and *arbeiten*, respectively, are inflected in a way that the agreement between the verb forms *works* and *arbeitet*, respectively, match the corresponding subject. Since the phrase pairs are extracted from parallel corpora containing grammatically correct sentences, we assume that the inflection of the words within a single translation phrase is correct. The difficulty however arises when translation pairs are combined with each other. In many cases, specific dependencies regarding inflectional variants are shared across the phrase boundaries. When combining phrases together, i.e., while generating the translation phrase-by-phrase, it may happen that phrases are chosen in which the inflectional dependencies are violated.

Take, for example, translation phrases given in Table 2.5. Given the small example parallel corpus, we may extract two different translation possibilities for the English verb *saw* as shown in (4) and (5). Assume we want to translate the English sequence *'that we saw the ball'*. We might split the English input into the following phrases: *'that we'*, *'the ball'* and *'saw'*. Given the target language sequence *'dass wir den Ball'*, it is hard for SMT to choose the correct translation for *saw*: in the given context, only the variant *sahen* is correct while the generation of *sah* leads to an incorrect German output.

Due to the limitation of the phrase length, many contextual dependencies cannot be considered by the SMT translation model. LMs are a helpful device for tackling the inflectional problems, however, they are also limited to a certain maximum number of

words building the n-grams which means that especially long-distance dependencies are not captured by the model.

Another problem for all SMT submodels is *data sparsity*: the training data is very unlikely to contain all target language inflectional variants. To recall, a TM is trained on parallel data. It thus contains only words and word sequences which have been seen in the training data. The translation model is not able to generate words, particularly inflectional variants, which were not present in the used training set. This is a big drawback of the standard SMT models: in such cases, one of the variants seen in the training data is chosen which with respect to the given context may (by chance) be correct or may not be correct.

2.4. Automatic evaluation of MT

BLEU (**bil**ingual **e**valuation **u**nderstudy) is a method for automatic evaluation of the quality of the MT outputs (Papineni et al., 2002). BLEU has been developed to speed-up the development of the SMT models by allowing for automatic estimation of the quality of the SMT outputs. BLEU relies on the n-gram similarity between a MT output and a reference translation. In other words, BLEU measures the overlap between machine and human translations.

BLEU is based on a *modified n-gram precision*. The MT output, as well as the reference translation are split into n-grams typically up to the length of 4 words. For each n-gram of the length n, the modified precision p_n is computed by considering the counts of the given n-gram found in all translations C of the test set $Candidates$. The final BLEU score p_n is computed as given in Equation (2.2). Thereby, the total count of an n-gram is *clipped* to the maximum number of that n-gram found in any of C.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{\text{n-gram} \in C} Count_{clip}(\text{n-gram})}{\sum_{C' \in \{Candidates\}} \sum_{\text{n-gram'} \in C'} Count_{clip}(\text{n-gram'})} \tag{2.2}$$

The modified n-gram precision is combined with the *brevity penalty* BP. As shown in Equation (2.3), BP compares the length of the candidate translation c with the length of the reference translation r and penalizes translations which are shorter than the reference.

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{\frac{1-r}{c}} & if \quad c \leq r \end{cases} \tag{2.3}$$

The resulting BLEU score is a combination of p_n and BP as shown in Equation (2.4) where N is the n-gram length and the weights w_n are uniformly distributed ($w_n = 1/N$).

$$BLEU = BP \cdot exp(\sum\nolimits^{N}_{n=1} w_n \; log \; p_n) \tag{2.4}$$

Although BLEU is still the most popular metric for the automatic evaluation of the machine translations, BLEU also has problems in appropriately scoring small, but meaningful improvements of the MT outputs. For example, in this thesis, we present methods for improving the position and inflection of only a small amount of words in a sentence, namely verbs. Such small differences between the different SMT outputs are barely measurable with BLEU: the improvements in terms of BLEU are very small. On the other hand, having a verb in an MT output or not has a very big impact on understanding the translations. Hence, our evaluations do not only include automatic evaluation with BLEU, but also manual analysis of the translations which is focused on examining the parts of the translations that the presented work aims at correcting, namely verbs.

2.5. Chapter summary

This Chapter presented components of standard phrase-based SMT with respect to modeling of the word order, as well as to modeling of the inflectional variants.

Both of the linguistic phenomena are captured within the different SMT submodels. For instance, the translation pairs which are part of a translation model contain sequences of the inflected source-target language words. The sequences reflect the order of the words typical for the target language which also needs to be given in the generated translations. The lexicalized reordering model explicitly models the position of a target side of the translation pairs by providing probabilities about their placements with respect to the previously translated phrase. Not only the order of the words is important, but also correct inflected forms of the words. Translation pairs are extracted from the training data and thus consist of sequences of the correctly inflected target language words. Finally, the target language model trained on large sets of well-formed mono-

lingual data helps to distinguish between appropriate and less appropriate translation options with respect to the validity of the generated target language words.

SMT however comes along with different limitations which decrease the ability of coping for different word order and inflectional variants between a source and a target language. For instance, the length of both the translation phrases, as well as of the target language n-grams is limited to a rather small number of words. This means that dependencies between words which span large number of words cannot be modeled properly. Even if the phrase/n-gram length is increased, the problems remain because the derived statistics are often inaccurate and can even lead to lower translation quality. The result is translations with erroneous word order or translations in which words are often omitted. Besides the limitations of SMT models, we also face the problem of data sparsity: SMT can solely generate target language words which have been seen in the training data. This means that inflectional variants which would be needed to generate well-formed translations are simply not available to the SMT model. This fact often leads to translations in which many word dependencies are violated.

This work presents methods which help to reduce errors in the German SMT outputs caused by specific linguistic properties of English and German. Both positional, as well as inflectional problems mentioned above are dealt with by reducing differences between English and German in terms of syntax and morphology. On the one hand, we modify English in a way that the positional differences of the English-German parallel words are minimized which is beneficial for SMT since considerably less reorderings, especially long-range reorderings, need to be performed. Details on this are given in Chapter 4. In addition to positional differences, we also reduce inflectional differences by *stemming* the German side of the data. The stemmed German outputs are inflected in a post-processing step which includes a context-sensitive derivation of the morphological features for the generated stems and a subsequent generation of the inflected German words. By doing this, SMT has considerably less target language variants to choose between for a given English word. Details on this method are given in Chapter 6.

The main focus of the present work are German verbs. Thus, the elimination of both positional, as well as inflectional differences is related to the verbs in English (as a source language) and German (as a target language).

3. Linguistic background

This Chapter presents linguistic properties of the verbs in English and German in the monolingual, as well as in the bilingual context. Particularly the differences between the verbs in the considered language pair are discussed in more detail since they often lead to the errors in English→German SMT which we aim to correct with methods described in Chapters 4 and 6. Note that the aim of this Chapter is to identify and to describe the most prominent verb-related differences: infrequent cases which often depend on semantic/pragmatic features to which the developed methods do not have access are not studied further in the present work.

In Section 3.1, we first introduce the most important linguistic terms used in this work. In Section 3.2, we take a deeper look into the English verbal complexes and define notions which are used in the following chapters to refer to the specific verbs and verb sequences. In Section 3.3, the positions of the verbs in English and German are discussed both in the monolingual, as well as in the bilingual context. In Section 3.4, we present inflectional properties of the verbs in the two languages. The discussion is focused on the inflection features person and number (agreement), as well as on tense and mood. Since the modeling of the inflectional features tense and mood includes deeper understanding of the use of tense and mood, Section 3.4 also includes a data driven analysis of the use of the syntactic tense and mood forms in English and German. Note that most of the discussed morphological and syntactic properties in German are based on (Eisenberg, 1998). To increase readability of the text, we however omit the repetitive usage of the respective citation throughout this Chapter.

3.1. Terminology

This Section introduces terms which are used throughout the present thesis. The terms are defined in a way that they allow to refer to specific parts of the English and German verbal complexes. The examples used to illustrate the terminological notions contain verbs highlighted in different colors: light blue refers to finite verbs, violet refers to

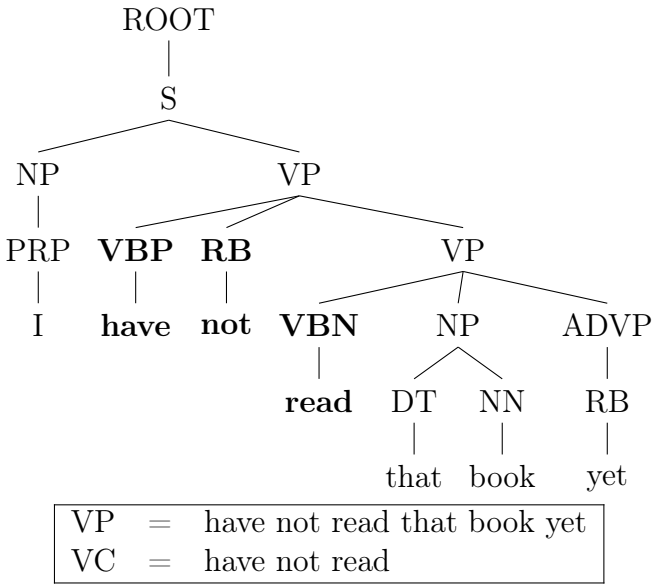


Figure 3.1.: The VP corresponds to the top VP node in the given parse tree. The corresponding VC of the type *composed* includes the verbs *have, read*, as well as the negation particle *not*.

non-finite verbal complexes, pink is used to denote non-finite verbs within a finite verbal complex. Finally, brown indicates negation particles.

3.1.1. Verbal phrase and verbal complex

In the linguistic theory, a verb phrase (VP) denotes the syntactic phrase of a sentence which contains at least one verb, as well as the verbal dependents such as objects, adjuncts and adverbials. A subunit of a verbal phrase which includes verbal elements of a VP is called a verbal complex (VC). In this work, we use the notion of a VC not only to refer to the verbs within a VP, but also to refer to three different types of particles:

(i) the negation (*not* and *'t* in English and *nicht* in German),

(ii) verb particles (e.g., *out, back* in English or *ab, vor* in German),

(iii) the infinitival particle (*to* in English and *zu* in German).

The difference between a VP and a VC is illustrated in the syntactic tree in Figure 3.1: the VP spans the words *'have not read that book yet'*, while the corresponding VC consists of the words *'have not read'*. An example of a VC containing a verb particle is given in Figure 3.2. A VC including an infinitival particle *to* is shown in Figure 3.3 where *to* is a part of the non-finite VC *'to buy'*.

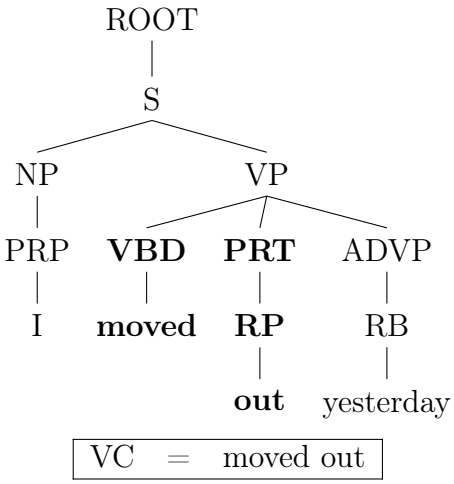
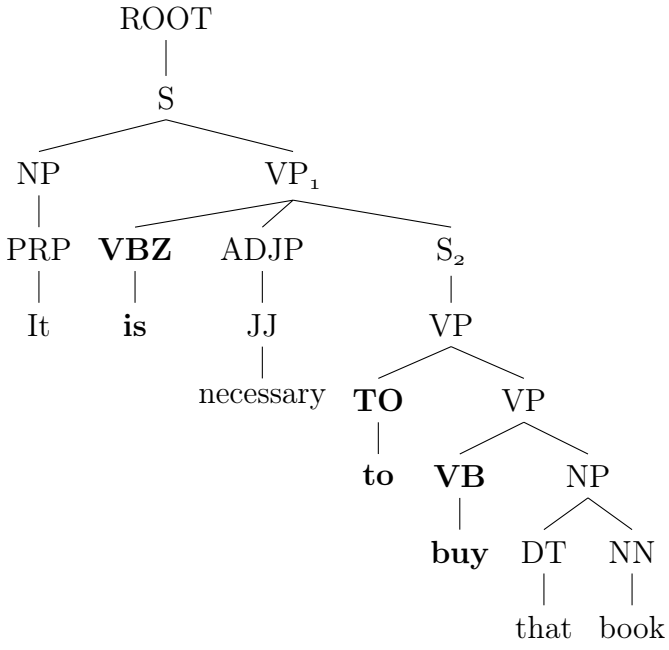


Figure 3.2.: Example of a simple VC with the verb *moved* and the verbal particle *out*.



VC	finite	non-finite	main
(VC_1) is	is	–	is
(VC_2) to buy	–	buy	buy

Figure 3.3.: Example of a composed sentence with two clauses each of them containing one verbal complex. VC_1 is a finite simple VC containing the finite verb *is*, while VC_2 is a non-finite composed VC with the infinitival particle *to* and the verb *buy*.

3.1.2. Finite, non-finite and main verb

Inflected verbs, i.e., verbs which expose agreement, as well as tense and mood features are referred to as finite verbs. For example, in the VC *'have not read'* in Figure 3.1, the auxiliary *have* is a finite verb, while *read* is a non-finite verb. Non-finite verbs include participles (e.g., *bought*), gerunds (e.g., *buying*) and infinitives (e.g., *(to) buy*).

A main (or full) verb is a verb which carries the meaning. It can be both finite, as well as non-finite. In the VCs with a single verb, the only occurring inflected verb is also the main verb which is the case for VC_1 in Figure 3.3. In VCs with more than one element, the main verb is typically non-finite. This also holds for the non-finite constructions as illustrated by the VC_2 in Figure 3.3.

3.1.3. Morphological and syntactic tense

In this work, we distinguish between two different categories of the notion of tense: morphological and syntactic tense. Morphological tense is a morphological feature of a finite verb, while syntactic tense refers to the tense expressed by the entire VC. Examples in (1) illustrate the difference between morphological and syntactic tenses in English and German. Given the English sentence in (1a), the morphological tense *present (Pres)* is marked on the auxiliary *has*, while the syntactic tense *present perfect progressive* refers to the whole VC. The same also holds for German. For instance, the finite verb *hatte* in (1b) has the morphological tense *past* while the syntactic tense of the VC *'hatte gelesen'* is *Plusquamperfekt*.[1]

(1) a. The boy has$_{Pres}$ been reading a book. → *present perfect progr.*
Der Junge hat gewesen lesend das Buch.
'Der Junge hat$_{Pres}$ ein Buch gelesen.' → *Perfekt*

b. The boy had$_{Past}$ been reading a book. → *past perfect progr.*
Der Junge hat gewesen lesend das Buch.
'Der Junge hatte$_{Past}$ ein Buch gelesen.' → *Plusquamperfekt*

Note that there is also the notion of a *logical* tense which refers to the time expressed by the given VC with respect to a specific reference time point. Logical tense is only marginally mentioned in this thesis and does not undergo deeper analysis (unlike morphological and syntactic tense).

[1] The full set of the morphological and syntactic tenses in English and German is presented in Section 3.4.3.

3.2. Definitions

In this Section, we introduce terms which we use to refer to the specific parts of the English verbal complexes. The main goal of the definitions is to establish the structural parallelism between the English and German VCs which is crucial information for the reordering method described in Chapter 4.

3.2.1. Main verb complex

In English, there are tense forms such as past progressive which take more than one auxiliary. An example is given in (2), where the English VC consists of the sequence *'had$_{AUX}$ been$_{AUX}$ reading'*. The auxiliary *had* is treated as a finite verb, the gerund *reading* is the main verb, while the non-finite auxiliary *been* belongs to a special group of verbs which are neither inflected (finite) nor considered to be main verbs.

(2) The boy had been reading a book.
Der Junge hat gewesen lesend das Buch.
'Der Junge hat ein Buch gelesen.'

We refer to this group of the verbs as *non-finite auxiliaries* and treat them as a part of the corresponding main verb, although they do not have a semantic meaning. We denote the English verb sequences consisting of a main verb and non-finite auxiliaries as a *main verb complex*. By doing this, we assume the correspondence between the parts of the English and German VCs as indicated in Example (2). The English finite verb *had* given in blue corresponds to the German finite verb *hat*, while the English main verb complex *'been reading'* indicated in violet corresponds to the German main verb *gelesen*.

3.2.2. VC and tense form types

In this work, we distinguish between *simple* and *composed* tense forms and VCs. Tenses[2] which are built with only one verb (e.g. simple present or simple past in English, and *Präsens* and *Präteritum* in German, respectively) are called *simple tenses*. Accordingly, the corresponding VCs are called *simple VCs*. An example of a simple VC is given in Figure 3.2 where the VC consists of a single verb, namely *moved*. Tenses which need at least one auxiliary along with a meaning bearing verb are called *composed tenses*, while

[2]Section 3.4.3 discusses the different tense forms in English and German.

the corresponding VCs are called *composed VCs*. Parse tree in Figure 3.1 provides an example of a composed VC which consists of the auxiliary *have* and the participle *read*.

Generally speaking, all VCs containing at least two verbs are composed VCs, regardless of the actually given tense form. However, we make a few exceptions to this rule. The first one is related to the English VCs in the present continuous tense as shown in Example (3).

(3) The boy is reading a book. → VC = *simple*
Der Junge ist lesend ein Buch.
'Der Junge liest ein Buch.' → VC = *simple*

In German, there is only one present tense, namely *Präsens*, which belongs to the group of simple tenses. To establish the structural parallelism of the English and German VCs in the present tense, we consider the English auxiliary in the present continuous VCs to be a part of the corresponding finite main verb complex. Hence, the English verb sequence in Example (3) *'is reading'* builds a simple VC which is assumed to correspond to the German finite verb *liest*.

The same definition of the English VCs and their verbs is also applied on the interrogative and negated VCs in the simple present tense. Both of these constructions in English take an auxiliary which does not exist in German. Hence, in Example (4) containing interrogative VCs, we assume the verb sequence *'does read'* to be the finite main verb complex. Similarly, the sequence *'does read'* in Example (5) with a negated VC is considered to be the finite main verb complex and as such to correspond to the German finite, main verb *liest*.

(4) Does the boy read a book? → VC = *simple*
∅ der Junge lesen ein Buch?
'Liest der Junge ein Buch?' → VC= *simple*

(5) Does the boy not read the book? → VC = *simple*
∅ der Junge nicht lesen das Buch?
'Liest der Junge nicht das Buch?' → VC = *simple*

3.2.3. Discussion

Definitions presented in the preceding subsections aim at aligning the English verbs with their German counterparts. The crucial question is how to split the English VCs into subparts which ensure the highest degree of the structural parallelism between the English and German VCs. The task directly implies the mapping of the English tenses

(1a)	The boy	had	a book	been reading.
	Der Junge	*hat*	*ein Buch*	*gelesen.*
(1b)	The boy	had been	a book	reading.
(1c)	The book	had	by a boy	read been.
	Das Buch	*war*	*vom Jungen*	*gelesen worden.*
(1d)	The book	had been	by a boy	read.

Table 3.1.: Possible splittings of two different English VCs to establish the structural equivalence with their German counterparts. The verbs in the English sentences are placed according to the German syntax to illustrate the equivalence between the verbs in English and German postulated in this thesis.

to the German ones because the tense forms in the two languages are reflected in the syntax (and morphology) of the verbs and the verbal complexes.

While the handling of the English present continuous VCs as shown in Example (3) is rather trivial since we expect the English present continuous to be translated into the German *Präsens*, there are also tenses which may be translated in different ways. Take, for example, the English negated simple past tense shown in Example (6): it may be translated into the the German *Präteritum* (simple tense), as well as into the *Perfekt* or *Plusquamperfekt* (composed tenses). The negated English simple past tense is composed of an auxiliary and an infinitive. As such, it has the same syntax as the German *Perfekt* tense. We might consider *did* and *read* as a main verb complex which would lead to the parallelism with the German tense *Präteritum*. However, we decide to keep the structural parallelism of the negated English simple past tense with the German composed tense *Perfekt* since *Perfekt* is the most often used past tense form in German. This might lead to the overgeneration of *Perfekt* in the German translations compared to *Präteritum* – a stylistic problem which we consciously take into account since both tense forms are valid translations of the English simple past tense.

(6) The boy did not read a book. → VC = *composed*
Der Junge hat nicht gelesen das Buch.
'Der Junge hat das Buch nicht gelesen.' → VC = *composed (Perfekt)*
'Der Junge las das Buch nicht.' → VC = *simple (Präteritum)*
'Der Junge hatte das Buch nicht gelesen.' → VC = *composed (Plusquamperfekt)*

In the preceding subsection, we introduced the notion of a main verb complex which denotes the sequence of a main verb and non-finite auxiliaries. Attaching the non-finite auxiliaries to the main verb aims at not only establishing the structural similarity

between the English and German VCs, but also at helping to disambiguate the highly ambiguous English main verbs. Let us consider again the sentence pair given in (2) on page 27. Possible splittings of the VC in the English sentence are shown in Table 3.1.[3] The splitting proposed in this section is given in (1a) and (1c), while the second possible splitting variant is given in (1b) and (1d), respectively. The main verbs in the variants (1b) and (1d) are highly ambiguous: they correspond to a number of different German forms of the verb *lesen (to read)*, some of them being finite, while the others may be non-finite.

In the context of the SMT, it is quite probable that *reading* and *read* in variants (1b) and (1d) are translated incorrectly due to the lack of the disambiguating context. On the other hand, given the splittings in (1a) and (1c), the SMT model learns that the sequence of a gerund and the participle *been* is very likely to be translated into the corresponding German participle. Furthermore, in (1c), it is required to generate the German participle *worden* which indicates the passive voice. It is very probable that *worden* is omitted when translating the variant (1d).

The same considerations may also be applied to the English finite verbs. For example, it might be problematic for a SMT model to translate *had* in (1c) into *war (was)*. If we wanted to disambiguate both the finite verb, as well as the main verb in examples (1c) and (1d), we might want to duplicate *been* and consider it both as a part of the finite verb, as well as of the main verb complex. In the context of reordering described in Chapter 4, such modifications of the English sentences would require a highly context-sensitive insertion of the words which we do not do in this work. We instead rely on the ability of an SMT model to generate correct auxiliaries for the German composed tense forms.

3.3. Position of the verbs in English and German

This section describes the positions of verbs in English and German. The discussion is focused on the most frequent verb placements. Alternative placements, especially in German, are mentioned but not discussed in detail since they are not considered in the development of the reordering rules which are presented in Section 4.3.1. The verb positions are illustrated with bilingual examples which indicate the differences regarding the verb placement between the two languages.

[3]The parts of the VCs are placed into the positions typical for German. The method for establishing the positional equivalence of the English and German verbs is described in Chapter 4.

3.3.1. English

English is a fixed word-order language in which the word order shows the relationship between the sentence constituents (**?**). It belongs to the group of the SVO languages with the word order corresponding to the pattern *subject-verb-object* as shown in Table 3.2.

$\mathbf{V}_{fin}$	**S**	$\mathbf{V}_{fin}$	**V**	**O**
∅	The boy	is	reading	a book.
Is	the boy	∅	reading	a book?

Table 3.2.: Position of the sentence constituents in English. V_{fin} = finite verb, S = subject, V = verb (complex), O = object.

In declarative sentences, the subject (S) is not placed directly in front on the main verb (V), but in front of the entire VC which may contain several verbs. In interrogative clauses, the finite verb (V_{fin}) is placed in front of the subject (S), while the remaining part of the VC (V) is placed after the subject. This word order is called *inverted word order* which is also possible with verbs of reporting (e.g. *ask*, *say*, *reply*, etc.) in the context of the direct speech where the quoted clause is put at the beginning of the sentence. This is shown in Example (7a) in which the verb *asked* is placed before the subject *'his mother'*. This kind of inversion is however not obligatory: the original SVO order is also possible as shown in Example (7b) where *asked* is placed after the subject.

(7) a. "Is the boy reading a book?" asked$_{V_{fin}}$ [his mother]$_S$.
"Ist der Junge lesend ein Buch?" fragte seine Mutter.
"'List der Junge ein Buch?", fragte$_V$ [seine Mutter]$_S$.'

b. "Is the boy reading a book?" [his mother]$_S$ asked$_{V_{fin}}$.
"Ist der Junge lesend ein Buch?" seine Mutter fragte.
"'List der Junge ein Buch?", fragte$_V$ [seine Mutter]$_S$.'

A verbal complex in English can be interrupted by adverbs which also enclose the negation. The negation *not* is placed within the VC, after the finite verb in declarative clauses as shown in Example (8a), or after the subject in clauses with inverted word order as illustrated in Example (8b).

(8) a. [The boy]$_S$ is$_{V_{fin}}$ not reading$_V$ a book.
Der Junge ist nicht lesend ein Buch.
'[Der Junge]$_S$ list$_V$ das Buch nicht.'

b. Is$_{V_{fin}}$ [the boy]$_S$ not reading$_V$ a book?
Ist der Junge nicht lesend ein Buch?
'List$_V$ [der Junge]$_S$ das Buch nicht?'

In English, the non-finite VCs follow the SVO rule as shown in Example (9): the subcategorizing clause, namely *'The boy wants'*, can be seen as the subject of the subordinated non-finite VC *'to read'*. Additional constituents are then placed after the verb according to the SVO rule.

(9) [The boy wants]$_S$ [to read]$_V$ a book.
Der Junge will zu lesen ein Buch.
'[Der Junge]$_S$ will$_{FinV}$ ein Buch lesen$_V$.'

3.3.2. German

In contrast to English, the verbs in German may have different positions within a sentence. To explain the differences, we make use of the *topological fields theory* which is commonly used to describe the German syntax. We concentrate on the verbs and give a bilingual comparison of their positions in German and English.

Topological fields The German syntax is usually described using the topological fields theory which originates in the description of the German sentence structure proposed by Drach (1963). Drach (1963) introduced the terms *Vorfeld (pre-field)* (VF), *Mittelfeld (middle-field)* (MF) and *Nachfeld (post-field)* (NF) which can be filled with specific sentential material. His theory has further been developed, mainly by introducing two additional terms, namely *linke Satzklammer (left sentence bracket)* (LSK) and *rechte Satzklammer (right sentence bracket)* (RSK) which are used to mark boundaries between the fields.

Table 3.3 illustrates the division of the German sentences into the topological fields. Note that the topological fields theory cannot directly be applied on English. However, we also give the corresponding English sentences in order to allow non-German readers to understand the division of a German sentence into topological fields and also to directly indicate the differences between English and German regarding the position of the verbs.

The VF may be occupied by any sentence constituent, although it is most commonly filled by the subject (Dürscheid, 2012). The MF can contain an arbitrary number of sentence constituents (noun phrases (NPs), prepositional phrases (PPs), adverbials).

	VF	LSK	MF	RSK	NF
(1)	Der Junge *The boy*	las *read*	ein Buch *a book*	– –	als ich nach Hause kam. *when I came home.*
(2)	Der Junge *The boy*	hat *has*	ein Buch *a book*	gelesen *read*	als ich nach Hause kam. *when I came home.*
(3)	– –	als *when*	er nach Hause *he home*	kam. *came.*	
(4)	– –	als *when*	er nach Hause *he home*	gekommen ist. *come has.*	

Table 3.3.: Topological fields in German. The main clause *'Der Junge las ein Buch'* analyzed in rows (1) and (2). The subclause *'als ich nach Hause kam'* placed in the NF can itself be split into the different fields similarly to the main clause. The analysis of the subclause is shown in rows (3) and (4).

Clause type	Context constraints		
	VF	LSK	RSK
V1	empty	finite verb	empty or filled with non-finite verbs
V2	not empty	finite verb	empty or filled with non-finite verbs
VE	empty	conjunction, relative pronoun, wh-word	finite/non-finite verbs

Table 3.4.: Type of the German clauses with respect to the position of the verbs.

The sentential complements are placed in the NF (subordinate and relative clauses).[4] The sentence brackets can only be filled with verbs and clause introducing conjunctions. The LSK must always be filled (cf. *las* in Example (1) and *als* in Example (3) in Table 3.3), while the RSK can also be empty as shown in Example (1) in Table 3.3.

The verbal elements are generally placed in one of the sentence brackets. The simple German VCs are placed either in the LSK (cf. Example (1) in Table 3.3) or in the RSK (cf. Example (3) in Table 3.3). A complex VC is either placed in the RSK (cf. Example (4) in Table 3.3) or the finite verb is in the LSK, while the non-finite verb(s) are in the RSK (cf. Example (2) in Table 3.3). In the latter case, we have a *discontinuous VC*, where the constituents placed in the MF (e.g., subject and objects NPs, PPs, etc.) are placed between the parts of a German VC.

[4]For more details on the placement of constituents in the topological fields, please refer to e.g. (Dürscheid, 2012).

Clause type	VF	LSK	MF	RSK	NF
V1	∅	Las	der Junge	ein Buch?	
	∅	Hat	der Junge ein Buch	gelesen?	
	∅	Lies!	∅	∅	
	∅	Lies	das Buch,	∅	weil es lustig ist.
V2	Der Junge	liest	ein Buch,	∅	weil es lustig ist.
	Der Junge	hat	ein Buch	gelesen.	
VE	∅	weil	der Junge das Buch	liest.	
	∅	weil	der Junge das Buch	gelesen hat.	

Table 3.5.: Syntactic structure of the different German sentence types.

Depending on the position of the finite verb in a sentence, we distinguish between different clause types in German which are given in Table 3.4. For example, the type *V1* is characterized by the verb in the sentence-initial position. In other words, V1 type denotes an interrogative clause as shown in Table 3.5. Indeed, the position of the verbs in German depends on the type of a clause, as well as on the type of a VC they occur in. In the following, we thus describe the German verb positions, as well as the difference in verb placement between English and German based on the different clause types.

Declarative clauses The German declarative sentences belong to the group of V2 sentences. The finite verb is always placed in the LSK, thus in the 2nd position in a clause. When a composed tense is given, we have a discontinuous VC in which the finite verb is in the LSK, while the non-finite verbs are placed in the RSK. This means that the main verb is either at the sentence end or directly before the beginning of a subclause. In the English declarative clauses, the entire verbal complex is placed in the 2nd position, directly preceded by the subject.

The German finite verb placed in the LSK may be preceded by a subject NP, but also by an extraponed phrase such as PP or ADVP. In the latter case, these are then placed in the VF, while the subject is placed *after* the finite verb in the MF. This kind of subject–finite verb inversion does not exist in the English declarative sentences. Given this analysis, we identify four different patterns regarding the position of the verbs in the German declarative clauses and summarize them in Table 3.6.

Subordinate clauses Subordinate clauses begin with a conjunction, a relative pronoun or a wh-word. In German, they belong to the group of the VE sentences where the entire

Pattern	Syntactic structure of a clause			
(d1)	**SUBJ**	**finite verb**	∗	
	Der Junge	liest	ein Buch.	
	The boy	reads	a book.	
(d2)	**SUBJ**	**finite verb**	∗	**non-finite verb(s)**
	Der Junge	hat	ein Buch	gelesen.
	The boy	has	a book	read.
(d3)	∗	**finite verb**	**SUBJ** ∗	
	Seit 3 Stunden	liest	der Junge ein Buch.	
	For 3 hours	reads	the boy a book.	
(d4)	∗	**finite verb**	**SUBJ** ∗	**non-finite verb(s)**
	Vor 3 Stunden	hat	der Junge ein Buch	gelesen.
	3 hours ago	read	the boy a book	read.

Table 3.6.: Position of the verbs in the German declarative clauses. Asterisks are placeholders for arbitrary sentence constituents. SUBJ refers to the subject NPs. Position of SUBJ is explicitly given since in many cases, the ordering of SUBJ and the verbs follows specific rules which need to be considered in the development of the reordering method described in Chapter 4.

Pattern	Syntactic structure of a clause		
(s1)	**conj/rel/wh**	**SUBJ** ∗	**verbal complex**
	weil	der Junge ein Buch	liest/gelesen hat.
	because	the boy a book	read/has read.
	weil	der Junge	geschlafen hat.
	because	the boy	was sleeping.

Table 3.7.: Position of verbs in the German subordinate clauses.

verbal complex is placed in the RSK (at the clause end) as shown in Table 3.7. The finite verb is in most cases placed after the main verb.[5] As a SVO language, English does not allow for the position of the verbs at the clause end, at least as long as the given sentence contains additional material such as objects and adjuncts. In case of short sentences without constituents usually placed after the verb(s), the verb(s) are placed at the end.

[5]There are some specific verbal complexes in which the finite verb is placed before the non-finite verbs. For example, in the VCs with at least three verbs, one of them being *lassen*, (e.g. *..., weil ich ihn habe warten lassen*), the finite verb is placed before the non-finite verbs (Eisenberg, 1998). This kind of a special ordering of the verbs in German is not considered in this work.

Pattern	**Syntactic structure of a clause**		
(inf1)	(main clause)	$*$	**verbal complex**
	(Ich habe etwas)	mit dir	zu besprechen.
	(I have something)	with you	to discuss.

Table 3.8.: Position of verbs in the German infinitival clauses.

Pattern	**Syntactic structure of a clause**		
(i1)	**finite verb**	**SUBJ** $*$	
	Liest	der Junge ein Buch?	
	Reads	the boy a book?	
(i2)	**finite verb**	**SUBJ** $*$	**main verb**
	Hat	der Junge ein Buch	gelesen?
	Has	the boy a book	read?

Table 3.9.: Position of the verbs in the German interrogative clauses.

Infinitival clauses The infinitival clauses are a subgroup of the subordinate clauses. As such, they have the same position of the verbs like finite subordinate clauses, namely at the clause end. Nevertheless, for the sake of completeness, we also explicitly give the position patterns for the German infinitival clauses in Table 3.8.

Interrogative sentences The German interrogative sentences belong to the V1 sentences which means that the finite verb is placed at the sentence-initial position, in front of the subject. As shown in Table 3.5, if in an interrogative clause a composed tense is given, the main verb is placed at the clause end, so we have a discontinuous VC. The verb placements in the German interrogative sentences are summarized in Table 3.9.

3.3.3. Discussion

Flexible word order in German The preceding linguistic analysis of the verb positions in English and German takes the most common patterns into account. Especially in German, the word/constituent order in a sentence is flexible to a certain extent. In some cases, the word order in German depends on specific pragmatic characteristics such as emphasis. For example, while in English, every word can be emphasized regardless of its position, in German, the emphasized words tend to move from their normal positions towards the clause beginning or the clause end (Curme, 1964). An example is given in (10). Here, the emphasis is on the action of *reading*. To express emphasis, the verb *gelesen (read)* is moved to the sentence-initial position. In a non-emphasized context,

sentence-initial placement of a participle in German is not possible.

(10) Gelesen hat der Junge das Buch.
Read the has boy the book.
'The boy has read the book.'

In some cases, the German subclauses may expose SVO word order instead of SOV which we assume to be valid word order in the German subordinate clauses (cf. Table 3.7). An example is given in (11).

(11) a. Ich denke, dass der Junge das Buch gelesen hat. → SOV
I think, that the boy the book read has.
'I think that the boy has read the book.'

b. Ich denke, der Junge hat das Buch gelesen. → SVO
I think, the boy has the book read.
'I think, the boy has read the book.'

Example (11a) shows the most common word order in the German subordinate clauses, namely SOV where all verbs are placed at the clause end. Example (11b), on the other side, exposes the SVO word order which is typical for the German declarative clauses. This word order is possible when the conjunction (in our example *dass (that)*) is omitted. SVO order in the German subordinate clauses is possible only in a combination with specific verbs in the main clause and with specific conjunctions. Since in these exceptional causes, the SOV order is also possible (as long as the conjunction is used), we decide not to consider SVO as a valid placement of the verbs in the German subclauses. In other words, we assume that the verbs in the German subclauses are always placed at the clause end.

There are also some exceptions which are related to the relative, as well as non-finite clauses. An example of the flexible placement of the German non-finite VCs is given in (12). In (12a), the non-finite VC *'zu kommen'* is placed after the finite verb of the governing clause, namely *versprochen*. This ordering corresponds to the ordering of the English VCs. On the other hand, in (12b), the non-finite German VC is placed before the main verb of the governing clause. This ordering is possible only in very specific contexts. Since the ordering given in (12a) is also correct and more frequent, we assume that all non-finite VCs are placed *after* the verbs, i.e., the governing clause.

(12) a. The boy has promised to his sister to come.
Der Junge hat versprochen seiner Schwester zu kommen.
'Der Junge hat zu seiner Schwester versprochen zu kommen.'

b. The boy has promised to his sister to come.
Der Junge hat versprach zu seiner Schwester zu kommen.
'Der Junge hat seiner Schwester zu kommen versprochen.'

Order of the verbs in the German VCs German VCs may include a number of non-finite verbs. Usually, in the VE clauses, the finite verb is placed at the end of a VC, but there are also exceptions of this rule as illustrated in Example (13). The finite verb *hat* in (13a) is placed after the participle *gelesen* which corresponds to the most common order of the German verbs in the RSK. However, in case of a combination of specific verbs such as non-finite auxiliaries and modal verbs as shown in (13b), the finite verb *hat* has to be placed in front of the following non-finite verbs *'lesen wollen'*.[6]

(13) a. ... dass der Junge das Buch gelesen hat.
... that the boy the book read has.
'... that the boy has read the book.'

b. ... dass der Junge das Buch hat lesen wollen.
... that the has boy the book has read want.
'... that the boy wanted to read the book.'

We do not consider special cases of the ordering of the verbs in the RSK. Within the context of the phrase-based statistical machine translation, we expect the translation system to be able to handle short-range reorderings which are required to choose the correct ordering of the verbs within a VC at the end of a clause.

Clausal negation In this work, we also aim at establishing the positional parallelism of a negation particle in the English and German sentences. Hereby, we solely consider the clausal negation which is in English realized by negating the verbal phrase. By doing this, we make an assumption that the German translations also expose clausal negation which must not always be the case. Consider, for example, the sentences in Example (14). The first German translation option has the negated VC *'hat nicht gegessen'* which is structurally equal to the negated English VC *'did not eat'*. In this case, both sentences expose the clausal negation. On the other hand, the second German translation option consists of a VC without the negation. The negation is expressed by *keine (none)* and represents the constituent negation.

[6] The ordering of the non-finite verbs in the RSK underlie the government relations between the verbs. For more details, please refer to Wöllstein (2010).

(14) The boy did not eat beans. → *clausal negation*
Der Junge hat nicht gegessen Bohnen.
'Der Junge hat die Bohnen nicht gegessen.' → *clausal negation*
'Der Junge hat keine Bohnen gegessen.' → *constituent negation*

Handling different kinds of negation is a complex topic and out of scope of this work. In the further discussion, we focus only on the clausal negation in English where the negation particle is enclosed within the English VCs.

3.4. Verbal inflection in English and German

The German finite verbs expose the following morphological categories: person, number, tense and mood (Hentschel and Vogel, 2009). These categories are presented and discussed in the following sections.

The morphological features tense and mood are related to the *syntactic tense* and *mood forms* which is why we study tense and mood forms in English and German in this Chapter in more detail. The syntactic analysis, as well as the usage of the different tense/mood forms are discussed from the monolingual (German), as well as the bilingual perspective. The analysis does not represent an extensive comparison of the tenses in English and German, but is instead focused on the phenomena interesting for the statistical machine translation. Our observations are based on the statistics derived from the English–German parallel corpora from different domains.[7] The corpora are annotated with tense and mood information by a tool presented in Chapter 6, Section 6.3.3.

3.4.1. Person and number

Person and number are referred to as verbal agreement features whereby the agreement needs to be established between a finite verb and the corresponding subject as shown in Example (15). The agreement features of the finite verb *liest* in (15a) and the finite auxiliary *hat* in (15b), respectively, agree with the agreement features 3rd person singular (3.*Sg*) of the corresponding subject NP *'der Junge' (the boy)*.

The German verbal morphology differentiates between three person values: first (*1*), second (*2*) and third (*3*), and two number values: singular (*Sg*) and plural (*Pl*). The

[7]The corpora used for the monolingual, as well as bilingual tense/mood analysis are presented in Section 5.2 and summarized in Table 5.1 on page 99.

correspondence of the person/number combinations with the German verb suffixes is not unique: while some of the person/number combinations are expressed by unique verbal suffixes, there are also verbal suffixes which correspond to more than one person/number combination. The uniqueness of the suffixes depends not only on the agreement features, but also on the tense of the finite verb: in Example (16), the verb forms for the first and third person singular are equal, namely *las*. They are however different when the morphological tense of the finite verb is present (*Pres*) as shown in Example (17).

(15) a. [Der Junge]$_{3.Sg}$ liest$_{3.Sg}$ ein Buch.
The boy reads a book.
'The boy is reading a book.'

b. [Der Junge]$_{3.Sg}$ hat$_{3.Sg}$ ein Buch gelesen.
The boy has a book read.
'The boy read a book.'

(16) a. [Der Junge]$_{3.Sg}$ las$_{3.Sg.Past}$ ein Buch.
The boy read a book.
'The read a book.'

b. [Ich]$_{1.Sg}$ las$_{1.Sg.Past}$ ein Buch.
I read a book.
'I read a book.'

(17) a. [Der Junge]$_{3.Sg}$ liest$_{3.Sg.Pres}$ ein Buch.
The boy reads a book.
'The is reading a book.'

b. [Ich]$_{1.Sg}$ lese$_{1.Sg.Pres}$ ein Buch.
I read a book.
'I am reading a book.'

While in German, the verbs differ to a certain extent in different contexts, English is a highly syncretic (or morphologically poor) language regarding the verbal morphology. This means that a single English verb form (cf. *read* in Example (16)) may correspond to numerous German verb forms. In the context of automatic translation, choosing a false German verb form, i.e., a form which does not agree with the corresponding German subject in person and number, results in a grammatically incorrect German sentence.

3.4.2. Syntactic and morphological tenses in English and German

In English and German, the tense is expressed morphologically, as well as morpho-syntactically. Accordingly, we distinguish between morphological and syntactic tenses.

While the morphological tenses are morphological features of the *finite* verbs, the syntactic tenses refer to the syntactic structure of the English and German VCs.

Morph. tense	**English**		**German**	
	Synt. tense	**Example**	**Synt. tense**	**Example**
present	present simple	(I) read	Präsens	(Ich) lese
	present progressive	(I) am reading		
	present perfect	(I) have read	Perfekt	(Ich) habe gelesen
	present perfect progressive	(I) have been reading		
	future I	(I) will read (I) am going to read	Futur I	(Ich) werde lesen
	future I progressive	(I) will be reading (I) am going to be reading		
	future II	(I) will have read	Futur II	(Ich) werde gelesen haben
	future II progressive	(I) will have been reading		
past	past simple	(I) read	Präteritum	(Ich) las
	past progressive	(I) was reading		
	past perfect	(I) had read	Plusquam-perfekt	(Ich) hatte gelesen
	past perfect progressive	(I) had been reading		
present*	conditional I	(I) would read	Konjunktiv II	(Ich) würde lesen
past*	conditional I progressive	(I) would be reading		
	conditional II	(I) would have read	Konjunktiv II	(Ich) hätte gelesen
	conditional II progressive	(I) would have been reading		
present*			Konjunktiv I	(Er) lese (Er) werde lesen

Table 3.10.: List of the tenses in English and German in active voice. The table indicates the tense correspondences in terms of their morpho-syntactic structure.

The inventory of the German and English morphological, as well as syntactic tenses is given in Table 3.10. Both languages distinguish between two morphological tenses: present and past. None of the languages has a morphological marking for the future tense. The morphological tenses *past** and *present** given at the bottom of the table are treated in this work as a special category of the morphological tense because they are closely related to the mood of the finite verbs which will be discussed in Section 3.4.4.

The English syntactic tense set consists of sixteen forms, while the German one includes eight forms. The richness of the English tense set is due to the fact that the English syntactic tenses also include explicit marking of the aspect (perfect, progressive). In German, the aspect is not given within the VCs, but is expressed with other linguistic means such as prepositional or adverbial phrases.[8] Table 3.10 indicates that the different tense forms in the two languages may be built morphologically, as well as morpho-syntactically. Simple tenses such as the English present simple tense or the German *Präsens* are expressed by means of the verbal morphology, while the composed tenses (cf. Section 3.2.2 for details about simple vs. composed VCs and tense forms) such as present perfect or *Perfekt*, respectively, are built morpho-syntactically, i.e., they require a combination of specific verbs (syntactic structure) with a finite verb, typically an auxiliary, carrying a specific morphological tense (morphology).

3.4.3. Tense in German

3.4.3.1. Morphological and syntactic tense forms

As shown in Table 3.10, the German language has eight different morpho-syntactic tense forms. Six of them are indicative tense forms which are introduced in this section, while two forms are related to the subjunctive mood which is discussed in Section 3.4.4.

The German indicative tenses can be grouped by the logical tense as shown in Table 3.11. The table also contains information about the morpho-syntactic structure of the German tenses.[9]. While the correlation between the logical[10] and the syntactic tenses in German is not unique, each of the morpho-syntactic tenses requires a specific morphological tense of the finite verb. For example, *Perfekt* needs an auxiliary in present tense

[8] Note this issue is not further explored in this work since it is not expressed within the German verbal complexes which are the main topic of this thesis.

[9] We use these morpho-syntactic patterns to automatically annotate German (and English) tense forms (cf. Section 6.3.3).

[10] The logical tense is out of scope of this work and will not be discussed further in this thesis.

Log. tense	Synt. tense	Pattern	Example
present	Präsens	V*FIN.Pres.Ind	Ich lese.
		VMFIN.Pres.Ind V(A|V)PP	Ich kann lesen.
past	Präteritum	V*FIN.Past.Ind	Ich las.
		VMFIN.Past.Ind V(A|V)INF	Ich konnte lesen.
	Perfekt	VAFIN.Pres.Ind V*PP	Ich habe gelesen. Ich bin gefahren.
		VAFIN.Pres.Ind V(A|V)INF VMINF	Ich habe lesen können.
	Plusquam-perfekt	VAFIN.Past.Ind V(A|V)PP	Ich hatte gearbeitet. Ich war gefahren.
		VAFIN.Past.Ind V(A|V)INF VMINF	Ich hatte arbeiten können.
future	Futur I	VAFIN.Pres.Ind V(A|V)INF	Ich werde arbeiten.
		VAFIN.Pres.Ind V(A|V)INF VMINF	Ich werde arbeiten können.
	Futur II	VAFIN.Pres.Ind V(A|V)PP VAINF	Ich werde gearbeitet haben. Ich werde gefahren sein.
		VAFIN.Pres.Ind V(A|V)PP VAINF VMFIN	Ich werde gearbeitet haben können.

Table 3.11.: The German indicative morpho-syntactic tense forms with examples of the different realization possibilities for the active voice. POS tags correspond to the German STTS tag set.

(defined by *VAFIN.Pres.Ind* in Table 3.11), while *Plusquamperfekt* needs an auxiliary in the past tense (defined by *VAFIN.Past.Ind*).

In German, there are two simple tense forms, namely *Präsens* and *Präteritum*. They are composed of a single verb with the corresponding morphological tense. The remaining four tense forms belong to the group of the composed tenses. The composed tenses may express past, as well as future actions. The corresponding tense forms differ morpho-syntactically in two perspectives: (i) choice of the finite verb and (ii) form of the main verb. While the composed past tenses take the auxiliaries *haben (to have)* and *sein (to be)* (only in the combination with the verbs of moving such as *fahren (to drive), laufen (to walk/run)*, etc.), the future tenses are built with the auxiliary *werden (to become)*. With respect to the form of the main verb, the composed past tenses require a participle, while the composed future tenses take an infinitive form of the main verb.

It is interesting to note that within the simple VCs, changing the morphological tense of the verbs leads to a change of the logical tense which has a major impact on the

meaning of the given utterance. On the other hand, changing the morphological tense of a finite auxiliary in the context of *Perfekt* and *Plusquamperfekt* leads to the switch between two different past tenses which in German bear almost equal semantic/pragmatic information (cf. following Section). Finally, changing the tense of the auxiliary *werden* in the future tenses from *present* to *past* leads to the non-grammatical German VCs.

Table 3.11 contains only the most frequent German indicative VCs in active voice. There are more combinatory possibilities which are listed in Appendix A.1.

3.4.3.2. Use of tense in German

This section discusses a few specifics of the usage of the tenses in German. The monolingual analysis aims at identifying the tense-related phenomena which need to be taken into account not only in the monolingual context, but also in the context of explicit modeling of the translation of tense from an arbitrary source language into German.

Tenses are used to establish specific temporal relation between events. Grouped by a logical tense as shown in Table 3.11, one might assume that there is a clear boundary between the morpho-syntactic tenses with respect to the logical tense. This is however for German not the case. In specific contexts, the logical tense may be switched. One of the most prominent examples is the use of the *Präsens* tense instead of *Futur I* to express a future event. The temporal information is then expressed with other textual material such as adverbials (Collins and Hollo, 2010). An example is shown in (18) where in (18b) the adverbial *morgen (tomorrow)* in a combination with the VC *kommt* in the *Präsens* tense specifies the time point of the respective action.

(18) a. Der Junge wird morgen kommen. → *tense = Futur I*
The boy will tomorrow come.
'The boy will come tomorrow.'

b. Der Junge kommt morgen. → *tense = Präsens*
The boy comes tomorrow.
'The boy is coming tomorrow.'

In addition the the *Präsens-Futur-I* **interchangeability**, it is also possible to use *Präsens* instead of one of the past tenses for an event that took place in the past. An example is given in (19). Assumed that the store has opened prior to the time point of reporting on it, one would usually use one of the past tenses as shown in Example (19a). But in specific contexts such as news article titles, it is possible to refer to the same event using the *Präsens* tense as shown in (19b). The information from the sentences

in the surrounding context needs then to contain clues which specify the time of the reported event. To those may belong rephrasing of the given sentence using one of the past tenses, use of an appropriate temporal adverbial, etc.

(19) a. Das Geschäft hat seine Pforten geöffnet. → *tense = Perfekt*
The store has its gates opened.
'The store has opened.'

b. Das Geschäft öffnet seine Pforten. → *tense = Präsens*
The store opens its gates.
'The store has opened.'

The most prominent case of the tense interchangeability in German is related to the different morpho-syntactic tenses expressing the logical tense *past*. For an example, consider sentences given in (20) in which one of the sentence alternatives is in *Präteritum*, the second one is in *Perfekt*, while the last alternative is in *Plusquamperfekt*. Despite the different tenses, the meaning of all alternatives with respect to the time point of the reported action is the same, namely that *reading* took place prior to the time point of the utterance being made.

(20) a. Der Junge las ein Buch. → *tense = Präteritum*
The boy read a book.
'The boy read a book.'

b. Der Junge hat ein Buch gelesen. → *tense = Perfekt*
The boy has a book read.
'The boy read a book.'

c. Der Junge hatte ein Buch gelesen. → *tense = Plusquamperfekt*
The boy had a book read.
'The boy had read a book.'

The choice of the past tenses in German is not always clear. There are some fine-grained differences between the respective tenses, but at least *Präteritum* and *Perfekt* are interchangeable in many contexts (Sammon, 2002). The difference is more a **question of style**. In addition to the stylistic preferences, there are also **lexical preferences** with respect to the choice of a past tense. For example, the German auxiliaries and modals are more often used in *Präteritum* than in *Perfekt* as shown in Example (21).[11]

[11]For example, in a set of 250k sentences from the news domain, we identified 190 occurrences of the auxiliary *sein (to be)* in one of the composed past tense forms in active voice in contrast to 10,247 occurrences in the simple past tense *Präteritum*.

(21) a. Der Junge war in der Stadt. → *tense = Präteritum*
The boy was in the city.
'The boy was in the city.'

b. Der Junge ist in der Stadt gewesen. → *tense = Perfekt*
The boy is in the city been.
'The boy was in the city.'

The use of tenses in German does not underlay any strict rules for tense combinations within a sentence or a sequence of sentences. Which of the tenses is used in a specific situation depends to a large extent on the register, speakers/writers preferences, domain, etc. Weinrich (2001) differentiates between two groups of the German tenses: (i) *discussing tenses* to which belong *Präsens*, *Perfekt*, *Futur I*, *Futur II* and (ii) *narrative tenses* containing *Präteritum*, *Plusquamperfekt*, as well as subjunctive tense forms *Konjunktiv I* and *Konjunktiv II*.[12] He observed that in most German texts, one of the tense groups clearly dominates. He refers to this fact as to *Tempus-Dominanz*, i.e., tense dominance. Moreover, Weinrich (2001) observed that within specific texts parts, very often a single tense form is used. He called this effect *Tempus-Nester*, i.e., tense nesting, which indicates the dominance of a single tense form within specific text parts (e.g., paragraphs, sections). The tense classification defined by Weinrich (2001) may be applied on the **type and genre** of the German texts (or speech). For instance, the narrative tenses, as the name already suggests, are mostly found in the written German language (e.g., literary works), while the discussing tenses are more often used in the spoken language.

