Masterarbeit

im Studiengang Computerlinguistik

an der Ludwig-Maximilians-Universität München

Fakultät für Sprach- und Literaturwissenschaften

Error Analysis in Machine Translation: Evaluating the Role of Morphology

vorgelegt von
Laura Isla Navarro

Betreuer:	Dr. Marion Di Marco
Prüfer:	Prof. Dr. Alexander Fraser
Bearbeitungszeitraum:	26. März - 08. August 2022

Abstract

Morphological correctness is key to the success of machine translation systems, yet it is a field which still needs to be further explored. In this thesis, we analyze to which extent linguistic features from the source side interact with those of the target side through several error prediction experiments. These experiments handle data from a shared task on Quality Estimation, which are annotated on the word-level as correct or incorrect depending on whether the words have been translated correctly, or whether there are missing or additional words on the target side. These datasets involve a morphologically poor language, English, and a morphologically rich language, German or Russian. For this purpose, we extend the work of Bollmann and Søgaard (2021) on error analysis and the role of morphology.

Firstly, the datasets for the different language pairs are annotated with their linguistic and non-linguistic features, to then be run with a random forest classifier. For our experiments, linguistic source-side features are extracted and transferred to the target side. From a general perspective, accuracy and F_1 scores are regarded to evaluate how well the classifier predicts errors. A more specific objective of this work is to use Feature Importance (FI) scores to identify which features are the most predictive of errors. Our findings show that combining linguistic source-side features with the target data enhances classifier performance when predicting errors. Additionally, the source-side features in the combined datasets are on the top of the ranking, which suggests that they are relevant for error prediction.

Zusammenfassung

Morphologische Korrektheit ist zwar entscheidend zum Erfolg von maschinellen Übersetzungssystemen, aber sie ist ein Bereich, der noch weitererforscht werden muss. In dieser Arbeit werden wir in mehreren Experimente zur Fehlervorhersage analysieren, inwiefern linguistische Merkmale von der Quellsprache mit denen von der Zielsprache interagieren. Diese Experimente umfassen Daten von einem *Shared Task* zu *Quality Estimation*, welche auf der Wortebene als korrekt oder inkorrekt annotiert werden, je nachdem, ob die Wörter richtig übersetzt wurden, oder es fehlende oder zusätzliche Wörter auf der Zielseite gibt. Diese Datensätze beinhalten eine morphologisch arme Sprache, Englisch, und eine morphologisch reiche Sprache, Deutsch oder Russisch. Zu diesem Zweck ergänzen wir die Arbeit von Bollmann and Søgaard (2021) zur Fehleranalyse und die Rolle der Morphologie.

In einem ersten Schritt werden die Datensätze für die verschiedenen Sprachkombinationen mit ihren linguistischen und nicht-linguistischen Merkmalen annotiert. Anschließend wird ein Random Forest-Klassifikator zusammen mit den neu annotierten Datensätzen ausgeführt. Für unsere Experimente werden die linguistischen Merkmalen aus der Quellsprache extrahiert und auf die Zielsprache übertragen. Um zu beurteilen, wie gut der Klassifikator Fehler vorhersagt, betrachten wir die Genauigkeit und F_1 Ergebnisse des Klassifikators. Ein spezifischeres Ziel dieser Arbeit ist die Verwendung von sogenannten Feature Importance (FI) Ergebnissen für die Identifizierung der fehlervorhersagendsten Merkmale. Unsere Erkenntnisse zeigen, dass die Kombination von linguistischen Merkmalen der Quellseite mit den bestehenden Merkmalen der Zielseite die Klassifikatorsleistung verbessert. Hinzu kommt, dass die linguistischen Merkmale der Quellseite in den kombinierten Datensätzen sich an der Spitze des Rankings befinden. Dies weist darauf hin, dass diese Merkmale relevant für die Fehlervorhersage sind.

Contents

List of Figures

List of Tables

Acronyms

ACL Association for Computational Linguistics. 1

AI Artificial Intelligence. 12

AQE Automatic Quality Estimation. 5, 15, 16, 21

ASR Automatic Speech Recognition. 13

BERT Bidirectional Encoder Representations from Transformers. 10

BLEU Bilingual Evaluation Understudy. 8, 9, 15

BPE Byte Pair Encoding. 9, 16

CoNLL Conference on Computational Natural Language Learning. I, 5, 18, 24, 25, 28–30, 33, 36–39, 41