Analysis of the texts used in this work indeed indicates differences in the tense usage across domains and text types. Direct comparison of the tense frequencies extracted from the German data from different domains illustrated in Figure 3.4 shows the differences in usage of the tenses in texts coming from different domains. For all domains, *Präsens* is the most frequent tense form, however, its relative frequency differs across domains. For instance, in News the relative frequency of *Präsens* is 10% lower than in Europarl (0.66 vs. 0.75), while *Präsens* represents 97% of the tense forms in Pattr (medical texts). There is also a mismatch in using the past tenses within the respective domains. While, for instance, in Europarl, *Präteritum* and *Perfekt* have almost equal relative frequency (0.08 and 0.10, resp.), News clearly prefers the narrative tense *Präteritum* over the discussing tense *Perfekt* (0.19 vs. 0.08).

Although the preceding discussion may suggest that the use of the tenses in German

[12]Subjunctive tense forms are discussed in Section 3.4.4.

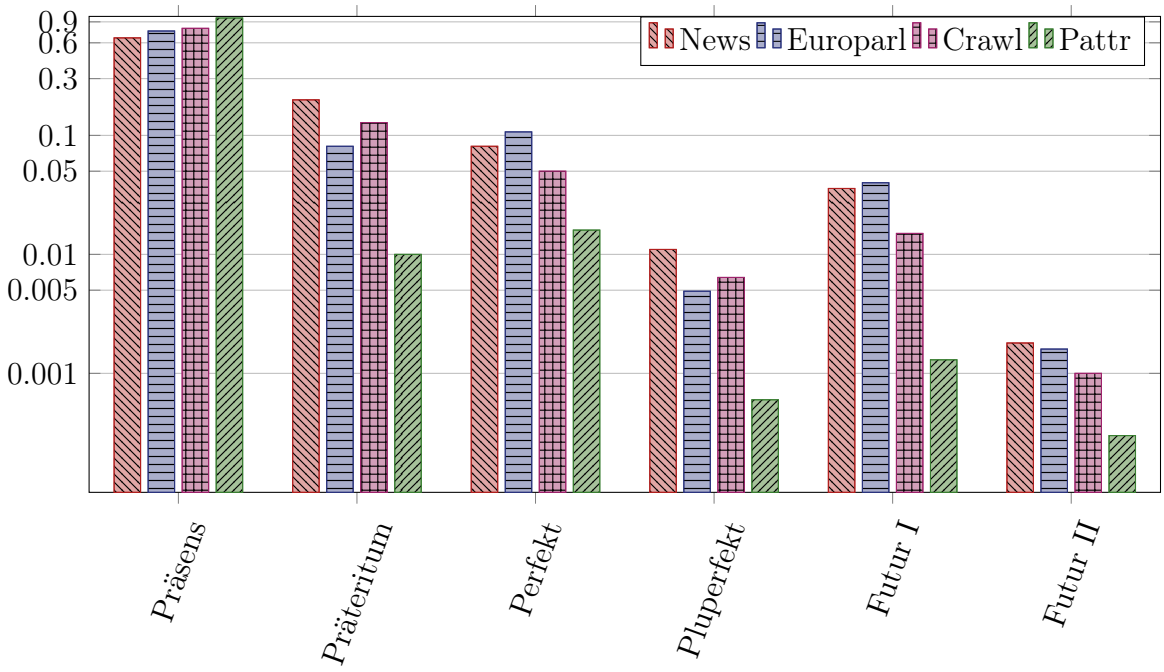


Figure 3.4.: Relative frequencies of the indicative active tense forms in four German corpora: (i) News, (ii) Europarl (political discussions), (iii) Crawl (mix-domain texts) and (iv) Pattr (medical texts).

is almost arbitrary, their usage does depend on specific contextual constraints. For example, PPs (e.g., *'vor zwei Tagen' (two days ago)*), NPs (e.g., *'letzte Woche' (last week)*) or adverbials referring to a time point in past can be combined only with VCs in one of the past tenses. Given for instance a temporal adverbial *gestern (yesterday)* in Example (22)), it is not possible to use *Präsens* as we saw it in Example (19b). In such a context, only the use of one of the past tenses leads to grammatically correct sentences.[13]

(22) a. Das Geschäft hat gestern seine Pforten geöffnet. → *tense = Perfekt*
The store has yesterday its gates opened.
'The store has opened yesterday.'

b. *Das Geschäft öffnet gestern seine Pforten. → *tense = Präsens*
The store opens yesterday its gates.
'*The store has opened yesterday.'

Usage of tense is a big research area. In this section, we mentioned some of the specifics

[13]Usage of *Präsens* in combination with appropriate temporal adverbials referring to an event in the past is possible, but it is mainly used to induce some specific rhetoric effects (cf. (Grewendorf, 1982) for more details).

of using tense in German which are directly related to the modeling of verbal inflection described in Chapter 6. To these belong the morpho-syntactic structure of the German tense forms, as well as a short description of the contextual and semantic/pragmatic factors which are very challenging for identifying the tenses in German and for modeling the tense (for translation).[14].

3.4.4. Mood in German

Mood is one of the devices to express modality of an utterance. Modality is related to the speakers attitudes and opinions, speech acts, subjectivity, non-factivity, non-assertion, possibility and necessity (Palmer, 1986). In German, mood is expressed in the morphology of the finite verbs. Different moods of the finite verbs also lead to specific syntactic tenses in German.

3.4.4.1. Morphological and syntactic mood in German

In this work, we differ between two moods in German: indicative and subjunctive as shown in Example (23).

(23) a. Der Junge $\text{liest}_{Pres.Ind}$ ein Buch. → *mood = indicative*
The boy reads a book.
'The boy reads a book.'

b. Der Junge $\text{läse}_{Pres.Subj}$ ein Buch. → *mood = subjunctive*
The boy would read a book.
'The boy would read a book.'

Depending on the tense of a German finite verb in the subjunctive mood, we distinguish between syntactic tense forms in German which we refer to as *Konjunktiv*. These are summarized in Table 3.12. The combination of the subjunctive mood with the morphological present tense is called *Konjunktiv I*, while the combination of the subjunctive mood with past is referred to as *Konjunktiv II*. These definitions, however, do not match with the syntactic tenses that *Konjunktiv* may build. For example, *Konjunktiv I* may also express an action in the past such as in example (4) in Table 3.12. The *Konjunktiv* forms related to the past events or actions are considered to have the syntactic tense *past*: there is no classification of *Konjunktiv* into past tenses *Präteritum*, *Perfekt* and *Plusquamperfekt*. The morpho-syntactic patterns of the German *Konjunktiv* tense forms are given in Table 3.13.

[14]Further discussion of tense and mood in the bilingual context is given in Chapter 9

Morph. tense	**Syntactic tense/mood**	**Example**
present	Konjunktiv I/Präsens	(1) Er lese ein Buch.
	Konjunktiv I/Futur I	(2) Er werde ein Buch lesen.
	Konjunktiv I/Futur II	(3) Er werde ein Buch gelesen haben.
	Konjunktiv I/past	(4) Er habe ein Buch gelesen.
past	Konjunktiv II/Präsens	(5) Er würde lesen. Er sollte lesen.
	Konjunktiv II/past	(6) Er hätte gelesen.
	Konjunktiv II/Futur II	(7) Er würde ein Buch gelesen haben.

Table 3.12.: Combination of the different morphological tenses with the German subjunctive mood.

Syntactic tense	**Pattern**	**Example**
Präsens	V*FIN.Pres.Subj	Er lese.
	VMFIN.Pres.Subj V(A|V)PP	Er könne lesen.
	V*FIN.Past.Subj	Er läse.
	VMFIN.Past.Subj V(A|V)INF	Er könnte lesen.
past	VAFIN.Pres.Subj V*PP	Er habe gelesen. Er sei gefahren.
	VAFIN.Pres.Subj V(A|V)INF VMINF	Er habe lesen können.
	VAFIN.Past.Subj V(A|V)PP	Er hätte gearbeitet. Er wäre gefahren.
	VAFIN.Past.Subj V(A|V)INF VMINF	Er hätte arbeiten können.
Futur I	VAFIN.Pres.Subj V*INF VAFIN.Pres.Ind V*INF VMINF	Er werde arbeiten. Er werde arbeiten können.
Futur II	VAFIN.Pres.Subj V(A|V)INF VAINF	Er werde gearbeitet haben. Er werde gefahren sein.
	VAFIN.Pres.Subj V(A|V)INF VMFIN	Er werde gearbeitet haben können.

Table 3.13.: The German subjunctive morpho-syntactic tense forms with examples of the different realization possibilities for the active voice.

3.4.4.2. Use of mood in German

Konjunktiv is syntactically required in a combination with specific German conjunctions such as *als (as), ob (if), 'als ob' (as if)* in sentences which express hypothetical comparisons. Examples are shown in (24).

(24) a. Der Junge weint, als habe er Schmerzen. → *Konj I/Präsens*
The boy cries, as has he pain.
'The boy is crying as if he had pain.'

b. Der Junge weint, als hätte er Schmerzen. → *Konj II/Präsens*
The boy cries, as has he pain.
'The boy is crying as if he had pain.'

Konjunktiv is often used in the reported speech. For expressing the non-assertion of the speaker related to the proposition of the given utterance, in most cases *Konjunktiv I* is used as shown in (25a), however, the same may also be achieved by using *Konjunktiv II* as shown in Example (25b). As a matter of fact, there are a few ambiguous forms of *Konjunktiv I* regarding indicative and subjunctive which are usually avoided by using *Konjunktiv II*. Consider the verb *haben* in (25a): since the subjunctive form in the third person plural of the verb *haben (to have)*, namely *haben*, equals to the indicative form in the third person plural of the same verb, *Konjunktiv II* is used instead to express the neutral attitude of the speaker towards the reported utterance as shown in (25b).

(25) a. Sie sagen, sie haben das Buch schon. → *Präsens* or *Konj I/Präsens*
They say, they have the book already.
'They say that they already have the book.'

b. Sie sagen, sie hätten das Buch schon. → *Konj II/Präsens*
They say, they have the book already.
'They say that they already have the book.'

Similarly to the interchangeability of tenses in German, there are no strict rules on how to use the different moods in German in reported speech:

> *"To signal reported speech, it is not obligatory to use Konjunktiv I, though: it is also acceptable to use Konjunktiv II or a plain indicative instead. ... Note further that Konjunktiv I, Konjunktiv II and indicative are often used interchangeably when used in reported speech."*
>
> Csipak (2015, p.9)

In contrast to the reported speech, *Konjunktiv II* is required in utterances about the non-factive events. To those belong, among others, the conditional sentences which are illustrated in Example (26). The sentence in (26a) is in indicative which indicates that the boy does have time to read. On the contrary, the sentence given in (26b) is in *Konjunktiv II* and thus indicates that the condition for the boy to read a book is not given, so the boy is very unlikely to read the book.

(26) a. Der Junge liest, wenn er Zeit hat. → *Präsens*
The boy reads, when he has time.
'The boy reads, when he has time to.'

b. Der Junge würde lesen, wenn er Zeit hätte. → *Konj II/Präsens*
The boy would read, if he time has.
'The boy would read, if he had time.'

Besides the subjunctive conditional sentences which are usually composed of at least two clauses as in Example (26), there are also subjunctive sentences in German which are called *free factive* subjunctives. To those may belong simple sentences which express politeness as shown in (27).

(27) a. Ich hätte gern ein Glas Wasser. → *Konj II/Präsens*
I have gladly a glass water.
'I'd like to have a glas of water.'

b. Das wäre nun geklärt. → *Konj II/Präsens*
That would be now clarified.
'That's clear now.'

3.4.5. Tense and mood in English and German

Modeling of the German verbal inflection, described in Chapter 6, involves inflecting the German verbs according to the agreement feature, as well as the tense and mood features. All of these features need to appropriately be transferred from English as a source language to German as a target language.

Especially the translation of tense and mood is a challenging task. In the preceding Section, we briefly discussed the use of tense and mood in German from the monolingual perspective. Thereby, we mentioned phenomena such as interchangeability, author preference and genre dependency which make the prediction of the appropriate tense in German hard. In this Section, we analyze English and German tenses in the bilingual context. The discussion does not aim at establishing rules for the tense and mood

translation, but rather to interpret the tense and mood distribution in English-German parallel texts.

3.4.5.1. Tense

Referring back to Table 3.10 on page 41, the first difference regarding the tenses in English and German that is striking is that English has considerably more tense forms than German. Particularly the absence of the German progressive tense forms is problematic because it leads to the translation of the English progressive tenses into a number of different German tenses. In addition to the differences in the tense sets of the two languages, the specifics of the tense usage in the monolingual context may lead to the non-trivial translations of the English tenses into German. For instance, given a future tense in English, it is very probable that the German translation is in *Präsens* due to the tendency of using *Präsens* in German to refer to future actions. On the other hand, it is also possible, or even required in specifics contexts, that the English future tense generates future tense in the German translation.

The tense translation between English and German thus represents a one-to-many relation where a single English tense may correspond to a number of the different German tenses. On the other side, it also represents a many-to-one relation where many different English tenses may correspond to a single German tense form. An interesting example for this case is the German *Konjunktiv I* which does not have a direct counterpart in English.

Both English and German have non-finite clauses which contain tenseless VCs. Similarly to the translation of the finite VCs, the translation between the English and German non-finite VCs is often not trivial. Especially the translation of the English non-finite VCs into the finite German VCs is interesting as it requires the generation of tense without obvious tense information in the source language.

Many-to-many relation One of the reasons for the many-to-many relation regarding the tense translation from English to German is the different granularity of the tense systems in the two languages. On the one hand, there are tenses in English which do not have a direct counterpart in German (progressive tense forms). On the other side, there are also tense forms in German which do not have a direct counterpart in English (*Konjunktiv I*). Hence, a single English tense may be translated into many different German tenses and a single German tense may be generated from different English tenses. In the following, we take a deeper look into a few interesting cases for

non-unique translations between English and German tenses.

Consider, for example, the English present perfect (progressive). It may correspond to the German *Präsens*, as well as to one of the past tenses as illustrated by example sentence pairs in (28) which we extracted from the English-German News corpus.

(28) a. ...policy response, which has been painfully slow. → *presPerf*
...politische Antwort, die hat gewesen quälend langsam.
'...Antwort der Politik, die schmerzhaft langsam ist.' → *Präsens*

b. Military reform has been reversed. → *presPerf*
Militäre Reform hat gewesen aufgehoben.
'Die Militärreform wurde auf Eis gelegt .' → *Präteritum*

c. Most people have forgotten that... → *presPerf*
Meisten Menschen haben vergessen dass...
'Die meisten Menschen haben vergessen, dass...' → *Perfekt*

The distribution of the translation of the English present perfect (progressive) into the different German tenses is shown in Figure 3.5. The graph shows that for both present perfect as well as present perfect progressive, there are three prominent translations into German, namely *Präsens*, *Perfekt* and *Präteritum*. For both English tense forms, *Perfekt* is the most prominent translation. If we consider the two German past tenses together, it becomes clear that both forms of the English present perfect tense more often correspond to one of the German past tenses compared to the present tense. Interestingly, the progressiveness has a large impact on this relation: the non-progressive present perfect tense corresponds in ca. 77% cases to one of the German past tenses, while this is the case for only 56% of its progressive form occurrences. In other words, the progressiveness still prefers to be translated into one of the German past tenses, however, it is more often translated into the German *Präsens* than the non-progressive present perfect. The choice of the German tense given the English present perfect tense is surely not arbitrary. Similarly to the interchangeability of the German past tense in the monolingual context, we assume that the choice between *Perfekt* and *Präteritum* is mainly a matter of style, domain., etc. However the choice between *Präsens* and one of the past tenses must be justified by the given context.

The study of the parallel data further reveals interesting cases of a *tense switch* caused, among other factors, by the interchangeability of tenses in German. As an example, we examine the translation of the English future tenses. According to the interchangeability of the German *Präsens* and the *Futur I* outlined in Section 3.4.3.2, we assume that the English future tenses are translated into both the German *Futur I*, as well as the *Präsens*.

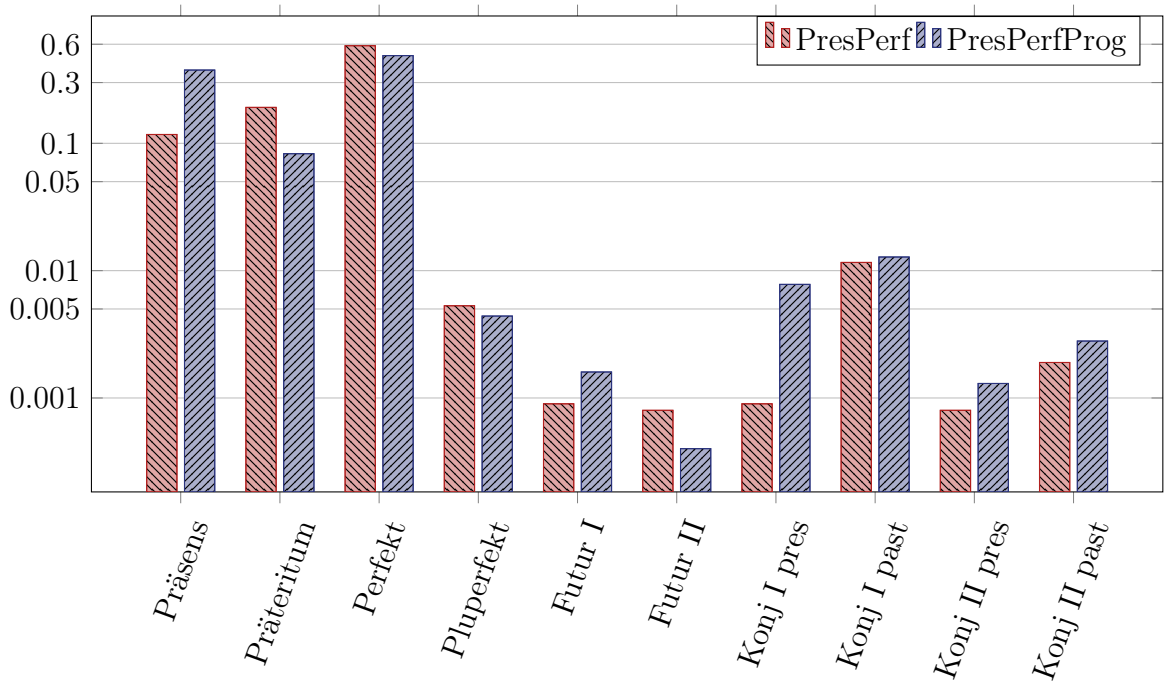


Figure 3.5.: Relative frequencies of the translation of the English present prefect (progressive) tense into German derived from the Europarl corpus.

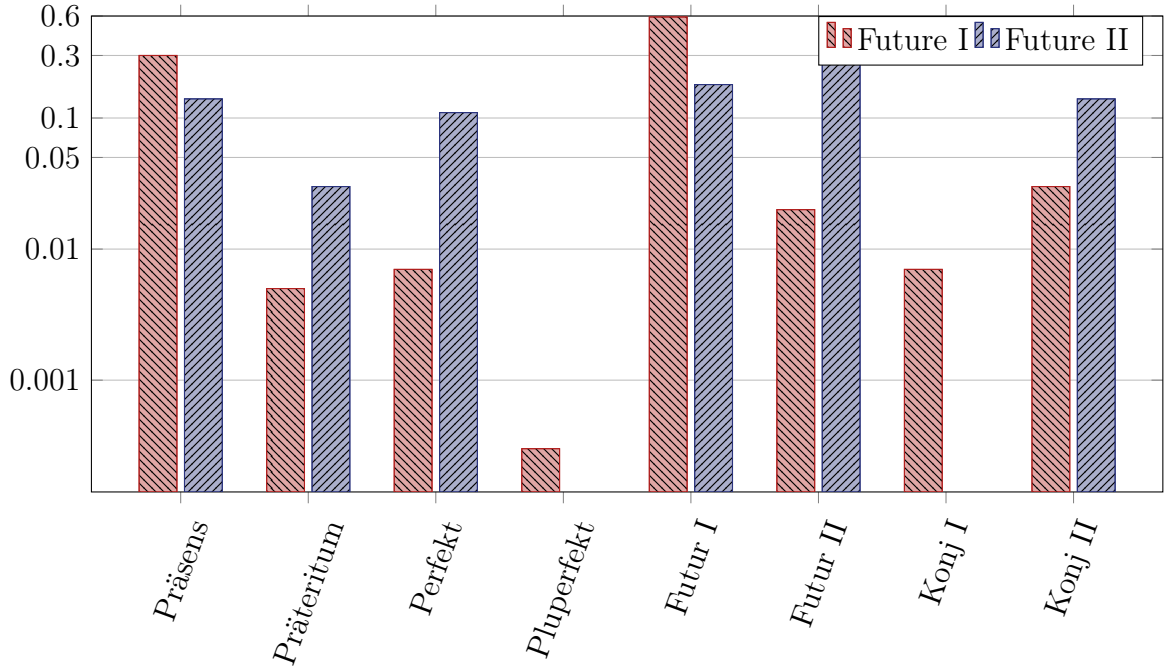


Figure 3.6.: Relative frequencies of the translation of the English future tenses into German derived from the News corpus.

	News		Europarl	
	EN	DE	EN	DE
Finite	83.3	92.1	81.8	93.8
Non-finite	16.7	7.9	18.2	6.2
Total VCs	285,472		3,112,160	

Table 3.14.: Distribution in % of the finite and non-finite parallel English and German VCs found in two different parallel corpora given in percent.

The assumption is supported by the distribution of the translations of the English future tenses shown in Figure 3.6. In the majority of the cases, future I is translated into the German *Futur I* tense. However, a relatively high portion of about 30% of the future I occurrences correspond to the German *Präsens*.

Tensed and tenseless VCs An even more interesting tense switch is related to the translation between finite and non-finite VCs. Besides the finite, tensed clauses (i.e., VCs), both English and German also contain non-finite, i.e., tenseless clauses, which enclose non-finite VCs. Statistics about the usage of the finite and non-finite VCs in both languages found in our parallel corpora are shown in Table 3.14. For both domains, the English texts contain considerably more non-finite VCs than German which implies that many of the non-finite English VCs correspond to the finite VCs in German.

Our data shows that the major part of the English non-finite VCs translate into finite VCs (cf. Figure 3.7) which is a very interesting problem in the context of the (statistical) MT. When translating from English to German, MT needs to generate a finite clause for the given non-finite source clause. Particularly, it needs to generate a finite German VC in a tense form for which there is no obvious evidence in the source. Examples of such translations found in our corpora are given in (29). In (29a), the non-finite clause *'applying force'* is translated into a finite clause *'wobei sie Gewalt anwenden'* in the *Präsens* tense. In (29b), on the other hand, the non-finite clause *'to rule Russia'* corresponds to the relative finite clause *'der Russland regierte'* in the *Präteritum* tense.

(29) a. ...applying force when a trade did not go well. → *gerund*
...anwendend Gewalt when ein Geschäft tat nicht gehen gut.
'...wobei sie Gewalt anwenden, wenn ein Geschäft
nicht in ihrem Sinne verläuft.' → *Präsens*

b. ...the only other KGB man to rule Russia... → *to-infinitive*
...der einzige andere KGB Mann zu regieren Russland...
'...der einzige andere KGB-Mann, der Russland regierte...' → *Präteritum*

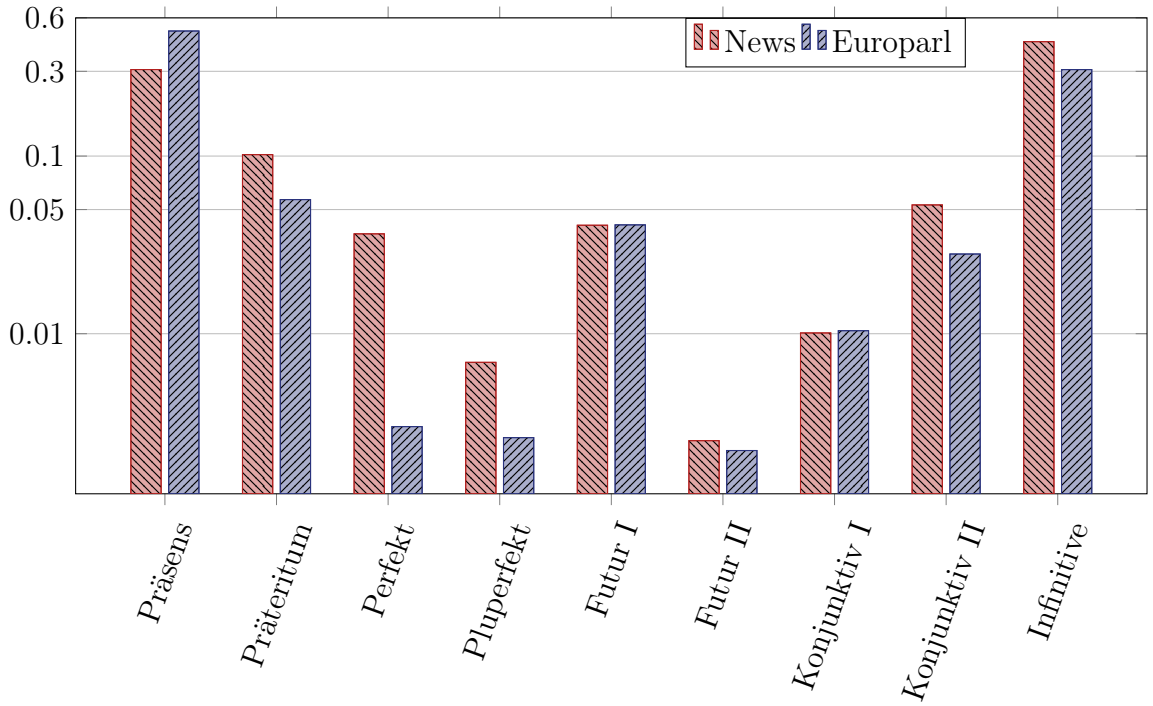


Figure 3.7.: Relative frequencies of the translation of the English non-finite VCs (gerunds and to-infinitives) into German.

In addition to the difference in the translation between finite and non-finite clauses, Example (29a) also nicely shows the tense switch from the English past tense to the German *Präsens* related to the translation of the English VC *'did not go'* into *'nicht verläuft'*. In contrast to the tense switch related to the English future tense discussed above, the tense switch from past to present is less intuitive. We assume that this kind of switch is valid only in certain contexts in which the tense, i.e., the time of a certain action plays a rather secondary role. The translation into *Präsens* indicates a general (timeless) validity of the uttered proposition.

Summary The corpus-driven analysis of the translation of English tenses into German indicates that all English tenses are ambiguous concerning their tense counterparts in German. This hypothesis is supported by Figures 3.8 and 3.9.[15]. The figures show that for each English tense, there is a preferred tense in German. However, there are also other translation possibilities which need to be considered. Despite one dominant tense translation for each of the English tenses, there might be contexts in which the dominant translation alternative is not the most appropriate one or it is even incorrect. In the

[15]The plots are based on the frequency distributions given in Tables A.12 and A.13 in Appendix A.3.

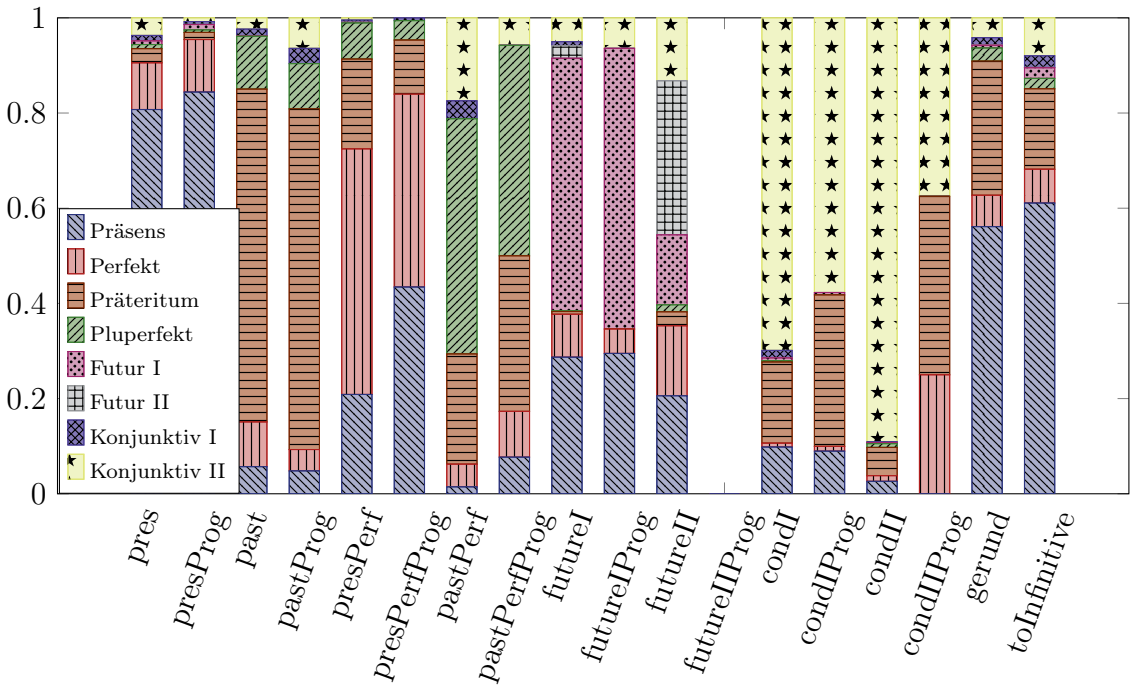


Figure 3.8.: Distribution of tense translation pairs derived from the News. The graph shows translations of the English VCs into *finite* German VCs.

context of translation, it is thus important to identify the contextual clues which restrict the choice of the German tense forms.

In addition to the factors such as contextual constraints, stylistic or genre preferences, the translation of tenses may also follow a *set of rules* defined for a specific translation project:

> *"Protokolle oder Berichte von Sitzungen werden in der deutschsprachigen Fassung stets im Präsens verfasst und zwar ohne Rücksicht auf das im Ausgangstext verwendete Tempus sowie den Umstand, dass die berichteten Ereignisse in der Vergangenheit liegen und in der Regel bereits abgeschlossen sind."*
> *In German, minutes or reports of the sessions are always to be written in the present tense regardless the tense given in the source text, as well as of the fact that the events being reported on happened in the past and are usually already completed.*
>
> The EC German style guide[16], p. 622.

[16] `https://ec.europa.eu/info/sites/info/files/german_style_guide_de_0.pdf`

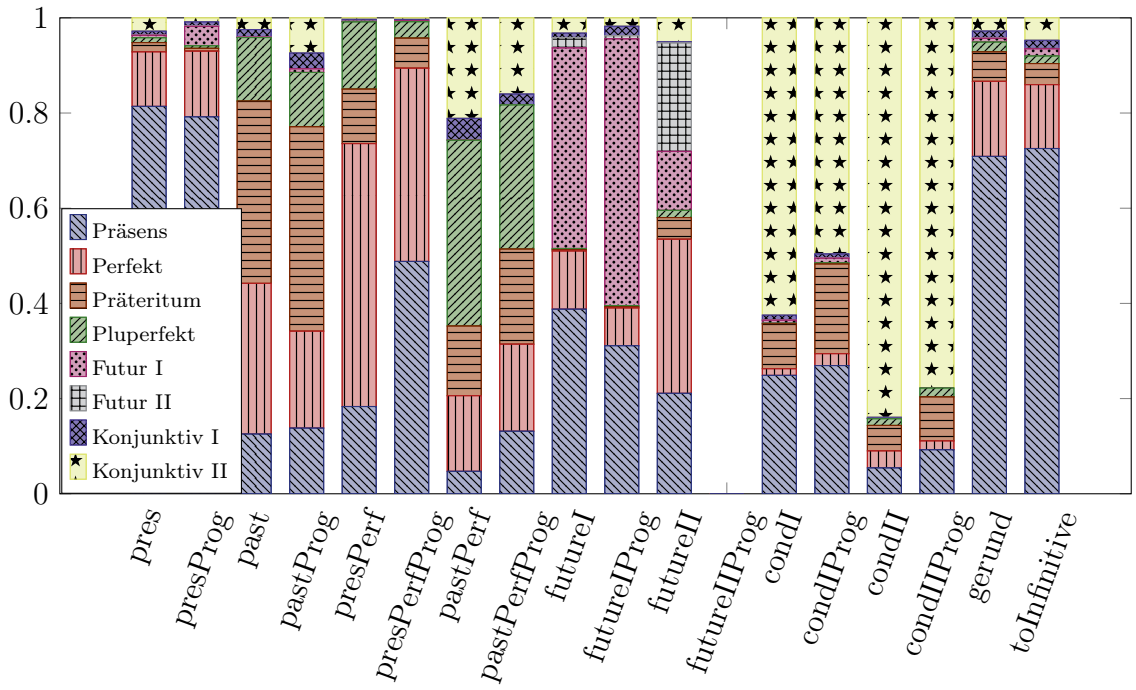


Figure 3.9.: Distribution of tense translations derived from the Europarl corpus.

3.4.5.2. Mood

Both English and German verbal inflection encloses the morphological feature *mood* which is expressed in the inflection of the finite verb. In this work, we distinguish between two mood values: indicative and subjunctive. Particularly the use of the subjunctive mood is of big interest for modeling verbal inflection described in Chapter 6 because of its divergent use in the two languages. Therefore, the following discussion concentrates on the differences between English and German with respect to the subjunctive mood.

The subjunctive mood in English is only present for the modal verbs: *shall-should, will-would, can-could, may-might*. Tense forms built with a modal in the subjunctive mood in English are referred to as conditionals as shown at the bottom of Table 3.10, page 41. In German, all verbs may take the subjunctive mood and as such build one of the *Konjunktiv* tense forms. The English conditional tenses may be seen as direct counterparts of the German *Konjunktiv II* in function for expressing non-factual events. Figure 3.10 supports this claim: the most frequent translation for all four English conditional tense forms is indeed *Konjunktiv II*. Further frequent translation options are *Präteritum* for the conditional I tense forms, and *Perfekt* for the conditional II tenses.

Examples of the translations of the English conditionals into one of the German indicative tenses found in the News corpus are shown in (30). The German translations

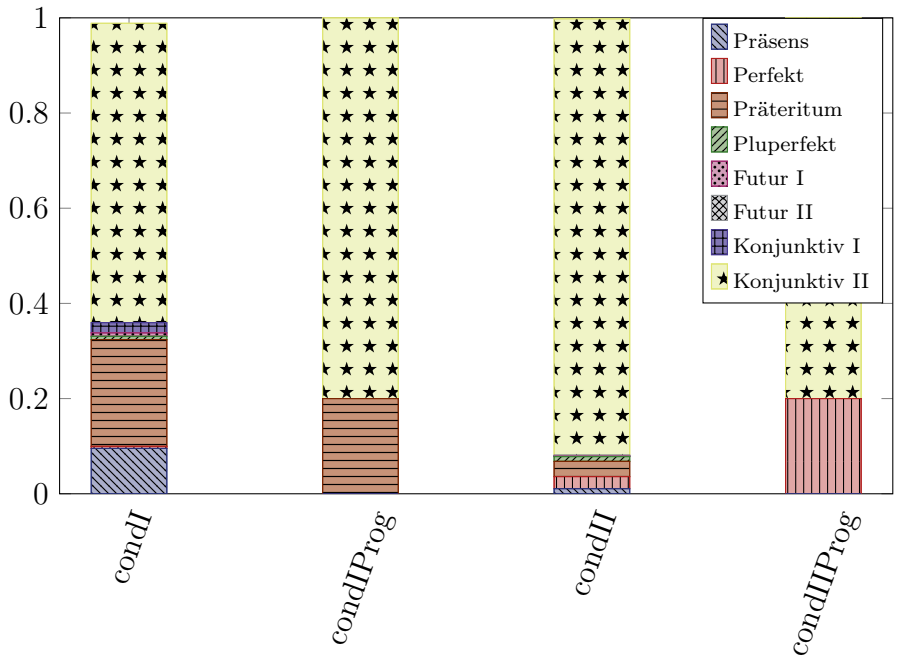


Figure 3.10.: Distribution of the translations of the English conditionals into German derived from the News corpus.

given in sentence pairs (30a) and (30b) contain other sentential material to express the meaning given in the conditional source sentence. In (30a), this is achieved by using the modal verb *kann*, while the uncertainty which is expressed in the English sentence in (30b) is conveyed within the adverbial *vielleicht (perhaps)*. The German sentence in (30c), on the other hand, seems to intentionally express stronger commitment to the proposition of *'building a Palestinian state'*. While the English VC *'should shift'* can rather be interpreted as a suggestion, the German VC *'müssen gerichtet sein (must be directed)'* allows no possibilities other than the proposed one.

(30) a. Of course, it could simply have been the weather;
von natürlich, es könnte einfach have gewesen das Wetter;
'Natürlich kann es auch einfach das Wetter gewesen sein, → *tense = Perfekt*

b. In the past, a poor African might have looked at his
In der Vergangenheit, ein armer Afrikaner dürfte haben geschaut zu seinen
compatriots...
Landsleuten...
'Früher hat ein armer Afrikaner vielleicht zu seinen Landsleuten geschaut...'
→ *tense = Perfekt*

c. ...the effort should shift to building a Palestinian state.
...die Mühe sollte verschieben zu bauen einen palästinensischen Staat.
'...die Bestrebungen müssen darauf gerichtet sein, einen palästinensischen Staat aufzubauen.' → *tense = Präsens*

The preceding examples deal with English clauses with conditional VCs which have direct conditional counterparts in German. However, the English conditional clauses usually occur in an indicative context: the subjunctive mood of the subordinated English conditional clauses (i.e., *if-clauses*) is expressed by a shift of the indicative tense forms as shown in (31) where the English main clause contains the indicative VC in the simple past tense (*had*), while the subordinate clause contains the conditional VC *'would read'*. In contrast, both clauses in the German translation have VCs in the subjunctive mood. Assumed the same non-factual proposition in both languages, the discrepancy of using the indicative mood in English compared to the German subjunctive mood as shown in Example (31) is always given. To resolve the ambiguity of translating the English indicative into the German subjunctive VCs, we need to identify appropriate contextual information which triggers the respective *mood switch*.

(31) The boy would read if he had time. → *tense = simple past*
der Junge würde lesen wenn er hatte Zeit.
'Der Junge würde lesen, wenn er Zeit hätte. → *tense = KonjII/Präsens*

Another very prominent case of the *mood switch* between English and German is related to the German *Konjunktiv I* in the context of the reported speech. Here, two facts play a role: (i) reported speech in English uses indicative tense forms, and (ii) the use of indicative/subjunctive mood in German is often interchangeable. Consider, for example, sentences in Table 3.15 which show several translation alternatives for a single English source sentence. All German sentences are valid translations, however they differ in a specific semantic aspect, namely epistemic commitment to the truth of the made proposition (**?**). Despite the semantic difference, the postulated German translations are interchangeable in the context of the reported speech as already discussed in Section 3.4.4.

3.5. Chapter summary

In this Chapter, we presented similarities, as well as the most striking dissimilarities between verbs in English and German. Particularly the described differences are of big

	Sentence	Tense
EN	The boy said that he already **visited** you.	past simple
(a)	Der Junge sagte, er **habe** dich schon besucht.	*Konjunktiv I/past*
(b)	Der Junge sagte, er **hätte** dich schon besucht.	*Konjunktiv II/past*
(c)	Der Junge sagte, er **hat** dich schon besucht.	*Perfekt*

Table 3.15.: Example translation pairs.

importance since they pose a great problem for the automatic translation from English to German.

In many cases, the verbs in the German SMT outputs are misplaced or even omitted. To cope with this problem, we apply a method described in detail in Chapter 4 which relies on the reduction of the positional differences between the verbs in English and German. The linguistic analysis given in Section 3.3 revealed that almost all types of clauses in English and German expose differences regarding the placement of the verbs. On the one hand, German is, depending on the clause type, both SVO, as well as a SOV language while English exposes solely the SVO word order. Furthermore, German may have discontinuous verbal complexes in which parts of a verbal complex may be disrupted by other sentential constituents such as objects and adjuncts. Typically, the non-finite part of a German verbal complex is then placed at the clause end which is not allowed in English. On the other hand, English VCs are often longer than their German counterparts which is quite challenging for the task of establishing syntactic parallelism between English and German VCs. Thus, in Section 3.2, we first defined which parts of the English and German VCs correspond to each other. We then used these definitions to develop a formal description of the positions of the verbs in German depending on the type of a clause they occur in. In Section 3.3.2, we identified eight positional patterns of the verbs in German. These patterns will be considered in the implementation of the *preordering* method for English→German SMT described in Chapter 4. It must though be mentioned that the position of the verbs in German may vary in specific contexts. A few of them were presented in Section 3.3.3. We do not consider these cases further due to two reasons: (i) the information needed to identify them is not directly accessible, and (ii) we aimed at gathering a general set of the positional differences in English and German which captures the most frequent cases and clause types. Description, as well as handling of infrequent exceptional cases would probably have only a minor impact on one of the aims of this thesis, namely correction of the positional errors of the verbs in the German SMT outputs.

The verbs in English and German differ not only with respect to their position, but also regarding their morphological richness. While English is a morphologically poor language, German belongs to a group of the morphologically rich languages. Regarding the verbs, this means that a single verb in English has only a very few inflectional alternatives. On the other hand, a single German verb has many different inflectional variants depending on the *person*, *number*, *tense* and *mood* it exposes. Explicit modeling of the verbal morphology which includes *generation* of the German inflected verbs (cf. Chapter 6) requires information about the mentioned verbal morphological features. In Section 3.4.1, we first explained the differences between English and German regarding the agreement features. Furthermore, we explored the contextual dependency between the verbs and subjects in German which is crucial for determining the agreement features for the German finite verbs.

While the verbal agreement features are determined by the morphological properties of the corresponding subject phrases, tense and mood depend on both contextual, as well as semantic/pragmatic factors. Morphological features tense and mood depend on the syntactic tense and mood of a VC they occur in. In Sections 3.4.3 and 3.4.4, we first presented tense and mood forms in German. In Sections 3.4.3.1 and 3.4.4.1, we described morphological, as well as syntactic properties of tenses in German. In Section 3.4.3.2, we outlined a few interesting facts about the usage of tense in German. Despite intuitive names of the different tense forms which indicate the time point that a specific tense form expresses, it became clear that the use of tense is a complex topic. The use of tense in German does not underlie syntactic constraints. In fact, it rather depends on the register, genre and also on the author preferences. Furthermore, we have seen that in many cases, tenses are interchangeable. In some cases, the interchangeability is justified by existence of temporal words such as adverbials which give information about the temporal location of a utterance. In other cases, the justification is not explicitly given but it is rather part of the *world knowledge* or derivable from the preceding context. Interchangeability is not only observed in the usage of the tenses in German, but also in the usage of the different moods which we discussed in Section 3.4.4.2.

After the monolingual discussion of the use of tense and mood in German, in Section 3.4.5, we carried out a data-driven analysis of tense and mood in English and German. Due to the different granularity of the tense/mood systems in English and German, the translation of tense and mood from English to German may be seen as a many-to-many relation. Even though there are clear preferences regarding the translation of a single English tense into German, the data analysis showed that the choice of a tense/mood

in German for a given English source tense is often not trivial. We identified several reasons for that: (i) a single English tense form has a number of different valid translation alternatives in German due to the different syntactic tense sets in the two languages, (ii) interchangeability of the tense/mood forms in German may lead to unexpected tense translations and (iii) author preferences or genre characteristics may require specific (non-trivial and less intuitive) tense/mood translations. Particularly, the phenomenon of a tense/mood switch is interesting since it often cannot be described by simple means of lexical information explicitly given in a source-target sentence pair:

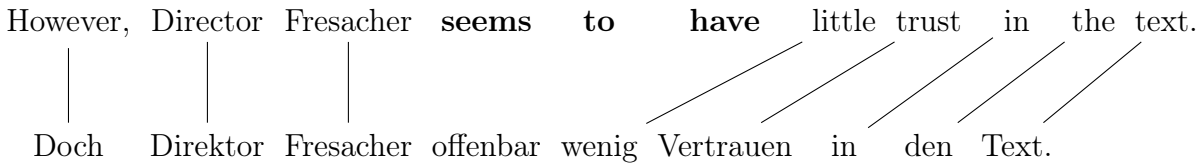
> *"Translation shifts may be due to cognitive factors, such as the translator's understanding, idiosyncratic preferences or constraints during the translation process, to contrastive differences between the languages involved or to different register characteristics."*
>
> Hansen-Schirra et al. (2012, p. 133)

4. Reordering

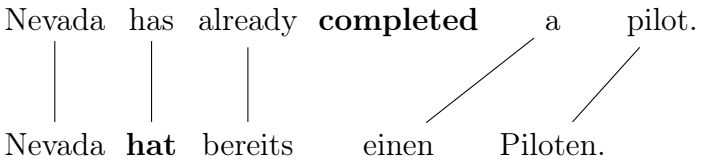
Phrase-based statistical machine translation (PBSMT) is known to have problems when translating between languages with considerable syntactic differences. For example, when translating from a subject-verb-object (SVO) language into a subject-object-verb (SOV) language, SMT needs to place the translation of the source verb(s) V into the target language specific position which is in this case *after* the object O. When the positional differences are large, the generated SMT outputs often show two problems: (i) the position of the words in the translations is erroneous or (ii) the target language words are not generated at all. These problems are also observed for the translation direction English→German which we discuss in Section 4.1. One of the most popular methods to deal with long-range reordering problems for SMT is the so-called *preordering approach*. The idea of preordering is to reduce syntactic differences between source and target languages prior to training and translation. We make use of this simple, yet effective approach to deal with positional problems of the verbs in the English→German SMT. Since 1990's, lot's of research has been done on the preordering approach which we summarize in Section 4.2. The adaptation of the preordering approach to the English→German translation direction is described in Section 4.3. Details about the reordering rules are given in Section 4.3.1, while the details about the implementation are described in Section 4.3.2. The Chapter summary is given in Section 4.4.

4.1. Verb positions in English→German SMT

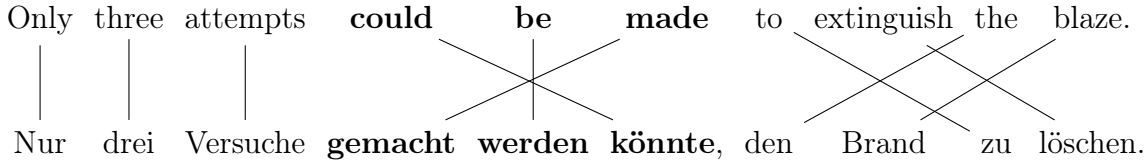
In the linguistic analysis of the verbs in English and German described in Chapter 3, we identified English as a SVO language, while German is both SVO, as well as a SOV language depending on the clause type. Additionally, German may also be seen as a $SVOV$ language in which a verbal complex may be discontinuous. In this case, parts of a German verbal complex are interrupted by numerous sentence constituents such as object noun phrases, adjunct prepositional phrases or adverbial constructions. In other words, the distance between the verbs of a single verbal complex in German can be very



(a) Missing translation of the English verbal complex *'seems to have'*.



(b) Missing translation of the English verb *'completed'*.



(c) Misplaced translations of the English verbal complex *'could be made'*.

Figure 4.1.: Examples of the German SMT translations along with word alignment between the English source words and their German translations.

large.[1]

In the context of a statistical machine translation, an English→German SMT model needs to appropriately model all the different verb placement combinations between the two languages. The translation step has to be able to place the German verbs into positions which often require movements over a large number of the words. As a consequence, many of the German SMT translations reveal errors such as missing verbs or wrong placement of the verbs as shown in the example translations in Figure 4.1.[2] In the translation of the English sentence given in 4.1a, all verbs in German are missing: the English verb sequence *'seems to have'* has not been translated at all. This also holds for the English verb *completed* in example 4.1b. Here, the auxiliary for the German tense *Perfekt* has been generated, however, the main verb, i.e., the direct translation of the English verb *completed* is missing. In the example 4.1c, the verbs have been generated but the placement of the verbs *gemacht, werden* and *könnte* is erroneous. The correct order, i.e. placement of the respective verbs is *könnte gemacht werden*.

[1] A detailed description and comparison of the English and German VCs in presented in Chapter 3.

[2] The SMT model which generated these translations will be presented in Section 5.2.

Systematic evaluations of the German SMT outputs generated from the English source sentences have been carried out only recently in the context of the comparison of errors in SMT and neural machine translation (NMT). Nevertheless, they show how problematic the SMT-based translation of the English verbs into German is. For example, Bentivogli et al. (2016) found out that 38.5 of all word order errors found in a standard English→German PBSMT system are related to the verbs.[3]. In other words, the verbs are the most often misplaced word category in the German SMT translations. Popović (2017) carried out an analysis of the German SMT (and NMT) outputs focusing on the language-specific issues. Regarding the verbs, different types of errors are considered: positional (i.e. order), inflectional and whether the verb is missing. The evaluation revealed that 24% of issues identified in the German SMT translations are missing verbs. 9.4% of the errors are related to the misplacement of the German verbs, while 9.4% of the errors are wrongly inflected verbs.[4]

Although the two mentioned analyses of the German SMT output include different PBSMT systems for English→German, as well as different test sets, they both show that SMT has big problems with the translation of the English verbs into German which are important to deal with. Certainly, there are additional issues in the German SMT outputs besides the verb-related errors, however, the problems of erroneously placed or even omitted verbs have a severe negative impact on understanding and acceptability of the generated translations. Wrong placement of the verbs negatively affects the fluency of the translations which makes it sometimes difficult to understand the translation (cf. Example 4.1c). On the other side, missing verbs lower adequacy of the translations meaning that some specific information present in the source sentence has not been transferred into the generated translation. Verbs bear central semantic information crucial for correct understanding of a sentence. If the verbs are missing in a translation, it is very likely that the translation cannot be understood properly. This may be seen in 4.1a 4.1b in which one is not able to understand the message of the translation without having access to the source sentence.

The examples presented in Figure 4.1 show that the difference in the position of the verbs in English and German is not only problematic for longer sentences in which the movements over a large number of words (i.e. long-distance or long-range reordering) is required. Even the short sentences in which the positional difference is only 2-3 words such as the one given in 4.1b are affected by the problems of missing or wrongly placed

[3]Detailed error analysis can be found in Table 5 in Bentivogli et al. (2016).

[4]For the complete error analysis, please refer to Table 2 in Popović (2017).

German verbs.

One of the aims of this work is to help SMT to generate more verbs in the German translations, as well as to place them into correct positions. Although we explicitly correct the translation of a single part-of-speech, namely verbs, both of the improvements will have a big impact on the quality of the German SMT output.

4.2. Previous work on preordering for SMT

The key idea of preordering for SMT is to make the source and the target language sentences syntactically more similar to alleviate statistical machine translation between the languages of a given language pair. Syntactic similarity is reached by transforming the data in such a way that the words in the language pair under consideration have (almost) the same positions in a sentence. Given a language pair $e - d$, the sentence transformation of e involves movements of the words into positions which are typical for d and vice versa. Preordering approach for SMT exists for almost 20 years. Since the early 1990's, the reordering approach has been further developed and adapted to many different language pairs, mainly in the context of the statistical machine translation.

The very first implementations of the preordering approach included transformations of both the source and the target language data prior to the translation step. A first work on preordering for machine translation was published by Brown et al. (1992). Brown et al. (1992) applied specific transformations of French as a source language and English as a target language to make the transfer-based translation between the two languages easier. The reorderings were carried out by a finite state recognizer which used the information about the POS tags and words. After the translation was performed, the transformations on the target language side were reversed. Brown et al. (1992) combined reordering of the data with additional pre-processing steps: they did not evaluate the reordering separately. However, together with other pre-processing steps, the reordering of the source and target data led to the improved English MT output generated by a transfer-based MT system. Encouraged by the simplicity and effectiveness of the preordering approach, Nießen and Ney (2000) and Nießen et al. (2001) combined it with the SMT. They defined a set of the transformation rules for German→English SMT and applied them on both languages. In contrast to Brown et al. (1992) who worked with POS-tagged data, the reordering rules proposed by Nießen and Ney (2000) were formulated for the parse trees. Nießen and Ney (2000) recognized the importance of handling the verbs when translating between German and English and defined a rule for

merging the German verbal prefixes with the corresponding verbs in order to increase the syntactic similarity between German and English. Indeed, this small modification of the German corpus lead to an improvement of the English SMT translations.

Starting with the work published by Collins et al. (2005), the research on the reordering approach for SMT has focused on the transformations only of the source language. The reordering method relies on the reordering rules, as well as on a specific representation of the data. Further research thus explored different ways to acquire the reordering rule sets, as well as on the utilization of the different kinds of the linguistic analysis of the source language data.

One of the first works on reordering based solely on the reordering of the source language data was published by Collins et al. (2005) for the German→English SMT. Similarly to the previous research on this topic, Collins et al. (2005) manually defined a set of deterministic reordering rules. Since the movements imply specific syntactic knowledge of the source sentences, they applied their rules on the German constituency parse trees. Most of the defined rules handled the different position of the verbs in the two languages. By reducing the syntactic difference between German and English, Collins et al. (2005) achieved impressive improvements of the English translations. Due to its simplicity and achieved improvements, this approach was rapidly adapted to other language pairs. For instance, Wang et al. (2007) developed a deterministic set of the reordering rules based on the constituency parse trees for the Chinese→English SMT. (Lee et al., 2010) presented their work on preordering for English→Japanese in which a deterministic set of manually acquired rules is applied on the English constituency parse trees. Both of the adaptations led to the improvement of the SMT outputs and thus confirmed the efficiency of the deterministic preordering approach for the SMT.

Preordering was also applied to English→French SMT in a similar way. Xia and McCord (2004) proposed a method for automatic learning of the reordering rules, the so-called *rewrite patterns*. The rules were automatically derived from the word-aligned source-target dependency trees and described the rearrangement of the child nodes for a specific mother node. The automatically induced set of the reordering rules was not deterministic. However, their application to the source language trees was forced to be deterministic by different heuristics. Thus, for each source sentence s, the pattern application generated a single reordered version s'. In this study, the authors made an important observation that the reordering may improve the translation quality even if the monotonic decoding is applied.

Instead of applying a set of rules to reorder the source data, Costa-jussà and Fonollosa

(2006) proposed to use a n-gram based SMT to translate the source language data S into S' with the target language-like word order. The reordering was thus defined as a translation step between the original source language data and its reordered variant. The source language data used to train the n-gram based SMT model did not consist of the words, but of the word classes in order to generalize over the source words. Tests on Spanish→English and Chinese→English translation directions showed improvements for both monotone and non-monotone decoding. The authors observed that the improvements were however higher when non-monotone decoding was used. This indicates that not all words in the source data have been moved to the target language specific positions and that the SMT model was able to perform missing reorderings during the translation step.

The syntactic similarity between two languages may not only be increased by moving the words, but also by merging the words as done by Nießen and Ney (2000) for German→English, and by splitting the words into their morphemes. The latter case is particularly interesting when one of the languages under consideration is an agglutinative and/or a morphologically rich language. Habash and Sadat (2006) experimented with different pre-processing schemes for the Arabic→English SMT. The aim was mainly to reduce the data sparsity problems caused by the complex Arabic morphology, but at the same time, also to achieve a more parallel word order between Arabic and English sentences. They split the Arabic words into morphemes which directly correspond to the specific English words. Similarly to the previous work on reordering, Habash and Sadat (2006) also reported on the improvements of the English SMT output when the syntactic differences between Arabic and English are reduced prior to the training and translation steps.

The approaches mentioned in the preceding discussion belong to the deterministic preordering methods in which for each source sentence s the reordering rules generate only one reordered variant s'. Obviously, already a single reordered variant led to the significant improvements of the SMT outputs for the different language pairs. However, a language may to a certain extent have a flexible word order. This means that for some languages, it might be more appropriate to generate a set of variants S' of a single source sentence s. This has been proposed by Li et al. (2007) for the Chinese→English SMT. Similarly to Xia and McCord (2004), they developed a non-deterministic method for automatic derivation of the preordering rules from the Chinese parse trees. The reordering knowledge consisted of tree nodes along with the reordering probabilities of the child nodes. These probabilities were estimated using a ME classification model

which did a binary classification for child nodes being reordered or not. The reordering rules were derived, as well as applied on the clause level. The decoding was a composition of the translations of the clauses of a given sentence. First, for each clause c in a sentence, a set of reordered clause alternatives C' was derived. Each of the clauses in C' were then translated independently of each other. After all clauses have been translated, the most probable sequence of translation alternatives was picked to generate the final SMT output.

Instead of automatically extracting the rules from the fully parsed corpora, Zhang et al. (2007) proposed to extract the rules from the POS-tagged, chunk-parsed source data. Similarly to Li et al. (2007), the extracted rule set was non-deterministic and was also applied in a non-deterministic way. Thus, for each input sentence s, many different reordering alternatives S' were computed. In contrast to Li et al. (2007) who translated the different clause alternatives independently and then glued the best clause translations together into the final SMT output, Li et al. (2007) represented S' as a *word lattice* with weighted paths and fed it into the phrase-based SMT decoder. The whole sentence with all the alternatives was thus translated in a single translation step. Due to a huge number of the reordering alternatives encoded within a lattice, the decoding processing was kept monotonic. The method was tested on the Chinese→English translation direction and great improvements of the English SMT output were achieved. An important outcome of the experiments carried out by Li et al. (2007) was that the source preordering in combination with monotonic lattice decoding not only improved the SMT output, but it was also faster than the non-monotonic decoding of the sequential input string.

Elming (2008) proposed a method for the integration of a preordering model into the SMT log-linear model. Their preordering model was used not only to preorder the source data, but also to score the target language hypotheses according to the relevant reordering rules. Elming (2008) followed the idea of the automatic acquisition of the reordering rules from the parsed data. However, his method did not extract all possible reorderings, but only the most general ones. Following the idea of Li et al. (2007), the different reordering variants of the English source sentences were encoded in a word lattice and then decoded monotonically. The results for the English→Danish language pair confirmed the improvements previously reported for other language pairs in the context of the preordering for SMT.

Tromble and Eisner (2009) applied machine learning methods to automatically induce the preordering model. They applied the averaged perceptron algorithm on the heuristically transformed German data into the English-like order in order to train the

reordering model. The transformation of the German data relied on the movements derived from the word alignment. Their reordering model incorporated the information about the words, their positions, POS tags and contexts. The model learned permutations of the words in the input string and was capable of assigning a different score to every possible permutation of a source-language sentence. The reordering model was used to preorder the source language data in a deterministic way allowing for the reorderings during the decoding. Again, preordering of the German source data prior to training and translation improved the English SMT output.

While the works presented in the preceding paragraphs considered a single language pair, Xu et al. (2009) dealt with a number of different language pairs. The aim was to use a single set of reordering rules based on the dependency trees to deal with English as a source language and a number of the different SOV languages as target languages. They manually defined a set of generic rules (e.g. moving verbal complex to the clause end, flipping the positions of the prepositional phrases and the modifying nouns, etc.) to transform the SVO (English) to the SOV word order and applied them to all language pairs under consideration. They observed that using a single set of the generic reordering rules led to better translations for all considered language pairs. Even when the monotonic decoding was applied, the translations improved significantly.

Niehues and Kolss (2009) explored the possibility of automatically learning the reordering rules on the POS-tagged sentences for multiple language pairs. The rules were based on POS tags, as well as on the source words. To cope for the long-range differences between English and German, mainly with respect to the verb positions, they allowed *discontinuous* reordering rules which allow for arbitrary words between fixed positions determined by the specific POS tags. Similarly to the previous approaches, the non-deterministic rules were automatically extracted from an aligned parallel corpus. Each of the rules got assigned a weight which indicated the relative frequency of the respective rule according to the corpus used to extract the rules. During reordering rule application, the different reorderings of an input sentence were derived and represented in a word lattice. The approach has been evaluated on English→German, as well as {German,French}→English. All of the MT outputs improved compared to the respective baseline translations.

Multilingual preordering was investigated further by developing language-independent methods for the automatic extraction of the reordering rules in contrast to Xu et al. (2009) who proposed a set of manually written rules which was applied on different language pairs. Genzel (2010) published work on the automatic extraction of the re-

ordering rules from a word-aligned parallel corpus, in which the source side data was shallow-parsed (containing information about POS, as well as dependency types). The automatically extracted rule set which described the permutation of the child nodes in a parse tree was non deterministic – the reordering alternatives of the source language test sentences were encoded within a weighted word lattice and decoded monotonically. The approach was tested on many different language pairs while the source language was always English. In addition, they also tested the method on German→English because of the interesting long-range reorderings related to the position of the verbs. As expected, the approach improved MT quality for all considered target languages. An interesting observation made for English→German was that allowing for the reordering during decoding led to the better German translations compared with those generated via monotonic decoding. Nakagawa (2015) reimplemented the method proposed by Neubig et al. (2012) and presented results for many different language pairs without constraining the source to a single language. His efficient reimplementation of the fully language-independent Bracketing Transduction Grammar (BTG) based preordering does not require any pre-processing of the training data. Instead, the words are enriched solely with their word classes obtained by using Brown clusters (Koo et al., 2008). His method led to the improvement of the MT quality for many different language pairs without significant speed overhead compared to the SMT systems without preordering.

All of the methods mentioned above make use of the automatically computed word alignment between the source and the target language training corpora. Neubig et al. (2012) not only presented a novel method for the automatic learning of a reordering model, but they also carried out experiments using different methods for automatic word alignment. Their discriminative preordering model was automatically learned from a word-aligned parallel using corpus by using an online large-margin method. They tested two different automatically induced preordering models, one of them containing no linguistic knowledge while the other one also incorporated knowledge about POS tags of the words. They tested the models on the English→Japanese, as well as on the Japanese→English SMT and reported on improvements for both translation directions. Furthermore, Neubig et al. (2012) explored the impact of the quality of the word alignment on the quality of the reordering models and experimented with training of the reordering models using a manually aligned corpus, as well as the automatically aligned training data. The experiments showed that even models trained on the automatically aligned data led to the significant improvements of the SMT output.

	Sentence
(EN)	The boy will read a book tomorrow.
(ENr)	The boy will a book tomorrow read.
(DE)	Der Junge wird morgen das Buch lesen.

Table 4.1.: Example of a reordered English sentence according to the German syntax. The original English sentence is denoted by *EN*, while its reordered variant is denoted by *ENr*.

4.3. Preordering for English→German

The handling of the word order problems in English→German SMT proposed in this work relies on the pre-processing of the source language data in order to eliminate syntactic differences between English and German. The source language sentences are transformed in a target-language specific way: we thus follow preordering approaches which solely include the transformation of the source language sentences. This source data modification, i.e. reordering of the source words, is performed prior to the training and testing of a SMT model which is why this method is referred to as *pre*-ordering. An example of the source data reordering for English→German is shown in Table 4.1. Note that the reordered English sentence involves only changing of the positions of the verbs. In this work, we focus only on the verbs since their differing positions are most problematic when it comes to the statistical machine translation from English to German.

In order to reorder the English sentences, we manually define a set of reordering rules which are applied on the English parse trees. The rule set is deterministic: for each English sentence e, only a single reordered alternative e' is generated. Since we limit the reordering rule set only to the verbal elements in an English sentence, the rule set is rather small: it consists of a total of nine rules which capture the most prominent positions of the verbs in English and German. We choose to formulate reordering rules on the basis of the constituency parse trees because the position changes of the verbs in English can most appropriately be described in terms of syntactic structures such as specific syntactic phrases or larger syntactic units like clauses.

The implemented reordering method does not delete any words from the source language nor insert the new words. In some cases, this might be desirable to eliminate specific tense-related differences between English and German. However, we decide not to delete any words (specifically the English auxiliaries which do not exist in German) but rather to reorder them in such a way that the SMT profits most from their (re-

ordered) positions. Adding the words (auxiliaries which exist in German but not in English) is problematic because it imposes a specific (sometimes appropriate, sometimes less appropriate) mapping of the syntactic tense forms between the two languages.

Since the positional differences between English and German not only include the verbs, but also other words, we allow the SMT to perform reorderings during the translation step if needed. While the verb-related positional differences often require the SMT to perform problematic long-range reorderings, the other differences mainly involve short-range reorderings which are well modeled within the SMT.

In Section 4.3.1, we give a detailed description of the developed reordering rules. Main concepts in the reordering rule presentation are *clause type* and *VC type* since the position of the verbs in German depends on them as outlined in Section 3.3.2. Most of the reordering rules imply movements towards the *clause-final position*. How this position is defined, as well as other implementation specifics of the preordering approach for English→German SMT are presented in Section 4.3.2.

4.3.1. Reordering rules

The linguistic analysis of the positions of the verbs in English and German outlined in Section 3.3 reveals that the positions of the verbs in the two languages depend on the type of a clause they are placed in. In other words, the different reordering rules which we aim at defining in this Section are triggered by the type of a given clause. The second constraint is the composition of the VCs. Reordering is not only applied to the verbs, but also the negation and verbal particles. Depending on the given VC, a single rule may thus contain a sequence of movements which reorder different elements of a given VC.

In the following subsections, we define the reordering rules which map the English syntax to the German in the verb position related aspects. The rules are presented separately for each of the clause types presented in Section 3.3.2.

4.3.1.1. Declarative main clauses

The German declarative main clauses belong to the V2 clauses in which the finite verb (VFIN) is placed in the 2nd position in a clause directly following the subject (SUBJ) placed in the first position. In case of a composed VC, the non-finite verbs are placed at the clause end, while the finite verb is in the 2nd position (discontinuous VC). The two positional patterns corresponds to the patterns (d1) and (d2) shown in Table 4.2.

Pattern	Syntactic structure of a clause			
(d1)	**SUBJ**	**VFIN**	∗	
	Der Junge The boy	liest reads	ein Buch. a book.	
(d2)	**SUBJ**	**VFIN**	∗	**main verb (complex)**
	Der Junge The boy	hat has	ein Buch a book	gelesen. read.
(d3)	∗	**VFIN**	**SUBJ** ∗	
	Seit 3 Stunden For 3 hours	liest reads	der Junge ein Buch. the boy a book.	
(d4)	∗	**VFIN**	**SUBJ** ∗	**main verb (complex)**
	Vor 3 Stunden 3 hours ago	hat read	der Junge ein Buch the boy a book	gelesen. read.

Table 4.2.: Position of the verbs in the German declarative clauses. Asterisks are place holders for arbitrary sentence constituents. SUBJ refers to the subject NPs, VFIN refers to the finite verbs, while *main verb (complex)* includes non-finite verbs as described in Section 3.2.1 on page 27.

In order to transform the English declarative sentences to correspond to the German syntax, the movement of the English non-finite verbs to the clause end has to be performed while no movements of the finite verbs are required. The reordering rules for the English declarative main clauses are given in Table 4.3. Example (32) shows an example English sentence along with its reordered version, as well as the corresponding German translation.

In order to transform the English declarative sentences to correspond to the German syntax, the movement of the English non-finite verbs to the clause end has to be performed while no movements of the finite verbs are required. The reordering rules for the English declarative main clauses are given in Table 4.3. Example (32) shows an example English sentence along with its reordered version, as well as the corresponding German translation.

(32) (EN) I have not carried out that experiment yet.
(ENr) I have that experiment yet not carried out.
(DE) I habe dieses Experiment noch nicht durchgeführt.

Rule	VC type	Reordering steps	Example
(Rd0)	Simple	No reordering required	*The boy reads a book.*
(Rd1)	Composed	1. The main verb (complex) is moved to the clause end	*I have not out that experiment yet carried.*
		2. In case of the negation particle, the negation is moved in front of the reordered main verb (complex)	*I have out that experiment yet not carried.*
		3. In case of a verb particle, move the particle after the reordered main verb.	*I have that experiment yet not carried out.*

Table 4.3.: Summary of the reordering rules for the English declarative clauses. Reordering steps for the composed VC are illustrated on an English sentence *'I have not carried out that experiment yet.'*

4.3.1.2. Declarative main clauses with a peripheral phrase

Main clauses with a peripheral phrase are sentences in which there is a prepositional or an adverbial phrase in the VF. The German main clauses with a peripheral phrase belong to the group of V2 sentences, however, in contrast to the declarative main clauses described in the preceding subsection, the finite verb is now placed in front of the subject. This position is captured by the patterns (d3) and (d4) given in Table 4.2.

Since the English verbs are placed after the subject in all type of the declarative clauses, they need to be moved after the subject according to the patterns (d3) and (d4). In case of a composed VC, the English main verb complex needs to be put at the clause end. We thus distinguish between two reordering rules for English which are summarized in Table 4.4, while example sentence pairs are given in (33). Note that the rule (Rd2) does not include the movement of the negation. This is due to the definition of a negated VC in English as a composed VC as explained in Section 3.2.2, page 27. As such, the negation is reordered with the rule (Rd3) defined for the composed VCs.

(33) a. During the break, I went to the canteen.
During the break, went I to the canteen.
Während der Pause ging ich in die Mensa.

b. Before you came, I had not eaten in the canteen.
Before you came, had I in the canteen not eaten.
Bevor du kamst, habe ich in der Mensa nicht gegessen.

Rule	VC type	Reordering steps	Example
(Rd2)	Simple	1. The finite verb is moved in front of the subject	*During the break, went I to the canteen.*
		2. In case of a verb particle, move the particle after the reordered finite verb.	
(Rd3)	Composed	1. The main verb (complex) is moved to the clause end	*Before you came, I had not in the canteen eaten.*
		2. In case of the negation, the negation is moved in front of the reordered main verb (complex)	*Before you came, I had in the canteen not eaten.*
		3. In case of a verb particle, move the particle after the reordered main verb	
		4. The finite verb is moved in front of the subject.	*Before you came, had I in the canteen not eaten.*

Table 4.4.: Summary of the reordering rules for the English declarative clauses with a peripheral phrase. Reordering steps for a simple English VC are illustrated on an English sentence *'During the break, I went to the canteen.'* while the reordering steps for a composed VC are shown on the sentence *'Before you came, I had not eaten in the canteen.'*.

4.3.1.3. Subordinate clause

Subordinate clauses typically begin with a conjunction, a wh-word or a relative pronoun. In the German subordinated clauses, the entire verbal complex is placed at the clause end, thus they belong to the group of the VE sentences. If a composed VC is given, the finite verb is placed after the main verb. The placement of the verbs in the German subordinate clauses corresponds to the pattern (s1) shown in Table 4.5. The transformation of the English clauses according to the pattern (s1) means that entire English VCs in the subordinate clauses always needs to be moved at the clause-final position. The reordering rules are given in Table 4.6 while the reordering steps are illustrated in Example (34). Given a composed VC for instance, we first move the main verb (complex) to the clause end. Subsequently, the negation is placed before the reordered main verb (complex). If the VC contains a verb particle, the particle is placed after the main verb

Pattern	Syntactic structure of a clause		
(s1)	**conj/rel/wh**	**SUBJ** ∗	**verbal complex**
	weil	der Junge ein Buch	liest/gelesen hat.
	because	the boy a book	read/has read.
	weil	der Junge	geschlafen hat.
	because	the boy	was sleeping.

Table 4.5.: Position of verbs in the German subordinate clauses.

(complex). The last movement involves the placement of the finite verb at the clause end, i.e. after the main verb (complex) or after the verb particle.

Rule	VC type	Reordering steps	Example
(Rs1)	Simple	1. The finite verb is moved at the clause end	*because the boy that book read.*
		2. In case of a verb particle, move the particle after the reordered finite verb	
(Rs2)	Composed	1. The main verb (complex) is moved to the clause end	*because the boy has not that book been reading.*
		2. In case of the negation, the negation is moved in front of the reordered main verb (complex)	*because the boy has that book not been reading.*
		3. In case of a verb particle, move the particle after the reordered main verb	
		4. The finite verb is moved at the clause end	*because the boy that book not been reading has.*

Table 4.6.: Summary of the reordering rules for the English subordinate clauses. Reordering rules for a simple VC are illustrated on an English subordinate clause *'because the boy read that book.'*, while the rules for a composed VC are shown on the clause *because the boy has not been reading that book.*

(34) a. (EN) because the boy read that book.
(ENr) because the boy that book read.
(DE) weil der Junge dieses Buch las.

Pattern	Syntactic structure of a clause		
(inf1)	(main clause)	∗	**verbal complex**
	(Ich habe etwas)	mit dir	zu besprechen.
	(I have something)	with you	to discuss.

Table 4.7.: Position of verbs in the German non-finite clauses.

b. (EN) because the boy has not been reading that book.
(ENr) because the boy that book not been reading has.
(DE) weil der Junge dieses Buch nicht gelesen hat.

Similarly to the definition of the rules (Rd2), the rule (Rs1) does not contain the movement of the negation particle. Since the negated simple VCs are treated as composed VCs, the negation in subordinate clauses is reordered according to the rule (Rs2).

In the German subordinate clauses, the finite verb is usually placed after the non-finite verbs (cf. Section 3.3.2). We do put the English finite verbs after the reordered main verb (complex), however, we assume that a possibly *false* order of the verbs at the clause end does not prevent a SMT model to place their translations into the correct order because the required small-range reordering can be well captured by SMT.

4.3.1.4. Non-finite clauses

Non-finite clauses do not have a finite verb, but consist of one or more non-finite verbs along with an optional infinitival particle (in English *to*, in German *zu*). The German infinitival clauses belong to the VE sentences in which the entire VC is placed at the clause end. The position of the verbs in the German non-finite clauses corresponds to the pattern (inf1) described in Table 4.7.

Due to the VE position of the verbs, the English non-finite clauses always need to be reordered. Since we do not distinguish between simple and composed non-finite VCs, there is only one reordering rule for the English non-finite clauses which is summarized in Table 4.8. An example of a non-finite reordered English clause is given in (35).

(35) (EN) (*The boy promised*) not to cheat during the exam.
(ENr) (*The boy promised*) during the exam not to cheat.
(DE) (*Der Junge hat versprochen*) während der Prüfung nicht zu schummeln.

Rule	VC type	Reordering steps	Example
(Rinf1)	Simple/ Composed	1. In case of the negation, the negation is moved at the clause end	(*The boy promised) to cheat during the exam not.*
		2. In case of an infinitival particle, the particle is moved after the negation (i.e. at the clause end)	(*The boy promised) cheat during the exam not to.*
		3. The main verb (complex) is moved after the reordered particle (i.e. at the clause end)	(*The boy promised) during the exam not to cheat.*
		4.In case of a verbal particle, the particle is moved after the reordered main verb (complex)	

Table 4.8.: Summary of the reordering rules for the English non-finite clauses. Reordering steps are illustrated on the English non-finite clause *'not to cheat during the exam'*.

4.3.1.5. Interrogative clauses

In the German and English interrogative clauses, the finite verb position is always equal. In the context of a composed VC, the German non-finite verbs are placed in the clause-final position. Examples of the interrogative sentences in both languages along with their placement patterns are given in Table 4.9. The table indicates that the finite verbs in both languages have the same position. The reordering of the English verbs is required only if a composed tense form is given. The reordering rules are given in Table 4.10 while an example of the reordering steps is given in Example (36).

(36) (EN) Has the boy not been reading the book?
(ENr) Has the boy the book not been reading?
(DE) Hat der Junge das Buch nicht gelesen?

4.3.1.6. Summary

The position of the verbs in German depends on the clause type, as well as on the type of a VC. Some of the possible positions correspond to the verb positions in English,

Pattern	Syntactic structure of a clause		
(i1)	**finite verb**	**SUBJ** $*$	
	Liest	der Junge ein Buch?	
	Reads	the boy a book?	
(i2)	**finite verb**	**SUBJ** $*$	**main verb**
	Hat	der Junge ein Buch	gelesen?
	Has	the boy a book	read?

Table 4.9.: Position of the verbs in the German interrogative clauses.

Rule	VC type	Reordering steps	Example
(Ri0)	Simple	No reordering required	*Is the boy happy?*
(Ri1)	Composed	1. The main verb (complex) is moved to the clause end	*Has the boy not the book been reading?*
		2. In case of the negation particle, the negation is moved in front of the reordered main verb (complex)	*Has the boy the book not been reading?*
		3. In case of a verb particle, the particle is moved after the reordered main verb	

Table 4.10.: Summary of the reordering rules for the English interrogative clauses.

however, the majority of the positions differ between the two languages. Most of the movements are towards the clause-final position. The non-finite elements of a VC are moved to the same positions for all clause types while the position of the finite verbs depends on the clause type. In total, we define nine reordering steps which are combined into six different reordering rule sets (i.e., reordering rules). Some of the reordering rules are the same (e.g., there is no distinction between the VC types for the non-finite VCs (Rinf1); (Rd1)=(Ri1)), while some of them are empty, i.e., no reorderings need to be performed: (Rd0), (Ri0).

Out of ten clause/VC type combinations defined in this work which are relevant for the preordering of English for the English→German SMT, only two of them do not require any movements of the English verbs. In other words, almost all of the English clauses have to be reordered to adapt the position of their verbs to the German syntax. This fact makes is very clear that explicit handling of the different verb positions between English and German in the context of the English→German SMT is a very important

task if the quality of the German SMT translations is to be considerably improved.[5]

Obviously, the rules are only moving the words within the corresponding sentence: there are no insertion or deletion rules in our rule set as already briefly mentioned in Section 3.3.3. There are many differences between the English and German tense forms which could be eliminated by adding or deleting words in the source language.

Consider, for example, the English sentence in (37). In (37a), the verb *bought* is translated into *kaufte*, while in (37b), it is corresponds to the verb sequence *'hat gekauft'*. The *Perfekt* tense is the most commonly used past tense in German. When translating the English simple past tense into the German *Perfekt* tense, the SMT model needs to generate an auxiliary which is not given in the source. Adding a pseudo auxiliary into the English sentence would increase the syntactic similarity to the German translation in *Perfekt* tense, however, it could also lead to unwanted verbose translations. Since both of the German tense forms are valid translations of the English simple past tense, we decide not to insert pseudo auxiliaries into the English sentences and let the SMT model to choose between the two German past tense forms.

(37) a. The boy bought a book yesterday. → simple past tense
Der Junge kaufte/gekauft ein Buch gestern.
'Der Junge kaufte gestern ein Buch.' → *Präteritum*

b. The boy bought a book yesterday. → simple past tense
Der Junge kaufte/gekauft ein Buch gestern.
'Der Junge hat gestern ein Buch gekauft.' → *Perfekt*

The syntactic similarity between the English and German VCs could in come cases also be increased by deleting auxiliaries in the source sentences. For example, the English progressive present tense is composed of an auxiliary and the main verb as shown in Example (38).

(38) The boy is reading a book. → present progressive
Der Junge ist lesend ein Buch.
'Der Junge liest ein Buch.' → *Präsens*

The German translation of the English verb sequence *'is reading'* is *liest*, thus, there is no direct German counterpart of the English auxiliary *is*. However, deleting the English auxiliary would lead to a loss of the very important contextual information which helps

[5]The problem of the different verb positions is not only interesting for the translation direction English→German, but also for German→English for which Collins et al. (2005) described pre-ordering rules similar to those presented in this section.

to generate the correct translation for the respective English VCs. We thus decide not to delete such auxiliaries but to treat them as a part of the main verb complex and reorder them along with the corresponding main verb.

4.3.2. Implementation details

The discussion about the differences regarding the position of the verbs in English and German given in Section 3.3 indicates which kind of knowledge is required to perform reordering of the English sentences in order to make them syntactically more similar to German. The definitions of the reordering rules outlined in Section 4.3.1 are based on the type of a clause that a VC to be reordered is placed in. The most prominent positions of the verbs in German can be set into the relation to the corresponding subject phrases, the clause-initial and the clause-final positions. Regarding the VCs, it is needed to determine their type which is dependent on the verbs, i.e. their parts-of-speech (e.g., finite verb (VBP, VBD, VBZ, MD), participle (VBN), auxiliary, etc.).

The rules require access to a specific syntactic information of English sentences. Especially, the representation of the sentences is required which reveals information about the phrase and clause boundaries. Due to this reason, we choose to apply the rules on English constituency parse trees. The constituency parse trees also include the POS tags of the words in the analyzed sentences which are required to determine the type of the VCs, as well as the verbs to be moved. Most of the information relevant for the reordering is directly read out from the tree. However, some information needs additionally be derived such as the number and type of the auxiliaries within a VC.

In this section, the implementation of the preordering approach for English→German is described. We first present some modifications of the trees that the reordering method relies on. We then explain the derivation of the not directly accessible knowledge about the English VCs and finally present details about the implementation of the reordering method.

4.3.2.1. Parsing

The English sentences are parsed using the constituency parser of Charniak and Johnson (2005). An example tree is given in Figure 4.2. The tree includes all the information needed to perform the reorderings:

(i) POS tags which indicate the type of a VC element (e.g. verb, negation, particle, finite, etc.),

(ii) sentence constituents such as NPs and VPs,

(ii) clauses (e.g., S and SBAR) which give information about the clause types, as well as about the span of the clauses within a sentence.

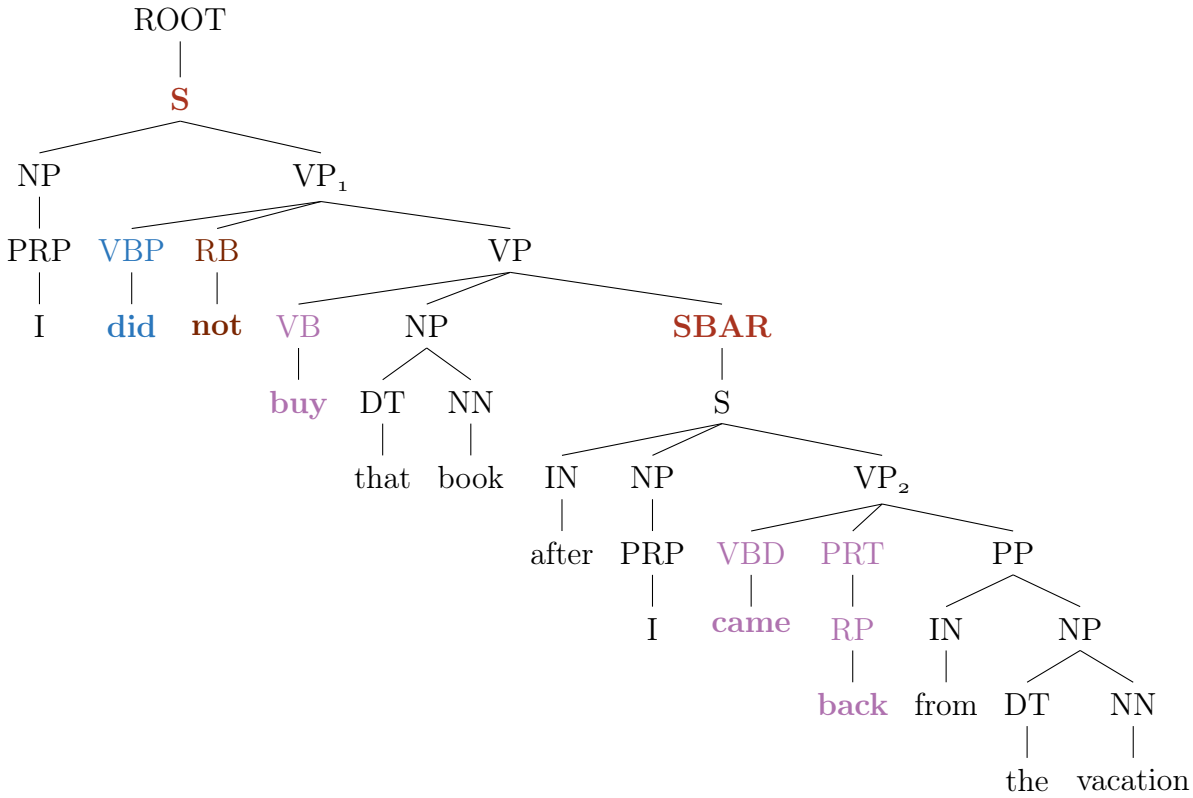


Figure 4.2.: Constituency parse tree for an Example English sentence. The sentence consists of two subclauses indicated by the nodes *S* and *SBAR*. The VCs are rooted in the nodes *VP1* and *VP2*, respectively.

The Charniak parse trees have different labels on the clause level. For example, the node label *SBAR* denotes a subcategorized clause, the label *SQ* denotes an interrogative clause, etc.[6] The availability of such clause type distinguishing labels is crucial for the application of the reordering rules. However, there is no special label for non-finite clauses as illustrated by the tree in the left panel of Figure 4.3 where both the declarative main clause under the node S_1, as well as the non-finite subcategorized clause under the node S_2 have the same node label, namely *S*. The same S-node label ambiguity is

[6]The list of the English POS tags used in the Penn Treebank on which the Charniak parser is trained can be found here: `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html`.

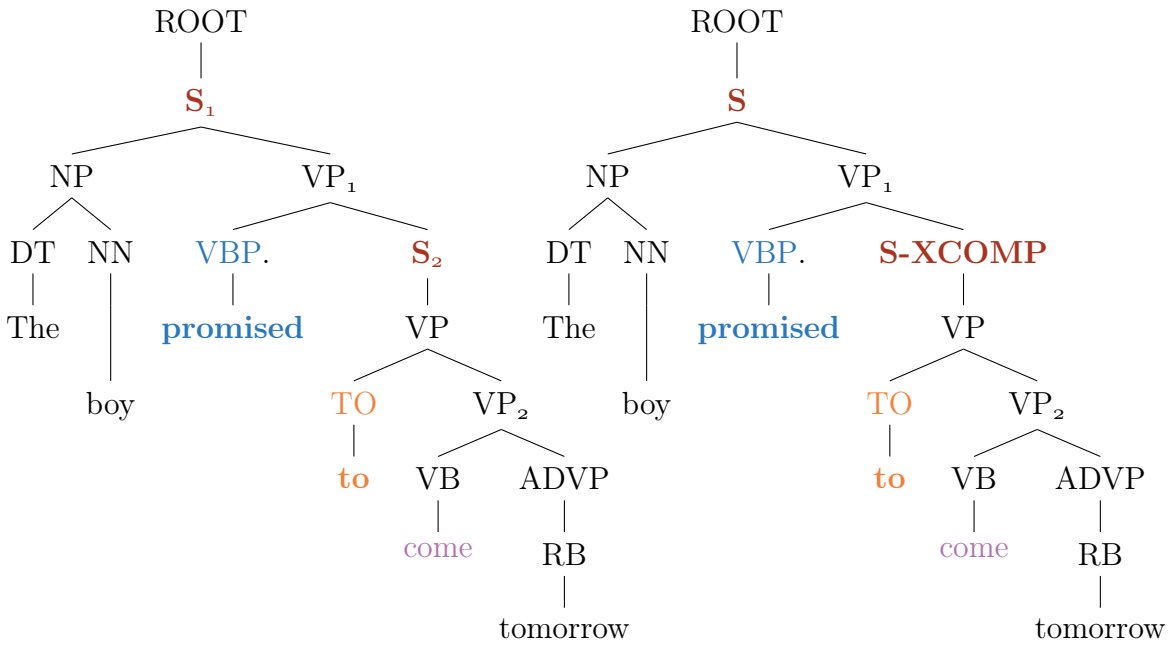


Figure 4.3.: Constituency parse tree for an example English sentence consisting of a non-finite subcategorized clause. The tree on the left side is the original tree, while the tree on the right shows the relabeling of the node S_2 to *S-XCOMP*.

also given for the declarative main clauses versus the declarative main clauses with a peripheral phrase.

In order to keep the rules as simple as possible, we implemented a script which relabels specific S-nodes in the Charniak trees which are interesting for the reordering. This clause labeling step is applied to the original parse trees. The output is then used in the reordering step. Node relabeling is described in the following paragraphs.

Non-finite subordinate clauses We enrich root S-nodes of the non-finite subordinate clauses with the label *XCOMP* as shown in the right panel in Figure 4.3. The identification of the respective subtrees is relatively easy: we assume that a non-finite clause is rooted in a S-node which is directly dominated by a VP node.

Declarative main clauses with a peripheral phrase Relabeling is also performed on the S-nodes enclosing a main clause with a peripheral phrase in front of the subject as shown in Figure 4.4. Root S-nodes of such clauses get the suffix *EXTR* attached. The context condition for the label *EXTR* is fairly simple: if in an English subtree rooted in a S-node, the NP node is not the leftmost child node of S (i.e., the subject NP is not the first constituent in the given clause), then the S-node is enriched with the label *EXTR*

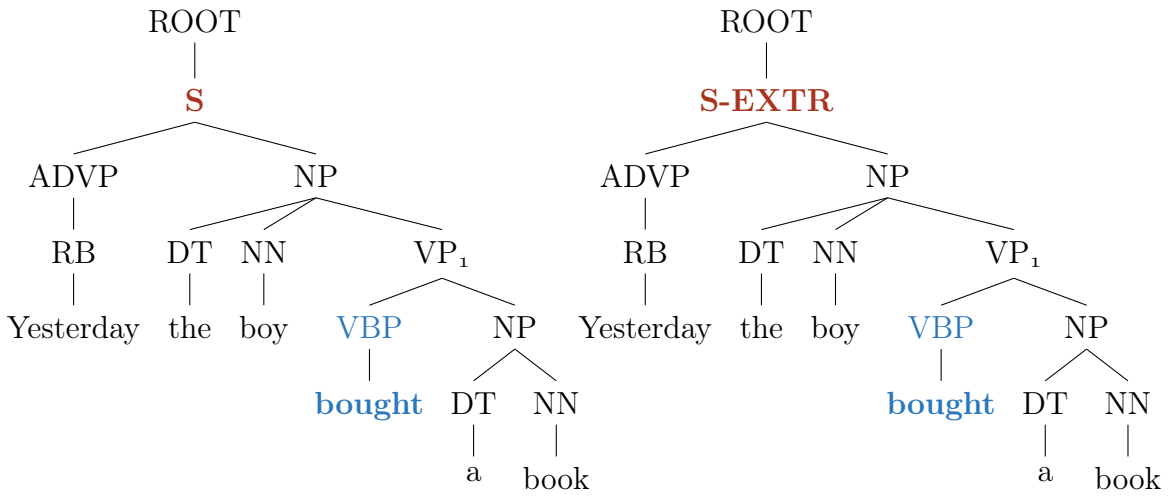


Figure 4.4.: Constituency parse tree for an example English sentence consisting of an adverbial in front of the subject *'the boy'*. The tree on the left side is the original tree, while the tree on the right shows the relabeling of the *S* node to *S-EXTR*.

as shown in the right panel in Figure 4.4.

4.3.2.2. VC types

The reordering rules defined in Section 4.3.1 depend on the clause type, as well as on the type of the given VC. We distinguish between simple and composed VCs (cf. Section 3.2.2). In some cases, these groups do not follow the actual composition of the VCs. For example, the English VCs in the present progressive tense consists of an auxiliary and a main verb. In a syntactic sense, such VCs are composed, however, in this thesis we treat respective VCs as simple VCs. In the context of the reordering rules presented in the preceding subsection, this means that verbs within the English VCs in the present progressive tense are not split in certain contexts. Instead, they are reordered according rules formulated for the simple VCs.

We distinguish between nine different English VCs, i.e., VC subtypes, which are grouped into simple, composed and non-finite VCs. The VC subtypes are listed in Table 4.11 while the reordering examples for the different VC subtypes are given in Table 4.12.

VC type	VC subtype	Examples
simple	simple	say, said
	simpleaux	am saying, does (not) say
composed	composed	have said, would say, have been
	modaux	would have said, would be saying
	modauxaux	would have been saying
	auxaux	have been saying, is being said
non-finite	auxtoinf	to be said
	gerund	saying
	toinf	to say

Table 4.11.: Categorization of the English VCs. VC subtypes refer to the syntactic composition of the English VCs: e.g., *modauxaux* refers to the following verb sequence: *modal + auxiliary + auxiliary*. Main verb complexes are indicated in pink and indicate which verb sequences are reordered jointly.

VC	Original English	Reordered English
simple	He usually reads in his room.	–
simpleaux	He is reading a book in his room. He does not read a book.	– He does read not a book.
composed	He has read a book. He did not read a book. He would read the book *(if he had time).* He has already been there.	He has a book read. He did the book not read. He would the book read *(if he time had).* He has already there been.
modaux	He would have read a book. He would be reading a book *(if he had time).*	He would a book have read. He would a book be reading *(if he had time).*
modauxaux	He would have been reading a book *(if ...)*	He would a book have been reading *(if ...)*
auxaux	He has not been reading a book. The book is being read by the boy.	He has a book not been reading. The book is by the boy being read.
auxtoinf	*(The book is)* to be read by the boy.	*(The book is)* by the boy to be read.
gerund toinf	Reading a book *(is a nice activity).* *(It is nice)* to read a book.	A book reading *(is a nice activity).* *(It is nice)* a book to read.

Table 4.12.: Reorderings of the different English VCs in declarative sentences (applied reordering rules are (Rd0)-(Rd3)). Verbs in blue are considered to be the finite verbs, while the verbs in pink indicate the main verb complexes.

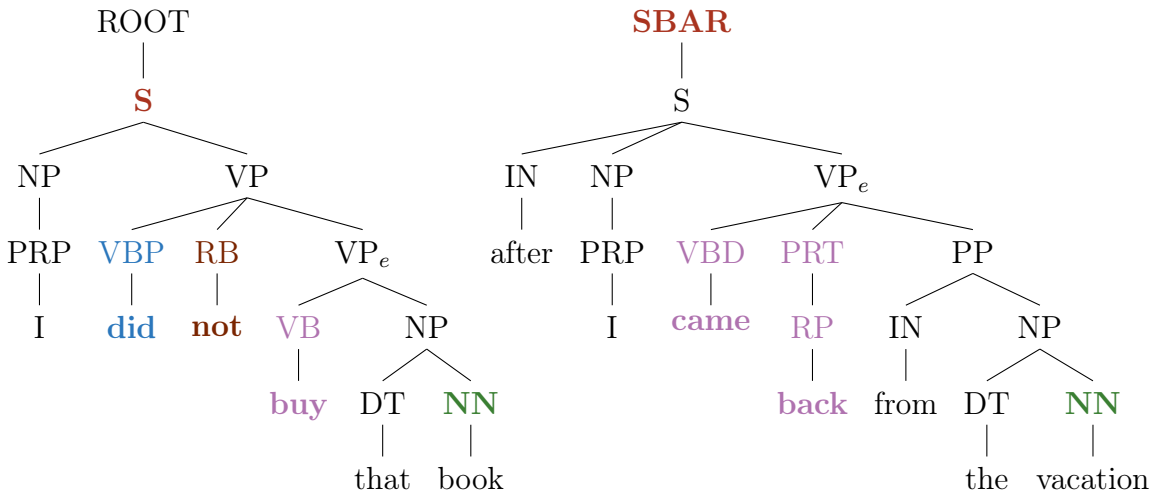


Figure 4.5.: Constituency parse from Figure 4.2 divided into two subtrees representing clauses of the given sentence. Clause-final positions are marked in green.

4.3.2.3. Clause-final position

Almost all reordering steps require knowledge about the end of the current clause. Within the English constituency parse trees, we define this position as the right most child of the last VP, namely VP_e in the VP chain of the given subtree (i.e., clause). If VP_e subcategorizes a S-node S_i, then the clause-final position is before S_i. An example of the division of a parse tree into subtrees according to the given definition of the clause boundaries is given in Figure 4.5.

The reorderings are carried out within the given subtree, i.e., clause. There are no movements across the subtrees. In most of the cases, this restriction corresponds to the German syntax, however, there are also some exceptions which are discussed in Section 3.3.3. An example of such an exception is again shown in (39). In (39a), the non-finite VC *'zu kommen'* is placed after the finite verb of the governing clause, namely *versprochen*. On the other hand, in (39b), the non-finite German VC is placed before the main verb of the governing clause. Given the parse tree in Figure 4.6 of the English sentence in Example (39), moving the English main verb *promised* to the position corresponding to the German verb ordering shown in (39b) would mean that the verb should be moved at the end of the subsequent clause. Since the ordering given in (39a) is also correct and in order to keep the reordering rules as simple as possible, and thus not to implement any exceptions of the restriction that the reorderings are carried out *within* the clause that the given VC (both finite and non-finite) is placed in.

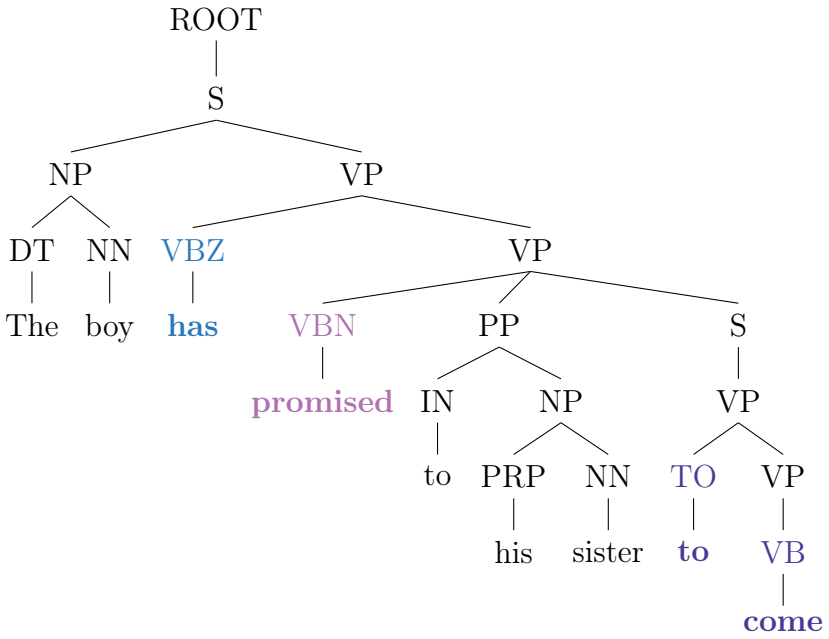


Figure 4.6.: Parse tree of the English sentence in Example (12).

(39) a. The boy has promised to his sister to come.
Der Junge hat versprochen zu seiner Schwester zu kommen.
'Der Junge hat zu seiner Schwester versprochen zu kommen.'

b. The boy has promised to his sister to come.
Der Junge hat versprach zu seiner Schwester zu kommen.
'Der Junge hat zu seiner Schwester zu kommen versprochen.'

4.3.2.4. Implementation

The movement of the verbs in the English trees is a recursive rearrangement of the subtrees. The reordering of a VC happens within the clause that the given VC is placed in. There are thus no verb movements across the clause boundaries. The subtree movements usually include the cutting of a branch rooted in the POS tag which triggers the reordering, and its attachment to a new mother node. Sometimes, the mother node remains the same, but the order of the daughter nodes changes. The reordered English parses have the same number of the terminal nodes like the original trees. They remain well-formed (i.e., it is possible to use them for further processing), although the English sentence itself becomes grammatically incorrect.

Figure 4.7 shows an example of an original parse tree and its reordered variant. The applied rule is (Rd1) (cf. Table 4.3 and Figure 4.8) where the finite verb (indicated by

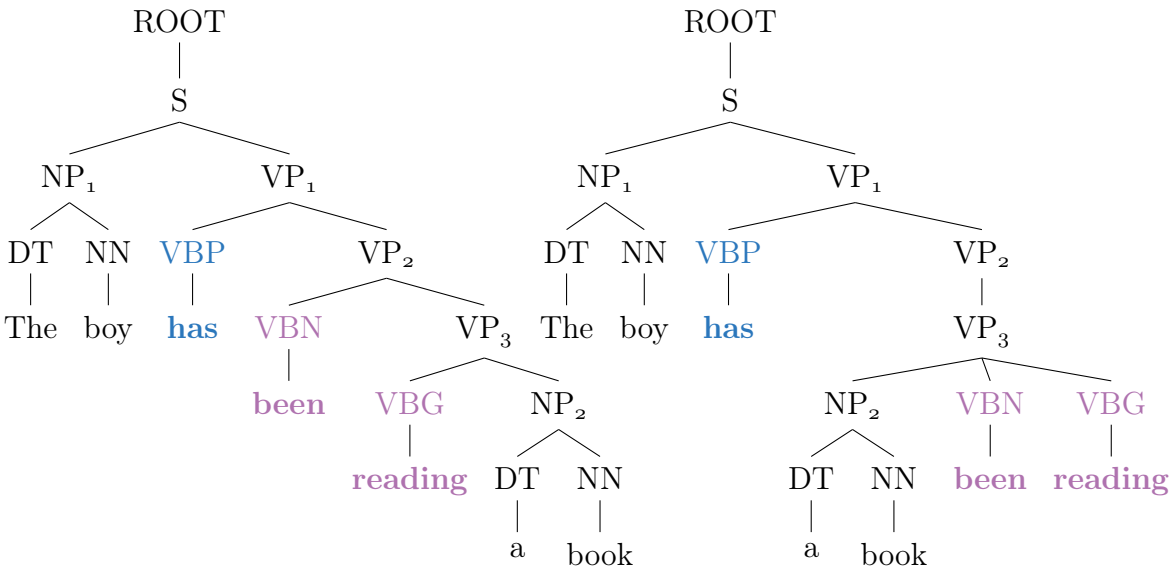


Figure 4.7.: Example for an original parse tree (left) and its reordered variant (right).

the POS tag *VBP*) remains at the original position, while the main verb complex *'been reading'* (indicated by the POS tags *VBN* and *VBG*) moves to the clause-final position (cf. line 10 in Figure 4.8). The branch *'VBG — reading'* is moved after its sister node NP_2 – its mother node remains unchanged. On the other hand, the branch *'VBN — been'* is attached to the new mother, namely VP_3. After moving *been*, the node VP_2 has only one child which is again a VP. We do not remove such nodes in order to reduce the redundancy within the reordered tree. Only if a node remains without children, it is subsequently deleted.

In most cases, the POS tags give clear information about the verb. However, in some cases, it is necessary to take a closer look to the elements of the given VC. For example, all English adverbials are tagged as *RB*. Since we only want to move the negation, we have to check the *word* which is tagged as RB. The word is further considered only if it is *not* of *'t*.

4.3.2.5. Pipeline

Preordering approach is applied on the source data prior to training and translation as shown in Figure 4.9. There is thus no interaction between the reordering of the English sentences and training/applying a SMT system.

Reordering is carried out as a pre-processing step of the English corpus. First, the English data parsed. Next, the reordering is applied on the English parse trees. Finally,

```
1: function REORDER(subtree)
2:     new_tree ← copy(subtree)
3:     vc ← get-VC(subtree)
4:     vcType ← get-vcType(vc)
5:     finV ← get-finV(vc)
6:     mainV ← get-mainV(vc)
7:     prtV ← get-prtV(vc)
8:     neg ← get-neg(vc)
9:     if vcType is composed then
10:        new_tree ← move-mainV(subtree, mainV)
11:        if neg not empty then
12:            move-neg(subtree, neg)
13:        end if
14:        if prtV not empty then
15:            move-prt(subtree, prtV)
16:        end if
17:    end if
18:    return new_tree
19: end function
```

Figure 4.8.: Function for reordering rule (Rd1). Then function is called after identifying the clause type as *declarative main clause*.

the leaf nodes of the reordered parse trees are collected in order to get the reordered version of the input English data. This pipeline is applied not only on the training data, but also on the English tuning, as well as testing data sets. As for the German data, it does not undergo any modification steps.

The SMT model is trained on a combination of the reordered English corpus with the non-modified German corpus. The SMT model is tuned on the reordered English dev set. The decoding of the English test data is applied on the reordered its reordered version.

4.4. Chapter summary

In this Chapter, we presented a detailed description of the preordering approach for SMT adapted to the English→German translation direction.