FI Feature Importance. I, III, IV, VII, 6, 18, 35–41, 43

HTER Human-targeted Translation Edit Rate. 15, 21

mBERT Multilingual Bidirectional Encoder Representations from Transformers. 9

METEOR Metric for Evaluation of Translation with Explicit Ordering. 9, 15

MFE Morphological Feature Entropy. 6

ML Machine Learning. 15

MQE Manual Quality Estimation. 16

MT Machine Translation. 1–3, 5–8, 11–17, 21, 22, 28, 41–43

NLP Natural Language Processing. 1–3, 5, 7, 11, 12, 15–17, 25, 36, 42

NMT Neural Machine Translation. 8, 9, 12, 13

OOV Out-Of-Vocabulary. 16

POS Part of Speech. I, 1, 5, 10, 23, 25, 28, 30, 31, 37, 39–41, 43, 49

QE Quality Estimation. VII, 3, 11, 15, 21, 43

RNN Recurrent Neural Network. 13

SEM Semantic Role Labelling. 5, 43

SMT Statistical Machine Translation. 8, 9, 12, 13

TER Translation Edit Rate. 15

UD Universal Dependencies. 5, 22, 24, 25

UDP Dependency Parsing. 5, 43

VMWE Verbal Multi-Word Expression Classification. 5

WALS World Atlas of Linguistic Structures. 2

WER Word Error Rate. 15

WMT Workshop on Machine Translation. I, III, VII, 21, 22, 28, 35, 41

CD Contents

- **ma_laura-isla-navarro.pdf:** Digital copy of this thesis.
- **figures:** Folder containing all the figures used in this thesis.
- **references:** Folder with PDF copies for most references.
- **experiments_bs:** Folder containing scripts by Bollmann and Søgaard.
- **experiments_lin:** Folder containing scripts designed for the source-feature extraction part of this thesis.
  - **original_data:** Folder containing the original datasets.
    - **wmt19_data:** Folder containing parallel and alignment data from the WMT19 shared task on QE.
    - **annotated_data:** Folder with data annotated with linguistic and non-linguistic features.
  - **pre-processed_data:** Folder containing the pre-processed datasets, that is, target data with their corresponding linguistic source-side features.
  - **results:** Folder containing the classifiers outputs, including accuracy, F_1, Feature Importance (FI) scores.
  - **en-zh:** Folder containing the data needed to run the experiment with the English-Chinese dataset. This folder also includes the results of the classifier for the individual files and the mixed dataset.

1. Introduction

Initially applied to determine the incidence, nature, causes and consequences of unsuccessful natural language in language learners (James, 1998), error analysis has become a crucial step in the evaluation of the performance of Natural Language Processing (NLP) systems to find better approaches to ongoing problems. The inflectional and derivational processes present in morphology makes it one of the hardest linguistic disciplines for NLP tasks to handle. A word can be inflected in many different ways, and this is even more notorious when we move away from languages that possess analytic morphologies. Furthermore, derivation is another paradigm to keep in mind. In many languages, derivation is the main tool to create words coming from the same stem resulting in tokens of the same family but with different Part of Speech (POS) tags. These phenomena are part of natural language, and as natural languages continue to evolve with time, it is paramount to research on error analysis methods to enhance the NLP systems available. Table 1.1 displays a high-level example of an inflected German verb in the present tense, *wohnen*, *to live* in English. In addition to this, Table 1.2 also offers some derivated forms from the Spanish noun *zapato*, *shoe* in English.

Person	**Conjugation**
1st singular	ich wohn**e**
2nd singular	du wohn**st**
3rd singular	er/sie/es wohn**t**
1st plural	wir wohn**en**
2nd plural	ihr wohn**t**
3rd plural	sie wohn**en**

Table 1.1.: German verb *wohnen* conjugated in the present tense.

Word	**Meaning**
zapat**ero**	shoomaker
zapet**ería**	shoe store
zapat**os**	shoes

Table 1.2.: Spanish noun *zapato* and some of its derivated forms.

Dichotomous in its nature, Machine Translation (MT) has to deal with linguistic phenomena from two different languages simultaneously which, in some cases, pertain to different language families. While the field of MT has grown by leaps and bounds in the current century with a recent shift from statistical to neural systems due to the enormous development of computational power and the increasing availability of data, there is still work to be done, especially when it comes to translating into morphologically richer languages. Since English is the most widely-spoken language in the world both by native and non-native speakers, and the most popular language in research venues, the majority of data sources for the development of systems are likewise available in English. English is a morphologically poor language, that is, it relies mostly on syntactic variation to provide meaning in context. By the same token, quality metrics for machine translation systems are still to better capture morphological correctness within the sentence.

For this matter, the objective of this thesis will be to evaluate to which extent the morphology from the source language interacts with that of the target language. The research carried out by Bollmann and Søgaard (2021) on error analysis and the role of morphology will play a central role in the development of the experimental setup for this dissertation. In their article, presented at the conference organized by the Association for Computational Linguistics (ACL) in 2021, they explored the influence of morphology in the

errors committed by different NLP systems such as dependency parsers and verbal multiword expression classifiers, and the importance of different linguistic and non-linguistic features, with case and gender being the most relevant across the different shared tasks. They part from the premises that morphology is predictive of errors and that the relevance of morphology increases with the morphological complexity of a language. The latter conjecture was disproved after running their experiments. It is also worth mentioning that the results are task-dependant, and that not all tasks introduced in their paper were studied to the same level, especially the MT task. Therefore, we will further address this task by adding source-side linguistic features to the target side to analyze the effect of this on the overall classifier performance, as well as on the importance of individual features.

1.1. Motivation

From newspapers to tweets, the worldwide increasing availability of machine-translated data across platforms makes it necessary to pay close to attention to the linguistic quality of these outputs, especially when we take into account the growing interest of multinational corporations to internationalize and localize their products.

Despite there only being 400 million of native English speakers, compared to the 1 billion of Chinese native speakers, the majority of data available on the Worldwide Web can only be found in English, making it the most present language in computational linguistics venues. Figure 1.1 shows the ten most widely-available languages online. Curiously enough, while the ranking is topped by English and Chinese, which are languages with analytic morphologies, that is, with poor inflection, the rest of the ranking is composed of synthetic languages. Languages with analytic morphologies tend to resort to derivative procedures for word formation. On the other hand, languages with synthetic morphologies have abundant complex words, and rely heavily on inflectional morphological processes. This fact drives the need for research in the context of morphologically rich languages, and especially of low-resource languages within this category.

The World Atlas of Linguistic Structures (WALS) (Haspelmath et al., 2005) lists around 192 typological features, 48% of which only exist in low-resource languages, which belong to categories 0, 1 and 2 in the taxonomy for the availability of data in each language (Joshi et al., 2020). It is essential to regard these features since they can be useful for generalization when creating language models, and to analyze what features these models are still able to capture. By the same token, working on languages beyond English can provide a better insight into the relationships between the languages of the world (Artetxe et al., 2020) (cf. Ruder (2020)).

Machine Translation, despite having experienced a grandiose development during the past decade with the shift from statistical to neural approaches, still has some battles to win. One of them is the central focus of this thesis: dealing with morphologically rich languages on the target side.

The addition of linguistic features from a morphologically poor source side language to those of a morphologically rich target language should provide us with a better understanding of whether the morphological output of the MT system is influenced to a greater or lesser degree by the morphology of the source language. This way, we can decide which areas of the MT system are to be enhanced.

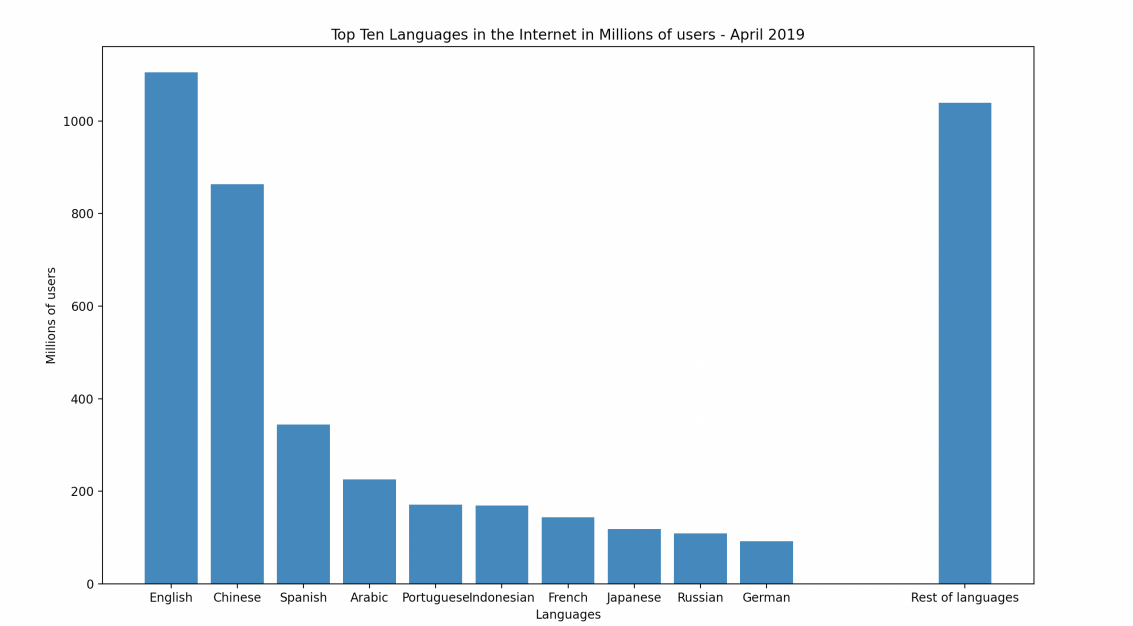

Figure 1.1.: The 10 most widely-used languages in the web. Graph adapted from Internet World Stats.