In Section 4.1, we first motivated the need for explicit handling of the position of the German verbs in the SMT outputs. According to two different evaluation studies of the German SMT (along with NMT) outputs, most of the word-order errors in the German translations are related to the verbs. Our example SMT outputs showed that

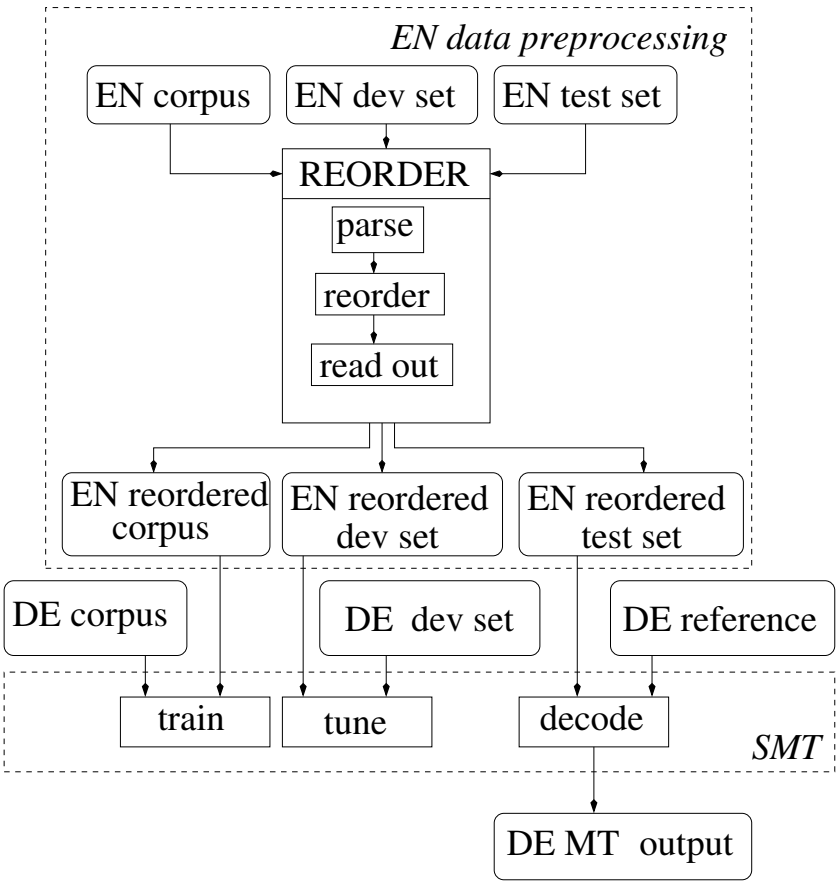


Figure 4.9.: Preordering of the English data as a part of the pre-processing step. The English data is reordered prior to the training, tuning and applying a SMT system.

SMT does not only place the German verbs into false positions, but it also often does not even generate them which has a negative impact on understanding the German SMT outputs.

In Section 4.2, we outlined previous work on the *preordering approach for SMT* which was also implemented in this work. Analysis of the previous research on this topic showed that preordering is a simple and effective way to cope with positional differences between many different language pairs in the context of statistical machine translation. There are many variants of preordering: implying modification only of a source language or of both source and the target language, deterministic vs. non-deterministic, using manually written or automatically derived sets of reordering rules, applying rules on the surface, POS-tagged or parsed sentences. All of them lead to better SMT outputs compared to those generated by the baseline SMT models.

In Section 4.3, we defined our preordering method as a deterministic modification of

the SL (i.e., English) data. The modification of the English sentences is based on a set of hand-crafted reordering rules which are applied on the English constituency parse trees. In Section 4.3.1, we gave a detailed description of the reordering rules developed for the English→German translation direction. We defined a total of nine reordering steps which are grouped into six rule sets. These are applied in a specific combination of the type of a VC, as well as type of a clause that the given VC occurs in.

Details about the implementation of our preordering approach were given in Section 4.3.2. First, in Section 4.3.2.1, we presented two modifications of the English parse trees in order to ensure simple and unique identification of the clause types. In Section 4.3.2.2, we gave a complete list of the English VCs which are considered for the definition of the reordering rules. Since many reordering rules imply movements of an English verb to the clause-final position, we defined that position in the used parse trees in Section 4.3.2.3. Finally, the implementation was described in Section 4.3.2.4, while the entire processing pipeline including its combination with SMT was sketched in Section 4.3.2.5.

5. SMT experiments with reordering

The success of the preordering approach depends on different factors which we aim to explore in the experiments described in this Chapter. In Section 5.1, we first give an overview of the different SMT experiments carried out with the preordering approach presented in the preceding Chapter. Our baseline SMT model, as well as the tools used to build SMT models are described in Section 5.2. In Section 5.3, we then present a number of different SMT models trained on the preordered version of the English training data. We perform both automatic evaluation of the different SMT outputs in terms of BLEU, as well as a manual inspection of the German translations which is described in Section 5.4. In Section 5.5, the findings of the performed experiments are discussed whereby we take a closer look to the applied reordering rules, to the performance of the used parser, the training speed and the type of the data used to train SMT models. Finally, the chapter summary is given in Section 5.6.

5.1. Overview of the experiments

We test the effectiveness of the preordering approach for English→German SMT in different experimental setups which are outlined in this Section. The experiments and the evaluation of preordering within the different setups described in the following paragraphs are presented in Section 5.3.

Lexicalized reordering Since preordering is deterministic and captures only the differences regarding the position of the verbs in English and German, we allow the SMT model to perform further required reorderings in the appropriate contexts during the decoding step. In the first set of experiments, we combine preordering with different lexicalized reordering models to find the one performing best for the English→German SMT. We consider three different variants of the lexicalized reordering model: word-based (Koehn et al., 2005), phrase-based (Tillmann, 2004) and hierarchical (Galley and Manning, 2008). The word-based model computes phrase orientation based on the word

alignment during the training, and based on the phrase alignment during the decoding step.[1] In contrast to this, the phrase-based model uses the phrases both during training and decoding. The hierarchical model uses a combination of several phrases to compute the orientation of a given phrase. By considering bigger context, the hierarchical model aims at better handling of the long-range reorderings to which also the problem of the correct placement of the verbs in the German SMT outputs belong. The experiments with the above mentioned reordering models are outlined in Section 5.3.1.

Word alignment We combine preordering with two different approaches for automatic computation of word alignment: *Giza++* (Och and Ney, 2003) and *FastAlign* (Dyer et al., 2013).[2] Giza++ implements IBM Models 1 and 4. Since it considers differences in the placement of the equivalent words in the source and target language sentence, it has a high alignment accuracy. FastAlign, on the other hand, is based on a modification of the IBM Model 2 which does not explicitly cope for positional differences between equivalent words. Consequently, the alignment accuracy is lower compared with Giza++.

The comparison of the performance of preordering in combination with the two word alignment methods is interesting with respect to the time needed to compute word alignment and the time needed to perform preordering. Particularly the use of FastAlign in combination with preordering is interesting because FastAlign is much faster than Giza++ and thus could compensate the additional training time caused by the preordering step. Furthermore, Ding et al. (2015) claimed that FastAlign, and thus also the MT quality, can be improved by preordering the corpus data. We test this hypothesis for the English→German SMT within the experiments described in Section 5.3.2.

Domain We train SMT models on two different data sets. The first set contains texts from the news domain, as well as from the domain of politics while the second data set includes texts from the medical domain. There are two reasons for this: (i) to test the interaction of parsing accuracy with the overall SMT improvement, and (ii) to measure benefit of preordering on different types of data.

Reordering is applied on the syntactic parse trees of the English sentences. The correctness of reordering directly depends on the parsing accuracy. Often, parsers are less accurate when they are applied on *out-of-domain* data, i.e., data from a domain

[1] More details about the lexicalized reordering models and the phrase orientation can be found in Section 2.2.5.

[2] For more details, please refer to Section 2.2.1.

which does not correspond to the domain that the data they are trained on comes from. Parsers are usually trained on texts from the news domain which is also the case for the parser used in this work. Nevertheless, we apply the same parser on the two above mentioned data sets to test whether reordering leads to the performance boost of the SMT models even when it is applied on data which is out-of-domain with respect to the used parser.

The presentation of the reordering rules given in Section 4.3.1 revealed that sometimes, reorderings are not needed since the verbs in English and German have equal positions. The respective contexts are related to the type of the given VC, as well as to the type of a clause that the given VC occurs in. Both of these contextual constraints are tightly related to the type/domain of the given text. For example, the most prominent reordering applies to the English subordinate clauses. In certain text types, it is possible that the proportion of such clauses is considerably lower than in some other text types. Furthermore, the VC types which are also important for reordering are directly related to the tense forms used in a certain text. We have seen that simple VCs (i.e., in simple present or past tense) in declarative clauses do not need to be reordered. Similarly to the clause types, there might be texts with very high proportion of this specific clause/VC type combination for which reordering would consequently not lead to higher quality of the German SMT outputs.

The experiments, as well as the evaluation of preordering applied on different data sets are presented in Section 5.4.

Sigtest filtering SMT models trained on a big amount of training data rely on very large phrase and reordering tables. Johnson et al. (2007) proposed a method for filtering the phrase and reordering tables not only in order to reduce their size, but also to increase performance of the SMT models by removing non-significant table entries.[3] We combine preordering approach with phrase and reordering table filtering in order to examine the effect of sigtest filtering on the performance of the baseline, as well as reordered English→German SMT models. The experiments are described in Section 5.3.3.

Evaluation The translations generated by the baseline and reordered SMT models are evaluated automatically, as well as manually. Automatic evaluation is given in terms of BLEU.[4] Automatic evaluation indicates the overall quality of the generated translations

[3]In the rest of the thesis, the significance filtering step is referred to as *sigtest*.

[4]Details about the BLEU metric are given in Section 2.4.

without taking into account specific word categories or linguistic phenomena. Therefore, in addition to the automatic evaluation, we also carry out manual evaluation of the German SMT outputs which is focused on errors regarding the verbs, i.e., generation, as well as the placement of the German verbs in the MT output. It thus represents qualitative analysis of the German SMT outputs.

5.2. General SMT settings

Toolkit All SMT models presented and discussed in this Section are built using the SMT toolkit *Moses* (Koehn et al., 2007). Moses enables training of all components of a standard SMT model: the (phrase-based) translation model, different kinds of the lexicalized reordering models mentioned in the preceding Section (word-based, phrase-based and hierarchical) and the language model. As for the word alignment, it integrates, among others, Giza++ and FastAlign which are both used in this work. Moses supports training and use of different kinds of language models. German language models used in this work are trained with the *KenLM Language Model Toolkit* which implements modified Kneser-Ney language model estimation method (Heafield et al., 2013).

SMT training settings Our SMT settings correspond to Moses' default training settings. The phrase-based translation model consists of the phrases up to the length of 5 words. We use 5-gram German language models for all experiments reported on in this work. We set the distortion limit to 6 words. The model weights are optimized using the Minimum Error Rate Training (MERT) (Och, 2003). The maximum sentence length is set to 80 words. The SMT models are trained on the tokenized data. While the English side of the corpus is lowercased, the German texts are truecased.

Data Preordering is tested on two different sets of data: (i) the WMT 2015 data (Bojar et al., 2015)[5] and (ii) texts from the medical domain made available for the WMT 2014 medical translation task (Bojar et al., 2014)[6]. The WMT data set is a concatenation of texts from three different domains: news (News Commentary), political discussions (Europarl) and mixed-domain texts crawled from the Web (Common Crawl). News and Europarl are typically clean data, i.e., they consist of complete, well-formed sentences.

[5] http://www.statmt.org/wmt15/

[6] http://www.statmt.org/wmt14/medical-task/, http://www.himl.eu/files/himl-test-2015.tgz

Training		Tuning		Testing	
Corpus	**Size**	**Corpus**	**Size**	**Corpus**	**Size**
News	272k			news2014	3003
Europarl	1.9m	news2013	3000	news2015	2169
Crawl	2.4m			news2016	2999

Table 5.1.: WMT data used for the reordering experiments. The size of the corpora denotes the number of the sentences.

On the other hand, the crawled data may also include foreign language material and different types of sentences: incomplete sentences, single NPs, sentences with many special characters, listings, etc. Examples of such training samples are given in (40). We emphasize this sort of training sentences since for preordering, the data undergoes parsing which is often inaccurate for non-standard sentences. The statistics about the used WMT corpus, as well as of the tuning and testing data are shown in Table 5.1.

(40) a. (EN) Apartments for rent in Marbella.
(DE) Ferienwohnung in Marbella in der Siedlung Las Chapas zu vermieten .

b. (EN) Presumably, Denobula is a [Class M] planet, since [Phlox] and other [Deno- bulans] have no difficulties with the [atmosphere] on board [Enterprise (NX-01) | "Enterprise"]
(DE) Denobula liegt demnach im [Alpha-Quadrant]. Das widerspricht jedoch der Tatsache dass im "Star Trek: Sternenatlas" "Iota Bootis" als der [Stern| Hauptstern] von Denobula Triaxa angegeben wird, dieser sich aber ca.

c. (EN) Holiday apartments | Hotels | Hostels | Camping, Dormis & Bungalows | Last Minute Offers!
(DE) Ferienwohnungen | Hotels | Pension | Camping, Dormis & Bungalows |* Last Minute Angebote!!

The medical data is a compilation of several subcorpora listed in Table 5.2 along with the details about their size, as well as the size of the tuning and testing sets used for the medical experiments. The medical subcorpora are of the different types. For instance, while the Muchmore corpus contains abstracts from the medical journals, UMLS is a compilation of medical terms. The medical corpus is thus a mixture of parallel full sentences and a bilingual terminology list. Examples are given in (41).

(41) a. The baseline PVR correlated inversely with its percentile value during PGI2 ($r=-0.76$, $p<0.05$) .

Training		Tuning		Testing	
Corpus	**Size**	**Corpus**	**Size**	**Corpus**	**Size**
EMEA	1m			Khresmoi	
Muchmore	29k	Khresmoi		medical	1000
PatTR	1.8m	summary	500	summary	
UMLS	2.3m			Cochrane	1583
Wiki titles	10k			NHS24	1258

Table 5.2.: Medical data used for the reordering experiments. The size of the corpora denotes the number of the sentences.

Data	Size	Experiment
WMT DE	4.5m	wmt
DE news mono	>120m	wmt, med
Medical DE	4.5m	med
DE medical mono	1.8m	med

Table 5.3.: Overview of the data used to build the German language models.

Der Ausgangs-PVR korrelierte jedoch invers mit dem prozentualen PVR-Wert unter PGI2 (r=-0,76, p<0,05).

b. 1,2 Dipalmitoyl Glycerophosphocholine
Dipalmitoyllecithin

c. Abadie 's sign of exophthalmic goitre
Abadie-Zeichen

Language models For the WMT (wmt) and medical (med) experiments, we use two different German language models (LMs) which differ in the composition of the training data. For the experiments from both domains, i.e., wmt and med, the German side of the WMT corpus in combination with a large collection of the German monolingual texts from the news domain is used.[7] In addition to this data, for medical experiments, we also use the German side of the English-German medical training corpus (cf. Table 5.2), as well as a collection of the monolingual German medical texts made available for the WMT 2014 medical translation task. The statistics about the data used to build the German LMs are given in Table 5.3.

We train two different LMs: (i) WMT+news LM and (ii) medical LM. The WMT+news

[7]The used German monolingual data can be downloaded from here: `http://www.statmt.org/wmt15/translation-task.html`.

LM is used for the wmt experiments. The LM used for the experiments in the medical domain is a combination of WMT+news LM and medical LM which are combined by interpolating the two LMs using weights optimized on medical development data (Schwenk and Koehn, 2008).

5.3. Combining preordering with SMT

In this section, we present numerous SMT experiments carried out in a combination with preordering which is applied as a pre-processing of the English data as described in Chapter 4. Various experimental setups correspond to the motivation outlined at the beginning of this Chapter, in Section 5.1. We refer to the SMT models trained on the non-modified training data as the baseline (BL) and to the models trained on the reordered version of the English data as reordered (RO).

5.3.1. Lexicalized reordering models

The first set of experiments explores the interaction between preordering and different lexicalized reordering models implemented within Moses: word-based (wbe), phrase-based (phrase) and hierarchical (hier). The SMT systems are built using the WMT data and tested on the test sets from the news domain (cf. Table 5.1 in the preceding Section). For each of the lexicalized reordering models, we build three models to ensure the stability of the results. The average BLEU score of the different SMT systems indicates the best performing lexicalized reordering model for the English→German BL and RO SMT systems.

Baseline The evaluation scores gained for the BL SMT models are summarized in Table 5.4. Although different training runs expose slightly different translation quality on different test sets, in general, the models seem to provide equally good translations when applied on English→German translation task. The hierarchical lexicalized reordering performs best, however the difference in terms of the BLEU scores between the different systems is not significant.

Preordering Reordering rules developed for English→German SMT are designed to primarily deal with long-range reorderings. Nevertheless, there are also local reorderings which need to be considered when applying SMT on the English→German translation direction. In order to allow for such reorderings, we perform non-monotonic decoding

Test set	Lexicalized reordering model								
	wbe			phrase			hier		
news2014	21.87	22.11	22.04	22.04	21.91	22.09	21.77	22.21	**22.25**
news2015	20.84	20.68	20.67	**20.98**	20.68	20.54	20.68	20.86	20.91
news2016	**25.77**	25.58	25.41	25.71	25.65	25.55	25.48	25.75	25.62
avg	24.21	24.12	24.06	23.87	24.26	24.24	**24.28**	24.11	24.13
		24.13			**24.12**			**24.17**	

Table 5.4.: Evaluation of the BL English→German SMT models using different lexicalized reordering models.

Test set	Lexicalized reordering model								
	wbe			phrase			hier		
news2014	22.73	22.64	22.45	22.04	22.59	22.49	**22.75**	22.46	22.71
news2015	21.11	21.41	21.13	20.94	21.21	**21.28**	21.18	21.06	21.22
news2016	26.35	26.22	26.30	26.06	**26.12**	26.28	26.14	26.19	26.11
avg	24.80	**24.83**	24.72	24.40	24.77	24.68	24.77	24.64	24.75
		24.78			**24.61**			**24.72**	

Table 5.5.: Evaluation of the RO English→German SMT models using different lexicalized reordering models.

of the reordered English sentences. Results of the combination of preordering with different lexicalized reordering models are given in Table 5.5. The phrase-based lexicalized reordering (phrase) seems to be the weakest one, while the word-based (wbe) and hierarchical (hier) models lead to the translations of almost the same quality.

Since the word-based model leads to the best results, it is used for all upcoming experiments.

5.3.2. Word alignment

We explore the interaction of preordering with Giza++ and FastAlign aiming at finding the best performing word alignment method for the English→German BL, as well as for the English→German RO SMT models. Furthermore, Ding et al. (2015) discovered that FastAlign can achieve results similar to Giza++ for the translation directions Japanese→English and German→English when applied on preordered data. Although they did not observe any improvement for English→German, we combine preordering with FastAlign to see whether SMT models build based on FastAlign may profit from our implementation of reordering for the English to German translation direction.

Test set	FastAlign BL	FastAlign RO	FastAlign gain	Giza++ BL	Giza++ RO	Giza++ gain
news2014	21.34	22.26	+0.92	21.87	22.64	+0.87
news2015	20.29	20.75	+0.46	20.84	21.41	+0.73
news2016	25.19	25.62	+0.43	25.77	26.22	+0.74

Table 5.6.: Performance of the models trained on data aligned with different word alignment methods. For all models, the word-based lexicalized reordering model is used.

The evaluation results of our experiments run on the WMT data set are shown in Table 5.6. Overall, all SMT models trained using word alignment computed with FastAlign have lower BLEU scores than the models trained with Giza++. On average, the BLEU drop for both BL, as well as RO models is about 0.4 BLEU points. Thus, our first conclusion is that preordering does not boost the performance of FastAlign to the level of Giza++ when applied on the English→German translation direction.

The comparison of the scores gained for RO models with those gained for the BL models shows that reordering improves SMT for both word alignment methods. Interestingly, preordering leads to a smaller increase of the BLEU scores when combined with FastAlign. While for Giza++, the average increase is 0.78 BLEU points, the average improvement for FastAlign is 0.60 BLEU points. Lower BLEU increase for FastAlign indicates that FastAlign does not profit from the increased syntactic similarity between English and German as much as Giza++. As a consequence, the improvement of the models trained relying on the FastAlign alignments are lower than those gained for the models which use Giza++ to perform automatic alignment of the training data.

5.3.3. Sigtest filtering

We combine both word alignment methods mentioned in the preceding section with sigtest filtering. The evaluation results are summarized in Table 5.7. Overall, the sigtest filtering lowers the quality of the German translations for both tested word alignment methods. A closer look to the results indicates that the improvements of the RO models over the BL systems are lower when sigtest filtering is applied compared to those without sigtest filtering. This indicates that the RO phrase/reordering tables contain low-probability phrase pairs which are filtered out by the sigtest filtering, which however, have a considerable impact on the translation performance of the RO systems.

	FastAlign				Giza++			
	no sigtest		sigtest		no sigtest		sigtest	
Test set	BL	RO	BL	RO	BL	RO	BL	RO
news2014	21.34	22.26	21.38	21.92	21.87	22.64	21.56	21.93
news2015	20.29	20.75	20.11	20.30	20.84	21.41	20.11	20.78
news2016	25.19	25.62	24.72	25.03	25.77	26.22	24.59	25.24

Table 5.7.: Performance of SMT models trained on data aligned with different word alignment tools in combination with sigtest filtering. For all models, the word-based lexicalized reordering model is used.

5.3.4. Summary

Experiments shown in the preceding sections show that different experimental setups lead to SMT models of different quality. In the following, we summarize the most interesting findings. Note that the following summary does not include the comparison between BL and RO: this is discussed in detail in the subsequent Section.

In terms of the different lexicalized reordering models, we discovered that hierarchical reordering leads to the best German outputs for the baseline English→German SMT models. In contrast, when SMT models are trained on the reordered English part of the corpus, the best lexicalized reordering model is the word-based model.

Word alignment is one of the crucial steps in building good SMT models. Our experiments showed that Giza++ is the better choice for building both the baseline, as well as the reordered English→German SMT models compared to the FastAlign which is, on the other hand, much faster compared to Giza++. Additionally, we observed that the improvement potential of reordering is lowered in combination with FastAlign compared with the combination with Giza++.

Experiments with sigtest filtering suggest that for English→German, sigtest leads to the lower SMT quality for both baseline, as well as reordered variants. Furthermore, sigtest leads to the lower improvement of the reordered SMT models compared to the non-filtered experimental setup which indicates that sigtest filtering removes table entries which are important for the reordered models to produce better translations compared with the baseline models.

In the following Section, we discuss another set of BL and RO SMT models. According to the above mentioned findings, all SMT models make use of Giza++ to compute word alignment. The models include word-based lexicalized reordering models. Sigtest filtering of the phrase and reordering tables is not applied.

5.4. Automatic and manual evaluation of preordering

In this Section, we present contrastive evaluation of the proposed preordering method for English→German SMT. The approach is tested on three different data sets: (i) full set of the WMT data, (ii) Europarl and (iii) medical data. Furthermore, we show results of the automatic evaluation of the different SMT models in terms of BLEU. We also present manual, qualitative evaluation of the generated translations to shed light into the differences between BL and RO SMT systems caused by preordering.

5.4.1. WMT data

This section presents evaluation of the preordering approach applied on the WMT data. We present two different sets of the experiments: (i) Europarl and (ii) full WMT set. The Europarl experiment includes smaller amount of training data, i.e., solely the Europarl data, while the full WMT experiment shows the results for SMT models trained on the full WMT data set (cf. Table 5.1 on page 99 for details about the used corpora). The aim of evaluating preordering on the data of the different size is to explore the effectiveness of preordering with respect to the size of the training data. Furthermore, we combine the Europarl experiment with LMs trained on different amount of the German monolingual data. Since the quality of the SMT systems considerably increases with the size of the used LMs, we aim at testing whether preordering improves the English→German SMT models even when a big LM is used.

Europarl The Europarl experiments are based on a relatively small amount of the data used to train the English→German SMT models. As indicated by the name, the Europarl SMT models are trained using only the Europarl corpus which consists of about 1.9 mio English-German sentences. The Europarl models are combined with two different language models: (i) Europarl LM which is trained only on the Europarl corpus, and (ii) WMT+news which is trained on the full WMT data set, as well as on a large amount of additional German monolingual texts which include more than 120 million sentences. The evaluation results for the Europarl experiments are given in Table 5.8.

Many studies have shown that the quality of an SMT model increases with the size of the utilized LM. Thus, as expected, the overall results gained for the models including the WMT+news LM are higher than those gained for the models with the smaller LM. For both experiment setups, the RO system outperforms the corresponding BL model which shows that preordering is beneficial regardless the size of the used LM. The average

	Europarl LM			WMT+news LM		
Test set	**BL**	**RO**	**gain**	**BL**	**RO**	**gain**
news2014	11.86	12.21	+0.35	18.10	18.37	+0.27
news2015	14.06	14.67	+0.61	17.63	17.98	+0.35
news2016	17.26	17.89	+0.63	21.59	22.24	+0.65

Table 5.8.: Automatic evaluation of the SMT performance using language models with considerable difference in the size of the data used to train them.

Test set	**BL**$_{hier}$	**RO**$_{wbe}$	**gain**
news2014	21.77	22.64	+0.87
news2015	20.68	21.41	+0.73
news2016	25.48	26.22	+0.74

Table 5.9.: Automatic evaluation of the SMT models trained on full WMT data set. The baseline system includes the hierarchical, while the reordered system includes word-based reordering model.

improvement gained for the model using the WMT+news LM is a little bit lower than for the model using the Europarl LM (0.42 vs. 0.53). This is due to the fact that the bigger LM is capable of capturing more of the target language sequences relevant for the correct placement of the words in the German translations. On the other side, the improvement reached by reordering of the English data shows that big LMs still do not capture many of the word sequences relevant for the long-range reorderings. These cases are successfully handled by applying the preordering approach.

Full WMT data set Table 5.9 shows the evaluation of the systems trained on the full WMT data set. We compare the best performing baseline model with the best performing reordered model.[8] The results show that preordering improves the quality of all considered test sets compared to the translations generated by the baseline SMT model. The improvements are between 0.74 and 0.87 BLEU points.

In addition to the automatic evaluation, we also perform manual quantitative and qualitative evaluation of the SMT outputs by checking the translations of the VCs in a set of randomly chosen sample test sentences. Starting from the evaluation set collected by Popović (2017), we take a closer look to the errors regarding the German verbs: order of the verbs (v_order), omission of the verbs (v_miss) and inflection of the verbs (v_infl). We compare our reordered translations with the translations produced by the

[8]Details about the SMT model combinations are described in Section 5.3.

	correct	v_order	v_miss	v_infl
BL	75	46	31	18
RO	115	18	21	17

Table 5.10.: Comparison of the verb-related errors in the BL and RO German translations. In total, 170 VCs from 154 test sentences taken from the news2016 test set are taken into account.

		MT input	MT output
(1)	BL	now he has registered his idea at the patent office in Munich .	nun hat er seine Idee beim Patentamt in München .
	RO	now he has his idea at the patent office in Munich registered .	nun hat er seine Idee beim Patentamt in München registriert .
(2)	BL	at the moment monsieur Bieber is in Berlin .	im Moment Monsieur Bieber ist in Berlin.
	RO	at the moment is monsieur Bieber in Berlin .	im Moment ist Monsieur Bieber in Berlin.
(3)	BL	now , six in 10 republicans have a favorable view of Donald Trump .	jetzt, sechs in zehn Republikaner haben ein positives Bild von Donald Trump .
	RO	now , have six in 10 republicans a favorable view of Donald Trump .	jetzt haben sechs von zehn Republikaner ein positives Bild von Donald Trump.

Table 5.11.: Example translations from the news domain. We show tokenized, lowercased English inputs and tokenized, truecased German SMT outputs.

baseline SMT model used by Popović (2017) for their evaluation work. The number of the errors found in the two sets of the German translations are given in Table 5.10. The results clearly show that preordering reduces both the verb order errors (by 61%), as well as the verb omission errors (by 32%). In sum, the reduction of these errors leads to 46% more correctly generated German VCs compared with the baseline. In other words, preordering reduces almost a half of the verb-related errors in the German translations.

Table 5.11 shows examples in which preordering corrects errors regarding the position of the verbs. A typical error of omitting the German verbs which need to be placed at the end of a sentence or a clause is shown in the first example. In the BL translation, the translation of the participle *registered* is missing. Preordering leads not only to the generation of the missing verb, but it also places it in the correct position. It thus has a big positive impact on both the adequacy as well as on the fluency of the generated translation compared with the baseline. Even small range reorderings such as the one

Test set	BL	RO	gain
Khresmoi	19.50	20.12	+0.62
Cochrane	50.48	53.30	+2.82
NHS24	24.90	25.12	+0.22

Table 5.12.: Evaluation of reordering on medical data. We show tokenized, lowercased English inputs and tokenized, truecased German SMT outputs.

shown in the second example may benefit from preordering. In the preordered version of the English sentence, the verb *is* is moved in front of the subject NP *'monsieur bieber'* which leads to the correct position of its translation *ist* before the subject *'Monsieur Bieber'* in the RO output. This position is required for sentences with adjuncts at the sentence-beginning position. The baseline fails to generate the German verb in the required position. The same error can also be found in the third example where the verb *haben* needs to be placed after the adverbial *jetzt*, before the subject NP *'six in 10 republicans'*.

5.4.2. Medical data

The evaluation results are shown in Table 5.12. Similarly to the results for the test sentences from the news domain, preordering also improves German translations of the texts from the medical domain. The average improvement for the used test sets is an improvement of 1.22 BLEU points.

Examples of the translation outputs are shown in Table 5.13. In the first example, reordering leads to the generation of the infinitival verbal complex *'zu untersuchen'* which is omitted in the baseline translation. In the second example, the baseline generates the verb *werden* in the sentence-initial position which is wrong. Preordering places *werden* into the correct position, namely after the subject *Herzstillstand*. It also puts the participle *bezeichnet* in the correct, i.e. clause-final, position which is not the case in the BL translation.

5.5. Discussion

In this Section, we take a deeper look into the specifics of the data used for the experiments discussed in the preceding section, as well as into the technical characteristics of preordering.

		MT input	MT output
(1)	BL	the aim of this study was to investigate the effect of in vivo inhibition of factor XI and TAFI in an experimental thrombosis model in rabbits .	Ziel dieser Studie war es die Wirkung von in vivo Hemmung von Faktor XI und TAFI in einer experimentelle Thrombose Modells in Kaninchen.
	RO	the aim of this study was the effect of in vivo inhibition of factor XI and TAFI in an experimental thrombosis model in rabbits to investigate .	Ziel dieser Studie war die Wirkung von in vivo Hemmung von Faktor XI und TAFI in einer experimentellen Thrombose Modells in Kaninchen zu untersuchen .
(2)	BL	cardiac arrests are sometimes referred to as cardiopulmonary arrest , cardiorespiratory arrest , or circulatory arrest .	werden manchmal Herzstillstand bezeichnet als Herz-Atem-Stillstand , Herz- und Atemstillstand , oder Kreislaufstillstand .
	RO	cardiac arrests are sometimes to as cardiopulmonary arrest , cardiorespiratory arrest , or circulatory arrest referred .	Herzstillstand werden manchmal als Herz-Atem-Stillstand , Herz- und Atemstillstand , oder Kreislaufstillstand bezeichnet .

Table 5.13.: Example translations from the medical domain.

In Section 5.5.1, we investigate importance of reordering rules. We relate their importance to the characteristics of the data, i.e., amount of the sentences/clauses and VC types interesting for reordering.

5.5.1. Applied rules

In Section 4.3.1, we present many different reordering rules. The interesting question now is which of them are the most important ones, i.e., which of them are most frequently applied on the used data. The answer to this question also provides some insights about the data, i.e., complexity of the sentences in the data used to train the SMT models.

Table 5.14 shows frequencies of the different VC types which are relevant for the English→German preordering approach (cf. Section 4.3.2.2). The data set denoted by *Europarl+News* consists of 250k randomly selected sentences from the Europarl and News Commentary corpus. The set consists of about 550k clauses which means that each English sentence consist of 2.2 clauses on average. 142,189 or 56% sentences contain at least one subordinate clause – a clause which always needs to be reordered.[9] In other

[9]The only exception of this rule are very short subordinate clauses without any additional constituents

VC type	VC subtype	Europarl+News	Medical
simple	simple	**228,063 (41%)**	**116,224 (46%)**
	simpleaux	5,294 (0.1%)	550 (0.1%)
composed	composed	**131,817 (24%)**	**39,259 (15%)**
	modaux	3,378 (0.1%)	133 (0.01%)
	modauxaux	520 (0.001%)	12 (0.001%)
	auxaux	6,558 (0.1%)	1,161 (0.1%)
non-finite	auxtoinf	7,136 (0.1%)	1601 (0.1%)
	gerund	56,060 (1%)	**49,419 (20%)**
	toinf	**82,660 (15%)**	21,716 (9%)

Table 5.14.: Frequencies of the different VC types extracted from the data from different domains. The Europarl+News set consists of 250k sentence pairs, i.e. 550,596 clauses. The medical set consists of the same set of sentences containing a total of 253,369 clauses. Three most frequent VC subtypes for each of the data sets are marked in bold.

words, more than a half of the English test sentences needs to be transformed.

Reordering depends not only on the clause type but also on the type of a VC within the given clause. The most frequent VC types in the *Europarl+News* data set are *simple* and *composed*. Simple VC type consists of a single verb which needs to be reordered only in the subordinate clauses, while the composed type involves verb movements for all types of clauses. Concretely, for the data set under consideration, 38% of the simple VCs, i.e., clauses with a simple VC, have been reordered. In contrast to this relatively low number, 91% of the clauses with a composed VC undergo reordering. Relatively low percentage of the simple VCs having been reordered might suggest that this type of reordering is not important for the English→German SMT. However, this is a misleading conclusion. Simple VCs need to be reordered in all kinds of subordinate clauses because the English and German subordinate clauses expose considerable differences in the placement of the verbs (SVO vs. SOV). The same also holds for non-finite VC types given at the bottom of the Table 5.14.

The comparison of the VC statistics extracted from the mixture of the Europarl and News data with those extracted from a subset of the medical data indicates an important difference between the two corpora. The total percentage of the simple VCs in Europarl+News is 41%, while 46% of the VCs found in the Medical corpus belong to the simple VC type. Consequently, the portion of the composed VCs in the Europarl+News

such as objects or adjuncts. In this context, the reordered position of the verbs corresponds to the original position.

	BL	RO		
		SR	PCFG	BLLIP
EN→DE	40.10	40.74	41.17	41.49

Table 5.15.: Evaluation of the RO models based on preordering applied on the output of three different English parsers: SR: Stanford shift-reduce parser, PCFG: Stanford PCFG parser, BLLIP: Charniak/Johnson parser.

corpus is bigger than in the Medical corpus. Furthermore, the average number of clauses per sentence in the Medical data set is 1.01. According to these numbers, the medical corpus under consideration consists of simple sentences with simple tense forms which are in most cases not interesting for the English→German preordering rules. This hypotheses is also supported by the fact that in the Europarl+News corpus, 76% of the sentences are reordered, while in the Medical corpus, 59% of the sentences are modified by the reordering rules.

5.5.2. Parsing

One of the most critical characteristics of the proposed preordering approach is that it relies on the automatically derived syntactic trees of the source sentences, in our case of English. Although the accuracy of the English parsers is quite high, they may still contain errors which lead to the incorrect reorderings and thus to the errors in the German translations.

In a case study carried out on the texts from the domain of politics, we explored the effect of applying reordering on the output of different constituency parsers for English. The experiments involved three different constituency parsers to parse the English data: (i) Stanford shift-reduce parser (Zhu et al., 2013), (ii) Stanford PCFG parser (Klein and Manning, 2003) and (iii) Charniak/Johnson parser (Charniak and Johnson, 2005). The Charniak/Johnson parser has been run with and argument *-T10* which speeds up the parsing process, however, it also lowers the parsing accuracy. Note that the parsing experiments presented in this section are not aiming at finding the best parser for English→German SMT based on preordering of the source language data, but primarily to examine if there is a difference of the MT quality when different parsers are used as a starting point for the preordering.

The results of our experiments are shown in Table 5.15. The table shows that the performance of the reordered systems indeed differs when the same set of the reordering rules is applied on the output of the different parsers. The best results are gained for

the BLLIP, i.e., Charniak/Johnson parser, which is also the most accurate parser in our experiment setup even after reducing its performance by lowering the value for the argument T from the default value 210 to 10.[10] On the other hand, the results for the PCFG parser are worse than those obtained with the SR parser which is much faster than the PCFG parser, but also a little bit less accurate than the PCFG parser.

These results leads to the following two conclusions: (i) performance of preordering depends on the used parser, and (ii) the effectiveness of preordering does not necessarily correlate with the accuracy of the used parser. Particularly, the second conclusion is interesting. Given the English-German language pair and reordering rules such es those described in Section 4.3.1, the parser which generates the most accurate parse trees on the clause level is also the most suitable one. Erroneous analysis of smaller sentence constituents does not play a big role for reordering rules for the English→German translation direction. On the other hand, for other language pairs for which smaller sentence constituents are considered in the reordering process, parsers are needed which have better accuracy on analyzing these sentence parts.

5.5.3. Speed

The biggest drawback of the preordering approach based on the parse trees is the processing speed. Prior to reordering, all training and testing data needs to be parsed, in our case with a constituency parser. Parsing is a time-consuming process. To keep the training time overhead as small as possible, two aspects of the training process may be optimized for speed: (i) parser and (ii) word alignment.

Parsers differ not only in their performance but also in their speed. Cer et al. (2010) measured both the parsing time, as well as the parsing performance of the different parsers. For example, in their experimental setup, the Charniak parser with the T value of 210 needed 11:09 minutes to parse the section 22 of the Penn TreeBank. On the other hand, with the T values of 50 and 10, it needed 2:06 and 0:14 minutes, respectively, to parse the same set of sentences. At the same time, the same parser with the T values of 50 and 10 had the unlabeled and labeled attachment F1 score of 79.7% and 75.7%, respectively. Generally spoken, it is possible to combine fast parsers with preordering approach to make the data preparation step faster. However, it might happen that the benefit of preordering becomes smaller in cases where a faster, but less accurate parser is used.

[10]For the parser evaluation, please refer to Cer et al. (2010).

BL		RO								
		SR			PCFG			BLLIP		
BLEU	train	BLEU	train	reor	BLEU	train	reor	BLEU	train	reor
40.10	187	40.74	97	254	41.17	372	579	41.49	1279	1468

Table 5.16.: Parsers: SR: Stanford shift-reduce parser, PCFG: Stanford PCFG parser, BLLIP: Charniak/Johnson parser. The total training time (train) and the time needed to reorder the training data (reor) are given in minutes.

The total SMT training time can also be reduced by taking the much faster FastAlign toolkit to perform word alignment instead of Giza++. Similarly to the choice of a parsing tool, it might thought depend on the context whether the reduced training time leads to the acceptable drop in performance of the SMT (cf. Table 5.6 on page 103). This question is particularly interesting for the commercial use of statistical machine translation where small improvements of the MT output often do not justify the considerably longer training time.

In our case study in which we explored the impact of the parser quality on the quality of the MT output based on the preordered source language data, we also kept track of the time required to preorder the training data, as well as of the overall time needed to train SMT models. Table 5.16 shows results for English→German.[11] The total training time of the English→German SMT models varies between 1.5 hours when the fastest parser is applied and 21 hours when the slowest parser is used. The difference in terms of BLEU between the two models when compared with the baseline is 0.64 and 1.39, respectively. The big difference in BLEU scores (and thus the quality of the reordered SMT models) emphasize the importance of carefully choosing the parsing software given a specific context in which the SMT based on preordering is used for the English→German language pair.

5.5.4. Data characteristics

Reordering rules described in this work are applied on the parse trees of the source language data. Most of the English parsers are trained on tree banks mainly consisting of the news articles. Such parsers may not be appropriate to parse, for instance, medical texts. Given the examples of the sentence pairs in Example (41), page 100, found in our medical data set, it would not be surprising if such sentences were not parsed correctly. Furthermore, the parsing accuracy drops with the length of a sentence. For example,

[11]For results acquired for two additional language pairs, please refer to Ramm et al. (2017b).

the Stanford SR parser used for some of our reordering experiments should not be used for sentences longer than 60 words because parsing of long sentences is slow, as well as inaccurate.

Howlett and Dras (2011) examined different factors which may have a negative impact on the rule-based preordering approach for German→English SMT. One of their conclusions was that the effort should be put into determining for which sentences preordering would lead to the improvement of their translations. Non-reordering of specific source sentences has some advantages, as well as disadvantages. The biggest advantage is that reordering errors are avoided which are quite probable to happen due to errors in the parsing step. At the same time, long sentences are very probable to consist of many subordinate clauses which in the context of preordering for English→German play the most important role. Finally, if specific sentences are not reordered, the training corpus used for the preordered models consists of the source sentences with source-side, as well as target-side specific word order which may have negative impact on the trained SMT models.

The above mentioned problems can be avoided by using a parser trained on the type of the data used for training the SMT models. However, gold parse trees needed to train the parsers are hardly available for all different domains. This leads to the conclusion that for such domains, i.e., text types, preordering probably does not lead to the improvement it potentially could to. Many errors found in the German SMT outputs can be corrected by applying preordering approach, however, it needs to be kept in mind that preordering can also lead to errors in the translations due to erroneous reorderings caused by errors in the pre-processing steps.

5.5.5. Summary

In the context of the parsed-based preordering for SMT, there is a strong relation between the data and the parsing performance. On the other hand, there is a strong correlation between the parsing accuracy and the reordering rules. If there are errors in the parse trees, it is very likely that the reordering will be also erroneous. The probability of performing wrong reorderings of the source data increases with the complexity and the domain specificity of the training data. The result may be in the worst case that preordering even hurts the translation quality.

Quality of the reordered English→German SMT models is not necessarily proportional to the accuracy of the used parser. Reordering rules rely on a very specific knowledge encoded within the parse trees. Parsers which have lower overall accuracy might be

more appropriate to use in combination with preordering since they provide sufficiently accurate analyzes of the sentence parts which are interesting for reordering.

Type of data is not only interesting in terms of parsing. It may also have an impact on the overall improvement of the translation quality due to its characteristics. For instance, if the training sentences are rather short and thus not interesting for the reordering rules which we defined for English→German, preordering may lead to a very small improvement of the German SMT outputs.

Despite the problems which may occur regarding the data, parsing, etc., our experiments clearly show that preordering improves the German SMT outputs for the different domains, different amount of the training data and different size of the LMs, word alignment method, lexicalized reordering models and post-processing of the phrase and reordering tables. The improvements are significant in a sense that they reflect the existence and placement of the words which are crucial for correct understanding of the sentences, namely verbs. Preordering helps SMT both to generate, as well as to place the verbs in the German translations in the correct positions. Particularly the generation of the verbs which are missing in the baseline systems is important since it has a big, positive, impact on the adequacy of the translations. Additionally, the correct placement of the verbs increases the fluency of the German SMT outputs.

5.6. Chapter summary

This Chapter presented detailed evaluation of the preordering approach introduced in Chapter 4. There are many parameters and factors which directly affect the quality of the SMT models. To these belong type of the lexicalized reordering models, method used to automatically align training data, type of the training data, etc. In Section 5.1, we first gave an overview of the SMT experiments run to find the best parameter settings for the English→German SMT, as well as to evaluate reordering in different experimental setups. In Section 5.2, we then presented the general SMT settings such as the used SMT tool *Moses*, various training parameters as well as the data used to train our SMT models.

In the first set of experiments described in Section 5.3, we discovered the following:

- In terms of the lexicalized reordering model, the baseline SMT models for English→German perform best when hierarchical reordering model is used. As for the reordered models, the best results are gained in a combination with the word-based reordering model. Details are given in Section 5.3.1.

- We explored two methods to perform word alignment of the training data: Giza++ and FastAlign. As expected, both baseline, as well as the reordered models perform better when Giza++ is used. Furthermore, preordering leads to higher improvement compared to the baseline when both systems are trained on the Giza++ word alignment. Lower improvement in combination with FastAlign indicates that FastAlign does not profit much from preordering for English→German as proven by Ding et al. (2015) for Japanese→English and German→English. The experiments are described in detail in Section 5.3.2.

- It is possible to reduce the size of the phrase and reordering tables by performing the so-called sigtest filtering. We applied sigtest filtering on both baseline, as well as reordered English→German SMT models and discovered that sigtest filtering lowers the quality of the respective SMT systems. Furthermore, the experiments showed that sigtest has a negative impact on the performance boost caused by preordering. Improvement of the sigtest-filtered reordered SMT system compared to the non-filtered baseline is smaller than the improvement gained when both systems are used without sigtest filtering. This suggests that sigtest filtering removes table entries which are needed to gain better results in the context of preordering. The experiments with sigtest are described in Section 5.3.3.

According to the above mentioned conclusions, we ran another set of experiments and evaluated them in terms of a direct comparison of the reordered SMT models with the corresponding baseline models. We carried out experiments with two different data sets: WMT and Medical. Furthermore, we divided WMT experiments into two subexperiments which rely on training data of the different size and on language models trained on different amount of the German monolingual texts. By doing this, we explored performance of preordering in different contexts: (i) with respect to the domain, (ii) regarding the size of the training data and (iii) in terms of the size of the used language model. Details about the respective experiments are given in Section 5.4.

The main finding of our experiments is that preordering helps to improve the baseline SMT models regardless the amount of bilingual training data used to trained translation models, as well as amount of target language data used to train the language model. Preodering also helps to improve SMT across domains. We gained improvements on a small corpus composed of political discussions (0.42-0.53 BLEU points), on a big corpus consisting of texts from a number of different domains (0.74-0.87 BLEU points, cf. Section 5.4.1) and on a corpus consisting of texts from the medical domain (0.22-2.82

BLEU points, cf. Section 5.4.2).

The absolute improvement in terms of BLEU varies between the data sets: we relate this to the characteristics of the implemented reordering approach as well as of the texts (i.e., test sets) used in our experiments. Those factors were discussed in Section 5.5. In Section 5.5.1, we first carried out an analysis of the reordering rules in terms of their importance. We related their usefulness with the number of contexts in which they are performed. Different types of texts may have different number of contexts, i.e., clause and VC types interesting for reordering. Hence, the benefit of reordering may vary between text types due to different complexity of the used sentences and tense forms. In our case, the medical data, for instance, consists of 1.01 clauses per sentence which means that there are considerably less subordinate clauses which always need to be reordered compared with the WMT data which on average consists of 2.2 clauses per sentence.

Complexity of the sentences, as well as the domain they come from have a direct impact on the parsing quality. Our parser is trained on tree banks from the news domain and it thus probably has lower parsing accuracy when applied on out-of-domain data such as medical texts which we used in one of our experiments. Parsing errors have a direct impact on the correctness of the reordering which itself has a direct impact on the quality of the reordered SMT models. In Section 5.5.2, we showed that the difference in the outputs generated by SMT models relying on the data reordered using different parser outputs is 0.74 BLEU points. This relatively big difference in performance indicates the importance of carefully choosing the parsing software for the reordering approach.

Constituency parsers are chosen for the English→German preordering approach because they contain all syntactic information needed to perform the required reorderings. However, parsing is slow, so we ran experiments with two additional parsers not only to see whether they lead to SMT models of the different quality, but also to relate their performance to the time needed to preorder the data. The experiments were discussed in detail in Section 5.5.3. As already mentioned in the preceding paragraph, different parsers lead to the SMT outputs of the different quality. In addition to this finding, we also observed that the quality of the final SMT outputs is not proportional to the accuracy of the used parser. A faster parser which may be less accurate than a specific slower parser may be more appropriate for reordering because it provides more accurate analyses of those sentence parts which are interesting for the set of the used reordering rules. A specific combination of parsers as proposed by Eckart and Seeker (2013) related to a corpus study regarding a specific linguistic phenomenon might lead to the best

results of the parsing-based preordering.

In Section 5.5.4, we took another look to the characteristics of the data used in our experiments and discussed them in the context of the parsing-based reordering for SMT. We briefly discussed the hypothesis of Howlett and Dras (2011) who said that more effort should be put into choosing which sentences are to be reordered. For instance, given that the used parser does not perform well on sentences with more than n words, we might want not to preorder sentence with more than n words in order to avoid preordering errors caused by errors in the parse trees. On the other hand, such long sentences are probably good candidates for preordering since they consist of a sequence of subordinate clauses which are of particular interest for preordering for the English→German translation direction.

The Chapter was concluded by a brief summary of the discussion of the above mentioned factors which have a big impact on the potential of the preordering approach for improvement of the English→German SMT (cf. Section 5.5.5). Despite the different aspects of the reordering approach related to the data, as well as the quality of the pre-processing steps, preordering leads to the improvement of the German SMT outputs in all experiments carried out within this work. The improvements are significant not only in the numerical sense in terms of BLEU, but also in a sense that they indicate that the German translations generated by the reordered SMT models more frequently include one of the most important words with respect to understanding of the generated translations, namely the verbs.

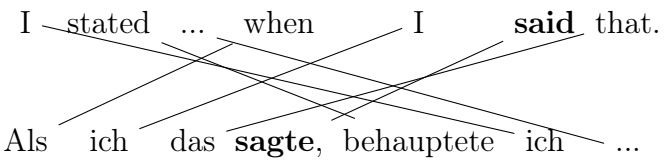
6. Inflection

SMT is known to have difficulties when translating from a morphologically poor language into a morphologically rich language. Morphologically poor languages such as English have considerably less forms of a single word. On the other side, morphologically rich languages such as German include many different forms of a single word. In the context of machine translation, a single source language word may thus be translated into many different target language word forms. For English→German, the discrepancy in the morphological richness also holds for the verbs. Erroneous verb forms in the German MT outputs lead to two problems: (i) erroneous subject-verb agreement, which makes the understanding of the translation hard, and (ii) inappropriate tense and mood, which may lead to false understanding of the generated MT output.

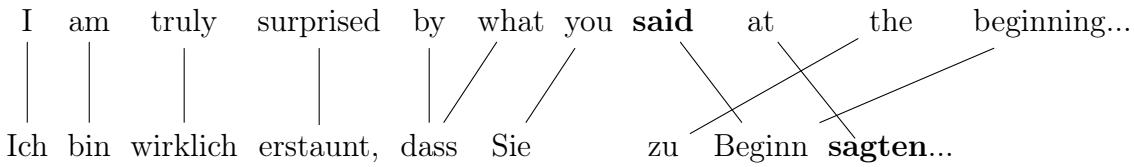
In Section 6.1, we first discuss in detail the morphological differences between English and German along with errors in the German SMT outputs that they cause. SMT problems caused by morphological richness of at least one of the considered languages are often handled by pre- or post-processing of the data in such a way that the inflectional variants are eliminated from the texts. In Section 6.2, we give an overview of the previous research on this topic. In Section 6.3, we then give a detailed description of our method implemented for English→German SMT. Hereby, we present the classification-based verbal feature prediction in Section 6.3.4 and its evaluation in Section 6.3.5 while the parsing-based approach to the correction of the agreement errors is described in Section 6.3.6. The Chapter summary is given in Section 6.4.

6.1. English↔German

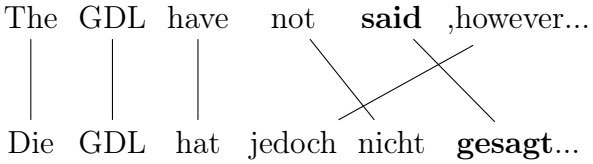
English belongs to the morphologically poor languages which means that it differentiates between only a few forms of a single lemma. On the other hand, German is a morphologically rich language with many different forms of a single word. In the context of machine translation, this means that a single English word form can be translated into many different German word forms. This also holds for the verbs which are handled in



(a) Translation of the English finite verb *said* into the German verb form *sagte*.



(b) Translation of the English finite verb *said* into the German verb form *sagten*.



(c) Translation of the English participle *said* into the German verb form *gesagt*.

Figure 6.1.: Different possibilities of translating the English verb form *said* into German.

this work.

An example for different translation possibilities of a single English verb form is given in Figure 6.1. The English verb *said* corresponds, among other, to the German verb forms *sagte*, *sagten* und *gesagt*. Which of the German verbs is correct, depends on the context, i.e., on the subject of the given verb, as well as on the tense form of the verb. For instance, the difference between the German alternatives in 6.1a and 6.1b includes solely the distinction in the agreement values: while *sagte* is first person singular in the given context, *sagten* is third person plural. Both verb forms are finite, indicative and in the *Präteritum* tense. The verbs in 6.1c, *said* and *gesagt*, differ from the preceding two forms in a sense that they are not finite verb forms, but participles used in composed tense forms in both languages. While the German verb form is unique, the English verb form is the same as in the preceding examples where it is used as a finite verb.

Inflection of the verbs in the German translations thus needs to meet the following requirements: (i) it needs to provide correct agreement with the corresponding subject in terms of person and number and (ii) it needs to reflect tense and mood given in the source language while providing the correct verb form within the generated VC. Both of these aspects are discussed in more detail in the subsequent sections.

6.1.1. Agreement

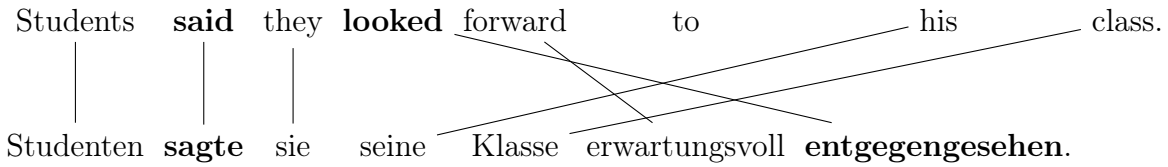
Person and number belong to the agreement features of the German finite verbs. For the present tense, the English verbal inflection solely differentiates between the 3rd person singular (e.g., *(he) says*) and non-3rd person singular (e.g., *(I, you, we, they) say*) forms. The past tense form of the English forms do not even have this distinction, i.e., there is a single past tense form for all person/number values (e.g., *(I, you, he/she/it, we, they) said*). It gets even more difficult when the participle and infinitive is considered: *'(I have) said', '(I need to) say'* where the non-finite verb forms are equal to the finite forms.