1.2. Research Question

As mentioned earlier in the introduction, the experimental setup for this thesis is based upon the research by Bollmann and Søgaard (2021) (Section 2.1) on error analysis and morphology. They carried out a series of experiments across four different NLP tasks, including MT.

Having said that, since the MT part of their work was not as extensive, in this thesis, we will provide a more thorough analysis of it, posing the following research question:

RQ: How do linguistic source-side features interact with the linguistic and non-linguistic features on the target side?

Originally, Bollmann and Søgaard (2021) regard the source and target side of the machine translation task independently, delivering this way separate results for each side. The objective of this thesis is to assess the source and target side jointly by incorporating the linguistic features from the source side into the target side. In this manner, we can then analyse whether the addition of linguistic source-side features to the target side impacts the results of the error analysis evaluation task.

1.3. Thesis Structure

Chapter 2 provides a brief overview into the current research on the field of morphology in NLP tasks, with a special focus on morphologically rich languages, MT, QE, and tokenization. Furthermore, the paper by Bollmann and Søgaard (2021) will be introduced to provide the reader with an overview of the original experiments before reaching the experimental setup of this thesis. Chapter 3 comprises theoretical background information with basic concepts that are needed to understand and perform the experimental part of this thesis. This includes an introduction to morphology and machine translation, and offers an insight into manual and automatic evaluation methods for machine translation systems, with a final mention of the challenges NLP tasks face when dealing with morphology.

Once the general knowledge is acquired, we proceed to Chapter 4, which exposes step-by-step the methods applied to convert the data into the required format for the error analysis task. Next, Chapter 5 describes the data used for the experimental part of the dissertation, as well as some pipelines necessary for the morphological pre-processing of the data presented. This part also introduces the manner in which we extract source-side features to be added to the target side. This chapter is succeeded by Chapter 6, which holds the experiments performed, as well as their results. These results are further discussed in Chapter 7 with a mention to its limitations. Finally, Chapter 8 concludes this work, offering some suggestions for future work.

2. Related Work

To support the relevance of this thesis, in this chapter, we review significant research within the field of error analysis and morphology in MT.

Since it serves as a foundation for this thesis, we will begin this chapter by introducing the paper by Bollmann and Søgaard (2021) on error analysis and the role of morphology. Afterwards, we will shift the focus to more specific topics within this category including a taxonomy for MT error analysis, the challenges morphology pose to NMT, and the issues tokenization methods encounter when dealing with morphologically rich languages, among others.

2.1. Error Analysis and the Role of Morphology

Languages with complex morphology are a recurrent challenge in the NLP world. With their research, Bollmann and Søgaard evaluated the extent to which morphology is actually predictive of errors, and whether there is a correlation between the importance of morphology and the morphological complexity of a language, which are two common conjectures in NLP. Rather surprisingly, they discovered that whereas adding morphological features such as case and gender improves error prediction across sundry tasks, the morphological complexity of a language does not play a big role. They conclude that this might be due to simpler morphology being more discriminative.

Their experiments were also task-dependant, hence, the analysis of these conjectures in three different monolingual NLP tasks, and in a bilingual one, MT. The authors obtained the data from the following publicly available shared tasks:

- Semantic Role Labelling (SEM): Joint parsing of syntactic and semantic dependencies in multiple languages (Hajič et al., 2009).
- Dependency Parsing (UDP): Combines the objectives of the CoNLL-2006/2007 tasks with those of the CoNLL-2008 task, parsing from raw text instead of gold standard annotations or POS tagging with consistent syntactic representations resulting from the application of the Universal Dependency (UD) framework. Additionally, the CoNLL-2018 task focuses more on morphological analysis and the incorporation of data from other languages (Zeman et al., 2018).
- Verbal Multi-Word Expression Classification (VMWE): Labelling English sentences in tweets, and other sources with MWEs and supersenses for nouns and verbs. In PARSEME 1.0, the goal was to identify MWEs in context. PARSEME 1.1 enhances and adds new languages to the available set (Ramisch et al., 2018).
- Automatic Quality Estimation (AQE) for Machine Translation: "The task of predicting the quality of the output of machine translation systems given just the source text and the hypothesis translations." (Fonseca et al., 2019). AQE can occur at several levels, but since their article is morphology-focused, the authors just regard word-level AQE in this case in point.

All these datasets belonging to the aforementioned shared tasks were annotated with their linguistic using the morphological and lexical parsers presented in Chapter 5. Besides these linguistic features, non-linguistic features such as frequency and length are also

extracted. These features are summarized on Appendix B. Additionally, the MT data contain annotation tags indicating whether a word was rendered correct or wrongly, or whether a word has been omitted. The former contributes to the general and specific results of the classifier. The latter case is not explored in the research carried out by Bollmann and Søgaard, and it is a task out of the scope of this thesis. Once all features are extracted, a random forest classifier (Breiman, 2001) is used to assess the prominence of individual features, measured by their Feature Importance (FI) metrics. This metric holds the degree to which different features are predictive of errors. Furthermore, accuracy and F_1 scores are regarded for evaluating the thoroughness of the classifiers. In the case of FI, the higher the FI score, the higher the importance of the feature category, whereas negative values imply that including these features has adverse effects in the F_1 score of the classifier. It should be noted that in the original paper other metrics such as Morphological Feature Entropy (MFE) (Çagri Çöltekin and Rama, 2018) are employed to analyze whether the differences in the F_1 scores are somewhat related to the morphological complexity of a language.

These shared tasks generally comprise a plethora of languages. However, this is not applicable to the data from the MT shared task. As mentioned earlier in the introduction (Chapter 1), MT is a dichotomous task, with which we pursue transferring meaning from the source language to the target language. In light of this, we attempt to integrate the linguistic source-side features to the annotated target file to evaluate the coexistence of both languages.

2.1.1. Findings

Bollmann and Søgaard (2021) employ random forest classifiers (Breiman, 2001) as implemented in Scikit-learn (Pedregosa et al., 2012) to study how important morphology is when predicting errors. Firstly, they evaluate the classifier including the full feature set, in other words, with the morphological and control features together, which at a later stage will be compared to the classifier when run only with the control feature set.

To evaluate the importance of morphological and control features learned by the random forest classifiers, the authors carry out a series of experiments. First, they assess the overall F_1 scores resulting of the individual classifiers. Next, they run the classifier without the morphological features to better estimate the importance of morphology. Finally, they evaluate the role of individual morphological and control features. Figure 2.1 shows the top 10 features for each shared task ranked according to their average FI across languages.

Category	FI
U:POS	32.65
FREQ	15.51
LEN	11.38
U:CASE	7.25
U:GENDER	6.96
EDIT	6.75
U:NUMBER	3.78
U:NAMETYPE	2.77
U:ANIMACY	2.73
U:ADPTYPE	1.96

(a) SEM

Category	FI
U:POS	34.74
FREQ	31.99
LEN	21.79
U:CASE	16.51
U:GENDER	10.98
EDIT	9.01
U:NUMBER	6.47
U:ANIMACY	3.78
U:ASPECT	2.28
SYNCRETIC	2.08

(b) UDP

Category	FI
FREQ	38.24
LEN	28.97
U:CASE	12.56
U:POS	12.47
U:GENDER	10.15
EDIT	9.78
U:ANIMACY	7.24
U:NUMBER	7.15
U:ASPECT	5.81
U:TENSE	2.68

(c) VMWE

Category	FI
FREQ	29.50
LEN	19.63
U:CASE	12.39
U:POS	10.93
U:GENDER	9.69
EDIT	5.91
U:NUMBER	3.94
U:ASPECT	3.25
SYNCRETIC	2.84
U:ANIMACY	2.30

(d) MT

Figure 2.1.: Top 10 features by average feature importance represented by the FI score. All FI scores given $\cdot 10^3$. Figure from Bollmann and Søgaard (2021)