In contrast to English, the German verb inflection is considerably more fine-grained. For the present tense, for example, the German verbal inflection differentiates between many different forms of a single verb: *'(Ich) sage', '(du) sagst', '(er) sagt'*, etc. The participles have unique forms, while the infinitives have the same forms such as 1st and 3rd person plural (e.g., $sagen_{Inf}$ vs. *(wir/sie)* $sagen_{3.Pl.Pres.Ind}$).

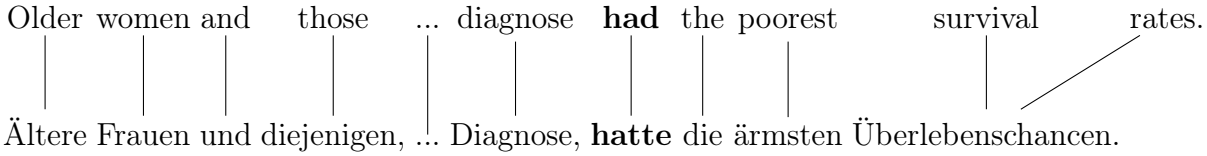
In the context of machine translation, we have a situation in which a single English verb form can be translated into many different German verb forms as already shown in Figure 6.1. Due to the ambiguity of the English forms, SMT often fails to choose the correct form as shown in Figure (6.2). In (6.2a), the English verb *said* is translated into the German verb *sagte*. The verb lemma is a valid translation, however, the verb form *sagte* does not agree with the subject *Studenten (students)*. A similar example is shown in (6.2b) where the German auxiliary *hatte* in singular does not match with the coordinated subject phrase in plural.

The contextual information is crucial for choosing the correct German verb regarding the agreement features. Referring back again to example 6.2a, the occurrence of the subject *Students* before the highly underspecified verb form *said* may help SMT to choose the correct verb form in German. However, this kind of local information does not always lead to the correct output as shown in the respective example. In more complex sentences as shown in 6.2b, the subject of the English verb *had* is very far away and not accessible to SMT as a disambiguation context. It is very probable that the SMT model has chosen a phrase pair including the German sequence *'Diagnose, hatte'* in which *hatte (had)* matches *Diagnose (diagnosis)* in terms of person and number which is in this context incorrect.

The preceding examples shows that SMT makes mistakes regardless of the existence of the appropriate lexical clues in the surrounding context of the source language verbs. In cases, where the appropriate contextual information is not available to the SMT



(a) Erroneous translation of the English finite verb *said* into the German verb form *sagte* in terms of person and number features. The subject *Studenten (students)* is 3rd person plural, while the German verb *sagte* is 1st/3rd person singular. Thus, the subject-verb agreement in the given translation is violated leading to an ungrammatical German SMT output.



(b) Erroneous translation of the English finite verb *had* into the German verb form *hatte*. The coordinated subject phrase *'Ältere Frauen und diejenigen (older women and those)'* requires the plural form of the German auxiliary, i.e., *hatten*. Different person and number features between the German finite verb and the corresponding subject lead to a grammatically incorrect German translation.

Figure 6.2.: Examples of the German SMT outputs with violated subject-verb agreement.

model, the situation is even more severe. By applying the preordering method described in Chapter 4, we actually increase the problem of the non-accessible disambiguation context. An example is shown in Figure 6.3. In the original English sentence, the subject *he* is placed right next to the finite verb *crossed* and can thus be used as a disambiguation context in the decoding step. In the reordered version of the same sentence where the verb is now placed at the end of the sentence, the information about the subject which guides the generation of the correct German form is not accessible in our reordered SMT model due to the distant placement of the two words. Thus, although reordering successfully deals with the placement problems of the German verbs, in certain clause types, it may lead to the loss of the contextual information which is needed to ensure the generation of the correct verb forms in the German SMT outputs.

To estimate in how many sentences the English subject is placed far away from the corresponding finite verb, we extracted subjects and verbs from the English corpora and derived their distances. The statistics are shown in Table 6.1. Although the average distance in words is rather small, there is a fair amount of subject–verb pairs with distance larger than 5 words (in Europarl 22%, in News 25%) which are problematic for SMT. As mentioned above, reordering even increases the problem. Particularly

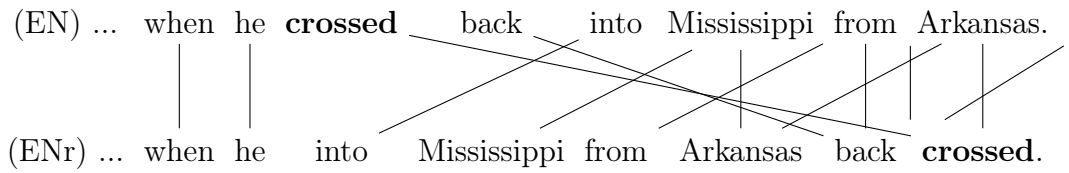


Figure 6.3.: Example of a non-reordered English sentence (EN) and its reordered version (ENr). In (ENr), the distance of the subject pronoun *he* and the reordered English finite verb *crossed* is problematic for our SMT model which takes into account phrases of the maximal length of 5 words as is the case for our SMT models.

Corpus	avg dist in words	>5 words
News	3.9	24%
Europarl	3.7	22%
Crawled	2.9	15%

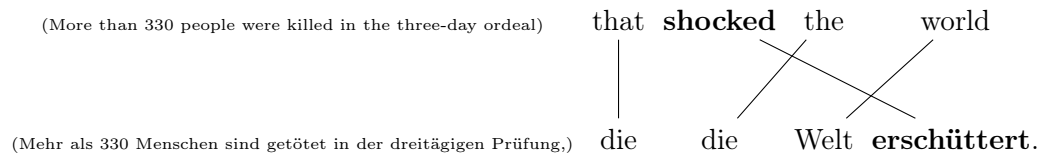
Table 6.1.: Statistics about the distance of the subjects and the corresponding finite verbs derived from the English corpora.

problematic are movements of the verbs in subordinate clauses where the entire German VP is placed at the clause end, while the subject is normally placed in the 2nd position (after the complementizer). In our training data, 20% of the clauses are reordered in a way that the distance between the reordered finite verb and the subject is more than 5 words. In addition to a big distance between the verbs and their subjects, we have also seen that SMT makes errors even if the corresponding words are placed next to each other as shown in the example translation in 6.2a which justifies the explicit handling of the agreement inflection of the verbs in the German translations.

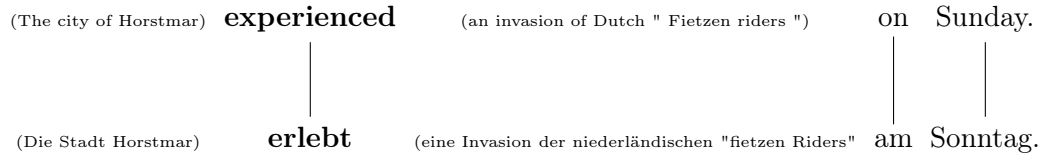
6.1.2. Tense and Mood

Choosing the correct form of the German verbs also implies the generation of the correct tense and mood. Similarly to the agreement errors, in some cases the tense/mood errors may lead to grammatically incorrect sentences. In addition, if the tense/mood features of the German finite verb are wrong, the adequacy of the translation is reduced since the information given in the source sentence is not correctly reflected in the translation. Examples of erroneous translations are given in Figure 6.4.

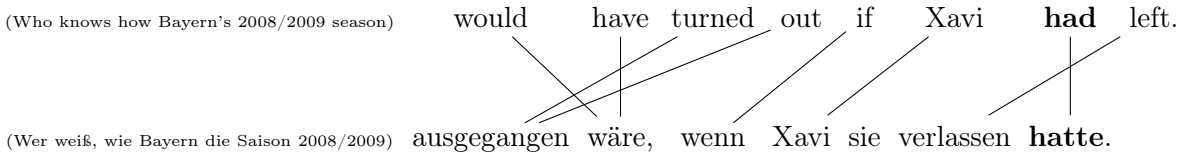
Consider the example translation in 6.4a where *shocked* is translated into *erschüttert*. The generated German verb may be finite verb in present tense or a participle. In the



(a) The English verb in the past tense *shocked* is translated into the German verb *erschüttert* which is either a finite verb in the present tense or a participle. Assumed that it is the finite verb, its tense does not correspond to the tense of the source verb. Assumed it is a participle, it has to occur with a finite auxiliary which is missing in the given translation.



(b) The English verb *experienced* in the simple past tense is translated into the German verb *erlebt* in the present tense. Since the English source suggests that the reported event happened on a Sunday in the past, the German translation may lead to misunderstanding of the proposition given in the source sentence.



(c) The English verb *had* is used in the conditional context and as such it requires the subjunctive translation into German (i.e.g, *hätte*). *hatte*, on the other hand, is indicative and thus does not transfer the information encoded in the source sentence into the generated German translation.

Figure 6.4.: Examples of the German translations with wrong choice of tense.

first case, it represents a well-formed VC, however, the information about the time point of the reported event does not correspond to that given in the source sentence which may lead to misleading understanding of the translation. In the case that the verb is a participle, if forms a non-grammatical VC in which the appropriate auxiliary is missing. A similar case is also given in example 6.4b where *shocked* is translated into the German verb *erlebt* in the present tense. In combination with the noun *Sonntag (Sunday)*, the reported event may be interpreted as a future action instead of something that already happened which is the interpretation of the English source sentence. Example 6.4c shows the wrong choice of mood. In the conditional context, the English indicative verb *had* needs to be translated in the German subjunctive auxiliary *hätte*. Although one is able to understand the translation, mainly due to the correct translation of the

English conditional VC in the main clause, the wrong mood of the auxiliary *hatte* does have a negative impact on understanding the translation. Furthermore, it leads to a grammatically incorrect German sentence.

Correct tense and mood translation involves two aspects. On the one hand, the syntactic structure of a specific tense/mood combination in terms of a combination of verbs needs to be generated. On the other, the verbs themselves need to be appropriately inflected. In this framework, we focus on a part of the second aspect, namely inflection of the German *finite verbs*. In other words, we let the SMT model generate the German VCs by considering the source VC, as well as contextual information encoded within the used translation and language model. By doing this, we let SMT provide the syntactic frame for tense/mood in a given German translation. We then identify the finite verbs in the German VCs and determine their tense and mood values by taking into account various contextual information. It is important to note that we are not aiming at generating verbs needed for a specific tense/mood combination, but rather at correcting the inflection of the finite verbs for tense/mood expected in the given bilingual context.

6.2. Related work

Morphologically rich languages are generally problematic for SMT. When translating from or into a morphologically rich language, SMT often has difficulties to choose the inflectional variant of a single lemma which is most appropriate in the given context. Since the choice is often incorrect, the SMT translations have many grammatical errors such as agreement errors between the words of a single phrase, typically noun or prepositional phrase. These errors may be avoided or at least reduced by incorporating linguistic knowledge into SMT.

Many researchers have recognized the potential of explicit handling of the inflectional errors in SMT and have proposed many different approaches which describe integration of different kinds of the linguistic knowledge into SMT. Some of the methods include knowledge integration into a SMT model, while the others are implemented as a pre- or post-processing steps. While the pre-processing steps typically change the training/translation data prior to training an SMT model or the decoding step, the post-processing steps aim at correcting the SMT outputs after the decoding step.

This Chapter is dedicated to the modeling of the verbal inflection for English→ German SMT. The German verbal inflection includes the information about the person and number, as well as tense and mood. Thus, in the following, we summarize previously

published work which explicitly concentrates on handling agreement, as well as tense and mood in the context of SMT.

6.2.1. Agreement

SMT provides a powerful device to integrate different kinds of the linguistic knowledge into the training and translation process. The so-called *factored* SMT models allow for adding an arbitrary word-based information to the training and translation data (Koehn and Hoang, 2007). An example of such knowledge might be POS tags, lemmas, etc. Avramidis and Koehn (2008) made use of the factored SMT models to cope for inflectional problems for English→Greek and English→Czech SMT. They focused on handling the case agreement, as well as the verb person errors which are, similarly to German, the most prominent errors in the Greek translations. They proposed to annotate the relevant linguistic information as a *factor* to the corresponding source words. Given, for instance, the English sentence *'I read a book.'*, the verb *read* was enriched with the information about the subject which was derived from the English parse tree. Thus, instead of using the English verb *read* in the training and the translation process, they made use of the enriched verb form *read*|*I*, where *I* indicated that *read* in the given context is in the first person singular. In case of Greek, the method reduced the verb agreement errors from 19% to 4.7%. The proposed method lead to the improved SMT outputs for both Greek, as well as Czech as a target language.

Minkov et al. (2007) proposed to cope with morphology-related problems when translating from a morphologically poor into a morphologically rich language in a post-processing step. They presented a method which involved transformation of the fully inflected SMT outputs (Arabic and Russian) into a sequence of stems. Subsequently, for each of the stems the inflected form was predicted using morphological and syntactic information both in the source sentence, as well as in the given translation. In their follow-up work, Toutanova et al. (2008) investigated different ways of combining the morphology integration method with SMT. They discovered that the most promising way was to train the SMT models on the *stemmed* target language part of the parallel data and to subsequently predict inflected words according to the different kinds of information available in the source and target language sentences.

Gispert and Mariño (2008) proposed a method for handling inflectional problems for English→Spanish n-gram based SMT. Their approach involved the translation of English into a *simplified* form of Spanish composed of the stems of the Spanish words instead of the fully inflected Spanish word forms. In other words, the SMT model was trained on

the original English data and stemmed Spanish side of the parallel corpus. The stemmed Spanish SMT output was then inflected in a post-processing step which included the prediction of the morphological features for each of the Spanish stems and a subsequent prediction of the inflected forms given the stem-morphology combination. Modeling of the verbal inflection included the same morphological features which are also required for German: person, number, tense and mood. Gispert and Mariño (2008) experimented with prediction of all four morphological features, as well as only of the agreement features. While the prediction of the agreement features lead to the improvement of the Spanish SMT output, the prediction of tense and mood did not improve the Spanish translations.

Formiga et al. (2012) extended the approach proposed by Gispert and Mariño (2008) to handle translation of the out-of-domain data for English→Spanish SMT. Similarly to Gispert and Mariño (2008), they translated English into simplified Spanish which consisted of the Spanish stems enriched with specific morphological information. The verbs, for instance, were stemmed, but they also carried information about finiteness, tense and mood. Person, number and gender features were then predicted in a post-processing step using a set of pretrained SVM (support vector machine) classifiers. The accuracy of the agreement predictions was 71% on the clean test data. The method lead to the improvement of the Spanish translations in different setups although the amount of improvements varied a lot depending on the data, i.e., domain that the testing data belongs to. Furthermore, the authors pointed to the fact that in terms of automatic evaluation (i.e., BLEU) the agreement between the verbs and the subject in the reference is evaluated and not between the verb and the actually generated subject phrase which may lead to the underestimation of the gained improvement.

Bojar and Kos (2010) proposed a two-step translation for English→Czech SMT. In a first step, the fully inflected English was translated into *simplified* Czech. Subsequently, the simplified Czech SMT output was translated into the fully inflected Czech. Both of the translation steps were performed using a standard phrase-based SMT model. The main difference lied in the data used to train the SMT models. The first model was trained on the Czech corpus which was lemmatized and enriched only with morphological information which also exists in English. The second model was trained on the simplified Czech as source and fully inflected Czech as a target language. Mareček et al. (2011) and Rosa et al. (2012) extended the method proposed by Bojar and Kos (2010) by combining it with a rule-based correction of the Czech SMT output. Their correction system called *Depfix* first parsed the Czech SMT output and then performed a search for and the

correction of specific errors to which also subject-verb agreement errors belong. The subsequent correction considerably improved the Czech translations.

Fraser et al. (2012) adapted the post-processing method to the morphology generation for SMT to the English→German translation direction. Similarly to the above mentioned approaches, their method also relied on the translation into the *stemmed* German sentences. In the next step, they predicted morphological features for the nominal stems (verbs were not handled in this work) using a set of pre-trained CRF (conditional random fields) classifiers. Finally, having the information about the lemmas, as well as morphological information, they ran a morphology generation tool for German to generate the inflected German words. In contrast to the related work, this approach also allowed the generation of the German words which have not been seen in the training data. The method lead to the improved German SMT translations.[1].

6.2.2. Tense and mood

Ye et al. (2006) presented an empirical study on the identification of the contextual features for the automatic prediction of the English tense given the Chinese source sentence. The classification task was formulated as a multi-class labeling problem: the label set consisted of the three tense labels, namely *present, future* and *past*. They used different kinds of information to train their CRF-based classifier: surface features such as adverbials, the phrase the source verb was embedded in, presence of quotation marks, etc., and *latent* features such as telicity, punctuality and temporal ordering of the events in a given sentence. Their tense classifier reached the prediction accuracy between 83% and 84%. The most important outcome of their work was that the latent features lead to the most accurate predictions. These features are however often not as easy to obtain as the surface features which can be gained from the data by text processing tools available for many different languages. Note that this work did not combine the tense prediction with machine translation, but is worth mentioning because of the findings regarding the features which are presumably needed for automatic prediction of tense in the bilingual context.

The already mentioned work for English→Spanish SMT described in Gispert and Mariño (2008) also included the prediction of tense and mood values for the verbs in the Spanish SMT outputs. Their classifiers were trained on features which Ye et al. (2006) refer to as *surface* features: translation phrase pair, presence of a full verb in the

[1]Further details about this work are given in Section 6.3.2.

source/target phrase, POS of the verbs, active/passive voice of the verbs, presence of an auxiliary, etc. They trained eight binary classifiers. Each of them predicted one of the Spanish tense/mood combinations, e.g., present-indicative, past-perfect-indicative, etc. The tense/mood classification reached accuracy of 82%. In order to increase the prediction accuracy, they also trained separate classifiers for tense and mood. Although the two dedicated classifiers had higher prediction accuracy, the combination of the features lead to the even lower accuracy of only 80% compared to the classifier which predicted tense and mood jointly. The prediction accuracy was unfortunately not sufficient in order to improve the Spanish SMT outputs.

An interesting work on prediction of tense (and aspect) was presented by Tajiri et al. (2012). They experimented with automatic prediction of tense and mood for English in the context of the automatic correction of the English texts written by the English learners. The classification task was defined as a sequence labeling problem. They trained a multi-class CRF-based tense/aspect classifier with a set of 12 labels representing combinations of three tenses, namely, *present, past* and *future* and 4 aspects: *perfect, progressive, perfect-progressive, nothing*. Their feature set includes local, as well as global contextual information: given tense/aspect, lemma of the verb, auxiliary, subject/object phrase, temporal adverbials, conjunction, etc. Their classifier successfully detected and corrected erroneous use of the present tense, while the detection of the erroneously used past tense was more difficult. Tajiri et al. (2012) presented an interesting comparison regarding the performance of different classification models with respect to the detection and correction of tense and aspect. They trained a SVN (support vector machine), a maximum entropy, as well as a CRF classifier for this task. The experiments showed that the CRF classifier performed best indicating that the tense (and aspect) modeling indeed can be seen as a sequence task where the sequence of the preceding tenses plays a role for the prediction of the tense/aspect for the current sentence.

Gong et al. (2012a) adapted the classification-based tense modeling to the Chinese→ English translation direction. They trained two SVM tense classifiers: one for the source language which reflected the translation of tense from Chinese to English, and the second one which modeled the use of tense in the target language. The models, i.e., their predictions were used as an additional feature in the standard phrase-based SMT model. The classifiers were trained on two kinds of information: (i) lexical information including information about the words within the given sentence, their POS tags as well as temporal adverbials, and (ii) semantic information including tense in the preceding sentence and the type of the given document. In the best set up, the classification reached

accuracy of 83%. The authors reported on significant improvements of the English SMT outputs when the tense models were used to rerank the translation hypotheses.

In the follow-up work, Gong et al. (2012b) proposed to build n-gram tense models and to integrate them into the standard phrase-based as an additional model. They applied their method on the Chinese→English translation direction. The tense models were learned on the tenses automatically extracted from the parsed English data. They train two tense models: one of them modeled the sequence of the tenses within a sentence, while the second one captured the sequence on tenses between the sentences. The label set consisted of four labels: *present, past, future* and *UNK (unknown)*. The classifiers were used within the decoding step: after all translation hypotheses have been generated, the tense models were applied on each of the hypotheses to obtain their *inter-* and *intra-sentential* tense scores. These were then used to rescore the translation hypotheses in order to find the best translation. Indeed, the integration of the tense models lead to the improved English SMT outputs.

Meyer et al. (2013) presented a method for improving the translation of the English past tense in one of the corresponding tenses in French. They proposed to use the so-called *narrativity* feature to help SMT to generate the correct tense in French. The narrativity feature was gained with a maximum entropy classifier trained on the English data manually annotated with the narrativity information. The classification features included information about the English verbs, their POS tags, syntactic category of the respective verb, temporal markers and the temporal ordering of the verbs (i.e., events) within the given sentence. The accuracy of the narrativity predictions in terms of F_1 *score* was 0.71. The narrativity feature was predicted for each of the English verbs in the past tense and added to the them as a factor in the framework of the factored SMT. This additional information lead to the improvement of French translations regarding the choice of tense.

Loáiciga et al. (2014) developed a 9-label maximum entropy classifier which for each English VP predicted the French tense. Similarly to the work proposed by Meyer et al. (2013), these predictions were added to the English verbs as factors in the factored SMT for English→French. The classification was based on the rule-based annotation of tenses in the English-French parallel texts.[2] The features used to train the classifier were mostly lexical: verbs, POS tags, neighboring words, temporal adverbials. In addition, also a few syntactic/semantic features were used such as dependency types, semantic roles and

[2]The tense annotation rules for French are integrated in our tool for the automatic annotation of syntactic tense, mood and voice for English, German and French which is described in Section 6.3.3.

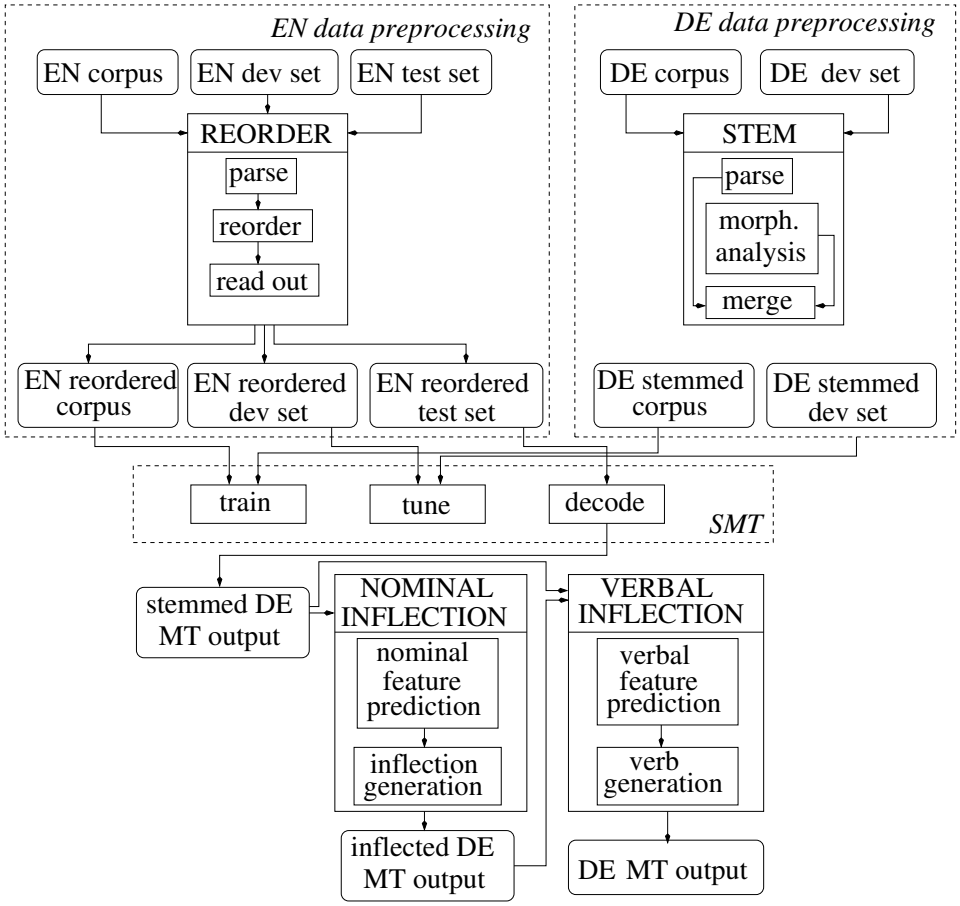


Figure 6.5.: Preordering of the English data is carried out as a part of the pre-processing of the English training data. German is stemmed prior to training which is, similarly to preordering, done as data pre-processing step. After the stemmed German SMT output is generated, it undergoes the nominal, as well as verbal inflection generation step which lead to the final, fully inflected German SMT output.

temporal ordering information. The classification accuracy was about 83% which has been proven sufficient to improve French translations.

6.3. Modeling of the verbal morphology

6.3.1. Architecture overview

Modeling of the verbal morphology for English→German SMT is a part of a combination of different systems which is shown in Figure 6.5. We combine verb inflection generation with the preordering approach presented in Chapter 4. Furthermore, we

combine it with the system for nominal inflection generation for English→German SMT proposed by Fraser et al. (2012) and implemented by Weller et al. (2013). Nominal inflection generation works with *stemmed* representation of the German data which represents German texts without inflectional variants. Details about this system are given in Section 6.3.2.

Verbal inflection is based on an English→German SMT system trained on the reordered English data and the stemmed German texts which produces stemmed German outputs. These are then inflected in a post-processing step. First, the nominal inflection generation is carried out which inflects nouns, adjectives, pronouns and determiners in the German outputs. Subsequently, the inflected German output undergoes the verbal inflection generation step which adapts the inflection of the finite verbs in the German sentences.

The verbal inflection generation step consists of a sequence of processing steps. First, the verbs in the given SMT output are identified by searching for the respective POS tags in the stemmed representation of the given sentence. After identifying finite verbs, we derive their morphological features (person, number, tense and mood) which are then used to generate final forms of the finite verbs. The feature derivation relies on two different approaches. We experiment with a classification-based feature prediction, as well as with a parsing-based derivation of the agreement features. The classification-based approach is presented in detail in Section 6.3.4 and evaluated in Section 6.3.5. For predicting tense and mood, we had to implement a programme which annotates English and German data with the tense/mood information. Details about the automatic tense, mood and voice annotation are given in Section 6.3.3. Due to many errors in the prediction of the agreement features, we implement and test a parsing-based method to gain information about the person and number which is described in Section 6.3.6.

6.3.2. Nominal inflection

Verbal inflection modeling described in this work incorporates nominal inflection modeling for English→German SMT proposed by Fraser et al. (2012). In order to eliminate the inflectional variants in the German data, Fraser et al. (2012) proposed to stem the German data. In other words, the German words are lemmatized and enriched with a specific linguistic information. An example of the German training data (and thus the German SMT output generated in this framework) is shown in the 2nd column in Table 6.2. Take, for instance, the stem of the German adjective *neu (new)*. The stemmed representation bears information about the POS, namely *ADJ* (adjective),

ENr input	SMT output with stem markup	Predicted features	Inflected forms
new	neu<+ADJ><Pos>	Neut.Nom.Pl.St	neue
drugs	Medikament<+NN><Neut><Pl>	Nom.Wk	Medikamente
may	konnte<VMFIN>		**konnte**
lung	Lungen-<+TRUNC>		Lungen-
and	und<KON>		und
ovarian			
cancer	Eierstockkrebs<+NN><Masc><Sg>	Acc.St	Eierstockkrebs
slow	verlangsamen<VVPP>		verlangsamen

Table 6.2.: Example of the nominal feature prediction procedure used in the framework of the verbal inflection correction.

and the comparison value – *Pos* (positive). In order to inflect *neu*, further information is needed which is shown in the 3rd column: gender: *Neut* (neutral), case: *Nom* (nominative), number: *Pl* (plural) and declension type: *St* (strong). Enriched with all the required, POS-specific morphological information, the stem is processed with the German morphology generation tool *SMOR* (Schmid et al., 2004) which generates the inflected form of the given stem according to the provided morphological features. In our case, the SMOR output of the stem *neu<...>* along with the above mentioned morphological features is *neue* which is shown in the 4th column.

The most important step in this pipeline is prediction of the morphological features. For this task, a set of pre-trained CRF-classifiers is used. The classifiers use different kinds of the contextual information and predict for every noun, adjective, pronoun and determiner in the German SMT output the corresponding morphological features. In this framework, there is no explicit handling of the verbs. As indicated in Table 6.2, the verbs are left inflected in the training data which means that they are already inflected in the German stemmed SMT outputs and are thus very likely to encounter the same errors as found in the fully inflected German SMT outputs discussed at the beginning of this chapter. Using the information that is generated from the stemmed representation, our goal is correcting the inflection of the finite verbs in the stemmed German SMT outputs.

6.3.3. Annotation of tense, mood and voice

Derivation of tense and mood for the finite verbs in the German SMT outputs includes modeling of the English and German tenses in the bilingual context. Although syntactic

ID	Word	Lemma	POS	Morphology	HeadID	SyntRel
1	Wir	wir	PPER	nom|pl|*|1	2	SB
2	wollen	wollen	VMFIN	pl|1|pres|ind	0	–
3	nicht	nicht	PTKNEG	_	2	NG
4	,	–	$,	_	3	–
5	dass	dass	KOUS	_	8	CP
6	wir	wir	PPER	nom|pl|*|1	8	SB
7	Ihnen	ihnen	PPER	dat|pl|*|3	8	DA
8	diktieren	diktieren	VVINF	pl|1|pres|ind	2	OC
9	,	–	$,	_	8	–
10	was	was	PWS	acc|sg|neut	13	OA
11	sie	sie	PPER	nom|sg|fem|3	13	SB
12	profitabel	profitabel	ADJD	pos	13	MO
13	macht	machen	VVFIN	sg|3|pres|ind	8	OC
14	.	–	$.	_	13	–

Table 6.3.: Example Mate output which is used to automatically annotate the syntactic tense, mood and voice for English, German and French.

tense, mood and voice belong to the most prominent features of a clause there are no tools which automatically provide this information for arbitrary English and German texts. We thus developed and implemented a tool which is able to annotate syntactic tense, mood and voice for English, German and French. Each of these languages have specific morphosyntactic rules to build syntactic tenses. For example, the German past tense *Perfekt* is built of an auxiliary *haben/sein* in the present tense and of a participle. Our tool makes use of such rules and automatically annotates tense, mood and voice to every VC in the given text.

The annotation relies on the dependency parses of the sentences, as well as on the morphological analysis of the verbs under consideration. We use the output of the Mate parser (Bohnet and Nivre, 2012) which is shown in Table 6.3. The first step includes automatic extraction of the VCs from the dependency trees. This is done by searching for verbal POS tags and by considering specific dependency relations between the verbs. In the example tree shown in Table 6.3, we extract three different VCs consisting of a single verb: *'wollen nicht'*, *diktieren* and *macht*. On each of the extracted VCs, we apply a set of hand-written rules based on the POS tag sequences and the morphological information to derive tense, mood and voice values. For instance, the rule which applies for the VCs from the example sentence is given in Table 6.4. The annotations for the respective German VCs are shown in Table 6.5.

The annotation procedure is same for all three languages. However, the knowledge

Conditions		TMV values		
POS	**Morphology**	**Tense**	**Mood**	**Voice**
V.FIN	pres.ind	present	indicative	active

Table 6.4.: An example of a TMV annotation rule: if a VC consists of a single finite verb (POS=V.FIN) in present tense and indicative mood (morphology=pres.ind), then the syntactic tense is present, mood is *indicative* and voice is *active*.

ID	VC	Finite	FinV	MainV	Tense	Mood	Voice	Neg	Coord
2,3	wollen nicht	yes	wollen	wollen	present	indic	act	yes	no
8	diktieren	yes	diktieren	diktieren	present	indic	act	no	no
13	macht	yes	macht	macht	present	indic	act	no	no

Table 6.5.: Tense, mood and voice annotation output of an example German sentence given in Table 6.3.

used within the annotation rules differs. While for German and French, the POS tags and the morphological analysis is required, the English rules mainly rely on the POS information. In all three languages, however, there are syntactically ambiguous VCs which cannot be correctly annotated without looking at the actual verbs. For example, the English VCs *'will come'* and *'would come'* expose the same POS sequence. However, they have different tense and mood values. Depending on the actual form of the used modal verb, the respective POS sequence is either *future I* in case the modal is *will* or *conditional I* in case the modal is *would*. The respective annotation rules are summarized in Table 6.6.

Other interesting ambiguous constructions are VCs expressing stative passive such as the German VC *'ist gesagt' (is said)*. Syntactically, such VCs are the same as some specific indicative tense forms, e.g., *'ist/is gegangen/went'* in the *Perfekt* tense. The same ambiguity is also given in French. To distinguish between certain active tense forms and the stative passive constructions in German and French, we use semi-automatically collected lists of the relevant verbs. Definition of the corresponding rules for German is given in Table 6.7, while the annotation examples are shown in Table 6.8.

Conditions		TMV values		
POS sequence	**Finite verb**	**Tense**	**Mood**	**Voice**
MD VB	will	future I	indicative	active
MD VB	would	conditional I	subjunctive	active

Table 6.6.: An example TMV annotation rule for English.

Conditions			TMV values		
POS	**Morphology**	***sein*-verb**	**Tense**	**Mood**	**Voice**
V.FIN	pres.ind	no	present	indicative	passive
V.FIN	pres.ind	yes	perfekt	indicative	active

Table 6.7.: TMV annotation rules which distinguish between ambiguous active and passive VCs in German. The condition *sein-verb* checks whether the main verb builds the *Perfekt* form with the auxiliary *sein*.

VC	Finite	FinV	MainV	Tense	Mood	Voice	Neg	Coord
ist gegangen	yes	ist	gegangen	perfekt	indic	act	yes	no
ist geschrieben	yes	ist	geschrieben	present	indic	pass	no	no

Table 6.8.: Tense, mood and voice annotation output of the German VCs *'ist gegangen* and *ist geschrieben*.

Tables 6.9 and 6.10 show the sets of the tense, mood and voice values annotated for English and German.[3] The tool outputs reflect the full syntactic tense, mood and voice sets in the two languages. While the English annotations are used as one of the features in the classification process, the German annotations are used as labels for the classifiers.

6.3.4. Classification-based verb correction

Following the idea of Fraser et al. (2012) for modeling nominal inflection, we develop classifiers which predict morphological features for the finite verbs in the German stemmed SMT outputs. The classifiers are trained with the toolkit *Wapiti* (Lavergne et al., 2010) which allows for training and applying of different kinds of the classification models. We experiment with maximum entropy Markov models, as well as with CRF-models (Lafferty et al., 2001).

The remaining of the section is structured as follows. First, we explain the extraction of the training samples from parallel data in Section 6.3.4.1. Next, we present features we use to train our classifiers in Section 6.3.4.2 and give a data-driven analysis of the prediction labels in 6.3.4.3. Finally, we evaluate the classification performance on clean data in Section 6.3.5.[4]

[3]Here, we show only English and German because these annotations are used in the remaining of the work. For French annotations, please see Ramm et al. (2017a).

[4]The performance on the German SMT outputs is evaluated in Chapter 7.

Finite	Mood	Tense	Voice	Example (active voice)
yes	ind	present	act pass	(I) work
		presProg		(I) am working
		presPerf		(I) have worked
		presPerfProg		(I) have been working
		past		(I) worked
		pastProg		(I) was working
		pastPerf		(I) had worked
		pastPerfProg		(I) have been working
		futureI		(I) will work
		futureIProg		(I) will be working
		futureII		(I) will have worked
		futureIIProg		(I) will have been working
	subj	condI		(I) would work
		condIProg		(I) would be working
		condII		(I) would have worked
		condIIProg		(I) would have been working
no	-	-	-	to work

Table 6.9.: Tense, mood and voice combinations for English.

6.3.4.1. Training samples extraction

Prediction of the verb morphology features includes the derivation of the contextual information not only related to a single word, but rather to the clause or the sentence that a particular verb occurs in. For this reason, we handle the verbs *clause-wise*: for each German clause d_i with a finite German verb dv_i, we first identify the parallel English clause e_i and then extract different contextual information related to d_i and e_i. This is illustrated in Figure 6.6. We start with the German sentence which contains three finite verbs for which the morphological features are to be predicted. Each of these verbs are related to a different English VC: *wollen* is aligned with *'do not want'*, *diktieren* is aligned with *'to dictate'* and *macht* is aligned with *makes*. The prediction of the morphological features of the German verbs depends on the properties of the verbs they are aligned with, as well as specific clausal information their English counterparts are placed in. For this reason, we talk about clause-wise extraction and prediction of the verbal features.

For the extraction of the features needed for the agreement classifier, we use the English reordered parse trees, the German stemmed sentences with full morphological annotation, as well as automatically computed word alignment of the English-German sentence pairs. For the training for which the agreement labels are additionally needed,

Finite	Mood	Tense	Voice	Example (active voice)
yes	ind	present	act	(ich) arbeite
		perfect		(ich) habe gearbeitet
		imperfect		(ich) arbeitete
		pluperfect		(ich) hatte gearbeitet
		futureI	pass	(ich) werde arbeiten
		futureII		(ich) werde gearbeitet haben
	konjI konjII	present		(er) arbeite/arbeitete
		past		(er) habe/hätte gearbeitet
		futureI+II		(er) würde arbeiten / gearbeitet haben
no	-	-	-	zu arbeiten

Table 6.10.: Tense, mood and voice combinations for German.

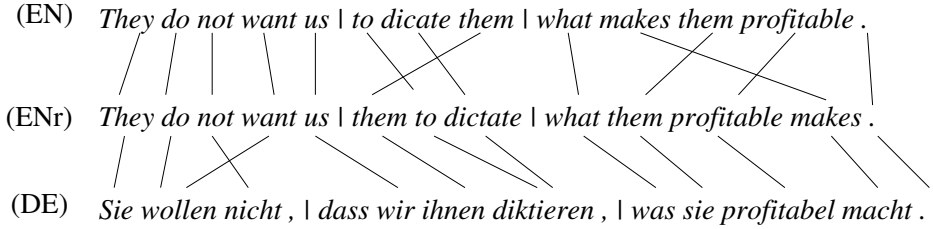


Figure 6.6.: Example of a word-aligned English-German sentence pair containing a sequence of clauses. Clause boundaries are indicated with vertical bars.

the output of the RFTagger is used (Schmid and Laws, 2008).[5] An example sentence pair along with the different data representations used for the feature extraction is given in Figure 6.7.

In the translation context, we use the morphological features derived in the nominal feature prediction step. In order to train the classifier on the data with similar correctness regarding the morphology annotation as in the SMT output for which the predictions are to be made, we first stem the German training data and then run the nominal feature prediction on it to acquire the morphological information for the stems.[6]

As already mentioned, the extraction of the classification training samples is clause-based. The English clauses correspond to the subtrees with S-* root nodes. The recognition of the clauses in the flat representation of the German sentences relies on the

[5]Note that the stemmed representation originates in the work presented by Fraser et al. (2012) in which the verbs are not stemmed. In order to get the morphological information of the verbs, we combine the RFTagger annotation with the stemmed representation of the German sentences.

[6]Another possibility is to derive case from RFTagger, however this might have negative impact on training/testing the classifier since the case annotation of the MT output is much more erroneous than of the well-formed DE.

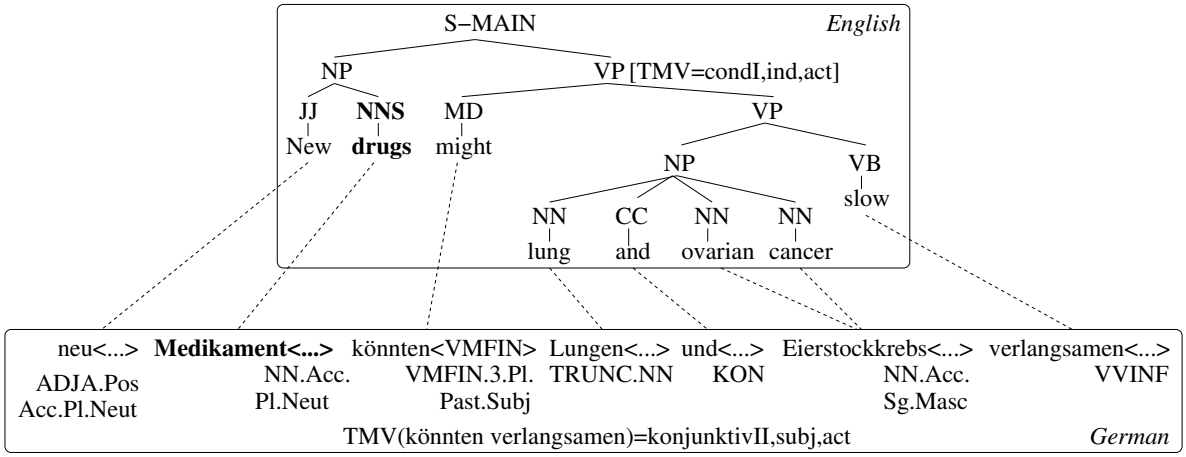


Figure 6.7.: Representation of a parallel English-German sentence pair used to derive features for the classification. The morphological features of the German verbs are attached to the stems. In the illustration above, they are split due to the limited space.

Input German sentence		
Sie möchten nicht, dass$_{KOUS}$ wir ihnen diktieren, was$_{REL}$ sie profitabel macht.		
Clauses		
Sie möchten nicht, *They don't want*	dass wir ihnen diktieren, *us to dictate them*	was sie profitabel macht. *what makes them profitable.*

Table 6.11.: Example of the segmentation of a German sentence into a list of clauses. For the readability reasons, the words are inflected: in the framework of the verbal inflection modeling, the German words are stemmed as shown in Table 6.2.

rule-based search for POS tags indicating relative pronouns (REL), conjunctions and complementizer (KOU*, KON*), a wh-word (PWAV) and a colon. Furthermore, the combination of the comma and a verb is considered to be a clause boundary in German. An example of a German sentence split into clauses is given in Table 6.11.

The identification of parallel clauses is driven by an automatically computed word alignment. For a given English clause, first, all verbs and their alignment links are identified. We then check whether there is a finite verb among the German words that the English verbs are aligned with. If so, the German clause containing that verb is considered to be the parallel clause of the given English clause. If there is no finite verb among the German words that the English verbs are aligned with, we consider the German clause with the most alignment links to be the parallel clause. The finite verb in that clause is considered to be the translation of the given English finite verb. An

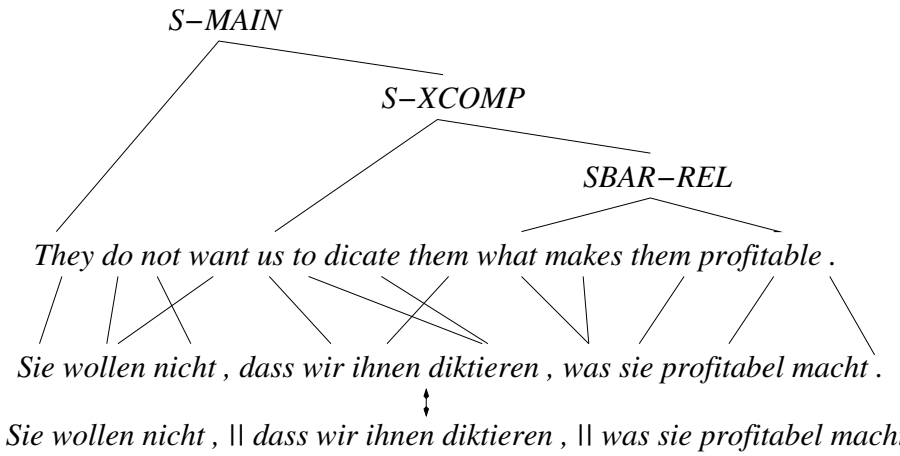


Figure 6.8.: Clause alignment based on the word alignment, the English parse trees and the German clause boundary annotation.

example is given in Figure 6.8. The English finite verb *makes* is not aligned with any finite verb in German, however its alignment link points to the German clause *'was sie profitabel macht'*. Thus, this clause is considered to be the parallel clause to the English clause *'what makes them profitable'*. Consequently, the verb *macht* is assumed to be the translation of the English verb *makes*.

When parallel clauses are identified, for each German finite verb a training sample is derived. If a finite German verb is aligned with more than one English clause (i.e., English finite verbs), multiple training samples are extracted where the English features are derived from the corresponding English clauses, while the German features are the same for all training samples.

6.3.4.2. Features

We train our verbal inflection classifiers on many different kinds of the contextual information derived from the English and German parallel data. In the following, we group the features by the language and explain them in detail.

The German verbal morphology includes person, number, tense and mood features. These features can be grouped into two groups: agreement and tense/mood features. We identify and extract different kinds of contextual information to model the morphological features of the respective group. In the following, we first outline the context information which is used to model the agreement and then present contextual clues used to model tense and mood.

Agreement features The agreement features of a German finite verb depend on the corresponding subject, i.e., on the morphological properties of the subject (pro)noun phrase. In the context of translation, we face the situation in which we need to account for possibly two different subjects playing a role in the prediction of the agreement features. On the one side, there is an English subject which we need to consider, on the other side, there is a German translation of the English subject which may have different morphological features than the source language subject. However, since we want to ensure the grammatical correctness of the German SMT outputs, the German subject needs to get a higher weight compared to the English source subject when predicting the agreement features for the German finite verbs.

One might wonder why do we use information about the English subjects if the agreement is to be established between the German subjects and finite verbs. We include information about the English subjects because in many cases, we are not able to (correctly) identify the subjects in the flat representation of the German sentences. The information about the source subjects are assumed to be more often correctly extracted because the extraction is based on the parse trees rather than a flat sentence representation which is used for German. Thus, the information about the English subjects may in many cases be helpful to overcome problems caused by erroneous identification of the German subjects in the SMT outputs.

Summary of the agreement features is given in Table 6.12. The features are presented in more detail in the following paragraphs.

English features

- Subject noun: lexical feature containing the inflected English subject head (pro)noun. The subject is assumed to be embedded within the first NP node (NP_{subj}) in the corresponding English subtree.

- Subject POS with values NN (common noun), NNS (common noun in plural), NP (named entity), PRP (pronoun): POS tag of the head (pro)noun in the NP_{subj}.

- Subject number with three possible values: *Sg*, *Pl* and *0*. If the POS tag of the most right noun, i.e., head noun, is *NNS*, then the number is *Pl*. Otherwise, it is *Sg*. In case the NP_{subj} does not contain a noun, but a pronoun, we apply a simple rule-based mapping of the English pronouns to the corresponding number values. In case that there is a coordination POS tag within the NP_{subj}, the number value is *Pl*. If no subject is found, the number value is *0*.

	Feature	Value
EN subj	noun	drugs
	POS	NNS
	pers	3
	num	pl
	coord	false
EN other	verb	might
	clause type	MAIN
DE subj	noun	Medikament
	POS	N
	pers	3
	num	pl
DE other	verb	könnten
Label		<3><Pl>

Table 6.12.: List of the contextual features used to train the agreement classifier. The features values are given for the verb *können* extracted from the parallel sentence pair given in Figure 6.7.

- Subject person with four possible values: *1*, *2* and *3* and *0*. In case of a pronominal subject, the pronouns are mapped to the corresponding person values (e.g. *I* → 1, *you* → 2), otherwise the value is *3*. If no subject is found, the number value is *0*.

- Coordinated subject, with two values: *1* and *0*. If the NP_{subj} is a coordination of NPs or nouns, i.e., if there is a word tagged with CC within the NP_{subj}, the coordination value is *1*, otherwise the coordination value is *0*.

- Clause type with values according to the clause type introduced in Section 4.3.2.2 in Chapter 4. Clause types are derived from the English parse tree. Particularly interesting is the distinction between finite and non-finite clauses since the English non-finite clauses no not have a local subject.

- Finite verb: lexical feature containing the English inflected verbs. In our case, the inflected verb is the leftmost (1st) verb in the given English VC.

German features

- Subject noun: lexical feature containing the stemmed German subject head (pro)noun. We apply a rule-based search for the nominative (pro)noun in the given stemmed German clause.

- Subject POS with values NN (common noun), NE (named entity) and PPRO (pronoun): POS tag of the subject head noun.
- Subject person with three values: *1*, *2*, *3*. If a subject pronoun is found, the person is derived from the stem markup. Otherwise, the person is *3*. The default value, e.g., when no subject is found (for instance, in the passive constructions), is *3*.
- Subject number with two values: *Sg*, *Pl*. The number is derived from the morphology annotation attached to the stemmed subject (pro)noun. In case no subject is found, the number value is *Sg*.
- Finite verb: lexical feature containing stemmed German verbs. They are identified using the POS tag information in the stemmed German sentences.

Tense and mood features The following enumeration of the features relevant for the prediction of tense and mood mainly uses the lexical information directly accessible from the text. A summary of the features is given in Table 6.13, while the feature description is given in the following paragraphs.

English features

- Finite verb: lexical feature containing the English inflected verbs. In our case, the inflected verb is the leftmost (1st) verb in the given English VC.
- Finite verb POS with values VBP (verb in simple present tense), VBD (verb in simple past tense), MD (modal). Represents the POS tag of the identified inflected English verb. It is used to generalize over the lexical values of the inflected verbs.
- Modal with values *1* and *0*. *1* indicates the existence of a modal within the given VC, while *0* means that there is no modal verb within the English VC. This information is interesting especially in combination with lexical information about the verbs in the subjunctive context.
- Infinitive with lexical value of an infinitive within the given English VC.
- Infinitive POS with values *VB* in case there is an infinitive and *0* otherwise.
- Participle with lexical value of a participle within the given English VC.
- Participle POS with values *VBN* in case there is a participle and *0* otherwise.

Feature	English	German
finite verb	may	können
finite verb POS	MD	VMFIN
modal	1	–
infinitive	slow	verlangsamen
infinitive POS	VB	VVINF
participle	0	0
participle POS	0	0
VC	may-slow	können-verlangsamen
VC POS	MD-VB	VMFIN–VVINF
main verb	slow	verlangsamen
prev. clause main verb	0	–
adverbial	0	0
complementizer	0	0
word -1	drugs	**Medikament**
word -2	new	**neu**
finite verb align	–	may
clause type	SMAIN	–
tense	pres	–
finite verb ends with -s	0	–
finite verb ends with -ed	0	–
VC quoted	0	–
sentence main verb	slow	–

Table 6.13.: Summary of the features used to predict tense and mood. Cell entries with a line indicate that these features are not defined for the respective language.

- VC: sequence of the verbs of the given English VC.
- VC POS: sequence of the POS tags in the given English VC.
- Main verb: lexical feature containing English main verbs. As a main verb, we define the most right verb in an English VC.
- Previous clause main verb: lexical feature containing the main verb of the preceding clause. The aim is to capture the dependencies between the verbs across clause boundaries.
- Adverbial with lexical values of an adverbial found within the given clause. In particular, temporal adverbials may be beneficial for predicting tense and mood. In case there is no adverbial, the value is *0*.

- Complementizer with a lexical value of a complementizer which introduced the given clause.
- Word -n: lexical feature which includes information about words preceding the English finite verb.
- Clause type with values according to the clause type introduced in Section 4.3.2.2 in Chapter 4. Clause types are derived from the English parse tree. Particularly interesting is the distinction between finite and non-finite clauses since the English non-finite clauses no not have a local subject.
- Finite verb ends with -s: binary feature which indicates whether the English finite verb ends with the suffix *-s*.
- Finite verb ends with -ed: binary feature which indicates whether the English finite verb ends with the suffix *-ed*.
- VC quoted indicates whether the given VC is enclosed within quotations. This information is interesting for distinguishing between direct and indirect speech in the context of indicate vs. subjunctive mood in German.
- Sentence main verb: lexical feature containing English main verbs. The sentence main verb is the main verb of the main clause in the given English sentence. This feature helps to capture the dependency between the verbs regarding their tense and mood.