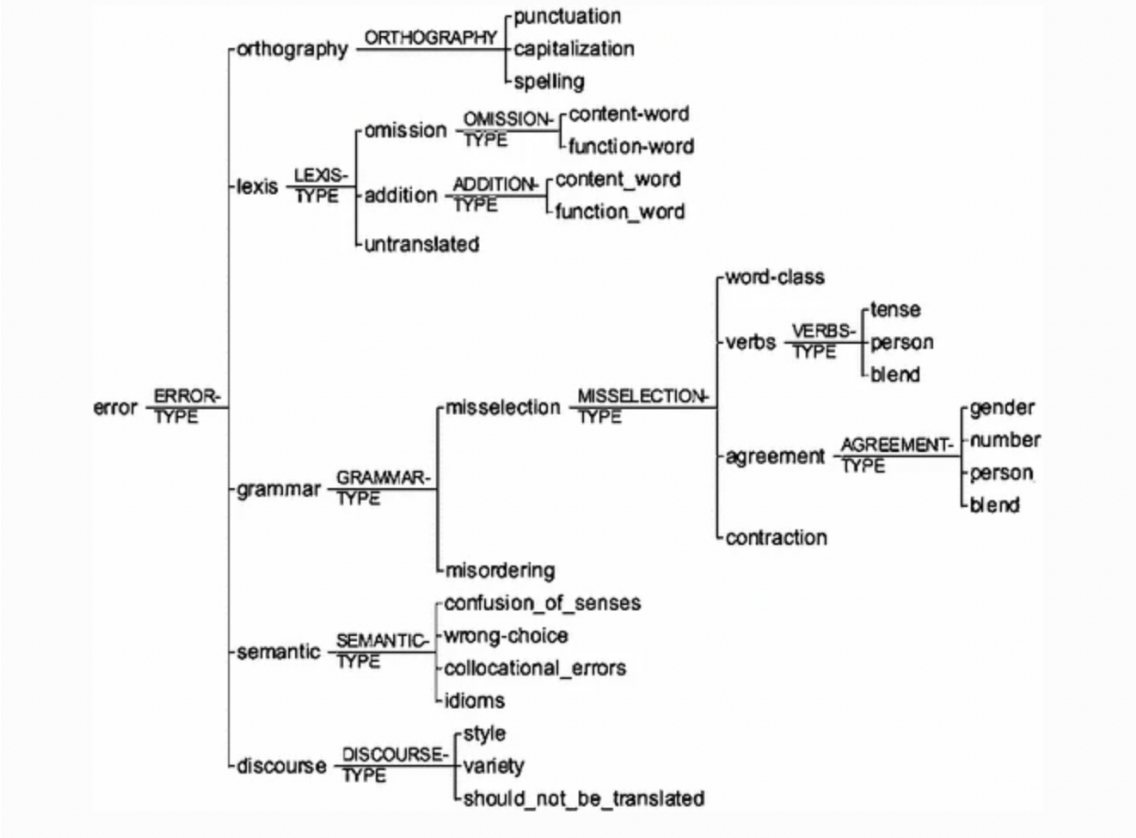

Figure 2.2.: Taxonomy of identified errors. Figure from Costa et al. (2015).

The length and frequency control features appear on the three most relevant features for each shared task, which can be explained by the fact that these are guaranteed to appear for every token.

2.2. Taxonomy of Machine Translation Error Analysis

Costa et al. (2015) state that "a detailed error analysis is a fundamental step in every natural language processing task, as to be able to diagnose what went wrong will provide cues to decide which research directions are to be followed.". For this matter, the authors performed a study to extend existing taxonomies for error analysis in NLP, with an emphasis on MT.

Since all errors are not easy to identify due to them being disguised within large units of text (James, 1998), this new taxonomy classifies errors regarding "the linguistic item which is affected by the error" (Dulay et al., 1983). Therefore, the authors established five different coarse categories: Ortography, Lexis, Grammar, Semantic and Discourse. Figure 2.2 displays a graphic with these categories, and their respective children. For our research, errors occurring at the lexis and grammar level are most interesting. At the lexis level, omitted, added, and untranslated words are taken into account, whereas the taxonomy regards missordering and misselection as grammar errors.

In morphologically rich langauges, agreement errors are of particular relevance. This sort of errors concern gender, number, person, and blend words. Omission errors are represented in the data used for this thesis due to words not being rendered into the target language, however, they are not explored in our experiments. Errors on the semantical or discoursive side will also not have a hand in this thesis since morphology is on the spotlight.

2.3. Morphology in Neural Machine Translation

In the past decade, there has been a breakthrough of neural approaches to MT. While Neural Machine Translation (NMT) systems produce high-quality outputs in comparison to their statistical counterparts, they are much less interpretable. In an attempt to comprehend to which extent morphology plays a role in the encoding part of NMT systems, Bisazza and Tump (2018) performed a fine-grained assessment of how different source-side morphological features are treated by the encoder at the word embedding and the recurrent state level with various target languages. Additionally, they also evaluated whether the relatedness between languages affects the morphological competence of the NMT encoder.

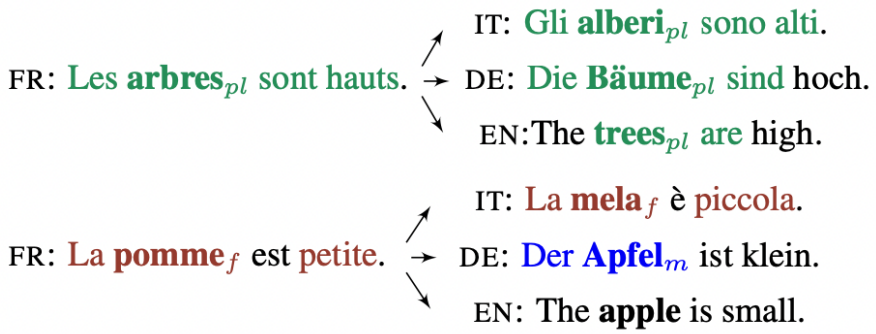

Figure 2.3.: Sample sentences to illustrate how number (top) and gender (bottom) morphological features are transferred into the target languages. Figure from Bisazza and Tump (2018)

.

Their findings confirm that source-side morphological features are only captured at the recurrent level where word representations are context-dependent, not considering word type properties as much. In their experiments, they found out that the mean accuracy of number feature was the highest, which can be explained by the fact of number being a morphological feature that remains mostly unchanged in the translation process. The results of the morphological features expressing number are closely followed by those of tense, which is a feature determined by semantics and specific language usage. Gender stays at the bottom of the ranking, which is justified by the little semantics involved in this feature, and the arbitrary manner in which it is assigned to a word.