German features

- Finite verb: lexical feature containing stemmed German verbs. They are identified using the POS tag information in the stemmed German sentences.
- Finite verb POS with values VAFIN (auxiliary), VMFIN (modal), VVFIN (full verb). Represents the POS tag of the identified inflected German verb. It is used to generalize over the lexical values of the inflected verbs.
- Infinitive with lexical value of an infinitive within the given German VC.
- Infinitive POS with values *VAINF, VMINF, VVINF* in case there is an infinitive and *0* otherwise.
- Participle with lexical value of a participle within the given German VC.

- Participle POS with values *VAPP, VMPPm VVPP* in case there is a participle and *0* otherwise.
- VC: sequence of the verbs of the given German VC.
- VC POS: sequence of the POS tags in the given German VC.
- Main verb: lexical feature containing German main verbs. Main verbs are identified by looking for specific POS tags.
- Adverbial with lexical values of an adverbial found within the given clause. In case there is no adverbial, the value is *0*.
- Complementizer with a lexical value of a complementizer which introduced the given clause. Complementizers are identified by searching for a specific POS tag at the clause-initial position.
- Word -n: lexical feature which includes information about words preceding the German finite verb.
- Finite verb align: lexical feature including the information about the English word(s) that the given German finite verb is aligned with.

6.3.4.3. Labels

We distinguish between agreement and tense/mood labels. The agreement features include person and number information, while the tense/mood labels include the set of the German syntactic tenses. The full set of the labels is shown in Table 6.14.

The agreement labels are derived from the output of the RFTagger (Schmid and Laws, 2008). Tense/mood labels are extracted from the output of the tool we developed for automatic annotation of syntactic tense, mood and voice already described in Section 6.3.3.

6.3.5. Classification performance

We experiment with a number of different classifiers. They differ in the label set, as well as in the used classification model as shown in Table 6.15. For the agreement features, we test whether it is more appropriate to predict them jointly with a single classifier (C2) or whether it makes more sense to predict each of the features separately (C3).

Morph. property	Values (labels)	Distribution		
		News	Europarl	Crawl
Person	1	0.027	0.19	0.06
	2	0.0005	0.0005	0.007
	3	0.94	0.77	0.91
Number	sg	0.64	0.65	0.60
	pl	0.33	0.32	0.37
Tense/mood	present	0.54	0.63	0.71
	perfect	0.11	0.14	0.12
	imperfect	0.19	0.06	0.09
	pluperfect	0.03	0.02	0.03
	futureI	0.01	0.03	0.01
	futureII	0.005	0.001	0.01
	konjunktivI	0.01	0.09	0.07
	konjunktivII	0.08	0.07	0.02

Table 6.14.: List of the tense/mood classification labels for the German finite verbs along with their distribution in the corpora used to train the classifiers.

For tense and mood, we experiment with two different classification methods aiming at discovering whether tense/mood prediction is a sequential or a linear problem.

All models are trained with Wapiti (Lavergne et al., 2010). As training data, we use a concatenation of the News, Europarl and Crawl texts. In total, we use 5,120,716 training samples which build in total 3,181,069 training sequences. For testing, we reserve 5,000 sentences from the respective corpora. Additionally, we test the classifiers on the news test set 2014 which is used in the SMT experiments (see Chapter 7).

All testing data has been annotated with our tool for automatic annotation of tense, mood and voice, as well as with the RFTagger (Schmid and Laws, 2008). While our tense, mood, voice annotation tool provides gold labels for the tense/mood predictions, the RFTagger provides gold labels for the agreement features.[7] The evaluation represents a comparison of the predicted labels with gold labels and is given in terms of precision, recall and F1 score.

[7]We are aware of the fact that there might be erroneous gold labels due to annotation errors. However, due to large test sets, we decided to use automatic tools to acquire gold labels rather than to carry out a manual annotation of the testing data.

Setup	Classifier	Labels
2 CRFs (C2)	person/number	1.sg, 1.pl, 2.sg, 2.pl, 3.sg, 3.pl
	tense/mode	present, perfect, imperfect pluperfect, futureI, futureII, konjunktivI, konjunktivII
3 CRFs (C3)	person	1, 2, 3
	number	sg, pl
	tense/mood	...

Table 6.15.: Classifier setups with the respective label sets.

6.3.5.1. Agreement

We give the performance of our classifiers with respect to the agreement predictions gained for the news test set 2014. We evaluate the performance of each of the classification setups shown in Table 6.15. Since the combination of person and number plays a role in the generation of the final verb from, we present evaluation results by looking at each of the person/number combinations. The evaluation results are given in Table 6.16.

Precision, as well as recall of the agreement predictions are between 0.81 and 0.83, depending on the classifier setup. The highest prediction accuracy in terms of F1 is gained with a classifier setup C2 which uses a single classifier to predict agreement feature combinations. This indicates that it makes more sense to predict person and number jointly than to predict each of the features with a separate classifier. There are large differences between prediction accuracy across the agreement features combinations. For instance, the best results are gained for the first person singular (*1.sg*) and the first person plural (*1.pl*). This is not very surprising since these agreement values, i.e., subjects, are most commonly used in our corpora. On the other hand, less frequent agreement combinations pose a big problem for the prediction of the agreement features. If we ignore the combinations including the second person which are very infrequent, it is striking that the prediction of the third person plural is quite problematic while the prediction of the third person singular reaches satisfactory prediction accuracy.

6.3.5.2. Tense and mood

For tense and mood, we train two different classifiers: a maximum-entropy classifier (*me*) and a CRF classification model. We evaluate the two classifiers on 5,000 randomly

Setup	Label	Precision	Recall	F1	Samples
C2	1.pl	0.94	0.86	0.90	111
	1.sg	0.92	0.83	0.87	98
	2.pl	0.00	0.00	0.00	1
	2.sg	1.00	0.50	0.67	2
	3.pl	0.62	0.76	0.68	1167
	3.sg	0.90	0.83	0.86	3216
	avg / total	0.83	0.81	0.82	4595
C3	1.pl	0.93	0.82	0.87	111
	1.sg	0.92	0.84	0.88	98
	2.pl	0.00	0.00	0.00	1
	2.sg	1.00	0.50	0.67	2
	3.pl	0.65	0.75	0.70	1167
	3.sg	0.90	0.86	0.88	3216
	avg / total	0.84	0.83	0.83	4595

Table 6.16.: Evaluation of the agreement feature predictions. Evaluation is carried out on the news test set 2014. The column *Samples* indicates the number of the test samples with the corresponding label.

selected sentences from the News corpus. The evaluation results are shown in Table 6.17. Somewhat unexpected outcome of the evaluation is that there is only a small difference between the two classifiers. The most striking difference in the prediction accuracy is related to the label *konjunktivI*, i.e., to the German *Konjunktiv I* which is primarily used in the reported speech. Here, the CRF model performs considerably better than the maximum entropy model. The reason for this is that *Konjunktiv I* is highly context-dependent. Its usage depends on the verb (usually a reporting verb) placed in the subordinating clause. The CRF model has access to this information, as well as to the tense/mood prediction of the preceding clause, while the maximum entropy classifier performs isolated predictions without considering already made predictions.

Following the discussion outlined in Chapter 3 about the use of tense and mood in German, we also evaluate the tense/mood classifiers on test sets from different domains. The aim is to see whether the domain switch has an impact on the performance of a single classifier trained on the mix-domain data. The evaluation results are shown in Table 6.18. We compare the classifier performance with a baseline classifier which for each English tense predicts the most frequent German tense/mood combination. The evaluation results of the baseline classifiers already indicate that the domain plays a role. The results for the news and crawl test sets are the same which indicates that the distribution (i.e., usage) of the tense/mood in the two domains are quite similar.

tense/mood	F_{1CRF}	F_{1me}
present	0.92	0.92
perfect	0.81	0.81
imperfect	0.85	0.85
pluperfect	**0.74**	0.73
future I	**0.84**	0.83
future II	0.50	0.50
konjunktiv I	**0.27**	0.17
konjunktiv II	0.83	0.83
overall	0.87	0.87

Table 6.17.: Performance of a CRF vs. maximum entropy classifier gained for a test set containing 5,000 sentence from the news corpus.

	F_{1CRF}		
	news	europarl	crawl
mostFreqTense	0.70	0.64	0.70
our model	0.87	0.90	0.88

Table 6.18.: Classifier evaluation using different test sets. Each of the test sets contain 5,000 sentences.

The performance of the baseline classifier on the europarl test set is much lower which indicates that the usage of tense/mood in that test set differs from the one found in news and crawl test sets. Interestingly, our classifier performs best on the europarl test set. It seems to learn the distribution which is more typical for europarl than for news and crawl data. Similarly to the baseline classifier, our classifier has almost the same performance on the news and crawl test sets.

6.3.5.3. Discussion of the agreement prediction

Although agreement features of the finite verbs are clearly constrained by the agreement features of the corresponding subjects, the classification makes many mistakes in predicting person and number. This contradictory result can however be explained by a deeper look into the data we work with. The prediction accuracy relies on the correct identification of a subject *within* the clause that the finite verb to be inflected is placed in. Our feature set includes knowledge about both English and German subjects. As will become more clear in the following discussion, both of these knowledge sources are problematic for the classification approach.

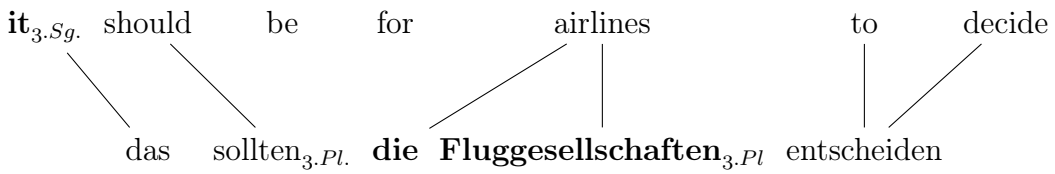


Figure 6.9.: Example of a subject mismatch between English and German. Subjects are given in bold.

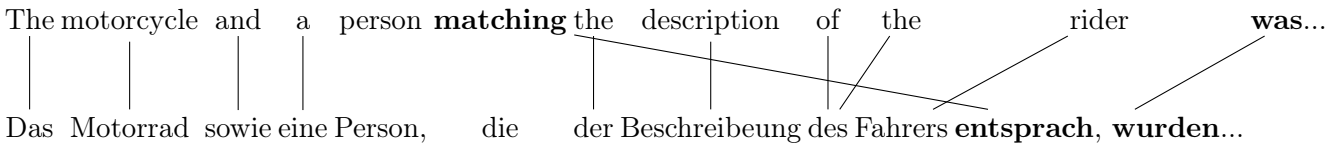


Figure 6.10.: Example for a clause mismatch in English and German. The interesting verbs are indicated in bold.

Use of the English subjects If we assume that the English subjects are translated into the German subjects with the same agreement features, we could simply use agreement features of the English subjects to predict agreement for the German finite verbs. However, a closer look to the parallel data (and the translations) reveals many cases for which this assumption is not sufficient or even wrong. The reason for this are structural differences between the two languages, as well as free translations.

Figure 6.9, for instance, shows an English-German parallel clause which has a subject mismatch. In English, the expletive *it* serves as a syntactic subject for *should*, while the PP *'for the airlines'* is actual subject in the German clause. Thus, the agreement features of the German verb *sollten* do not correspond to the subject of the corresponding English verb *should*.

Another frequent structural mismatch is given when in English non-finite clauses are used as shown in Figure 6.10. The English non-finite VC *'matching'* is translated as a finite German clause consisting of the finite verb *'entsprach'*. The subject of *entsprach* is the relative pronoun *die* which refers to the the noun *Person (person)*. The English word *person* and the gerund *matching* are not situated in the same clause in terms of the parse tree of the given English sentence and it is thus not accessible to the subject extraction procedure developed for the English parse trees. In other words, due to the given syntactic difference between English and German, we are not able to derive the English subject features needed for the agreement feature prediction for the German finite verb *entsprach*. Let us now consider the German verb *wurden* which is translation of *was*. *Wurden* is third person plural and as such, it matches the coordinated subject

EN source	When this country counselled other countries on how to forge civil and democratic societies, Americans **explained** that the right...
DE MT output	Wenn dieses Land andere Länder auf beratschlagen, wie die zivilen und demokratischen Gesellschaften zu schmieden, **erklären**, dass die Amerikaner das Recht,...

Table 6.19.: An example English sentence with its German SMT output. The verbs for which the agreement features, as well as their English counterparts are to be predicted are given in blue.

phrase *'der Motorrad sowie eine Person...'*. Assuming the parse tree of the given English sentence is correct, we may be able to derive correct agreement features of the English coordinated subject. However, our classifier also uses information about the English verbs and, in this case, *was* bears contradictory information in terms of person and number to that which we derive for the complex English subject.

Such structural mismatches between English and German pose a problem for finding the correct English subject (within the identified parallel English clause) in order to use it for the prediction of person and number of the corresponding German finite verb. Even if the subject is correctly identified, it might be ambiguous regarding the number (cf. relative pronoun *which* in Figure 6.10). Finally, not only the information about the subjects is used: we also use the verbs which might be confusing for the classifier since they are paired with unexpected subject agreement values.

Use of the German subjects Given the fact that we cannot always use an English subject to predict agreement for the German finite verbs, we could – or even should – use the German subject for this task. While in English, we have parse trees from which in many cases we can correctly extract the subjects, the German sentences are flat consisting of stemmed words enriched with morphological information. Searching for a subject relies on the clause marking of the German sentences (which is done using a few heuristics) and on the morphological annotation of the words. Both of these knowledge sources might have errors which lead to the identification of an erroneous (or no) subject.

Consider example sentences in Table 6.19. The generated translation is quite different from the English source sentence, especially when the subject of *erklären* is considered for which we need to predict the agreement features. In fact, it is even hard to identify the subject of *erklären*: without knowing the source sentence, we would tend to say that *Länder (countries)* is the subject. The correct subject is though *Amerikaner*, which is

in the given translation unfortunately the subject of the subsequent clause with its own finite verb. Using the subject extraction from German described in Section 6.3.4.1, we are not be able to find the correct subject for *erklären*. Its agreement could thus not be predicted using information about the German subject.

Use of the English and German subjects Sentence pairs shown above lead to the conclusion that we need information from both English and German to predict agreement for the German finite verbs which we also do in our experiments. However, we have seen that in both languages, the subject extraction may be erroneous. In other words, the information given in the training samples might be erroneous which might be confusing for the prediction model. Furthermore, the subjects might also not be near the verb so that using context words to decide on the agreement would often lead to erroneous predictions. In case that for a given verb subjects in both languages are identified, but differ in their agreement features, the model would have to decide which of the subjects is *more* correct and thus used to predict person and number.

Simple contextual features do not seem to be appropriate for the task of the agreement prediction. In may be helpful to enrich the classification feature set with additional information saying, for example, how reliable the identified subjects are. Instead of using constituency trees or a flat representation of the sentences, it would probably be more appropriate to use dependency trees. Avramidis and Koehn (2008), for instance, proposed to use dependency trees with semantic role labels of source side data, while Rosa et al. (2012) were successful in identifying target language subjects by using dependency parse trees of the MT outputs. We adapt the latter method to our language pair and show in Section 6.3.6 that the method also leads to moderate improvements of the German verbs with respect to the agreement between verbs and subjects in the SMT outputs.

6.3.5.4. Discussion of the tense and mood prediction

Results for the tense and mood predictions given in Section 6.3.5.2 are obtained with classifiers which use only a small subset of the features outlined in Section 6.3.4.2. Table 6.20 shows the full set, as well as the set of the actually used features.

Although classifiers trained on the full set of features get a very high accuracy on our well-formed test sets, their classification performance considerably lowers the quality of the German SMT outputs compared to the presented classifiers. We assume that classifiers trained on all defined features overfit to the training data and thus do not

Feature	English	German
finite verb	may	können
finite verb POS	MD	VMFIN
modal	1	–
infinitive	slow	verlangsamen
infinitive POS	VB	VVINF
participle	0	0
participle POS	0	0
VC	may-slow	können-verlangsamen
VC POS	MD-VB	VMFIN–VVINF
main verb	slow	verlangsamen
prev. clause main verb	0	–
adverbial	0	0
complementizer	0	0
word -1	drugs	**Medikament**
word -2	new	**neu**
finite verb align	–	may
clause type	SMAIN	–
tense	pres	–
finite verb ends with -s	0	–
finite verb ends with -ed	0	–
VC quoted	0	–
sentence main verb	slow	–

Table 6.20.: Summary of the features used to predict tense and mood. The features used for the final tense/mood classifier which is also applied on the German SMT outputs are given in bold.

generalize well when applied on unseen, slightly different data which is not well-formed.

In fact, all features used for prediction of tense and mood are a bit problematic in a sense that they have been defined to give clues about the usage of very specific labels, i.e., tense/mood combinations. For instance, *Konjunktiv I* is mostly triggered by reporting verbs which are placed in the subordinating clause. The feature which should account for this fact is called *prev. clause main verb*. This feature is presumably very interesting for *Konjunktiv I* in combination with the main verb of the previous clause such as *say, report, claim*, etc. All other verbs do not contribute to prediction of tense and mood, at least according to our intuition. On the other hand, Lee (2011) defined the relation between the verbs as anaphoric: if a given verb is anaphoric to the preceding verb, is tends to keep the time expressed by the preceding verb. Otherwise, it tends to change the time. We allow our model to cope for this fact by having access to both the current

verb, as well as to the verb placed in the preceding clause in terms of their lexical values.

As a second example, take, for instance, adverbials. There are many different adverbials, but our model may profit only from temporal adverbials such as *yesterday, lately, tomorrow*, etc. However, we extract all adverbials which might bring too much noise into the model. Furthermore, if we consider the adverbial *tomorrow*, it does point to a time point in future, but due to tense interchangeability given in both English and German with respect to present and future, it is very probable that it triggers present tense, also in contexts in which future tense would be more appropriate or even required.

One of the features which we introduced to help predict tense/mood for non-finite clauses is *clause type* which we combine with the tense of the subordinating clause which in most cases is transferred to the non-finite clause, i.e. its German translation as a finite clause. Again, this information is mostly interesting for non-finite constructions: for tensed VCs, the information is either not important or it even leads to false predictions which is why this feature is not included into the final set of features.

Obviously, our classifier has access mainly to the lexical features. If we refer back to previous work on modeling of tense in the context of machine translation, we come to the conclusion that more abstract features, i.e., *latent* features as defined by Ye et al. (2006), are indeed required to build classification models of a reasonable quality. This is also proven by Meyer et al. (2013) and Loáiciga et al. (2014) who used features such as *narrativity*, dependency types, semantic roles and temporal ordering information. Specially, the temporal ordering information is interesting to which we did not have a direct access. Instead, we indirectly provided this information by providing access to tenses in the preceding clause(s). Unfortunately, this was also one of the features that did not prove to be beneficial for our tense/mood classification model.

6.3.6. Parsing-based approach to correct agreement

The parsing-based correction of the German verbs is an adaptation of the method proposed by Rosa et al. (2012). In their tool called *Depfix*, they make use of the parse trees of the Czech SMT outputs to identify the subject-verb pairs and then to ensure that the verbs are inflected according to the agreement features of the corresponding subject.

We apply the same method for English→German SMT. After the nominal inflection step which generates fully inflected German SMT output, we first parse the German sentences. Subsequently, for each finite verb v_i in a parse tree, the corresponding subject s_i is determined. The person and number features of s_i are then transferred to v_i. Along with the tense and mood features, the verb v_i is inflected and inserted into the final

idx	token	lemma	POS	morph	head	syntRel
1	neue	neu	ADJA	nom|pl|neut|pos	2	NK
2	Medikamente	Medikament	NN	nom|**pl**|neut	3	**SB**
3	**konnte**	können	**VMFIN**	sg|3|past|ind	0	–
4	Lungen-	Lunge	TRUNC	–	5	CJ
5	und	und	KON	–	6	CD
6	Eierstockkrebs	Eierstockkrebs	CD	acc|sg|masc	3	MO
7	verlangsamen	verlangsamen	VVINF	–	3	OC

Table 6.21.: Dependency parse tree of an example German SMT output. Information used to correct the German subject-verb agreement is highlighted in bold.

SMT output.

In this work, the German dependency parser *Mate* (Björkelund and Nivre, 2015) is used to parse the German translations. An example of the Mate output is given in Table 6.21. The Mate parser provides different information which we need for the agreement correction: POS tags of the words in the parsed sentence (*POS*), their morphological features (*morph*), index of the head word (*head*), as well as the syntactic relation to the respective head word (*syntRel*).

Finite verbs are searched by looking for the corresponding POS tags, i.e., tags which end with the suffix *-FIN*. For every finite verb v_i, we subsequently look for the noun s_i whose syntactic relation to the given v_i is subject (i.e., *SB*). Finally, the agreement features agr_i of s_i are transferred to v_i. In the example sentence given in Table 6.21, the only finite verb is *konnte* with the subject head noun *Medikamente*. The number of the subject phrase is plural (*pl*) which is then copied to *konnte*. The person information is implicitly given in the Mate trees: all nouns are the 3rd person. After the generation step, the corrected form of the modal verb *konnte* exposes morphological features 3rd person plural (instead of the 3rd person singular generated by the SMT model) and thus matches its subject in terms of person and number.

Parsing-based agreement correction may in some cases be problematic. The SMT outputs are rarely grammatically fully correct which may lead to incorrect syntactic analyses of the SMT outputs. With respect to the agreement correction, we face with two problems: (i) incorrect recognition of the finite verbs and (ii) incorrect assignment of the subject relation. In German, there are verb forms which are ambiguous with respect to the finiteness (e.g. $arbeiten_{Inf}$ vs. $arbeiten1/3.Pl.$). We observed that sometimes, the Mate parser is not able to correctly analyze such verbs leading to the assumptions that a finite verb is an infinitive, or that an infinitive is a finite verb. In the first case, the finite

verb is not considered for the agreement correction due to the false POS tag, while in the latter, the infinitive gets inflected which usually leads to the incorrect clauses which have more than one finite verb.

Another problem we encountered in the Mate parses of the SMT outputs is multiple attachment of the subject relation to a single finite verb. In such cases, we use a simple heuristic based on the distance between the subject candidate and the finite verb to choose between the subjects. We assume that the nearest subject candidate is the correct subject of the given finite verb.

Despite these problems, in Chapter 7, we will show that parsing-based handling of the agreement errors is quite beneficial for correcting agreement errors in the German translations.

6.4. Chapter summary

This Chapter outlined our approaches to model verbal inflection for English→German SMT. In Section 6.1, we first gave an overview of the relevant morphological differences between the verbs in English and German. As a main problem for SMT, we defined high degree of syncretism given in the English verbal morphology. In contrast to English, German verbal morphology differentiates between a number of different inflectional variants of a single verb lemma. The inflectional variants contain information about person, number, tense and mood. SMT often makes mistakes when translating ambiguous English verbs regarding the choice of the German inflectional variants. On the one hand, this leads to the cases of erroneous subject-verb agreement. On the other side, wrong choice with respect to tense and mood may provide misleading information in the generated translations.

Inflectional problems are typical for SMT when translating between languages with different morphological richness. Accordingly, there is already lot's of work done in the past related to this topic. An excerpt of the approaches which are also related to our methods was given in Section 6.2. Many of the methods rely on pre-/post-processing of the data which simplifies the words of the morphologically rich language. When translating into a morphologically rich language, data pre-processing is combined with an additional step which transforms simplified SMT outputs into fully inflected set of sentences. This additional, *inflection generation* step may be carried out in different ways. For instance, the inflected words may be *predicted* with a classification model, simplified sentences can be *translated* with an SMT model into inflected sentences, etc.

One of the wide spread approaches is however to predict morphological features and then to generate the corresponding inflected word forms. The biggest strength of this approach is the ability to generate unseen target language words. Another way to reduce inflectional errors is to directly integrate specific linguistic information into SMT, for instance via factors in the factored approach to SMT.

Our approach to model verbal inflection was presented in Section 6.3. We first gave an overview of the processing pipeline in Section 6.3.1 and characterized our approach as a post-processing step to the translation. In other words, the methods aim at *correcting* finite verbs in the German SMT outputs. They may thus be seen as automatic post-editing step. Our methods are combined with the approach to handle nominal inflection for English→German SMT which was outlined in Section 6.3.2. As mentioned above, the German verbal morphology also includes information about tense and mood. For modeling of these morphological features, it was necessary to annotate the data with the respective information. In Section 6.3.3, we presented a tool that was developed to fulfill this task. To our knowledge, this tool is a first (open-source) tool for automatic annotation of tense, mood and voice for German, English (and French).

The first approach to model verbal inflection was described in Section 6.3.4. In line with the approach for modeling nominal inflection for English→German, as well as many previously proposed works, our approach is based on the prediction of the verbal morphological features. For the prediction, we use a CRF classification model trained on various contextual information. The features are predicted for every finite verb in a German SMT output. Subsequently, a verb form corresponding to the predicted features is generated with a morphology generation tool and inserted into the German SMT output which is to be corrected. The success of this method largely depends on the accuracy of the predictions which was presented in Section 6.3.5. For the agreement features, we obtained prediction accuracy in terms of F_1 score of 0.8, while for tense and mood, the prediction accuracy was between 0.87-0.9 depending on the classification method and the used feature set. Somewhat disappointing results with respect to agreement may be explained by the availability and correctness of the contextual features used to train the classifier. The features mainly bear information about the English and German subjects. The extraction of this information is often problematic due to syntactic divergences between English and German, as well as non-literal translations which come along with contradictory morphological information of the subjects in an English source and a German translation. Therefore, we implemented another method to handle agreement errors which is based on the identification of the subject-verb pairs in the German

translations by looking at their parse trees.

In addition to agreement, we also provided a discussion of the prediction problems with respect to tense and mood. Here, the problems are however much more complex. On the one hand, the extraction of the contextual information may be erroneous. But a more severe problem is the actual identification of the relevant contextual information.

The classifier presented in the current Chapter is a first attempt to model tense and mood for English and German using a set of directly accessible contextual features. In Chapter 9, we provide a further theoretical analysis of tense and mood in the bilingual context which indicates that a much larger set of not only morpho-syntactic, but also semantic and pragmatic features is required to account for the different characteristics of tense and mood in both mono-, as well as bilingual context.

7. SMT experiments with verbal inflection

Modeling of the verbal inflection is tested in two different scenarios. The methods developed within this work and described in detail in Chapter 6 intend to correct the German finite verbs in the SMT outputs in a post-processing step. In addition to this approach, we also test whether explicit tense/mood information may lead to better translation quality when it is used in the context of the factored SMT for the English→German translation direction.

In Section 7.1, we first specify data, as well as training settings which are used to build the phrase-based SMT models for English→German. In Section 7.2, the experiments with post-processing of the German phrase-based SMT outputs are described. Hereby, in Section 7.2.1, we first present an oracle experiment which helps to get an intuition of improvements which may be gained with the automatic correction of the German finite verbs. Subsequently, in Section 7.2.2, the results gained for the automatic correction of the verbs in the German translations are presented. In addition to the experiments with post-processing of the German SMT outputs, we also experiment with integration of tense and mood information into SMT via factors. The experiments with factored SMT models are described in Section 7.3. Findings of our experiments are summarized in the chapter summary in 7.4.

7.1. General SMT settings

Toolkit Similarly to the models build to test the performance of preordering, all models presented and discussed in this Section are as well built using the SMT toolkit *Moses* (Koehn et al., 2007). Moses enables training of both phrase-based, as well as factored SMT models which are used for the experiments described in this Chapter.

Training		Tuning		Testing	
Corpus	**Size**	**Corpus**	**Size**	**Corpus**	**Size**
News	259k			news2014	3003
Europarl	1.9m	news 2008-13	16,071	news2015	2169
Crawl	2.3m			news2016	2999

Table 7.1.: WMT data used for the verbal inflection modeling experiments. The size of the corpora denotes the number of the sentences.

Data	Size
WMT DE	4.5m
DE news mono	>120m

Table 7.2.: Overview of the data used to build the German language model.

SMT training settings Our SMT settings correspond to Moses' default training settings. The phrase-based translation model consists of the phrases up to the length of 5 words. We use 5-gram German language models for all experiments reported on in this Chapter. We set the distortion limit to 6 words. The model weights are optimized using the Minimum Error Rate Training (MERT) (Och, 2003). The maximum sentence length is set to 80 words. The SMT models are trained on the tokenized data. While the English side of the corpus is lowercased, the German texts are truecased.

Data Verbal inflection is tested on the WMT 2015 data (Bojar et al., 2015)[1]. The WMT data set is a concatenation of texts from three different domains: news (News Commentary), political discussions (Europarl) and mixed-domain texts crawled from the Web (Common Crawl). The statistics about the used WMT corpus, as well as of the tuning and testing data are shown in Table 7.1.

Language model For building the German language model, the German side of the WMT corpus in combination with a large collection of the German monolingual texts from the news domain is used.[2] The statistics about the language model data are given in Table 7.2.

[1] `http://www.statmt.org/wmt15/`

[2] The used German monolingual data can be downloaded from here: `http://www.statmt.org/wmt15/translation-task.html`.

7.2. Post-processing approach

The methodology to ensure German outputs with correctly inflected finite verbs described in Chapter 6 may be seen as an automatic post-editing of the German translations. The automatic post-processing of the German SMT outputs solely affects the German finite verbs. Words belonging to other word categories remain unchanged in our post-processing method.

In this Section, we aim at answering several questions. First, we try to estimate the correctness on the finite verbs in our baseline SMT model. A clearer picture about the amount of incorrectly inflected German finite verbs gives us a better intuition of how much improvement may be gained using our automatic correction of the finite verbs found in the German baseline SMT outputs. We present an oracle experiment to answer these questions in Section 7.2.1. We will see that only a small fraction of the finite verbs lexically match the reference translations which is not only problematic for computing upper bound of the expected improvement, but also for the general automatic evaluation of our post-processed outputs.

The main aim of the experiments described in Section 7.2.2 is to assess the performance of the methods for automatic correction of the agreement, as well as tense/mood errors in the German SMT outputs. We evaluate the German translations not only automatically in terms of BLEU, but also manually which gives more insights into contexts in which our methods lead to the expected results, as well as into cases which are particularly problematic for generation of verb forms with correct tense and mood properties.

7.2.1. Oracle

In this Section, we present an oracle experiment which helps us to estimate how many of the finite verbs in the German SMT outputs are already correctly inflected and how much improvement may be gained when inflectional differences regarding the finite verbs in the SMT outputs and the reference translations are removed. Additionally, we determine the number of the finite verbs in the German translations which lexically match the verbs in the corresponding reference translations. Since finite verbs represent only a small subset of the verbs in the German translations, this number may help us to estimate whether the post-processing of the German translations can even be captured with BLEU.

To answer the above mentioned questions, we create two different variants of an SMT outputs and their reference translations, respectively. One of the variants is the original data with inflected verbs, while the second variant consists of lemmatized finite verbs.

Data type	Example
original	Die Geschichte **ist** ein großer Lehrer.
vlemma	Die Geschichte **sein** ein großer Lehrer.

Table 7.3.: Examples of the two variants of data used for the oracle experiment. *original* denotes original, fully inflected reference sentence, while *vlemma* shows a reference sentence in which the verbs are finite verbs which are stemmed. In the shown example, the *original* sentence includes the verb form *ist (is)*, while the *vlemma* representation includes the stem *sein (to be)*.

By comparing inflected SMT outputs with inflected reference translations, we assess the actual quality of our baseline SMT system which we aim at correcting. By evaluating SMT outputs with lemmatized finite verbs against references with lemmatized finite verbs as well, we assume that all finite verbs in the SMT outputs are correct and can thus compute the maximum improvement which may be gained for the automatic correction of the German finite verbs. The two data representations (i.e., data types) are shown in Table 7.3. The lemmas of the finite verbs are gained by processing both SMT outputs, as well as reference translations with the RFTagger (Schmid and Laws, 2008).

For the oracle experiment, we make use of the news2015 test set. The baseline SMT outputs which are evaluated against two variants of the respective test set are generated with SMT models which include reordering (RO), as well as nominal inflection (ni).[3] The evaluation is performed on lowercased, tokenized data in order to eliminate side effects caused by detokenization and truecasing steps. The results in terms of BLEU are shown in Table 7.4. The results nicely show that for both baseline, as well as the reordered German SMT outputs, further quality improvements can be gained by handling the inflection of the finite verbs. The reordered German SMT output indeed has more errors and thus higher potential for the improvement regarding inflection of the finite verbs compared to the baseline (0.7 BLEU vs. 0.65 BLEU points). The potential improvement is however not as big as expected: direct verb comparison between the translations and the references revealed that in total, about 83% finite verbs in the outputs are already correctly inflected.

It needs though be noted that there is a problem with this estimation. The computed BLEU scores rely on the lexical comparison of the verbs in the SMT outputs and reference translations. A closer look to the computation of the number of the correctly inflected verbs showed that only 21% of the finite verbs in the translations match the

[3]These models will be used to generate translations which are post-processed in order to correct finite verbs. Details about the models will be given in the following Section.

	BL-ni	**RO-ni**
original	24.24	24.98
vlemma	24.89 (Δ 0.65)	25.68 (Δ 0.7)

Table 7.4.: BLEU scores of the German SMT outputs gained for different data representations.

reference at all. In other words, BLEU does not consider the major part of the finite verbs. This leads us to the conclusion that the potential improvement is probably even higher than the scores in Table 7.4 suggest. Furthermore, it needs to be noted that the RO-ni output contains 3,639 while the BL-ni output contains 3,408 finite verbs. The difference in the number of the actually generated finite verbs may explain why the improvement potential for BL-ni is somewhat lower than the improvement which can be reached for the considered RO-ni output. A very low portion of the finite verbs being considered by BLEU raises the question of what is the best way to (automatically) evaluate the translations that are output of our verb handling method. Scores given in Table 7.4 indicate that we may expect that the improvements are reflected in the BLEU evaluation. However, a large number of verbs not matching the reference also indicates that manual evaluation is needed to assess the actual benefit of the correction of the German finite verbs.

7.2.2. Automatic correction of the finite verbs

Correction of the German finite verbs is implemented as a post-processing step to the translation. In this section, we present evaluation of the German post-processed outputs by applying post-processing on the outputs of the SMT models which are trained on the reordered English data (cf. Chapter 4) and stemmed German outputs as described in Section 6.3.2. The models are trained using the SMT settings given in Section 7.1. We test the models on the news2015 test set and give the evaluation results in Table 7.5. If we compare the scores for BL and BL-ni, respectively, we conclude that reordering in a combination with nominal inflection leads to higher quality of the German SMT outputs. The post-processing of the RO-ni outputs regarding the verbs leads to an additional improvement of 0.05 BLEU points. Hereby, most of the improvement comes from the agreement correction (0.08 BLEU) while the tense/mood prediction unfortunately lowers the BLEU score of the BL-ni output by 0.05 BLEU points.

	BLEU$_{ci}$
BL	21.59
RO-ni	22.00
RO-niV	22.05
+ agreement	22.08
+ tense/mood	21.95

Table 7.5.: BLEU scores of the different German SMT outputs. *BL* refers to the baseline SMT model without any pre-processing of the data. *RO-ni* denotes the SMT model trained on the reordered English and stemmed German data including the inflection generation step. *RO-niV* is a model which includes a post-processing step for the correction of the verbal morphology.

	Grade			
	1	2	3	nA
RO-ni	**29**	19	4	19
RO-niV	17	31	4	19

Table 7.6.: Results of human evaluation. 1 = better, 2 = worse, 3 = don't know, nA = no majority vote.

Manual evaluation of tense and mood Small improvements in terms of BLEU do not provide meaningful insights into the differences between the different German translations. Therefore, in addition to the automatic evaluation, we also carry out manual evaluation of the German SMT outputs. We let four human evaluators to annotate a total of 70 sentence pairs consisting of a RO-ni and a RO-niV sentence. We were particularly interested in the correctness of the German verbs with respect to tense and mood. Agreement is ignored in this evaluation setup.

The evaluators had to make a binary decision: (i) which of the translation alternatives is better (1 is assigned to the better, 2 to the worse alternative) and (ii) whether it is possible at all to make a judgment (both alternatives are assigned the grade 3 if that was not the case). The results of the human judgments in terms of the majority votes for each of the sentence pairs are given in Table 7.6. The Table indicates that human evaluators prefer the choice of tense (expressed in verbal inflection) made by the BL-ni model. Only one third of the RO-niV alternatives is considered to be better than the corresponding RO-ni translation. In other words, our tense/mood prediction are not appropriate for about 66% of the finite verbs. However, 33% of the finite verbs in the given SMT outputs can be corrected via inflection generation according to the predicted tense and mood values. An interesting fact is that the annotator agreement in terms of

	Model	Output
RO-niV correct	EN1	The claimants **presented** proof of extortion.
	RO-ni	***Legte**$_{3.Sg}$ die Kläger Beweise von Erpressung.
	RO-niV	**Legten**$_{3.Pl}$ die Kläger Beweise von Erpressung.
	EN2	Then he **put** his finger on it.
	RO-ni	Dann ***legt**$_{Pres.Ind}$ er seinen Finger auf sie.
	RO-niV	Dann **legte**$_{Past.Ind}$ er seinen Finger auf sie.
	EN3	I **fear** I may need more surgery.
	RO-ni	Ich **fürchte**, ich ***kann**$_{Pres.Ind}$ eine Operation nötig.
	RO-niV	**Ich fürchte**, ich **könnte**$_{Past.Subj}$ eine Operation nötig.
RO-niV incorrect	EN4	Maybe his father **intended** to be cruel.
	RO-ni	Vielleicht **soll**$_{Pres.Ind}$ seine Vater grausam zu sein.
	RO-niV	Vielleicht ***sollte**$_{Past.Subj}$ seine Vater grausam zu sein.
	EN5	I **have rung** Mr. Pifffl and suggested that we get together.
	RO-ni	Ich **habe**$_{Pres.Ind}$ geklingelt Herr Pifffl und schlug vor, dass wir gemeinsam.
	RO-niV	Ich ***hatte**$_{Past.Ind}$ geklingelt Herr Pifffl und schlug vor, dass wir gemeinsam.
	EN6	No word **could get** beyond the soundproofing.
	RO-ni	Kein Wort **konnte** üer die Schalldämmung.
	RO-niV	Kein Wort ***könnte** über die Schalldämmung.

Table 7.7.: Example of the SMT outputs with improved (upper part) and incorrect verbal inflection (lower part).

Kappa was only 0.33 which means that the annotators often disagreed which translation alternative was better.

Translation examples To see what kinds of improvements, as well as errors are made by the correction of the German verbal morphology, we take a deeper look to the translations which are given in Table 7.7. We discuss the example translations from the perspective of the human evaluation, as well as in the context of SMT, parse-based agreement correction and the tense/mood prediction accuracy.

The RO-niV translation of EN1, for instance, shows a case of the corrected subject-verb agreement, in this example between the plural subject *Kläger (claimants)* and the finite verb *legten (presented)*. In the RO-ni translation, the agreement in terms of number is violated: while the subject is plural, the generated verb form *legte* is singular.

The translations of EN2 and EN3 show examples in which tense/mood prediction leads to the correction of the German finite verbs with respect to tense and mood. In EN2, the English verb *put* in past tense is translated by RO-ni as *legt* which is *Präsens*.

On the other hand, the post-processed RO-niV translation leads to the generation of the verb form *legte* and thus corrects the translation of *put* with respect to tense. In EN3, the German translation of the subordinate clause should be past subjunctive as generated by RO-niV (i.e., *könnte*). The English verb *may* does not directly indicate the subjunctive context. In German, it is typical to use the verb *fürchten (fear)* in a combination with *Konjunktiv* in the subordinated clause. Obviously, our tense/mood classification model recognizes this contextual dependency as it suggests to inflect *können* according to the German subjunctive tense form *Konjunktiv II*.

We now take a closer look to a few examples with erroneous verbal inflection caused by drawbacks in our approach. For instance, the RO-niV translation of *intended* in EN4 retains the tense in the source sentence. The human evaluators, however, prefer the RO-ni translation which switches to present tense. The RO-ni translation of the English VC *'have rung'* given in EN5 is *Perfekt* (*'habe/have geklingelt/rung'*) while the RO-niV translation is *Plusquamperfekt* (*'hatte/had geklingelt/rung'*). Both tense translations refer to an event happened in the past and even for a human, it is hard to decide which of the translations is *better*. Finally, the translation of EN6 shows a problem with English modal verbs such as *could* which expose functional ambiguity. As subjunctive, *could* almost always translates into subjunctive German modal *könnte*. Thus the classifier generally predicts *Konjunktiv II* for English modal verbs for which the past indicative form equals to the subjunctive form.

7.3. Factored-SMT

Experiments with post-processing of the German verbs with respect to tense and mood discussed in the preceding Section show that the prediction accuracy of our tense/mood classifier is not sufficient to improve the German SMT outputs. Therefore, we did another attempt to directly integrate tense/mood knowledge into SMT by using a factored approach to SMT. Factored SMT allows to integrate linguistic information by representing each word in a sentence as a combination of different *factors*. Typically, such factors include lemmas, POS tags, morphological information, etc. In our case, we experiment with adding a tense/mood factor which indicates the syntactic tense form of a VC in which a specific verb occurs.

We aim at investigating whether factored SMT is a better way to deal with verbal inflection regarding tense and mood. We examine two ideas: (i) annotation of the source syntactic tenses to the source verbs and (ii) annotation of the target-language syntactic

tenses to the source words. In the first scenario, we check the possibility of omitting the often inaccurate step of tense/mood prediction. In the second one, we check how beneficial the information about syntactic tense and mood in the target language is.

7.3.1. Monolingual tense/mood factors

English verbs expose a high degree of syncretism which is problematic when they need to be translated into an appropriate German verb form. Consider, for instance, an example sentence given in (42a). The VC *'have reacted'* consists of two verbs both of which may correspond to a number of different German inflectional variants. For instance, *reacted* is in this context a participle, however, its form is the same as the finite form of the lemma *react* in the past tense. While in the original English sentences, SMT will probably be able to capture the dependency between *have* and *reacted*, in the context of reordering, this is more problematic as the distance between the verbs may become much larger as shown in (42b). Access to an additional information which indicates in which context such verbs occur may help SMT to correctly interpret the English context and thus to generate appropriate German verb forms. For *reacted* in ENsyncVerbs, this information would be the tense form, namely *present perfect*.

(42) a. Several companies **have** thus far **reacted** cautiously when it **comes** to hiring.

b. Several companies **have** thus far cautiously **reacted** when it to hiring **comes**.

The factored representation of the sentences in (42) is shown in (43). The syntactic tense form of the VC *'have reacted'* is present perfect. We explicitly make this information available by enriching respective verbs with the factor *presPerf*. We assume that SMT may now be able to differentiate between ambiguous English verb forms and thus to to more often generate correct German verbs forms.

(43) a. Several companies **have**|**presPerf** thus far **reacted**|**presPerf** cautiously when it **comes**|**pres** to hiring.

b. Several companies **have**|**presPerf** thus far cautiously **reacted**|**presPerf** when it to hiring comes|pres.

7.3.2. German tense/mood factors

When translating English verbs, i.e., tenses into German, we aim at preserving the tense information given in the source sentence, however we also aim at providing the respective information in a German-specific way. In other words, we aim at generating tense forms which are typical for German in a specific context. We test whether we can reach this by enriching English verbs with tense/mood factors which correspond to their German counterparts. Consider the sentence pair given in Example (44) where the English VC *'have reacted'* in the present perfect corresponds to the German simple VC *reagierten* in the *Imperfect* tense. We take the German tense/mood information and add it as a factor *imperf* to the corresponding English verbs as shown in Example (45).

(44) Several companies **have** thus far **reacted** cautiously when it comes to hiring. ⇔ Denn zahlreiche Betriebe **reagierten** bislang verhalten bei Einstellungen.

(45) a. Several companies **have|imperf** thus far **reacted|imperf** cautiously when it comes|pres to hiring.

b. Several companies **have|imperf** thus far cautiously **reacted|imperf** when it to hiring comes|pres.

7.3.3. Experiments with tense/mood factors

We train a number of different SMT models using the WMT data set and test their performance on three news test sets (cf. Table 7.1, page 162 for details on the used data). The experimental setups are summarized in Table 7.8. We experiment with both non-reordered (*Baseline (BL)*), as well as with reordered English data (*Reordered (RO)*). We contrast the standard phrase-based SMT models (indicated by *Surface*) with the factored models. We test two kinds of factored models. The first one uses English tense/mood factors (i.e., *monoTM*) while the second relies on the factors derived from German (i.e., *deTM*). It should be noted that the aim of the presented experiments is to explore the *potential* of tense/mood factors for the English→German translation direction. For the *deTM* experiments, we thus use gold tense/mood labels derived from the reference translations instead of tense/mood values predicted with our tense/mood classifier which did not lead to expected results (cf. Section 7.2 for more details). For the factored models, only the verbs in the English texts are enriched with tense/mood factors. All other words are annotated with a dummy factor *null*. The German side of the corpus remains unchanged. The translation factors are word-to-word and tense/mood-to-word.

Model	PBSMT	Reordered	Factored	monoTM	deTM
BL	+	-	-	-	-
BL-monoTM	-	-	+	+	-
BL-deTM	-	-	+	-	+
RO	+	+	-	-	-
RO-monoTM	-	+	+	+	-
RO-deTM	-	+	+	-	+

Table 7.8.: Overview of the SMT experiments with tense/mood factors. *PBSMT* refers to a standard phrase-based SMT, while *Factored* denotes factored SMT models. *monoTM* includes tense/mood factors derived from English, while *deTM* makes use of tense/mood factors derived from the parallel German sentences. The models are partially trained on reordered English data which is indicated by the label *Reordered*.

	Baseline			Reordered		
	PBSMT	monoTM	deTM	PBSMT	monoTM	deTM
dev	18.00	18.01	18.00	18.69	18.82	18.83
news2014	18.37	18.29	18.42	18.82	18.95	19.17
news2015	19.98	20.25	20.05	20.63	20.41	20.71
news2016	24.64	24.82	25.02	25.38	25.36	25.63
avg	20.99	21.12	21.16	21.61	21.57	21.83

Table 7.9.: BLEU scores of the different German translations generated by phrase-based, as well as factored SMT models.

By comparing scores across different experiments, we provide answers to the following two questions:

(i) Do the tense/mood factors lead to improved German translations (*PBSMT* vs. *Factored*)?

(ii) Which type of factors is more appropriate (*monoTM* vs. *deTM*)?

The evaluation results in terms of BLEU gained for the different experiments are summarized in Table 7.9 and discussed in the following paragraphs.

PBSMT vs. Factored Our English→German baseline SMT model can indeed be improved by adding tense/mood factors to the English side of the data. While the monolingual tense/mood factors lead to an average improvement of 0.13 BLEU points, the German factors improve over the baseline by 0.17 BLEU points. The improvements however behave differently in the context of reordering. When combined with preordered

English data, the monolingual factors lead to a small decrease of BLEU by 0.04 points. On the other side, the German factors lead to an improvement of 0.22 BLEU points which is also the highest score gained for the current set of experiments.

Unfortunately, the results for the baseline factored model using monolingual tense/mood factors are not stable in a sense that the improvement is gained for all test sets. The model leads to better translations for the news2015 and news2016, but it fails to improve for the test set news2014. Similarly, the reordered factored model with monolingual factors performs well for news2014 and news2016, but it does not improve the translation of the test set news2015. These differences may indicate that the type of data plays a role for the factored tense/mood SMT. On the one hand, it might be the complexity of the sentences which causes problems regarding the automatic pre-processing and annotation of the data. On the other hand, the translations themselves may expose different tense/mood properties which do not match the training data well. In contrast to the monolingual factored model, the model with German factors improves for all test sets.

Monolingual vs. German factors Our experiments indicate that providing German tense/mood factors leads to greater improvements compared with models which use monolingual, i.e., English factors. However, the difference between the scores obtained for the baseline models is very small indicating that both monolingual, as well as German factors lead to almost the same improvement compared with the corresponding phrase-based model. In the context of reordering, the difference is however much bigger (0.26 BLEU) and indicates that German factors lead to higher translation quality compared with the quality of the translations generated by a factored model with monolingual tense/mood factors. One of the explanations might be the fact that establishing syntactic parallelism between English and German in a combination with explicit information about the verbs, i.e., tense forms, on the German side is very helpful for SMT with respect to the generation of verbs in the German SMT outputs.

Translation examples Table 7.10 shows a few example translations generated with the a phrase-based (RO), as well as with a factored reordered model (RO-deTM).

An example of a disambiguation of English verbs via tense/mood factors is given in the translations of the English input sentence EN1. The English word *struggle* may be both noun, as well as a verb. The RO model assumes that *struggle* is a noun and generates *Kampf* as a translation which is in this context wrong. On the other hand, RO-deTM correctly interprets *struggle* as a verb and generates the German infinitive

Model	Output
EN1	And there are many reasons he would **struggle** in a general election.
RO	Und es gibt viele Gründe, er in einer allgemeinen Wahl **Kampf** würde.
RO-deTM	Und es gibt viele Gründe, die er bei den Parlamentswahlen **kämpfen** würde.
EN2	It is possible that the event was observed by witnesses or residents **may** have heard something.
RO	Es ist möglich, dass die Veranstaltung von Zeugen beobachtet wurde oder Einwohner **kann** etwas gehört haben.
RO-deTM	Es ist möglich, dass die Veranstaltung von Zeugen beobachtet wurde oder Einwohner **könnte** etwas gehört haben.
EN3	Up to now, the parties to proceedings **have** agreed on the arbitrator amongst themselves...
RO	Bis jetzt, die Parteien Verfahren ist auf dem Schiedsrichter unter- einander vereinbart...
RO-deTM	Bis jetzt, die Verfahrensparteien **haben** sich auf die Schiedsrichter un- tereinander vereinbart...
EN4	Small and medium sized enterprises in particular **could be disadvantaged**, they said.
RO	Kleine und mittlere Unternehmen **benachteiligt werden könnten**, sagte sie .
RO-deTM	kleine und mittlere Unternehmen insbesondere **könnte benachteiligt werden**, sagten sie.
EN5	He has also said he could **skip** some sessions...
RO	Er hat auch gesagt, er könne einige Sitzungen...
RO-deTM	Er hat auch gesagt, er könne einige Sitzungen **überspringen**...

Table 7.10.: Example SMT outputs.

kämpfen. Besides the disambiguation of the English verbs according to their finiteness, agreement, tense and mood features, tense/mood factors can thus also help to distinguish between words which are ambiguous with respect to the part-of-speech.

An adaptation of the verb (i.e., tense form) translation to German specifics is shown for EN2 which contains the modal verb *may*. The RO system generates verbose translation, namely *kann*[4] which is typically not used in this specific combination of verbs in German. In fact, we would expect the German modal to have the subjunctive mood (i.e., *könnte*) as is the case in the output of RO-deTM.

Tense/mood factors may also help to choose correct auxiliary for the composed Ger-

[4]Note that the present analysis is focused on tense and mood. Other errors, also the subject-verb agreement errors, are ignored in the current discussion.

man tense forms. Given the VC *'have agreed'* in EN3, RO system generated *'ist vereinbart'* which corresponds to the stative passive. This translation option may be valid in certain contexts. However, in the given sentence, we would instead expect to have *Perfekt*, i.e., active VC in the past tense which is indeed generated by the RO-deTM model (*'haben vereinbart'*).

The RO translation of EN4 shows an incorrect order of the German verbs, namely *'benachteiligt werden könnten'*. This order is valid in subordinate clauses, however, in main clauses, the finite verb must be the first verb in the given VC. Tense/mood factors might have been helpful to generate the correct order of the verbs, namely *'könnte benachteiligt werden'* as shown in the translation output by the RO-deTM model.

Reordering does not always guarantee that the verbs in the German outputs are actually generated. For instance, the translation of the verb *skip* given in EN5 is missing in the RO output. In the RO-deTM translation, we do find the the translation of *skip*, namely *überspringen*, which is in addition placed in the correct position.

7.4. Chapter summary

This Chapter presented SMT experiments in which we evaluate performance of the method for modeling verbal inflection for English→German SMT presented in Chapter 6.

The proposed method is applied as a post-processing step to the translation a simple 2-step-pipeline. First, the German SMT outputs are generated using a SMT model trained on reordered English sentences and a stemmed representation of their German counterparts. Details about the data used to train the models, as well as training settings are summarized in Section 7.1. The experiments with the post-processing approach are presented in Section 7.2. The Section began with an oracle experiment described in Section 7.2.1 which aims at assessing the amount of improvement that may be gained by correcting inflection of the finite verbs in the German translations. By eliminating the inflectional differences regarding the finite verbs between the considered SMT outputs and the corresponding reference translations, we compute upper bound of the improvement which may be obtained by applying our methodology to correct inflection of the German finite verbs. The experiment revealed that ca. 80% of the German finite verbs are already correct. Nevertheless, the correction of the remaining verbs in the considered data set may lead to a considerable improvement of the German translations of 0.7 BLEU points. The experiment however also revealed that many of the verbs are

not considered in the BLEU evaluation since they do not match the verbs in the reference translations on the lexical level which may indicate that the computed potential improvement is probably somewhat underestimated.