As for the source-target language relatedness, they state that the impact of the target language is especially significant when it comes to gender. For instance, their classifier obtained relatively high Bilingual Evaluation Understudy (BLEU) scores for the French-Italian language pair in terms of gender. Both languages belong to the romance language family, where gender is usually consistent across languages. These results where compared to those of the French-English language pair for the same feature, which did not obtain BLEU scores as high.

2.4. Morphological Competence of Machine Translation Systems

Despite the enormous development of neural approaches to MT in the past lustrum (Burlot and Yvon, 2017), the metrics that are widely used to evaluate the performance of Statistical Machine Translation (SMT) do not provide enough insights into the competence of Neural Machine Translation (NMT) systems. For that reason, Burlot and Yvon (2017) proposed a new evaluation approach to assess the morphological competence of

these systems when translating from English into a morphologically rich languages.

In their research, Burlot and Yvon (2017) compare the performance of phrase-based SMT with that of numerous approaches to NMT systems such as word-based NMT, and Byte Pair Encoding (BPE)-based NMT on English to Czech and English to Latvian data. Furthermore, they also explored NMT modeling target morphology, which also involves BPE segmentation. The systems are optimized to enhance target morphology. To evaluate the overall rendition of all these distinct systems, metrics such as BLEU are put into place. Additionally, the morphological accuracy of these systems is also considered. For this purpose, the translation output of the system is aligned to its reference translation using the Metric for Evaluation of Translation with Explicit Ordering (METEOR), which is introduced together with BLEU in Section 3.2.3. These alignments are then to be analyzed word-wise to find whether they share the same form. The authors assume that two different tokens sharing the same lemma are two different inflections of the same lexeme.

Their research concludes that BPE-based NMT outperforms phrase-based SMT and word-based NMT approaches, most likely due to the fact that in word-based NMT, the vocabulary size is much smaller containing many closed class words. Nevertheless, when it comes to predicting the morphology of closed class words, word-based NMT systems show better results than their BPE counterparts at the time of conveying the target form for Czech pronouns.

As for the morphological accuracy of the output, morphological phenomena such as agreement are better modeled by sequence-to-sequence systems using BPE tokenization as opposed to phrase-based SMT or word-based NMT systems (Burlot and Yvon, 2017). NMT focused on target morphology modelling did not perform as well in this respect.

Factored NMT modeling target morphology is less sensitive than other NMT systems in the face of lexical variability, making more stable morphological predictions. SMT engines scored rather low, which is presumed to be a consequence of the concatenation of phrases to constitute an output sentence. This technique does not help to predict morphology in various contexts.

2.5. Adequacy of Word-Piece Modelling for Complex Morphologies

Morphologically rich languages belonging to the fusional and non-concatenative subtype, as it is the case of Semitic languages, are inherently challenging for tokenization methods such as Byte Pair Encoding (BPE) and WordPiece. Non-concatenative morphology resorts to apophony, transfixation, reduplication, and truncation to alter the form of a word. In concatenative morphologically rich languages such as Turkish and Russian, the morphemes of a word are linearly connected to the stem, which makes segmentation logical. On the other hand, in non-concatenative morphologically rich languages such as Hebrew or Arabic, the morphemes of a word can splitted in many different forms depending on its semantics. The challenge resides here for language models segmentation methods. For instance, the Hebrew word **בצלם** can be segmented in different ways depending on its meaning. If it means *in their shadow*, the segmentation would be as follows: **ב** - Preposition, **צל** - Noun, and **שהב** - Possesive. For the meaning *their onion*, it would be: **בצל** - Noun, and **שהב** - Possesive.

Klein and Tsarfaty (2020) examine how word-pieces capture morphology by investigating the segmentation resulting from multi-tagging in Modern Hebrew. For this purpose, they utilize the built-in tokenizer of mBERT for the segmentation of the input sentences

into word-pieces. Next, each word gets mapped to a multi-tag feature, comprising all the different POS tags the word is related to. They carried out their experiments in different settings. In their so-called Oracle setting, they segmented the word before being tokenized by BERT, assigning a different POS tag in the multi-tag to its corresponding morpheme. Despite this approach being rather costly, the language model performed better than when compared against other settings. Other settings do not tokenize the word before it is tokenized by BERT, which does not capture the morphological richness and variations of the words. Another drawback of this setting is that BERT fails to generalize to an unseen composition of tagged-morphemes into a new multi-tag. Figure 2.4 summarizes all tokenization alternatives for the individual settings.