In Section 7.2.2, the evaluation of the proposed approach to the correction of the German finite verbs was presented. German finite verbs expose different morphological features: person and number (agreement), tense and mood. We make use of two different approaches to gain the verbal morphological features. Agreement features are gained by parsing the German translations which allows for the identification of the subject-verb pairs. Given those pairs, person and number of the a finite verb are adapted to the person and number features of the corresponding subject. Tense and mood features are obtained by a classification model which considers different contextual information and predicts the corresponding tense and mood values. The evaluation of the German SMT outputs is given in terms of an overall improvement compared with the translation that we want to correct, as well as regarding the two above mentioned feature groups. This allows us to assess the appropriateness of the respective methods. Our experiments showed that the overall improvement is rather small. Compared with the original translations, the post-processed translations are better by 0.05 BLEU points. We observed that the agreement correction comes along with an improvement of 0.08 BLEU points, while the tense/mood handling lowers the quality of the original translations by 0.05 BLEU points. In other words, the implemented approach to agreement correction leads to translations in which the subject-verb agreement errors are not as frequent as in the original translations. Prediction of tense and mood unfortunately does not lead to better German translations according BLEU. This result has also been confirmed by a manual evaluation of the translations which includes human judgments about the quality of the original vs. post-processed translation pairs. Human evaluators preferred the original translations in 66% of the cases, while the post-processed translations were considered better in 33% of the cases.

In Section 7.3, we explored another possibility to improve German translations with respect to tense and mood, namely by explicitly providing information about tense to and SMT model via tense/mood factors annotated to the English verbs (i.e., factored SMT). We experimented with two different types of tense/mood factors: (i) monolingual which include information about syntactic tense in English and (ii) German which provide information about syntactic tense of the German verbs that are translations of a given English verbal complex. The evaluation showed that monolingual factors may lead to better translations, however, they failed to improve all of the used test sets. On the other

hand, the German factors lead to better translations in all experiments whereby the improvement in the context of preordering is considerably higher that the improvement gained for the baseline model trained on non-reordered English (0.26 vs. 0.17 BLEU points, respectively).

Both the oracle experiment presented in Section 7.2.1 as well as the factored SMT experiments described in Section 7.3 indicate that SMT has problems with tense and mood, i.e., with generating correct verb forms in the German translations. The proposed post-processing approach aims at using a dedicated model to correct at least some of these errors, however the classification model that we used is not accurate enough to successfully fulfill this task. From the theoretical point of view, one might argue that this kind of post-processing is not appropriate to model tense and mood translation. It may indeed be problematic to cope with this issue completely independently from the translation process. Therefore, we started another attempt to tackle the problem in which we provide tense/mood information in form of factors. They are then used in the translation process and indeed often lead to generally better translations compared with the German phrase-based (post-edited) translations. We observed that German (i.e., target language) factors seem to be more informative as they point to the tenses which are typical for the target language. This fact however brings us back to the question of how to obtain these factors for the testing data. Our classifier provides the methodology, however further research is needed to come up with features which ensure predictions of sufficient accuracy (cf. Chapter 9).

8. Verbs in English→German NMT

Since 2016, the neural machine translation (NMT) is the new state-of-the-art approach to the MT. The studies which assess the quality of the NMT translations, as well as the main differences between SMT and NMT outputs indicate that NMT leads to considerably better translations compared with SMT for many different language pairs, English-German being one of them. Bentivogli et al. (2016) and Popović (2017), for instance, report that NMT is very successful in dealing with problems which are very difficult for SMT such as the long-range reordering. Inflectional errors are reduced as well. Both of these findings, among other, also affect verbs in the German translations which are the main interest of this thesis. We thus take a closer look to the verbs in the German NMT outputs and analyze them with respect to their position (cf. Section 8.1), as well as inflection (cf. Section 8.2).

In Section 8.1.1, we first present data that we use for the analysis and then summarize evaluation results in Section 8.1.2. In Section 8.1.3, we outline an experiment which combines preordering introduced in Chapter 4 with NMT. Finally, in Section 8.1.4, we summarize conclusions derived from the manual analysis of the translations, as well as from the experiments with preordering. Section 8.2 includes an evaluation of the German NMT outputs with respect to the inflection of the verbs. A special attention is given to tense and mood which are discussed in Section 8.2.1. The Chapter summary is given in Section 8.3.

8.1. Positional issues

Bentivogli et al. (2016) report on an extensive comparison of the SMT and NMT outputs for the English→German translation direction. They compare different SMT systems with an NMT system developed by Luong and Manning (2015) which was the best performing MT system in terms of BLEU in the WMT 2015 news translation task (Bojar et al., 2015). The comparison is focused on three categories of errors: morphological, lexical and word order errors. They observed that particularly the verb order errors are

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

reduced by 70% in the examined NMT translations compared with the considered SMT outputs. Similar findings are also reported by Popović (2017). In the examined German NMT translations generated by the system developed by Sennrich et al. (2016b), the verb order errors are almost negligible: they occur in only 1.5% of the investigated test sentences. The author however notes that the used test sentences are not longer than 36 words, thus the evaluation does not include a discussion about the errors in longer sentences which are known to be problematic for NMT.

In this section, the verb order in the German NMT outputs of the different sentence length is discussed. First, an overview of the errors found in the sentences of the different lengths are given. Subsequently, a detailed analysis of the errors found in the sentences with more than 50 words is presented.

8.1.1. Evaluation data

In order to examine the position of the verbs in the German NMT outputs, we select a random set of sentences of the different length (i.e. number of tokens) from the WMT 2017 (Bojar et al., 2017) news test set[1]. Their translations are obtained with the English→German NMT system developed by Sennrich et al. (2017). 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 VCs, the evaluation is focused on the VCs found in the chosen test set. The statistics about the test set are given in Table 8.1.

Examination of the test sentences indicates that there are two different types of the long English sentences, i.e., sentences with more than 50 tokens. To the first type belong sentences which are a sequence of short clauses. Assumed that the verbs are not moved outside the given clause, this means that for this kind of long sentences the difference in the position of an English source verb and its German translation is often not very

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

big. The second type of the sentences includes the sentences with many noun phrases or prepositional adjuncts within a single clause. For certain clause and VC types (cf. Section 3.3), this means that the position of a source verb and its translation often differ considerably. Examples of the two sentence types are given in Table 8.2.

8.1.2. Results of the manual evaluation

The evaluation is to make 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 8.3.

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 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).

Although the numbers in Table 8.3 indicate that 50% of the sentences from the length range 50-59 have at least one verb order related error, these numbers are a bit misleading because they may give an impression that many of the long German translations are not good. 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 8.4. The results show that only 20% of the VCs in the respective English sentences are 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).

How severe are these errors? Let us consider the sentence pair in Table 8.5. The English source sentence consists of many short non-finite clauses. NMT needs to carry out two difficult tasks to generate the correct German translation for the given English sentence: (i) it needs to translate English gerund clauses into finite German clauses and (ii) it has to put the verbs in the generated finite clauses into the correct positions. As the colors indicate, all of the English verbs have been translated into German. With the exception of *stopping*, i.e., *Stoppen*[2], all of the gerund clauses have indeed been translated as finite clauses. The German sentence starts with the conjunction *wenn (if)* which requires the verbs in the subordinate clauses to occur in the clause-final positions. This also holds for *kuppte (cupping)* which should have been placed after

[2]Here, the English gerund *stopping* is translated into the German noun *Stoppen* which is a valid translation although the synonym *Anhalten* is more frequently used.

Words	EN source	DE reference	Type
81	Employers were hopeful that the continued positive engagement on other important topics - such as deployment , flexibility in training , additional training for those returning from career breaks , costs of training , mutual recognition of syllabus , study leave and the gender pay gap in medicine - were a sign of how serious employers , Health Education England and the Department of Health were about honouring the agreements reached with the BMA in November , February and May .	Die Arbeitgeber zeigten sich hoffnungsvoll, dass das anhaltende positive Engagement für andere wichtige Themen - wie der Einsatz, die Flexibilität in der Ausbildung, die Zusatzausbildung für diejenigen, die aus beruflichen Pausen zurückkehren, Ausbildungskosten, die gegenseitige Anerkennung von Lehrplänen, Studienurlaub und das geschlechtsspezifische Lohngefälle in der Medizin - ein Zeichen dafür waren, wie ernst Arbeitgeber, Health Education England und das Gesundheitsministerium die mit dem BMA im November, Februar und Mai erzielten Vereinbarungen einhalten.	Sequence of NPs and/or adjuncts
56	If this blog was a televised news report , the camera would follow me as I walked down the middle of a busy Soho street , wearing a modest grey suit and gesturing wildly before stopping , cupping my hands and saying something authoritative like : " So , let 's take a look . "	Wenn dieser Blog ein Nachrichtenbericht im Fernsehen wäre, würde mir die Kamera folgen, während ich inmitten einer geschäftigen Straße in Soho laufe, einen bescheidenen grauen Anzug trage und wild gestuliere, bevor ich stehenbleibe, meine Hände halte und etwas Bestimmendes sage wie: "Also, lassen Sie uns einen näheren Blick darauf werfen."	List of clauses

Table 8.2.: Examples of English sentences with more than 50 words.

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 8.3.: 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 8.4.: Number of the erroneously translated English VCs in sentences with token number greater than 50 words having at least one verb order related error.

EN source	NMT output	Errors
If this blog was a televised news report , the camera would follow me as I walked down the middle of a busy Soho street , wearing a modest grey suit and gesturing wildly before stopping , cupping my hands and saying something authoritative like : " So , let 's take a look . "	Wenn es sich bei diesem Blog um einen Fernsehbericht handelte, würde mir die Kamera folgen, als ich die Mitte einer belebten Soho Street hinunterging, einen bescheidenen grauen Anzug trug und wild vor dem Stoppen, kuppte meine Hände und sagt etwas maßgebliches wie: "Also nehmen wir mal einen Blick".	kuppte, sagt

Table 8.5.: Example of the German NMT output. The source sentence contains 56 tokens. Verbs in the source and the translation are indicated with different colors.

the object NP *'meine Hände' (my hands)* and for *sagt (saying)* which should have been put after *maßgebliches (authoritative)*. To summarize, out of 9 difficult English verbal complexes, 7 of them have been translated correctly while the position of two of the verbal translations is wrong.

The low number of the erroneously translated English VCs into German suggests that NMT has solved the problem of the verb placement in English→German NMT. On the other hand, it also shows that there still is (a small) room for improvement. As Table 8.4 indicates, the problem of the not generated verbs is more frequent than the problem of the misplaced verbs. Not having the verbs in the translations presents a big problem for adequate understanding of the generated German translations.

8.1.3. Preordering for NMT

Since preordering is very simple to combine with different MT paradigms, we also tested its influence on NMT. The combination of preordering with NMT was tested on three language pairs, one of them being English→German handled in this work. In the fol-

	BL		RO	
	BLEU	Human	BLEU	Human
EN→DE	38.26	49.2	36.74	50.08

Table 8.6.: Evaluation results for the preordering combined with English→German NMT. *BL* denotes the model trained on the non-modified parallel corpus, while *RO* refers to a model which has been trained on the reordered English part of the training data.

lowing, we only concentrate on English→German, for other language pairs, please refer to (Ramm et al., 2017b).

Preordering for NMT was tested on an English-German corpus from the legal domain consisting of about one million sentences. The pipeline is the same as for SMT: the source language data is preordered (see Section 4.3.2 and Figure 4.9) and then used to train the English→German NMT model. We used the open-source toolkit OpenNMT (Klein et al., 2017) to train a single RNN (Recurrent Neural Network) encoder-decoder model (Cho et al., 2014), (Sutskever et al., 2014) with attention mechanism (Bahdanau et al., 2014). We used a word-segmentation with byte pair encoding (BPE) (Sennrich et al., 2016c) of 25,000 operations. We built the BPE dictionary from normal-cased (i.e., lower- and upper-cased) tokens. Each model was trained for a maximum of 15 epochs, using the ADAM (Kingma and Ba, 2014) learning optimization function with initial learning rate of 0.005.

The results for English→German are shown in Table 8.6. In terms of BLEU, the RO system performs worse compared to the BL system. That means that preordering hurts the NMT performance. However, the human evaluation which consists of a simple judgment which of the translations alternatives is better, indicates that the RO translations are slightly better than the BL translations. However, the results of the human evaluation do not correlate with the results gained for other language pairs for which preordering is combined with NMT (cf. (Ramm et al., 2017b), (Du and Way, 2017)). Hence, we cannot see them as a reliable indicator that preordering is helpful for English→German NMT.

8.1.4. Discussion

NMT is able to solve many of the reordering problems which SMT has difficulties to deal with. Particularly for the English→German translation direction, the NMT is very successful in generating German translations with considerably better word order compared

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

to SMT. Nevertheless, our evaluation shows that longer sentences may be problematic with respect to the placement of the verbs in the German NMT outputs. The errors do not occur very often, however, in cases in which they are observed, they are related to the non-generation of the verbs which has a negative impact on the understanding of the translations.

The combination of preordering with NMT was already tested on different language pairs. The evaluation results reported on the respective experiments (cf. (Ramm et al., 2017b) and (Du and Way, 2017)) are in favor of the BL NMT models indicating that preordering is not beneficial for NMT. On the other hand, the manual evaluation shows that not all of the RO translations are considered to be worse than their BL counterparts. On the contrary, for instance, for the English→Chinese language pair, the human evaluation shows that the majority of the BL translations is considered to be better than their RO alternatives, however, 30% of the RO sentences are judged to be better than BL. Obviously, preordering may help to improve some of the NMT translations. Future work would be to examine possible combinations of the BL and RO models in the context of NMT to get the best of both worlds.

8.2. Verbal inflection

Similarly to the conduction of the verb placement errors in the German NMT translations, we also evaluated the German NMT outputs with respect to the inflection of the finite verbs (details on the evaluation set are given in Section 8.1.1). The errors regarding the verbal morphology are counted separately for each of the German verbal morphological features: agreement (person, number), tense and mood.

The evaluation results shown in Table 8.7 indicate that the total amount of errors regarding the inflection of the German finite verbs is very small. In a total of 74 sentences consisting of 138 VCs on the source side, only 10 VC translations (7%) have an error

EN	The police patrolled the area **closing** parts of the street (...) and **preventing** visitors from accessing house
DE NMT	Die Polizei patrouillierte den Bereich (...,) die Teile der Straße **schließen** und die Besucher vom Zugang zu Haus **hindern**.
DE REF	Die Polizei patrouillierte den Bereich (...,) **sperrte** Teile der Straße und **hielt** Besucher davon **ab**, das Haus zu betreten.

Table 8.8.: Example of erroneously translated English non-finite VCs (given in bold).

regarding the verbal inflection. Out of these, there are no errors regarding the subject-verb agreement: all of the observed errors are related to tense and mood of the generated finite verbs. Note that the errors are spread across the different sentence lengths. While the verb placement errors indeed can be put into the relation with the sentence length, tense and mood errors depend on sentence properties which are not necessarily related to the amount of words in a sentence. In the following, we discuss contextual properties in which tense and mood errors in the German NMT outputs occur.

8.2.1. Tense and mood errors

Manual inspection of the German NMT translations revealed a few contexts in which NMT tends to make mistakes regarding tense and mood of the German finite verbs. We relate the observed errors to the following linguistic issues:

1. non-finite (tenseless) English VC translating into finite German VCs;
2. ambiguous English VCs with respect to tense;
3. indicative English VCs translating into subjunctive German VCs;
4. translation into the German subjunctive mood.

These issues are discussed in the following paragraphs. We also show examples of NMT outputs for each of the above mentioned aspects.

Non-finite English VCs English often uses non-finite VCs in subordinate clauses which are usually translated into the German finite constructions. An example is given in Table 8.8 in which *closing* and *preventing* are translated into finite German VCs *sperrte* and *'hielt ab'* both of which have specific tense/mood values (tense = past, mood = indicative). In the English source VCs, these features are not overtly given which, in this example, leads to false tense of the verbs in the German NMT output *schließen* and *hindern*. Both verbs are namely in the present tense.

EN	Heavy rain and widespread flooding in Louisiana **lead** the governor to declare a state of emergency on Friday ...
DE NMT	Starker Regen und weit verbreitete Überschwemmungen in Louisiana **führen** den Gouverneur dazu, am Freitag ...
DE REF	Heftige Regenfälle und großflächige Überschwemmungen in Louisiana **zwangen** den Gouverneur dazu, am Freitag ...

Table 8.9.: Example of an erroneously translated English ambiguous verb.

EN	Donald Trump wouldn't really mind if he **lost** the US presidential election in November.
DE NMT	Donald Trump würde es nicht wirklich ahnen, wenn er die US-Präsidentschaftswahl im November **verloren hat**.
DE REF	**Verliert** Donald Trump die US-Präsidentschaftswahlen im November, wäre ihm das relativ egal.

Table 8.10.: Example of an erroneously translated English ambiguous verb.

Ambiguous English verbs English verbs are highly ambiguous with respect to their morphological features. Consider, for instance, the example given in Table 8.9. The English verb *lead* is lexically correctly translated as *führen*, however the tense of the generated verb form is wrong. It indicates present tense, while the source, as well as the reference translation *zwangen (forced)* are in the past tense. An interesting fact about this example is that we are able to decide whether the translation is incorrect only by looking at the context in which *lead* occurs. *'Heavy rain and flooding'* obviously happened first which was the reason for *'declaring a state of emergency'*. The time point of the declaration is determined by a PP *'on Friday'*. The respective temporal PP is underspecified as it could also point to a time point in the future. However, in the combination with something that already happened in the past, it is very likely that it refers to a time point in the past as is the case in our example. These very briefly sketched contextual dependencies are very complex. NMT obviously failed to capture them properly and generated the German verb in the present tense.

Indicative → subjunctive An interesting case of a mood-related error has been found in the conditional context as shown in Table 8.10. The English finite verb *lost* is indicative of past tense. In the given context, we however expect it to be translated into the German subjunctive tense form *Konjunktiv II* (i.e., *'verlieren würde'*) which is typically used in the conditional clauses beginning with a conjunction *wenn (if)*. Our NMT model translated *lost* as *'hat verloren'* which is indicative past tense (*Perfekt*)

EN	Angela Rochelle Liner, Stefanie R. Ellis and Ellis' daughter Maleah **were shot** (...,) authorities said.
DE NMT	Angela Rochelle Liner, Stefanie R. Ellis und Ellis 'Tochter Maleah **seien** am 12. Juni **erschossen worden** (...,) teilten die Behörden mit.
DE REF	Angela Rochelle Liner, Stefanie R. Ellis and Ellis Tochter Maleah **wurden** am 12. Juni **erschossen** (...,) sagten die Behörden
EN	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 ...
DE NMT	Nach Grundstück, Parcelling und Barrierefreiheit **seien** Fragen **beantwortet worden**, und Anträge für Messungen **seien gestellt worden**, es **stehe** dem Verkauf ...
DE REF	Nachdem die Grundstücks-, Parzellierungs- und Erschließungsfragen **geklärt** und die Anträge auf Vermessung **gestellt werden konnten**, **steht** dem Verkauf ...

Table 8.11.: Example of translations into German *Konjunktiv I* tense forms.

and thus generated a grammatically incorrect German translation. Note that the reference translation indeed contains the indicative verb form *verliert*, however, this is a rather exceptional usage of a mixture of an indicative and a subjunctive mood within a conditional sentence.

Translation into German subjunctive mood The subjunctive mood in German, particularly *Konjunktiv I*, is commonly used to indicate indirect speech. However, not only subjunctive mood is syntactically allowed, but also indicative. The choice is often described as a matter of author's preference and genre/domain specifics.[3]

In our test set, we found two English sentences for which NMT seems to be confused about the German subjunctive mood. The sentences, as well as their translations are given in Table 8.11. In the first example, NMT translated *'were shot'* as *'seien erschossen worden'* while the reference translation suggests the indicative translation in the past tense *'wurden erschossen'*. In the second example, the English VC *'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* sounds however rather strange in this context.

[3]This issue is discussed in more detail in Chapter 3.

For both sentences, we count mood as erroneous since they do not match mood given in the reference translations. However, for both sentences there might be contexts in which subjunctive is a favorable translation option. Assuming that the first sentence occurs in a news article, it is quite common to use *Konjunktiv I* which NMT also generated, although indicative is grammatically also correct. There is a small pragmatic difference between the two moods: subjunctive mood indicates the non-assertion of the author regarding the proposition of the utterance while this is not clearly expressed in the indicative form. Subjunctive translations for the English VCs in the second example would be fine in a context in which in the preceding sentence(s) somebody is writing about findings reported by someone else.

Similarly to the discussion of translating ambiguous English verbs with respect to tense, the choice of the mood depends on different factors which are beyond the lexical information expressed in the words of the given sentences. Furthermore, particularly for *Konjunktiv I*, access to the preceding sentence(s) is often required, even for a human translator/evaluator, to decide whether *Konjunktiv I* is appropriate or not.

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 *say, answer*, etc. (cf. examples in Table 8.11) or with quotation marks. However, it is sometimes not able to distinguish between context in which these contextual clues should not lead to the generation of subjunctive German VCs.

8.3. Chapter summary

NMT is a very powerful device for automatic translation. The difference between the quality of the NMT and SMT translations is huge for many different language pairs. Studies about the quality of the NMT translations showed that NMT generates syntactically, as well morphologically correct translations in considerably more cases compared with SMT. The quality improvement also affect the English→German translation direction where particularly errors regarding generation, placement and inflection of the verbs are significantly reduced. Nevertheless, NMT outputs are yet not perfect: the errors are reduced compared with SMT, but they are not completely eliminated. We took a deeper look to the German NMT outputs and examined them with respect to the verbs.

In Section 8.1, we carried out a manual evaluation of the German NMT outputs with

respect to the generation and placement of the verbs. After introducing the evaluation data in Section 8.1.1, we presented the results of our analysis in Section 8.1.2. Since it is known that NMT has problems with translating long sentences, our evaluation results are provided in relation with sentence length. Indeed, we observed that positional problems start to occur with sentence length of 50 words. Shorter sentences seem not to be problematic for NMT. Although we found several verb-related errors in the set of sentences with more than 50 words, when the number of errors is put into relation with the total number of the verbal complexes in the entire test set under consideration, it came out that only a small fraction of the English VCs of 7% was not correctly translated into German. More errors are made with respect to omission than placement of the verbs. This might be seen as a problem worth dealing with since sentences (or clauses) without verbs are often difficult or not possible to understand. These errors in the context of SMT are successfully handled by preordering the source language data. We examined whether NMT may similarly profit from the preordering approach. The results of our experiments were given in Section 8.1.3. The outcome or our experiments is sobering: preordering considerably lowers the quality of the NMT outputs. Despite this negative result, manual evaluation revealed that about one third of the reordered translations are considered to be better than their non-reordered counterparts indicating that for certain sentences, it might indeed be desirable to combine NMT with preordering.

The analysis of the verbs in the German NMT outputs also includes issues regarding inflection. Section 8.2 summarized outcomes of our evaluation with respect to the agreement, as well as tense and mood properties of the German finite verbs. First, the evaluation revealed that NMT makes no errors regarding the subject-verb agreement and only a few errors with respect to tense and mood. These errors are not related to the sentence length. In fact, they are spread across all considered lengths. In Section 8.2.1, we gave a detailed analysis of the identified tense/mood errors. We observed that errors occur when translating English non-finite VCs into finite German constructions, as well as in cases when English VCs contain an ambiguous verb with respect to tense. Furthermore, we observed certain difficulties when translating into German subjunctive mood. On the one hand, subjunctive constructions are not generated in cases where they need to be used. On the other, particularly the German *Konjunktiv I* seems to be overgenerated which indicates that NMT has troubles to distinguish between reported speech and other contexts. Often, very complex contextual dependencies need to be taken into account, even by a human, to decide on tense and mood when translating English tense forms into German.

9. Revisiting tense and mood in (machine) translation

In Chapter 3, many different linguistic aspects with respect to tense and mood have been mentioned. While some of them were taken into account within different, mainly lexical, features derived from the English-German parallel sentences, a few of the tense/mood related aspects were not considered further in the present work, until now. This Section presents a more detailed discussion of the tense/mood in the bilingual context. The discussion provides more insights into this complex topic which may help to explain the insufficient quality of the tense/mood classifier presented in Chapter 6. The discussion will furthermore point the way for further research on theory and modeling of tense and mood.

The success of classification-based modeling of tense and mood for machine translation depends on the features used to train a tense translation model. Our attempt to build such a model for English→German using mainly lexical features has proven that lexical features do not provide sufficient information which would be required to achieve good prediction accuracy. In their work on tense translation for Chinese→English, Ye et al. (2006) observed that features such as *telicity, punctuality* and *temporal ordering* are more informative than the lexical features to which our feature set belongs. On the other hand, Loáiciga et al. (2014) have shown that their tense classifier trained mostly on lexical features is able to improve the French SMT outputs with respect to tense. The classifier developed by Loáiciga et al. (2014), however, also used information about temporal ordering of the events in a sentence. Furthermore, it was trained on a rather small set of parallel texts which has been partially manually cleaned with respect to specific annotations. The two successful attempts to model translation of tense point to two important aspects: (i) important features are related to semantic and pragmatic properties of a text and (ii) choice of the training data, i.e., correctness of the annotations play an important role.

Statistical modeling of translation of tense (and also mood, voice and aspect) is mostly

used in the context of SMT and is based on specific contextual information. In other words, we provide some information to computational models and assume that they are able to learn how the tense is to be translated within a specific language pair. The opposite approach to this is manual definition of tense translation as needed for rule-based MT systems. A deeper look into such rules provides information about theoretical description of tense in the context of *human translation* and how that theory can be formalized for rule-based MT. The analysis provides a concrete idea about which textual properties need to be considered to model tense/mood translation.

The remaining of this Section is structured as follow. Section 9.1 gives a brief overview of the relevant differences between English and German with respect to the *meaning* of the tense forms given in each of the languages. In Section 9.2, the use of tense and mood regarding domain and register is discussed. Section 9.3 discuss the factor *human translator* with respect to tense and mood translation, while Section 9.4 outlines difficulties regarding the evaluation of tense and mood. An overview of an attempt to formalize the translation of tense in the context of rule-based MT is given in Section 9.5. The main findings of the Chapter are summarized in the Section 9.6. Finally, the Chapter is concluded in Section 9.7.

9.1. Linguistic aspects

In Chapter 3, we have shown that the tense systems in English and German differ from each other. The English tense system includes different aspects (progressive, perfect) which results in a set of 16 different tense forms (see Table 3.10 on page 41). On the other side, the German tense system does not have an explicit information about aspect but it includes morphosyntactically expressed subjunctive mood which does not exist in English. These differences lead to a non-unique mapping between English and German tense forms (see, for instance, Table 3.8 on page 57).

English and German tenses differ in their meaning as given in Table 9.1, taken from (König and Gast, 2012, p. 92). Interesting cases are those which include what we refer to as *tense switch*. For instance, the German present tense can be used to refer to a future time reference (*futurate* use in Table 9.1) while in English, the reference to a future time point by means of a simple present tense is possible only in very specific contexts determined, for instance, by schedules (*e.g., 'Mary starts her new job on Tuesday.'* (König and Gast, 2012, p. 85)). Note that the description of tense usage refers to different aspects: (i) time reference (*past, futurate, future*, etc.), (ii) relation to the moment of

Use	German	English
Präsens/present tense		
non-past	Ich schlafe von 12 bis 7.	I sleep from midnight to seven.
futurate	Morgen weiß ich das.	→ future tense (*I will know that tomorrow.*)
Präteritum/simple past		
past time	Ich schlief den ganzen Tag.	I slept the whole day.
Futur I/future tense		
future time	Ich werde schlafen.	I will sleep. I am going to sleep.
Perfekt/present perfect		
resultative	Jemand hat mein Auto gestohlen.	Someone has stolen my car.
existential	Ich habe (schon mal) Tennis gespielt.	I have played tennis.
hot news	Kanzler Schröder ist zurückgetreten.	Chancellor Schröder has resigned.
universal	→ Präsens (*Ich lebe hier seit 2 jahren.*)	I have lived here for two years.
narrative	Ich bin gestern im Theater gewesen.	→ past tense (*I was in theater yesterday.*)
Futur II/future perfect		
future results	Ich werde das bis morgen erledigt haben.	I will have done this by tomorrow.
Plusquamperfekt/past perfect		
pre-past	Ich hatte geschlafen.	I had slept.

Table 9.1.: Use of tenses in English and German.

utterance (*resultative, universal, narrative*, etc.). In other words, the parallelism or non-parellism of specific German and English tenses can be established with respect to a set of specific semantic properties of a verb and the utterance it occurs in.

Different aspects in the English tense system have an impact on the use of a specific tense. For instance, in contrast to the simple present tense, the English present progressive can be used in the futurative context: *'I am going out tonight'* (König and Gast, 2012, p. 95). This suggests that our data contains the tense translation pair *'present simple ↔ Futur I'* which may be used in almost all contexts, as well as the translation pair *'present progressive ↔ Futur I'* which on the other hand can be used in only very specific contexts.

Finally, there is a difference in the grammatical mood in English and German. The subjunctive mood in German is expressed in the verbal morphology and as such, it interacts with the German tense system. The German subjunctive moods (in this thesis, we treat them as a part of the German tense forms) have impact on the time of an utterance. Furthermore, they come along with specific usage contexts such as indirect speech (e.g., *'Er sagt, er sei/wäre krank' (He says that he is ill)*, non-factual and condi-

tional statements (e.g., *'Wenn ich Zeit hätte, würde ich kommen' (If I had time, I would come)*), as well as to signal politeness. For subjunctive mood in English, König and Gast (2012) rather use the term *quasi-subjunctive* since subjunctive mood in English exists only for the verb *be*. Other forms used in the subjunctive contexts correspond to the infinitives (e.g., *'I demand that he go there'*).

Preceding paragraphs show that, although English and German share a common ground of six tenses (present, simple past, present perfect, past perfect, future I and future II), they also expose many differing linguistic properties such as aspectual information, modality (i.e., existence of a modal verb), grammatical mood, etc. These properties lead to great differences between the tense systems in the two languages. Furthermore, combinations of the different linguistic properties of the verbs, as well as of the verbal complexes lead to a language-specific usage of tense forms which makes the mapping between the tense forms in English and German even more complicated. Ultimately, linguistic, as well as tense usage specifics result in a many-to-many relation regarding tense (mood and aspect) as illustrated in Figure 9.1. A formal description of the respective many-to-many relation requires knowledge on different linguistic levels: lexical, syntactic and semantic/pragmatic.

9.2. Influence of the domain/register and author

The characteristics of usage of specific tenses and moods are often used as linguistic indicators which point to a specific domain or register. Neumann (2013), for instance, presents a corpus-based study, in which the frequency of tense, among other textual properties, in different texts is used to induce the *goal type* of the text: argumentation, narration, instruction, etc. Her studies support the initial claim. For example, she observed that the frequency of the present vs. past tense across texts from different domains expose different (i.e., domain-specific) distributional specifics. One of her findings is that past tenses are rather typical for narrative texts, i.e., while the verbs in the present tense are more typical for argumentative texts such as political essays, popular science articles, etc.

These findings come along with a theoretical study presented by Weinrich (2001) which was briefly mentioned in Chapter 3. For German, he differentiates between two groups of the German tenses: (i) *discussing tenses* to which belong *Präsens*, *Perfekt*, *Futur I*, *Futur II* and (ii) *narrative tenses* containing *Präteritum*, *Plusquamperfekt*, as well as subjunctive tense forms *Konjunktiv I* and *Konjunktiv II*. The distribution of the

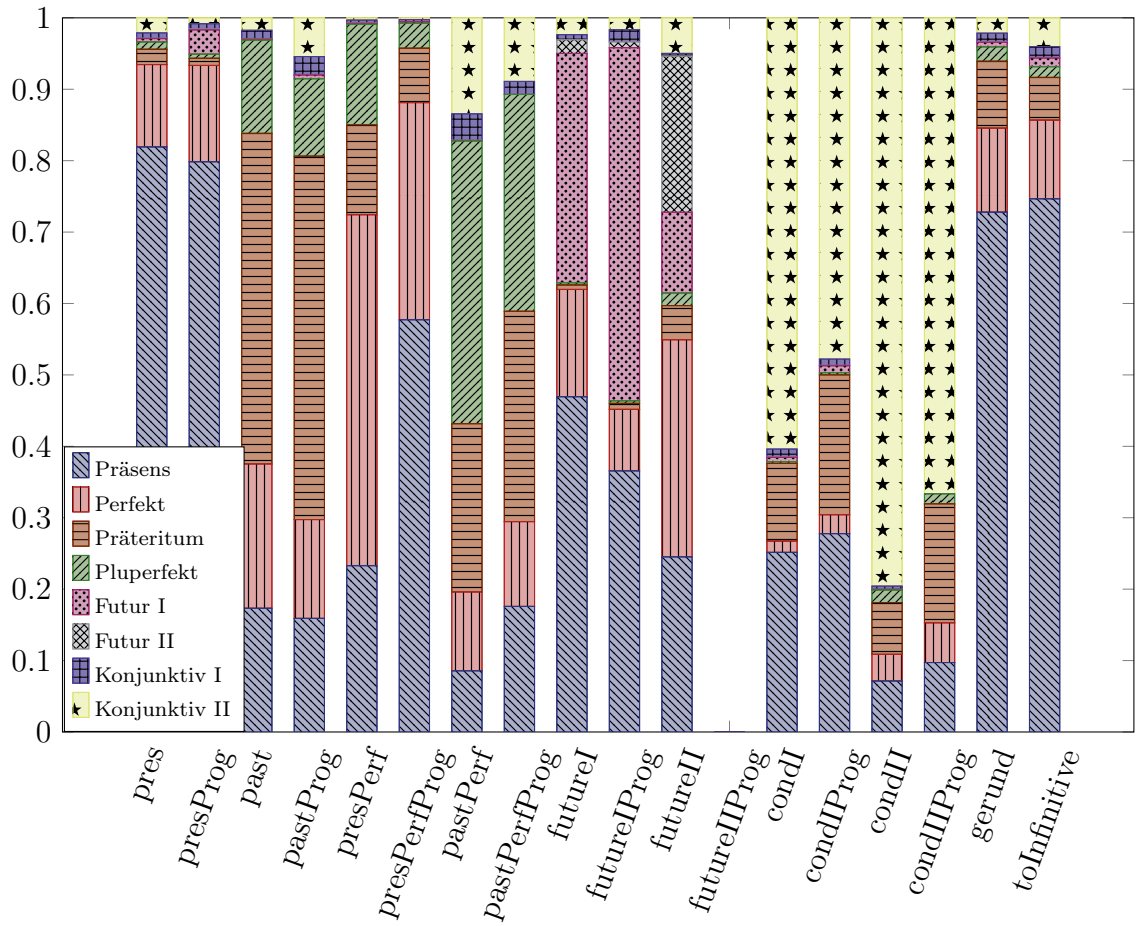


Figure 9.1.: Distribution of tense translations derived from the News, Europarl and Crawl corpus.

German tense forms in data that we used in the experiments presented in this thesis is shown in Figure 9.2. The picture is not as clear as the statement made by Weinrich (2001) might suggest. However, it must be noted that our corpora are not necessarily clean in a sense that each of them strictly belongs to the domains discussed by Weinrich (2001). Nevertheless, we do observe differences in the tense distribution. For instance, the present tense is most frequently used in Crawl. There is also a considerable difference in the frequency of *Präteritum*. Although, *Präteritum* is the 2nd most frequent tense form in all corpora, it has highest relative frequency in News, while its lowest relative frequency is found in Europarl (0.23 in News vs. 0.11 in Europarl, respectively). Also interesting is the frequency of *Konjunktiv II*. Its highest relative frequency is given in News (0.076), while the lowest relative frequency is found in the Crawl corpus (0.013).

Obviously the tense usage (reflected in the frequency of the different tense forms)

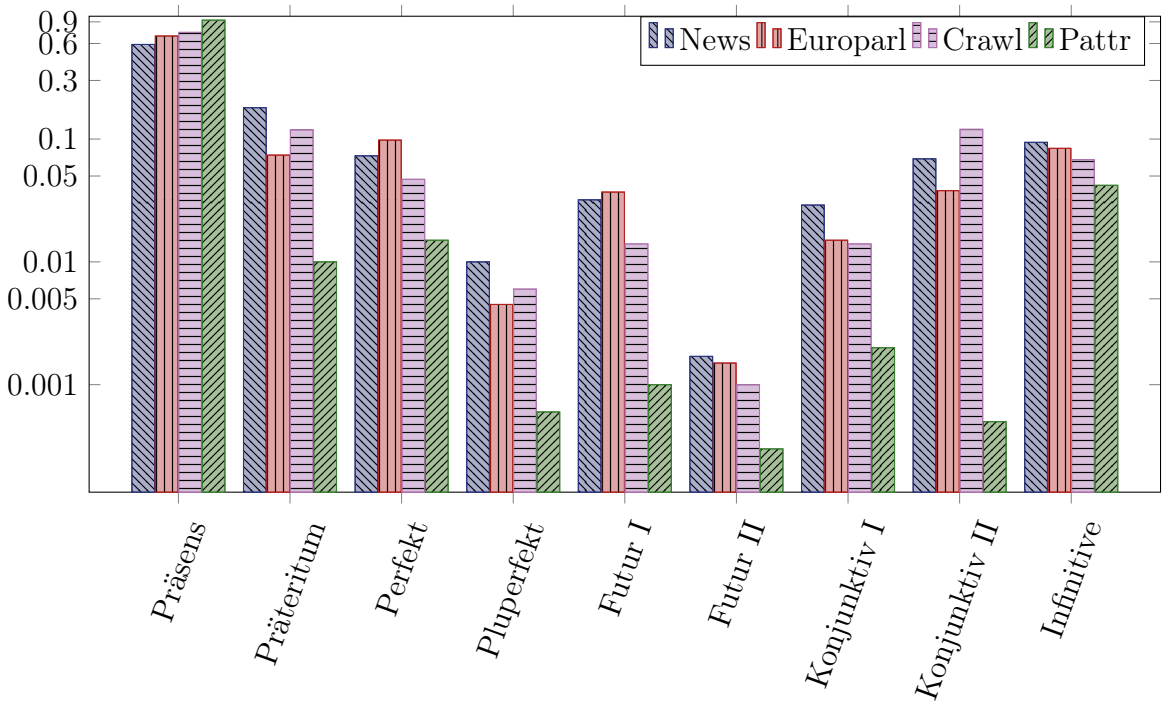


Figure 9.2.: Overall distribution of the active tense forms in the German corpora used throughout this thesis. In addition to tense forms, the graph also shows the proportion of the non-finite VCs found in the used corpora.

differs across domains. This fact has also impact on the translation of tense and mood. It raises the question whether it is possible to train a *universal* model for translating tense and mood. Domain specifics in the monolingual context as shown in Figure 9.2 have a direct impact on matching specific source language tenses onto their target language counterparts. Due to domain specifics, a model, for instance a classification model similar to that described in Section 6.3.4 of Chapter 7, would learn different possible tense/mood correspondences. The correct choice would then require access to the knowledge of the domain/register that both the test and the training instances belong to.

9.3. Context of (machine) translation

Corpus-based approaches to machine translations such as SMT and NMT require parallel texts from which the translation models are automatically learned. Typically, those texts are produced by humans, i.e., professional translators who translate texts from a source language (SL) into a target language (TL).

The translation process leads to TL texts which typically differ from the TL texts

which are originally written in the respective TL. König and Gast (2012) mention and empirically verify two aspects which are also important with respect to bilingual modeling of tense and mood: (i) *SL shining through* and (ii) *TL normalization*. The former aspect indicates the closeness of the translation to the source, while the latter one is related to the adaptation of the source text properties to the TL. König and Gast (2012) observed that *SL shining through* is less prominent when translating into a language which has fewer options with respect to a specific grammatical system. In such cases, the *TL normalization* is more prominently used.

König and Gast (2012) do not consider tense in their study, however their hypotheses can also be applied to this specific feature. *SL shining through* would indicate that the tenses in English are preserved in the German translations. Additionally, their observation that *SL shining through* is less prominent when translating into a language with fewer possibilities to express a specific linguistic information also indicates that our parallel texts may expose great variation in the tense translation. On the one hand, there is lots of parallelism with respect to tense, on the other hand, due to the smaller set of the German tenses, in many cases, the TL specific usage of tense can be found which may considerably differ from a form given in the source. We do not further study this hypothesis, however the statistics about the tense translation between English and German extracted from our corpora indeed indicate that both issues are important (cf., for instance, Tables 3.8 on page 57 and 3.9 and 58). On the one hand, each of the English tense forms is most frequently translated into the respective German tense. On the other hand, we observe many cases in which different German tense forms are used with a single English tense form which indicates a certain degree of adaptation of tense to the specifics of the target language, namely German.

The study presented by König and Gast (2012) explains the differences in the translations by looking at linguistic properties of the languages involved in the translation process. The differences may, however, also be explained by taking the factor *human translator* into account. The translation process is tightly connected with specific characteristics of the translator such as his/her understanding of the source text, proficiency in the target language, stylistic preferences with respect to diverse linguistic aspects, one of them also being tense. Large parallel text collections such as parallel corpora used throughout this thesis are a concatenation of the translations produced by many different translators and thus very probably reflect also differences in tense translations caused by preferences of a specific translator.

9.4. Evaluation issues

In our work, we automatically evaluate MT outputs for which we adapt tense and mood. Automatic evaluation is based on a comparison of our MT outputs with a single, human generated reference translation. Both of these aspects may lead to problems. A *single* reference translation may be inappropriate because of high degree of interchangeability in German concerning specific tenses. If our SMT model does not generate the tense form given in the reference, it gets punished in terms of the overall BLEU score. On the one hand, the SMT tense choice may indeed be false, on the other hand, it might also be correct due to tense interchangeability. Related to this issue is also the fact that our reference translations are written by human translators. We discussed in the preceding section that translators have their own preferences with respect to different linguistic phenomena including tense and mood. Particularly in the case of tense interchangeability, it might thus happen that specific SMT outputs perform better or worse in terms of automatic evaluation depending on the used reference translation set.

In Section 7.2, we presented a manual evaluation of our SMT outputs with modified tense and mood. The inter-annotator agreement in terms of Kappa was only 0.33 which means that the human annotators involved in the manual evaluation very often disagreed on which translation alternative regarding tense and mood was better. Here again, factors such as author/translator/evaluator preference play a large role. Furthermore, our annotators were shown the SMT outputs in isolation, i.e., without surrounding context which additionally complicates the decision whether a translation is correct or wrong. In Section 8.2.1 of Chapter 8, we have discussed cases in which the judgment of whether a specific tense is correct or wrong depends on the preceding context, i.e., information given in the preceding sentences.

9.5. Rule-based tense translation in EUROTRA

Statistical modeling of translation of tense (mood, voice and aspect) is mostly used in the context of SMT and is based on specific contextual information. In other words, we provide some specific information to a computational model and assume that the model is able to learn how the tenses are to be translated within a specific language pair. The opposite approach to this is a manual definition of tense translation as needed for rule-based MT systems. A deeper look into the respective rules provides information about the theory of tense translation in the context of a *human translation* and how that theory

can be formalized for (rule-based) MT. In the following paragraphs, we summarize the tense translation rules used by the rule-based MT system EUROTRA[1].

Tense/aspect form vs. tense/aspect meaning Tense translation in EUROTRA includes knowledge about three categories of a verbal complex: tense, aspect and Aktionsart.[2] There is a distinction between *tense/aspect (TA) forms* which denote grammatical tense and aspect categories expressed by verbal affixes, as well as auxiliaries, and a *TA meaning*. The translation of TA is a three-step process: (i) TA form in the source language is mapped to a universal (language-independent) context-specific TA meaning, (ii) source language TA meaning is mapped to a target language context-specific TA meaning and (iii) generation of the target language TA form. Thus, there is no direct translation of tenses between source and target language. Instead, the tense translation is modeled monolingually as mapping between the TA forms and an interlingual representation of the TA meanings.

This approach to tense translation is based on the assumption that there is no one-to-one mapping of the TA forms between the languages. A single TA form in one language can correspond to a number of appropriate TA forms in another language. The choice of the appropriate one often depends on the context which makes it impossible to enumerate all form correspondences.

Representation of TA meaning A sentence can be seen as a combination of a tenseless clause with some temporal information such as the tense form (*'had eaten': past perfect*), temporal adverbial (*yesterday*), temporal PP *(e.g., '3 hours ago'*), etc. For a correct interpretation of TA meaning, the *time of reference* R is needed since TA meanings are defined as relations between R (when did something happen?) and the time of speech S. There are four relations between R and S: anteriority, posteriority, identity and simultaneity (i.e., proper inclusion). These relations correspond to the concepts of past, present and future. In order to derive the relation between R and S, different groups of temporal modifiers are made use of:

- Locational (answer to *when*- questions):
  - simultaneous: *now*, ...
  - anterior: *yesterday, two weeks ago,* ...

[1] `http://www-sk.let.uu.nl/stt/eurotra1.htm`

[2] A detailed description of tense translation in EUROTRA is given in Durand et al. (1991).

Tense form	Tense class
Present	{simul, posterior}
Past	{anterior}
Future	{posterior}
Conditional	∅

Table 9.2.: Mapping of the English tense forms to tense classes. ∅ refers to no temporal meaning in isolated clauses.

Aspect form	Aspect class	Example
simple	{perfective}	*(I) do/did*
perfect	{retrospective, terminative}	*(I) have/had done*
progressive	{durative,perfective}	*(I) am/was doing*
perfect progressive	{terminative}	*(I) have/had been doing*

Table 9.3.: Mapping of the English aspect forms to the aspect classes.

  - posterior: *tomorrow, next summer, ...*

- Relational:
  - simultaneous: *at the same time, at that moment, ...*
  - anterior: *two weeks before, previously, ...*
  - posterior: *one week later, then, ...*
- Aspectual:
  - durational: *for three minutes* (process), *in three minutes* (event)
  - boundary: PPs with *since, from, until, till, from...till.*

Mapping between TA forms and TA meanings TA meanings are described in terms of a tense and aspect class, as well as of an Aktionsart. The sets of the values of the respective features are given in Tables 9.2 and 9.3.

Mapping of an aspect form to an aspect class involves knowledge about the Aktionasart of the sentence. Aktionsart concerns the temporal properties of the situation which is denoted by a basic tenseless clause: it specifies whether the time of an event E is bounded or not. Aspect, on the other hand, concerns the temporal relation between time of E and a specific R. Aktionsart properties of E may influence the relations E can have with R. In EUROTRA, Aktionsart can have the following values: *event*, *state* and *process*. Depending on the Aktionsart, the relations between aspect and Aktionsart are defined as follows:

1. Event: time of E is bounded and its relation to R is more likely to be:
   - perfective than durative,
   - retrospective than terminative,
   - prospective than inchoative,
2. State/process: time of E is unbounded and its relation to R is more likely to be:
   - durative than perfective,
   - terminative than retrospective,
   - inchoative than prospective.[3]

Computation of the Aktionsart takes into account the lexical properties of a main verb of the given clause, as well as semantic properties of its arguments. As for the main verb, the relevant property is whether the verb is stative (Aktionsart = state) or dynamic (Aktionsart = event or process). As for the arguments, the relevant property is whether the arguments are bounded (i.e., non-additive and non-homogeneous) or unbounded (i.e., additive and homogeneous). If all arguments of a dynamic verb are bounded, then the clause is bounded. If at least one of the arguments is unbounded, the entire clause is unbounded. (Un-)boundedness of an argument is derived by looking at, among others, lexical properties of a noun (mass vs. count), the number of a noun phrase (singular vs. plural) and semantic properties of the determiners/quantifiers of a noun phrase.

The resulting TA meaning is a composition of the meaning of the given tense form with the meaning of the given aspect form. Many of the TA forms can be mapped to more than one TA meaning. Temporal modifiers such as adverbials help to disambiguate between the different meanings. For instance, *'He is coming tomorrow'* has present tense with meaning {posterior,simul}, while the adverbial *tomorrow* has the meaning {posterior}. The union of {posterior, simul} and {posterior} leads to the resulting meaning {posterior}.

9.6. Discussion

Tense and mood are pragmatic features of texts. Their usage largely depends on the linguistic properties of the involved languages, register/domain specifics, author's/ translator's preferences, as well as on general characteristics of the human translation process.

[3]Definitions of the Aktionsart values are given in Durand et al. (1991). Here, we do not go into a deeper discussion of these categories.

König and Gast (2012)'s analysis of English and German tense systems reveals that the German and English grammatical tense systems differ not only in terms of the information they explicitly enclose, but also with respect to the time a specific tense form may refer to. As for the information, English, for instance, has an explicit marking of the progressive aspect, while in German this is not the case. On the other hand, German has explicit tense forms for the subjunctive mood, while these forms are mostly missing in English. In addition to these differences, the German and English tenses also expose different ability of being used with respect to a specific logical tense. For example, the German present tense can generally be used to express future actions while the respective usage of the English simple present tense is almost excluded. However, in a combination with the progressive aspect, the English present tense may indeed be used to refer to future, at least in a combination with appropriate temporal modifiers. In addition to these rather semantic meanings of the tense forms, in English, there are, at least in the context of the conditionals, clearly defined sequences of tense forms such as 'simple past tense + conditional I' in *'If I had time, I would come'*. Such grammatical rules may lead to highly context dependent, sometimes rather non-intuitive, tense/mood translations between English and German where the English indicative verb tenses translate into a German subjunctive tense form.

Not only linguistic properties of a language play a role when it comes to the use of tense and mood. The usage greatly depends also on the register and domain. In fact, the respective distributional differences are often used to (automatically) derive the domain or the purpose of a text as described by Neumann (2013). A direct comparison of tense distribution in the texts used in this work (i.e., news wire, political or medical domain) supports this hypothesis. The domain specific tense usage and distribution, respectively, are tightly related with the classification of tenses into narrative and discussing as proposed by Weinrich (2001). For instance, *Präteritum* is considered to be a narrative tense, while *Perfekt* is considered a discussing tense. Assuming that news articles are quite narrative, we would expect that *Präteritum* is more dominant than *Perfekt*. As shown in Figure 9.2 (page 194), *Präteritum* is indeed most frequently used in the news corpus compared with other corpora. On the other hand, Europarl which is a collection of political discussions prefers *Perfekt* rather than *Präteritum*.

The domain dependency raises an interesting theoretical question about automatic modeling of tense and mood in the context of translation. The differences between domains and registers may suggest that models, for instance classification-based models, for tense and mood translation need to be trained on a domain which matches the domain

of the texts for which tense and mood translation modeling is of interest. One might also consider the possibility to directly include this information into the training data, so that the model may learn tense/mood translations along with the information about the given domain.

The main topic of the present thesis is *machine translation*. The statistical (as well as neural) approach to MT make use of parallel texts which have been created by human translators. Therefore, it makes sense to discuss the process of the *human translation* when looking for explanations of how pragmatic information such as tense (i.e., time) and mood is transferred from a source into a specific target language. We mentioned two interesting observations regarding the process of human translation. On the one hand, the specifics of the source language are often kept in the target language, on the other hand, the translator may also perform the so-called *TL normalization* which may lead to rather non-literal translations which, however, obey the grammar rules or other characteristics of the target language. In terms of tense and mood translation, we thus draw the following conclusion. Modeling of tense/mood translation needs to consider both bilingual tense/mood correspondences, as well as monolingual characteristics of the target language. From the theoretical point of view, it might be interesting to examine the degree of the two above mentioned translation process factors and its variability related to the factor *human* translator.

A very concrete idea of how the tense translation can be described in a formal way may be obtained by looking at the tense translation in the context of rule-based MT. Tense and mood translation in EUROTRA, for instance, relies on an interlingua representation of tense to which the source sentence is mapped, and which is mapped to the syntax of the target language respectively. The mappings to and from the tense/aspect meaning is rule based. It follows a set of manually defined rules which make use of different kinds of information such as verbs and their syntactic/semantic properties, existence of temporal expressions, as well as their subparts, arguments such as subjects and objects along with specific semantic information about them. The rules also show that tense cannot be considered in isolation, but rather in a combination with other related linguistic features such as aspect and Aktionsart. Although not explicitly mentioned in the preceding discussion, tenses comes also along with specific modality, as well as voice properties which makes the problem even more complex.

In the following, we map different textual characteristics discussed in the preceding sections to the lexical/syntactic level in order to point to the information directly accessible from a sentence which could (or should) be used to develop tense translation models.

Text property	Lexical/syntactic level	Tool availability
Tense	VC	POS tagging and parse trees
	main verb	POS tagging + parse trees
	tense, mood, voice	TMVannotator (Ramm et al., 2017a)
	temporal expressions (NPs and PPs):	TARSQI (Verhagen et al., 2005)
	head noun	
	preposition	POS tagging + parse trees
	adjective	
	adverb	
	temporal ordering	TARSQI (Verhagen et al., 2005)
Aspect	auxiliary (combination)	POS tagging and parse trees + mapping rules
Aktionsart	event/state/progress	sitent (Friedrich and Palmer, 2016)
	subject NP:	parse trees
	determiner	semantic properties
	quantifier	POS tagging
	number	WordNet
	mass	
	count	
Domain/ genre		-
Reported speech		QSample (Scheible et al., 2016)
Conditional clauses		-

Table 9.4.: Mapping of the different textual properties to the corresponding lexical/syntactic levels. Column *Tool availability* lists tools for automatic annotation of the English texts with the respective information.

The respective contextual features are summarized in Table 9.4. Many of the features can be derived from parsed and POS tagged data, however, some of them require access to other annotation tools, as well as lexical databases which include semantic properties of the English words (e.g., WordNet). Automatic annotation of the temporal ordering, for instance, can be done with the tool TARSQI (Verhagen et al., 2005). Information about tense, mood and voice of the VCs in the English texts can be obtained with the TMVannotator (Ramm et al., 2017a) which has been developed within the present work. Information about Aktionsart in terms of state, event and progress can be gained from the output of the tool sitent (Friedrich and Palmer, 2016).

As of January 2018, there are no publicly available tools for automatic annotation of

texts with genre and/or domain information, although there has been ongoing research in this area, see for instance (Santini, 2007), (Sharoff et al., 2010) and (Biber and Egbert, 2016).

Automatic identification of conditionals in English is important for the translation into the German subjunctive mood. Similarly to genre annotation, no tools are publicly available which perform this task, however the set of syntactic rules described in Olivas et al. (2005) can be re-used to easily identify the respective contexts in English.

The summary of the features and tools given in Table 9.4 reveals two important facts about tense and mood with respect to the (classification-based) modeling of their translation. Textual properties that need to be considered represent on their own subtasks of the natural language processing. Tools developed to annotate the respective information are mostly based on classification models which use many different, sub-task related, information. In many cases, the predicted annotations are correct. However they may also be erroneous which might have negative impact on using those annotations to train a tense/mood classification model. Instead of using outputs of many different tools (which would require a quite complex processing pipeline) one might train a tense/mood classifier directly on the features which are used to train models for predicting each of the relevant textual properties. The future attempts to model tense and mood translation via a classification-based method need to carefully choose the training data. The distribution of the tenses in the German data is highly imbalanced which poses a big problem for training a well-performing classification model.

9.7. Chapter summary

In this Chapter, several aspects related to the automatic modeling of tense and mood translation were discussed. In Section 9.1, we gave a brief comparison of the English and German tense forms with respect to the ability of using them in a combination with a specific logical tense (i.e., present, past, future) originally elaborated by König and Gast (2012). In Section 9.2, we have discussed the influence of the domain, register and author on the tense use in terms of the frequency distributions within a specific text or even part of a text. In Section 9.3, we discussed two interesting facts regarding the process of the human translation. On the one hand, SL characteristics may be transferred to the TL resulting in rather literal translation, i.e., literal tense/mood translation. On the other hand, a translator may prefer to generate translations which are more TL specific. Both of the facts lead to diverse translation pairs with respect to tense and mood as indicated

by the frequency distribution of the tense pairs in our corpora illustrated in Figure 9.2 on page 194. Factors such as domain/register characteristics and author/translator preferences are also problematic for the automatic evaluation of tense and mood as discussed in Section 9.4. In addition, correct, or rather incorrect tense/mood forms can sometimes be identified as such only by looking at the preceding context, which is not yet accessible to MT systems. In Section 9.5, we examined the tense translation rules developed for the rule-based MT system EUROTRA. The analysis revealed that tense should not be considered as an isolated phenomenon, but in a combination with many different contextual properties such as existence of temporal expressions, semantic characteristics of the clausal arguments, Aktionsart, etc. In Section 9.6, we presented an attempt to map these properties to the lexical/semantic level and gave an overview of the tools which automatically provide the respective information.

10. Conclusion

The statistical approach to machine translation is a powerful device to obtain automatic translation of large amounts of texts in a small amount of time. Furthermore, statistical machine translation (SMT) is language-independent and can thus be applied on arbitrary language pairs. However, the performance of the SMT models depends, among others, on the closeness of the languages between the translation is being performed. The closeness refers to the syntactic structure of the languages, as well as to their morphological richness. The smaller the differences within two languages regarding these aspects, the better the SMT outputs.

This work deals with SMT for the English→German translation direction. English belongs to the group of morphologically poor languages, while German is a morphologically rich language which means that a single word in English corresponds to many different German word forms. Furthermore, while English is a SVO language with rather strict word/constituent order within a sentence, German is both SOV and SOV language depending on the type of a given clause. Additionally, German exposes a certain degree of the word/constituent freedom. The outlined morphological, as well as syntactic differences lead to specific verb-related errors in the German outputs: (i) due to syntactic differences, the verbs are either misplaced or not generated at all, and (ii) the inflection of the verbs is often erroneous. Both of these problems are handled within this work.

The main idea of the proposed approaches is to eliminate syntactic differences between English and German and to remove inflectional variants in the German data which are then added to the German translations in a post-processing step which includes explicit morphology generation. In the following, we take a closer look to both approaches. We discuss their drawbacks and present ideas for the future work. Furthermore, we examine the usefulness of the presented methods for the new state-of-the-art approach to machine translation, namely neural machine translation (NMT).

10.1. Preordering

10.1.1. Preordering characteristics

Preordering is a well-known method for dealing with syntactic differences within the framework of SMT. Is has been proven to work well for many different language pairs. However, the translation direction English→German has not successfully been handled before. In Chapter 4, we present our adaptation of preordering for English→German. We define several reordering rules which are applied on the English constituency parse trees and which aim at placing the English verbs into the German-specific positions. By establishing the syntactic parallelism between English and German, we help SMT in a sense that the SMT model needs not to cope for problematic long-range reorderings needed to ensure correct placement of the verbs in the German translations. As numerous experiments presented in Chapter 5 show, preordering is very effective for handling verb-related problems for English→German SMT. Compared to the baseline SMT systems, our preordered SMT models generate German translations with considerably more verbs which are additionally more often correctly placed compared to the baseline SMT translations.

Our implementation of preordering is deterministic and is based on parse trees. Furthermore, the reordering rules were developed manually through a thorough study of the English and German syntax presented in Chapter 3. Each of these characteristics may be seen both as an advantage as well as a drawback of the presented implementation of preordering which we discuss in the following paragraphs.

Deterministic preordering Previous research on preordering did not only introduce deterministic preordering, but also non-deterministic alternatives. While deterministic preordering generates a single reordered variant of a given source sentence, the non-deterministic preordering may generate several reordered alternatives. These are then usually encoded within a word lattice and monotonically translated. One of the biggest advantages of non-deterministic preordering is the fact that it can cope with word order freedom as it considers different possible placement of a specific sentence constituent. Deterministic preordering, on the other hand, requires one choice for the reordered position of the specific sentence constituents which is not necessarily the only grammatically correct position as discussed in Section 3.3.3.

This work aims at correcting generation and placement errors regarding the verbs in German SMT outputs. We thus deal with a single word category for which we

are able to identify unique positions within the English/German sentence pairs. The exceptional positions are not only rare, but also grammatically correct only in very specific contexts which are hard to identify, even for the non-deterministic preordering. Other syntactic differences between English and German may be seen as local differences which are in most cases handled well within SMT. Another reason for using deterministic preordering is that we can control the preordering process and adapt the rule set if required. Particularly in the commercial use of preordering, it is desirable to be able to track and handle errors in the entire training/translation process, also those related to preordering.

Parse trees Preordering may not only be defined on the basis of parse trees, but also on the level of POS tags or even on the surface data which does not include any kind of linguistic abstraction. The problem with preordering based on linguistically processed data is error propagation. Whether POS tagging or parsing, both of these sentence representations may contain errors which have a direct impact on the quality of the reordering and thus of the final SMT outputs.

If we consider the type of the syntactic differences between English and German related to the verbs, we come to the conclusion that English parse trees is the most appropriate representation to perform the required reorderings. Flat structures such as POS tags or surface words cannot provide unique information which we need to put English verbs into the German-like positions. An attempt with POS-based preordering for English→German has been done by Niehues and Kolss (2009). Their experiments showed that POS-based preordering is not able to improve German SMT outputs.

In our implementation of preordering, we use English parse trees keeping in mind that the trees might be erroneous. Indeed, we have also observed errors in the German SMT outputs caused by error propagation. Nevertheless, the trees come along with parsing accuracy which is sufficient to significantly improve the German SMT outputs.

Manually developed rules Non-deterministic preordering is typically based on a huge reordering rule set which is automatically learned from parallel texts. Deterministic preordering, on the other side, is based on a rather small set of the reordering rules which are manually formulated by taking into account specific linguistic characteristics of the given language pair.

Nakagawa (2015) presented a fast implementation of a language-independent non-deterministic preordering method and tested it successfully on many different language

pairs (note that English→German was not among the considered language pairs). The main strength of his approach is the language-independent automatic extraction of the reordering rules which is particularly interesting in the multilingual context since the rule induction does not require previous linguistic analysis of the language pairs under consideration. The present work is focused on handling of verb errors in the English→German SMT. We showed that the position of the verbs in German is usually not flexible which was the main reason for implementing a deterministic preordering based on a small set of hand-crafted preordering rules.

A difficulty with non-deterministic, automatically conducted reordering rule sets is tuning of the SMT models. While SMT (i.e., Moses which is used in this work) supports lattice-based training and decoding, it does not support lattice-based tuning. Thus, for tuning, we need to choose between the reordering alternatives. Daiber et al. (2016), for instance, proposed a method to acquire the reordered data for the tuning set in the context of non-deterministic preordering. They applied the method also to the English→German translation direction. They did report on the improved German SMT outputs, however, in terms of BLEU, their improvements were lower than those gained with our preordering approach.

10.1.2. Preordering for NMT

Previous studies on the quality of the German NMT outputs revealed that NMT makes almost no errors regarding the verb placement (Bentivogli et al., 2016), (Popović, 2017). Our own inspection of the German NMT translations generated by the best NMT model submitted to the WMT 2017 news translation task showed that NMT is indeed able to correctly translate almost all types of the English verbal complexes, as well as to put almost all German verbs into the correct positions. Even large positional differences given in our evaluation data set do not pose a problem for NMT.

Yet, we observed that the error rate increases with increasing sentence length. In long sentences with more than 50 words, which are often a sequence of clauses, it occurs occasionally that the German verbs are either placed incorrectly or not generated at all – problems which we successfully handled in the context of SMT. Thus, the question was raised whether preordering may also help NMT to avoid such errors.

We applied our method to English→German NMT and observed that preordering actually hurts the NMT performance. Similar finding were also reported by Du and Way (2017) for other language pairs. We assume that especially erroneous reordering brings noise into the training data and thus into the NMT model which obviously lowers

their quality. Since we observed that NMT only makes mistakes in ordering for very long sentences, a question raises itself as to whether preordering restricted only to very long sentences might be more appropriate than preordering all sentences, even those which are not problematic for NMT. The benefit of such a restricted reordering is however not ensured: on the one hand, long sentences are also problematic for parsing which is the basis for applying our reordering rules. On the other hand, mixing reordered and non-reordered English sentences might be confusing for NMT and ultimately lead to the same negative results gained for non-restricted preordering of the source language data. It remains for future work to examine these considerations and also to estimate as to whether the time-consuming pre-processing of the data leads to improvements which justify the effort of preordering the data.

10.2. Inflection generation

Our verb inflection generation method is incorporated into an already existing approach for nominal inflection generation for English→German SMT proposed by Fraser et al. (2012) and further refined by Weller et al. (2013). By means of classification, we aim at correcting the inflection of the finite verbs in the German SMT outputs. The German verbal morphology includes morphological information about person, number, tense and mood. While person and number are constrained by the corresponding subject phrases, tense and mood relate to semantic and pragmatic aspects which are often not overtly expressed: a fact that makes the prediction of tense and mood in the bilingual context, as we discussed in Chapter 6, quite problematic. Even the prediction of the agreement features proved problematic due to syntactic divergences in English and German, as well as free translations which make the extraction of the features for training the agreement classification model hard.

As for the agreement, we therefore decided to implement a parsing-based agreement correction similar to that proposed by Rosa et al. (2012) for English→Czech SMT. Our experiments showed that this is an efficient method to correct agreement errors although parsing of an often not well-formed German SMT output may lead to many errors regarding the syntactic analysis of the German sentences which have a negative impact on the agreement correction task.

With respect to the prediction of tense and mood, we experimented with many different contextual features. According to the definition of Li et al. (2007), almost all our features are surface features. However, not only Li et al. (2007) observed that the so-

called latent features such as temporal ordering information are more appropriate for the task of tense/mood prediction. In the recent studies on English→French SMT, Meyer et al. (2013) and Loáiciga et al. (2014) showed that more sophisticated features such as narrativity indeed lead to better French translations regarding the choice of tense.

The respective works on English→French integrated the tense-related features into an SMT model via factors. Another previous successful attempt to model tense translation for Chinese→English carried out by Gong et al. (2012b) incorporated the tense-related knowledge as an additional model into the SMT. All of these works as well as our experiments with factored SMT which include tense/mood factors raise the question whether it makes more sense to allow SMT to handle tense and mood translation within the training/decoding steps in contrast with our two-step pipeline which separates the translation from the subsequent tense/mood prediction. Indeed, our rather unsatisfactory results may be seen as a proof for this hypothesis. Negative results for the two-step procedure similar to that applied for English→German have also been reported by Gispert and Mariño (2008) for English→Spanish SMT.

Our explicit handling (i.e., prediction) of tense and mood for the verbs in the German translations assumes that we not only have theoretical knowledge of tense and mood translation between languages, but also that we have access to that knowledge. Both of these assumptions are problematic since tense and mood translation often follows regularities which are beyond the words in the texts as shown in Chapter 9. For instance, author preference, genre specifics, as well as human-defined rules may play a role for how exactly a specific tense in the source language is translated into the given target language. Furthermore, characteristics of using specific tense forms in the target language, in our case in German, may lead to unexpected tense translations which are difficult to cope with via a classification-based post-processing step.

In addition to the difficulty of identifying features relevant for the prediction of tense and mood, it is also an open question which classification method is most appropriate for prediction of tense and mood in the context of translation. For instance, the work described by Tajiri et al. (2012) where tense is modeled in the monolingual context. The authors compared three different classification methods, namely SVM, maxent and CRF, for predicting tense in the English texts. They found that CRF performs best indicating that tense prediction is a sequential problem in which previously made decisions need to be taken into account. On the other hand, experiments done by Ye et al. (2006) lead to the conclusion that *"sequential dependencies between tenses of adjacent verbs in the discourse may be rather week"*. Although the literature, as well as our intuition

are in favor of findings reported by Tajiri et al. (2012), apparently, in the bilingual context, tense sequences are not significant. One of the possible explanations for this fact is that tense usage in the target language in the context of translation does not necessarily follow the tense usage rules in the respective language in the monolingual context. It thus remains for the future work to bring more insights into the problem of translation of tense and mood, primarily related to the process of the human translation which can then be used to identify more reliable context features to model the problem automatically.

Verbal inflection for NMT A study of integrating linguistic knowledge into NMT for English→German presented by Tamchyna et al. (2017) has already shown that NMT profits from access to linguistic knowledge. This study also included morphological information about the verbs. Each verb (as well as other word categories) is represented as a sequence of a verb stem and a morphology markup. For example, the verb *denkst (you think)* is represented as follows: *'denken +V.2.Sg.Pres.Ind'*. Similarly to our morphology generation framework, the German NMT outputs are generated in a post-processing step by considering the generated stem+morphology pairs. Obviously, this approach does not aim at adapting the tense/mood in the German translation, but rather to simplify morphology of the German verbs, among others, and so to ensure that the finite verbs are correctly inflected.

Adding information about the verbs into the texts used to train SMT models may also be used in another scenario. For instance, Sennrich et al. (2016a) included tags into the source language data which indicate contexts in which the polite form of address is to be used. In German, polite form of address requires a specific pronoun, as well as a specific verb form (3rd person plural). Such kind of controlling the NMT output with respect to a specific characteristic might also be combined with tense/mood. An interesting case is indirect speech. While, for example, in the German news articles, very often *Konjunktiv* forms are used to indicate indirect speech, in other, less formal text sorts, one might want to have indicative forms in the same context. Annotating the source texts with appropriate labels could be seen as adaptation of tense/mood translation to the genre in the context of NMT.

In Chapter 8, we discussed examples for which it was difficult even for a human to derive the appropriate tense and mood for a given sentence because some of the crucial information came from the outside of the given sentence. Not only the NMT models which have access to the respective context (Tiedemann and Scherrer (2017), for

instance, propose a method for NMT using segments beyond a single translation unit (i.e., sentence)) might profit in terms of a more correct choice of tense and mood, but also a classification-based annotation of the tense/mood labels might be more accurate if it considered specific inter-sentential dependencies.

Bibliography

Avramidis, E. and Koehn, P. (2008). Enriching Morphologically Poor Languages for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL-HLT)*, Columbus, Ohio.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Banff, Canada.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, Austin, Texas, USA.

Biber, D. and Egbert, J. (2016). Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:3–36.

Björkelund, A. and Nivre, J. (2015). Non-Deterministic Oracles for Unrestricted Non-Projective Transition-Based Dependency Parsing. In *Proceedings of the 14th International Conference on Parsing Technologies*, Bilbao, Spain.

Bohnet, B. and Nivre, J. (2012). A Transition-Based System for Joint Part-of-Speech Tagging and Labeled Non-Projective Dependency Parsing. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jeju, Korea.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT)*, Baltimore, USA.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 Conference on Machine Translation (WMT). In *Proceedings of the Second Conference on Machine Translation,*

Volume 2: Shared Task Papers, Copenhagen, Denmark.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal.

Bojar, O. and Kos, K. (2010). Failures in English-Czech Phrase-based MT. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR (WMT)*, Uppsala, Sweden.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J., Mercer, R. L., and Rossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Lafferty, J. D., and Mercer, R. L. (1992). Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Transaltion (TMI)*, Montréal, Canada.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Comput. Linguist.*, 19(2):263–311.

Cap, F., Weller, M., Ramm, A., and Fraser, A. (2014). CimS - The CIS and IMS joint submission to WMT 2014: translating from English into German. In *Proceedings of the 9th Workshop on the Statistical Machine Translation (WMT)*, Baltimore, Maryland, USA.

Cap, F., Weller, M., Ramm, A., and Fraser, A. (2015). CimS - The CIS and IMS Joint Submission to WMT 2015 addressing morphological and syntactic differences in English to German SMT. In *Proceedings of the 10th Workshop on the Statistical Machine Translation (WMT)*, Lisbon, Portugal.

Cer, D. M., de Marneffe, M.-C., Jurafsky, D., and Manning, C. D. (2010). Parsing to Stanford Dependencies: Trade-offs between Speed and Accuracy. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, Malta.

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

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–

Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Collins, M., Koehn, P., and Kučerová, I. (2005). Clause Restructuring for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan.

Collins, P. and Hollo, C. (2010). *English grammar. An introduction.* Palgrave macmillan.

Costa-jussà, M. R. and Fonollosa, J. A. R. (2006). Statistical Machine Reordering. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Csipak, E. (2015). *Free factive subjunctives in German.* Doctoral thesis. Niedersächsische Staats- und Universitätsbibliothek Göttingen.

Curme, G. O. (1964). *A grammar of the German language.* Frederick Ungar publishing co., 2 edition.

Daiber, J., Stanojevic, M. S., Aziz, W. A., and Khalil, S. (2016). Examining the Relationship between Preordering and Word Order Freedom in Machine Translation. In *Proceedings of the 1st Conference on Machine Translation. Volume 1: Research Papers (WMT)*, Berlin, Germany.

Ding, C., Utiyama, M., and Sumita, E. (2015). Improving *fast align* by Reordering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Lisbon, Portugal.

Drach, E. (1963). *Grundgedanken der deutschen Satzlehre.* Wissenschaftliche Buchgesellschaft Darmstadt, 4 edition.

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

Durand, J., Bennett, P., Allegranza, V., van Eynde, F., Humphreys, L., Schmidt, P., and Steiner, E. (1991). The eurotra linguistic specifications: An overview. *Special Issue of Machine Translation on Eurotra*, 6(2).

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of the the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, USA.

Dürscheid, C. (2012). *Syntax. Grundlagen und Theorien.* Vandenhoeck & Ruprecht, 6 edition.

Eckart, K. and Seeker, W. (2013). Task-based Parser Output Combination. In *Pro-*

ceedings of the 25th European Summer School in Logic, Language, and Information (ESSLI), Workshop on Extrinsic Parse Improvement (EPI). Düsseldorf, Germany.

Eisenberg, P., editor (1998). *Duden. Grammatik der deutschen Gegenwartssprache*, volume 4. Dudenverlag, 6 edition.

Elming, J. (2008). Syntactic Reordering Integrated with Phrase-based SMT. In *Proceedings of the Second Workshop on Syntax and Structure in Statistical Translation (SSST)*, Columbus, Ohio.

Formiga, L., Hernández, A., Mariño, J. B., and Monte, E. (2012). Improving English to Spanish Out-of-Domain Translations by Morphology Generalization and Generation. In *Proceedings of the Monolingual Machine Translation – 2012 Workshop – AMTA 2012*, San Diego, USA.

Fraser, A., Weller, M., Cahill, A., and Cap, F. (2012). Modeling Inflection and Word-Formation in SMT. In *Proceedings of the the European Chapter of the Association for Computational Linguistics (EACL)*, Avignon, France.

Friedrich, A. and Palmer, A. (2016). Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Galley, M. and Manning, C. D. (2008). A Simple and Effective Hierarchical Phrase Reordering Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Honolulu, Hawaii.

Genzel, D. (2010). Automatically Learning Source-side Reordering Rules for Large Scale Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China.

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

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

Gong, Z., Zhang, M., Tan, C., and Zhou, G. (2012a). Classifier-based tense model for SMT. In *Poster at the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India.

Gong, Z., Zhang, M., Tan, C., and Zhou, G. (2012b). N-gram-based tense models for statistical machine translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, Korea.

Grewendorf, G. (1982). Zur Pragmatik der Tempora im Deutschen. *Deutsche Sprache*, 10:213–236.

Habash, N. and Sadat, F. (2006). Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York, USA.

Hansen-Schirra, S., Neumann, S., and Steiner, E. (2012). *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German.* Text, translation, computational processing. De Gruyter Mouton.

Heafield, K., Pouzyrevsky, I., Clark, J. H., and Koehn, P. (2013). Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Hentschel, E. H. and Vogel, P., editors (2009). *Deutsche Morphologie: (De Gruyter Lexikon).* de Gruyter, Berlin.

Howlett, S. and Dras, M. (2011). Clause Restructuring for SMT Not Absolutely Helpful. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT): Short Papers - Volume 2*, Portland, Oregon, USA.

Johnson, J. H. J., Martin, J., Foster, G. F., and Kuhn, R. (2007). Improving translation quality by discarding most of the phrase table. In *In Proceedings of the joint conference on Empirical Methods in Natural Language Processing and conference on Computational Natural Language Learning (EMNLP-CoNLL)*, Prague, Czech Republic.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diego, USA.

Klein, D. and Manning, C. D. (2003). Accurate Unlexicalized Parsing. In *41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints.*

Koehn, P. (2004). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, Washington, DC, USA.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *The tenth Machine Translation Summit*, Phuket Island, Thailand.

Koehn, P. (2010). *Statistical machine translation.* Cambridge University Press.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *IWSLT*, Pittsburgh, PA, USA.

Koehn, P. and Hoang, H. (2007). Factored translation models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Prague, Czech Republic.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL) - Volume 1*, Edmonton, Canada.

Koo, T., Carreras, X., and Collins, M. (2008). Simple Semi-supervised Dependency Parsing. In *Proceedings of Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics and the Human Language Technology Conference (ACL-HLT)*, Columbus, Ohio.

König, E. and Gast, V. (2012). *Understanding English-German constrasts*. Number 29 in Grundlagen der Anglistik und Amerikanistik. Erich Schmidt Verlag.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Factored translation models. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, Williamstown, USA.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, Uppsala, Sweden.

Lee, J. (2011). Verb tense generation. *Procedia. Social and Behavioral Sciences*, 27:122–130.

Lee, Y.-S., Zhao, B., and Luo, X. (2010). Constituent Reordering and Syntax Models for English-to-Japanese Statistical Machine Translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*.

Li, C.-h., Zhang, D., Li, M., Zhou, M., Li, M., and Guan, Y. (2007). A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*

ACL).

Loáiciga, S., Meyer, T., and Popescu-Belis, A. (2014). English-French Verb Phrase Alignment in Europarl for Tense Translation Modeling. In *Proceedings of the The 9th Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Luong, M.-T. and Manning, C. D. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, Da Nang, Vietnam.

Mareček, D., Rosa, R., Galuščáková, P., and Bojar, O. (2011). Two-step Translation with Grammatical Post-processing. In *Proceedings of the 6th Workshop on Statistical Machine Translation (WMT)*, Edinburgh, Scotland.

Meyer, T., Grisot, C., and Popescu-Belis, A. (2013). Detecting Narrativity to Improve English to French Translation of Simple Past Verbs. In *Proceedings of the 1st DiscoMT Workshop at 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Minkov, E., Toutanova, K., and Suzuki, H. (2007). Generating Complex Morphology for Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL)*, Prague, Czech Republic.

Nakagawa, T. (2015). Efficient Top-Down BTG Parsing for Machine Translation Preordering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, Beijing, China.

Neubig, G., Watanabe, T., and Mori, S. (2012). Inducing a Discriminative Parser to Optimize Machine Translation Reordering. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Jeju Island, Korea.

Neumann, S. (2013). *Contrastive register variation. A quantitative approach to the comparison of English and German.* Trends in Linguistics. Studies and Monographs. De Gruyter Mouton.

Niehues, J. and Kolss, M. (2009). A POS-based Model for Long-range Reorderings in SMT. In *Proceedings of the Fourth Workshop on Statistical Machine Translation (WMT)*, Athens, Greece.

Nießen, S. and Ney, H. (2000). Improving SMT Quality with Morpho-syntactic Analysis. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany.

Nießen, S., Ney, H., and Vi, L. F. I. (2001). Morpho-Syntactic Analysis for Reordering

in Statistical Machine Translation. In *Proceedings of MT Summit VIII*, Santiago de Compostela, Spain.

Och, F. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics ACL (ACL)*, Sapporo, Japan.

Och, F. J. and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.

Olivas, J. A., Puente, C., and Tejado, A. (2005). Searching for causal relations in text documents for ontological application. In *Proceedings of the 2005 International Conference on Artificial Intelligence (ICAI)*, Las Vegas, Nevada, USA.

Palmer, F. (1986). *Mood and Modality*. Cambridge University Press.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, Philadelphia, Pennsylvania.

Plitt, M. and Masselot, F. (2010). A productivity test of statistical machine translation postediting in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16.

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

Ramm, A. and Fraser, A. M. (2016). Modeling verbal inflection for English to German SMT. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers (WMT)*, Berlin, Germany.

Ramm, A., Loáiciga, S., Friedrich, A., and Fraser, A. (2017a). Annotating tense, mood and voice for English, French and German. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (ACL), system demonstrations*, Vancouver, Canada.

Ramm, A., Superbo, R., Shterionov, D., O'Dowd, T., and Fraser, A. (2017b). Integration of a Multilingual Preordering Component into a Commercial SMT Platform. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation (EAMT)*, Prague, Czech Republic.

Rosa, R., Mareček, D., and Dušek, O. (2012). DEPFIX: A System for Automatic Correction of Czech MT Outpus. In *Proceedings of the Seventh Workshop on Statistical*

Machine Translation (WMT), Montréal, Canada.

Sammon, G. (2002). *Exploring English grammar*. Cornelson Verlag.

Santini, M. (2007). *Automatic Identification of Genre in Web Pages*. Doctoral thesis. University of Brighton.

Scheible, C., Klinger, R., and Padó, S. (2016). Model Architectures for Quotation Detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

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

Schwenk, H. and Koehn, P. (2008). Large and diverse language models for statistical machine translation. In *In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNP)*, Hyderabad, India.

Sennrich, R., Birch, A., Currey, A., Germann, U., Haddow, B., Heafield, K., Miceli Barone, A. V., and Williams, P. (2017). The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, Copenhagen, Denmark.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NACL-HLT)*, San Diego, California.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*.

Sennrich, R., Haddow, B., and Birch, A. (2016c). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, Berlin, Germany.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web Library of Babel: evaluating genre collections. In *Proceedings of the European Language Resources Association (LREC)*, Malta.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to Sequence Learning with

Neural Networks. In *Advances in Neural Information Processing Systems 27 (NIPS)*, Montreal, Canada.

Tajiri, T., Komachi, M., and Matsumoto, Y. (2012). Tense and aspect error correction for ESL learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL): Short Papers - Volume 2*, Jeju Island, Korea.

Tamchyna, A., Weller-Di Marco, M., and Fraser, A. (2017). Modeling Target-Side Inflection in Neural Machine Translation. In *Proceedings of the 2nd Conference on Machine Translation (WMT)*, Copenhagen, Denmark.

Tiedemann, J. and Scherrer, Y. (2017). Neural Machine Translation with Extended Context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark.

Tillmann, C. (2004). A Unigram Orientation Model for Statistical Machine Translation. In *Proceedings of Human Language Technology conference/North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL), Short Papers*, Boston, Massachusetts.

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying Morphology Generation Models to Machine Translation. In *Proceedings of 46th Annual Meeting of the Association of Computational Linguistics (ACL-HLT)*, Columbus, Ohio, USA.

Tromble, R. and Eisner, J. (2009). Learning Linear Ordering Problems for Better Translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore.

Verhagen, M., Mani, I., Saurí, R., , Knippen, R., Littman, J., and Pustejovsky, J. (2005). Automating Temporal Annotation with TARSQI. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Ann Arbor, Michigan, USA.

Wang, C., Collins, M., and Koehn, P. (2007). Chinese Syntactic Reordering for Statistical Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, Prague, Czech Republic.

Weinrich, H. (2001). *Tempus. Besprochene und erzählte Welt.* C.H.Beck, 6 edition.

Weller, M., Fraser, A., and im Walde, S. S. (2013). Using Sub- categorization Knowledge to Improve Case Prediction for Translation to German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

Wöllstein, A. (2010). *Topologisches Satzmodell*. Universitätsverlag WINTER Heidelberg.

Xia, F. and McCord, M. (2004). Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, Geneva, Switzerland.

Xu, P., Kang, J., Ringgaard, M., and Och, F. (2009). Using a Dependency Parser to Improve SMT for Subject-object-verb Languages. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Boulder, Colorado, USA.

Ye, Y., Fossum, L., Victoria, and Abney, S. (2006). Latent Features in Automatic Tense Translation between Chinese and English. In *Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing*, Sidney, Australia.

Zhang, Y., Zens, R., and Ney, H. (2007). Chunk-level Reordering of Source Language Sentences with Automatically Learned Rules for Statistical Machine Translation. In *Proceedings of the NAACL-HLT 2007/AMTA Workshop on Syntax and Structure in Statistical Translation (SSST)*, Rochester, New York, USA.

Zhu, M., Zhang, Y., Chen, W., Zhang, M., and Zhu, J. (2013). Fast and Accurate Shift-Reduce Constituent Parsing. In *Proceedings of the The 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sofia, Bulgaria.

A. Supplementary material

A.1. German syntactic tense patterns

This Section lists the German syntactic tense patterns which served as a basis for developing the TMV annotator, a tool for automatic annotation of tense, mood and voice for English, German and French. The details about the tool are given in Section 6.3.3 on page 133, as well as in (Ramm et al., 2017a). The patterns are given in terms of the German verbal POS tags taken from the STTS set.[1] The verbal POS tags, as well as used morphological features are given in Table A.1.

Verbal POS tags	
VAFIN	Finite auxiliary (*sein, haben, werden*)
VMFIN	Finite modal (*wollen, sollen, mögen, können, dürfen, müssen*)
VVFIN	Finite full verbs (e.g., *lesen, arbeiten*, etc.)
VAINF	Infinitive auxiliary
VMFIN	Infinitive modal
VVFINF	Infinitive full verb
VAPP	Participle auxiliary
VMPP	Participle modal
VVPP	Participle full verb
Morphological annotation	
Pres	present
Past	past
Ind	indicative
Subj	subjunctive

Table A.1.: Verbal POS tags and the morphology annotation used to describe the German syntactic tense patterns.

[1] `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-table.html` retrieved on January 23rd, 2018.

The German indicative active tense patterns are given in Tables A.2 and A.3, while the passive patterns are listed in Table A.4. The subjunctive patterns are given in Tables A.5 and A.7. The passive subjunctive syntactic patterns are listed in Tables A.6 and A.8.

Note that the subjunctive tense patterns do not have all of the German indicative tense forms. For instance, there is no distinction between the past tense forms *Präteritum*, *Perfekt* and *Plusquamperfekt*. Instead, all of the past subjunctive forms are considered to be *Past*. Furthermore, according to Eisenberg (1998), the German subjunctive mood may or may not be mapped to *Future I* depending on the function. For instance, in the context of reported speech, *'würde lesen'* is the subjunctive mood of *Future I*, while used in other contexts, it is considered to be *Präsens* which is also our analysis of the respective subjunctive forms. For more details about mapping of the German subjunctive mood to the respective tense, please refer to (Eisenberg, 1998, p.168).

Synt. tense	Pattern	Example
Präsens	V*FIN.Pres.Ind	Ich lese.
	VMFIN.Pres.Ind V(A|V)PP	Ich kann lesen.
	VVFIN.Pres.Ind VVINF	Die Bitte bleibt bestehen.
	VVFIN.Pres.Ind VVINF VVINF	Er lässt mich studieren gehen.
	VMFIN.Pres.Ind VVINF VVINF VMINF	Du kannst spielen lernen wollen.
Präteritum	V*FIN.Past.Ind	Ich las.
	VMFIN.Past.Ind V(A|V)INF	Ich konnte lesen.
	VMFIN.Past.Ind VVPP VAINF VVFIN.Past.Ind VINF	Was mochte geschehen sein? Die Bitte blieb bestehen.
	VVFIN.Past.Ind VVINF VVINF	Er ließ mich studieren gehen.
	VMFIN.Past.Ind VVINF VVINF	Du konntest es spielen lernen.
Perfekt	VAFIN.Pres.Ind V*PP	Ich habe gelesen. Ich bin gefahren.
	VAFIN.Pres.Ind V(A|V)INF VMINF	Ich habe lesen können.
	VAFIN.Pres.Ind VVINF VVPP	Ich habe es spielen gelernt.
	VAFIN.Pres.Ind VVINF VVINF VMINF	Ich habe es spielen lernen wollen.
	VAFIN.Pres.Ind VVINF VVINF	Ich habe es fallen lassen.
	VAFIN.Pres.Ind VVINF VVINF VVINF	Er hat mich studieren gehen lassen.
	VAFIN.Pres.Ind VVPP VVPP	Er ist verloren gegangen.
Plusquam-perfekt	VAFIN.Past.Ind V(A|V)PP	Ich hatte gearbeitet. Ich war gefahren.
	VAFIN.Past.Ind V(A|V)INF VMINF	Ich hatte arbeiten können.
	VAFIN.Past.Ind VVINF VVPP	Ich hatte es spielen gelernt.
	VAFIN.Past.Ind VVINF VVINF VMINF	Ich hatte es spielen lernen wollen.
	VAFIN.Past.Ind VVINF VVINF	Ich hatte es fallen lassen.
	VAFIN.Past.Ind VVINF VVINF VVINF	Er hatte mich studieren gehen lassen.
	VAFIN.Past.Ind VVPP VVPP	Er war verloren gegangen.

Table A.2.: Full list of the German indicative active morpho-syntactic tense patterns (part 1).

Synt. tense	Pattern	Example
Futur I	VAFIN.Pres.Ind V*INF	Ich werde lesen.
	VAFIN.Pres.Ind V*INF VMINF	Ich werde lesen können.
	VAFIN.Pres.Ind VVINF VVINF	Ich werde es spielen lernen.
	VAFIN.Pres.Ind VVPP VVINF	Es wird verschwunden bleiben.
Futur II	VAFIN.Pres.Ind V(A|V)INF VAINF	Ich werde gelesen haben. Ich werde gefahren sein.
	VAFIN.Pres.Ind VVPP VAINF VMINF	Ich werde gelesen haben können.
	VAFIN.Pres.Ind VVINF VVPP VAINF	Ich werde es spielen gelernt haben.
	VAFIN.Pres.Ind VVPP VVPP VAINF	Es wird verschwunden geblieben sein.
	VAFIN.Pres.Ind VVINF VVINF VAINF VAINF	Er wird mich studieren gehen lassen haben.

Table A.3.: Full list of the German indicative active morpho-syntactic tense patterns (part 2).

Synt. tense	Pattern	Example
Präsens	VAFIN.werden.Pres.Ind V(A|M|V)PP	Es wird gelesen.
	VAFIN.sein.Pres.Ind V(A|M|V)PP*	Es ist gelesen.
	VMFIN.Pres.Ind V(A|V)PP VAINF.werden	Es kann gelesen werden.
	VMFIN.Pres.Ind V(A|V)PP VAINF.sein*	Es kann gelesen sein.
Präteritum	VAFIN.weden.Past.Ind V(A|V)PP	Es wurde gelesen.
	VAFIN.sein.Past.Ind V(A|V)PP*	Es war gelesen.
	VMFIN.Past.Ind V(A|V)PP VAINF.werden	Es konnte gelesen werden.
	VMFIN.Past.Ind V(A|V)PP VAINF.sein*	Es konnte gelesen sein.
Perfekt	VAFIN.sein.Pres.Ind VVPP VAPP.werden	Es ist gelesen worden.
	VAFIN.sein.Pres.Ind VVPP VAPP.sein*	Es ist gelesen gewesen.
Plusquam-perfekt	VAFIN.sein.Past.Ind V(A|V)PP VAPP.werden	Es war gelesen worden.
	VAFIN.sein.Past.Ind V(A|V)PP VAPP.sein*	Es war gelesen gewesen.
	VAFIN.haben.Past.Ind V(A|V)INF VAINF.werden VMINF	Es hatte gelesen werden können.
	VAFIN.haben.Past.Ind V(A|V)INF VAINF.sein VMINF*	Es hatte gelesen sein können.
Futur I	VAFIN.werden.Pres.Ind V(A|V|M)PP VAINF.werden	Es wird gelesen werden.
	VAFIN.werden.Pres.Ind V(A|V|M)PP VAINF.sein*	Es wird gelesen sein.
	VAFIN.werden.Pres.Ind V(A|V)PP VAINF.werden VMINF	Es wird gelesen werden können.
	VAFIN.werden.Pres.Ind V(A|V)PP VAINF.sein VMINF*	Es wird gelesen sein können.
Futur II	VAFIN.werden.Pres.Ind V(A|V)PP VAPP.werden VAINF.sein	Es wird gelesen worden sein.
	VAFIN.werden.Pres.Ind VVPP VAPP.werden VAINF.sein VMINF	Es wird gelesen worden sein können.

Table A.4.: Full list of the German indicative passive morpho-syntactic tense patterns.

Synt. tense	Pattern	Example
Präsens	V*FIN.Pres.Subj	Er lese.
	VMFIN.Pres.Subj V(A|V)PP	Er könne lesen.
	VVFIN.Pres.Subj VVINF	Die Bitte bleibe bestehen.
	VVFIN.Pres.Subj VVINF VVINF	Er lasse mich studieren gehen.
	VMFIN.Pres.Subj VVINF VVINF VMINF	Er könne spielen lernen wollen.
Past	VAFIN.Pres.Subj V*PP	Er habe gelesen. Er sei gefahren.
	VAFIN.Pres.Subj V(A|V)INF VMINF	Er habe lesen können.
	VAFIN.Pres.Subj VVINF VVPP	Er habe es spielen gelernt.
	VAFIN.Pres.Subj VVINF VVINF VMINF	Er habe es spielen lernen wollen.
	VAFIN.Pres.Subj VVINF VVINF	Er habe es fallen lassen.
	VAFIN.Pres.Subj VVINF VVINF VVINF	Er habe mich studieren gehen lassen.
	VAFIN.Pres.Subj VVPP VVPP	Er sei verloren gegangen.
Futur I	VAFIN.Pres.Subj V*INF	Er werde lesen.
	VAFIN.Pres.Subj V*INF VMINF	Er werde lesen können.
	VAFIN.Pres.Subj VVINF VVINF	Er werde es spielen lernen.
	VAFIN.Pres.Subj VVPP VVINF	Es werde verschwunden bleiben.
Futur II	VAFIN.Pres.Subj V(A|V)INF VAINF	Er werde gelesen haben Er werde gefahren sein.
	VAFIN.Pres.Subj VVPP VAINF VMINF	Er werde gelesen haben können.
	VAFIN.Pres.Subj VVINF VVPP VAINF	Er werde es spielen gelernt haben.
	VAFIN.Pres.Subj VVPP VVPP VAINF	Es werde verschwunden geblieben sein.
	VAFIN.Pres.Subj VVINF VVINF VAINF VAINF	Er werde mich studieren gehen lassen haben.

Table A.5.: Full list of the German *Konjunktiv I* active morpho-syntactic tense patterns.

Synt. tense	Pattern	Example
Präsens	VAFIN.werden.Pres.Subj V(A|M|V)PP	Es werde gelesen.
	VAFIN.sein.Pres.Subj V(A|M|V)PP*	Es sei gelesen.
	VMFIN.Pres.Subj V(A|V)PP VAINF.werden	Es könne gelesen werden.
	VMFIN.Pres.Subj V(A|V)PP VAINF.sein*	Es könne gelesen sein.
Past	VAFIN.sein.Pres.Subj VVPP VAPP.werden	Es sei gelesen worden.
	VAFIN.sein.Pres.Subj VVPP VAPP.sein*	Es sei gelesen gewesen.
Futur I	VAFIN.werden.Pres.Subj V(A|V|M)PP VAINF.werden	Es werde gelesen werden.
	VAFIN.werden.Pres.Subj V(A|V|M)PP VAINF.sein*	Es werde gelesen sein.
	VAFIN.werden.Pres.Subj V(A|V)PP VAINF.werden VMINF	Es werde gelesen werden können.
	VAFIN.werden.Pres.Subj V(A|V)PP VAINF.sein VMINF*	Es werde gelesen sein können.
Futur II	VAFIN.werden.Pres.Subj V(A|V)PP VAPP.werden VAINF.sein	Es werde gelesen worden sein.
	VAFIN.werden.Pres.Subj VVPP VAPP.werden VAINF.sein VMINF	Es werde gelesen worden sein können.

Table A.6.: Full list of the German *Konjunktiv I* passive morpho-syntactic tense patterns.

Synt. tense	Pattern	Example
Präsens	V*FIN.Past.Subj	Ich läse.
	VMFIN.Past.Subj V(A|V)PP	Ich könnte lesen.
	VVFIN.Past.Subj VVINF	Die Bitte bliebe bestehen.
	VVFIN.Past.Subj VVINF VVINF	Er ließe mich studieren gehen.
	VMFIN.Past.Subj VVINF VVINF VMINF	Du könntest spielen lernen wollen.
	VAFIN.Past.Subj V*INF	Ich würde lesen.
	VAFIN.Past.Subj V*INF VMINF	Ich würde lesen können.
	VAFIN.Past.Subj VVINF VVINF	Ich würde es spielen lernen.
	VAFIN.Past.Subj VVPP VVINF	Es würde verschwunden bleiben.
Past	VAFIN.Past.Subj V*PP	Ich hätte gelesen. Ich bin gefahren.
	VAFIN.Past.Subj V(A|V)INF VMINF	Ich hätte lesen können.
	VAFIN.Past.Subj VVINF VVPP	Ich hätte es spielen gelernt.
	VAFIN.Past.Subj VVINF VVINF VMINF	Ich hätte es spielen lernen wollen.
	VAFIN.Past.Subj VVINF VVINF	Ich hätte es fallen lassen.
	VAFIN.Past.Subj VVINF VVINF VVINF	Er hätte mich studieren gehen lassen.
	VAFIN.Past.Subj VVPP VVPP	Er wäre verloren gegangen.
Futur II	VAFIN.Past.Subj V(A|V)INF VAINF	Ich würde gelesen haben. Ich würde gefahren sein.
	VAFIN.Past.Subj VVPP VAINF VMINF	Ich würde gelesen haben können.
	VAFIN.Past.Subj VVINF VVPP VAINF	Ich würde es spielen gelernt haben.
	VAFIN.Past.Subj VVPP VVPP VAINF	Es würd verschwunden geblieben sein.
	VAFIN.Past.Subj VVINF VVINF VAINF VAINF	Er würde mich studieren gehen lassen haben.

Table A.7.: Full list of the German *Konjunktiv II* active morpho-syntactic tense patterns.

Synt. tense	Pattern	Example
Präsens	VAFIN.werden.Past.Subj V(A|M|V)PP	Es würde gelesen.
	VAFIN.sein.Past.Subj V(A|M|V)PP*	Es wäre gelesen.
	VMFIN.Past.Subj V(A|V)PP VAINF.werden	Es könnte gelesen werden.
	VMFIN.Past.Subj V(A|V)PP VAINF.sein*	Es könnte gelesen sein.
Past	VAFIN.sein.Past.Subj VVPP VAPP.werden	Es wäre gelesen worden.
	VAFIN.sein.Past.Subj VVPP VAPP.sein*	Es wäre gelesen gewesen.
Futur I	VAFIN.werden.Past.Subj V(A|V|M)PP VAINF.werden	Es würde gelesen werden.
	VAFIN.werden.Past.Subj V(A|V|M)PP VAINF.sein*	Es würde gelesen sein.
	VAFIN.werden.Past.Subj V(A|V)PP VAINF.werden VMINF	Es würde gelesen werden können.
	VAFIN.werden.Past.Subj V(A|V)PP VAINF.sein VMINF*	Es würde gelesen sein können.
Futur II	VAFIN.werden.Past.Subj V(A|V)PP VAPP.werden VAINF.sein	Es würde gelesen worden sein.
	VAFIN.werden.Past.Subj VVPP VAPP.werden VAINF.sein VMINF	Es würde gelesen worden sein können.

Table A.8.: Full list of the German *Konjunktiv II* passive morpho-syntactic tense patterns.

A.2. English syntactic tense patterns

The English patterns are given in terms of the English verbal POS tags taken from the Penn Treebank POS set.[2] The verbal POS tags are explained in Table A.9.

Verbal POS tags	
MD	Modal verb (*will/would, shall/should, may/might, must, can/could, need/ought*)
VB	Infinitive verb
VBD	Verb, past tense
VBG	Verb gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Table A.9.: Verbal POS tags used to describe the English syntactic tense patterns.

Active tense patterns are given in Table A.10, while the passive patterns are listed in Table A.11.

[2] `https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html` retrieved on January 23rd, 2018.

Synt. tense	Pattern	Example
Present	VB(Z|P)	He writes. I write.
	MD	I can.
	MD VB	I can write.
Pres prog	VB(Z|P) VBG	He is writing. I am writing.
Past	VBD	I wrote.
Past prog	VBD VBG	I was writing.
Pres perfect	VB(Z|P) VBN	He has writing.
Pres perf prog	VB(Z|P) VBN.be VBG	He has been writing.
Pluperf	VBD VBN	I had written.
Pluperf prog	VBD VBN.be VBG	I had been writing.
Future I	MD VB	I will write.
Future I prog	MD VB.be VBG	I will be writing.
Futur II	MD VB VBN	I will have written.
Futur II prog	MD VB VBN.be VBN	I will have been writing.
Cond I	MD.subj VB	I would write.
Cond I prog	MD.subj VB.be VBN	I would be writing.
Cond II	MD.subj VB VBN	I would have written.
Cond II prog	MD.subj VB VBN.be VBG	I would have been writing.

Table A.10.: Full list of the English active morpho-syntactic tense patterns.

Synt. tense	Pattern	Example
Present	VB(Z|P).be VBN*	The letter is written.
	MD VB.be VBN	The letter can be written.
Pres prog	VB(Z|P) VBG.be VBN	The letter is being written.
Past	VBD.be VBN*	The letter was written.
Past prog	VBD VBG.be VBN	The letter was being written.
Pres perfect	VB(Z|P) VBN.be VBN	The letter has been written.
Pres perf prog	VB(Z|P) VBN.be VBG.be VBN	The letter has been being written.
Pluperf	VBD VBN.be VBN	The letter had been written.
Pluperf prog	VBD VBN.be VBG.be VBN	The letter had been being written.
Future I	MD VB.be VBN	The letter will be written.
Future I prog	MD VB.be VBG.be VBN	The letter will be being written.
Futur II	MD VB VBN.be VBN	The letter will have been written.
Futur II prog	MD VB VBN.be VBN	The letter will have been being written.
Cond I	MD.subj VB.be VBN	The letter would be written.
Cond I prog	MD.subj VB.be VBN	The letter would be being written.
Cond II	MD.subj VB VBN.be VBN	It would have been read.
Cond II prog	MD.subj VB VBN.be VBG	The letter would have been being written.

Table A.11.: Full list of the English passive morpho-syntactic tense patterns.

A.3. Frequency tables of the English-German tense pairs

Table A.12 shows frequency of the German tense forms found in the News corpus, while Table A.13 shows the frequencies derived from the Europarl corpus.

	pres	perf	imperf	pluperf	futI	futII	konjI	konjII	-
pres	114683	3359	2074	91	1204	177	1226	2641	838
presProg	7201	244	139	3	101	8	47	56	78
presPerf	3059	13005	4650	145	9	9	131	132	80
presPerfProg	220	315	63	3	0	1	5	5	6
past	2620	5483	36248	1468	93	4	718	1056	178
pastProg	62	49	636	39	0	0	32	53	12
pastPerf	15	133	448	1106	0	1	96	364	9
pastPerfProg	1	6	16	31	0	0	1	0	1
futureI	3604	84	64	4	7112	309	92	474	113
futureIProg	1	0	0	0	0	0	0	0	0
futureII	12	10	3	0	16	32	0	12	0
futureIIProg	0	0	0	0	0	0	0	0	0
condI	1630	73	3846	120	126	8	367	10907	197
condIProg	0	0	1	0	0	0	0	3	0
condII	15	39	47	15	1	1	2	1359	3
condIIProg	0	2	0	0	0	0	0	8	0
gerund	4581	732	2014	195	297	41	150	743	2584
toInfinitive	10090	1012	2857	138	1656	83	334	1796	18470

Table A.12.: Contigency matrix of the tenses in parallel English and German VCs extracted from the News corpus.

	pres	perf	imperf	pluperf	futI	futII	konjI	konjII	-
pres	1352169	38553	31584	1142	10276	1632	14899	11333	11870
presProg	93204	3100	3473	93	5798	285	1073	592	2614
presPerf	30081	149825	49121	1365	250	217	3277	786	961
presPerfProg	2263	2938	499	27	10	3	124	25	45
past	28666	92365	131889	6730	284	149	6009	4991	1114
pastProg	600	1259	2585	238	43	2	216	377	91
pastPerf	241	1758	2067	3254	1	3	517	1895	39
pastPerfProg	15	51	34	62	0	1	5	26	1
futureI	62063	1546	1336	59	72841	2433	1940	1978	1282
futureIProg	2	0	0	0	3	0	0	0	0
futureII	198	424	65	1	191	384	7	60	8
futureIIProg	0	0	0	0	0	0	0	0	0
condI	112939	2874	34066	1083	1510	131	3083	49317	3140
condIProg	1	0	0	0	0	0	1	8	0
condII	328	347	637	117	8	6	47	9360	25
condIIProg	3	1	5	1	0	0	0	50	2
gerund	40825	8009	6777	511	6177	324	1852	3612	18627
toInfinitive	246405	15855	25524	1005	17133	961	4088	12381	155602

Table A.13.: Contigency matrix of the tenses in parallel English and German VCs extracted from the Europarl corpus.