	Before Tokenization:		After Tokenization:	
Nickname	Word	label	WP	label
Oracle	ל	IN	ל	IN
	ה	DEF	ה	DEF
	משטרה	NN	מש	NN
			##טר	NN
			##ה	NN
Word-Level	למשטרה	IN-DEF-NN	ל	IN-DEF-NN
			##משטרה	IN-DEF-NN
Word-Level Host	למשטרה	NN	ל	NN
			##משטרה	NN
Word-Level Prefix	למשטרה	IN-DEF	ל	IN-DEF
			##משטרה	IN-DEF
Decomposed	ל	IN	ל	IN
	ה	DEF	ה	DEF
	משטרה	NN	מש	NN
			##טר	NN
			##ה	NN
Decomposed Informed	למשטרה	IN-DEF-NN	ל	IN-DEF
			##משטרה	NN

Figure 2.4.: Labeled data crafted for model fine-tuning. Figure from Klein and Tsarfaty (2020).

They conclude that constraining word-pieces to reflect their morphological functionalities might enhance the results of models trained to predict multi-tags for word-pieces instead of complete words, boosting the performance of these models in the face of morphologically rich languages.

3. Theoretical Background

This chapter lays the foundation for the rest of this thesis. Since the experimental part revolves fundamentally around MT, morphology, and QE, the following sections introduce and explain these areas thoroughly.

3.1. Morphology

Although the concept of morphology is generally attributed to Goethe (1790), who coined the term with biological connotations, it was not until 1859 that the linguist Schleicher proposed the linguistic approach to morphology.

In the NLP pyramid (Figure 3.1), morphology is depicted as the foundation for the rest of linguistic subdisciplines, that is, syntax, semantics, and pragmatics. In particular, morphology focuses on the internal structure of words. Every word is composed by the following parts:

Stems Stems are theoretical constructs that stands for the unitary and shared syntactic properties of a group of word forms (Andreou, 2019). An example of a stem for an English would be *wait* for the verb *to wait*.

Morphemes Morphemes are the smallest elements of a word with grammatical function and meaning. Words can be formed by a combination of so-called bound morphemes or by a single free morpheme. Depending on the use of these morphemes, there are two main typologies for morphology:

- **Inflectional morphology:** This type of morphology conveys grammatical properties such as number or tense, and it is realized through the addition of bound morphemes to a particular stem. For instance, when forming the past tense of regular English verbs, such as *walk* or *paint*, a suffix *-ed* is appended to the word (*walked*, *painted*). In other languages such as Spanish, it is also common to have this sort of bound morpheme to express the future tense, however, this is not the case for English, which has a rather poor inflectional morphology.

  On the other hand, in German, nominal and adjective inflection are characterized by the addition of suffixes that express not only gender, but also case. For instance, *die schöne Blume* is the translation for *the beautiful flower*, however, if we were to speak about *the beautiful flowers*, we would have to add a suffix *-n* to both the adjective and the noun obtaining the following sentence *die schönen Blumen*. This is even more notorious when we handle morphological features such as case, which is nonexistent in English. The genitive case expresses possession as can be seen in *der schönen Blumen* in the phrase *die Farbe der schönen Blumen*. This phenomenon would be tackled in English by combining a preposition with a determinant obtaining the phrase *the color of the beautiful flowers.*. Likewise, whereas in English direct objects do not assume any special case, phrases turn into the accusative case in German as in *sie schenkt den schönen Blumen.*, which renders as *she gives away the beautiful flower.* in English.

- **Derivational morphology:** The process of creating a new word from an existing word is the task of the derivational morphology. The new derived word is related to its root, but has acquired a different part of speech, resulting in a semantic turn. In English, the suffix *-er* is appended to a verb to designate the agent carrying out the action described by the verb, as it is the case for *teacher*, *baker*, *dancer*, or *singer*. Similar to derivation is composition which is the process by which new words are created by combining meaningful stems as it would be the case with the German word for *umbrella*, *Regenschirm*. *Regen* means *rain*, whereas *Schirm* means *shield*.

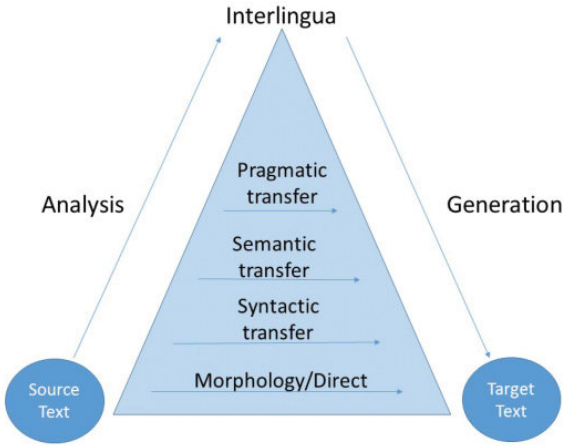

Figure 3.1.: Vaquois Triangle. Figure from Wiriyathammabhum et al. (2016).

3.2. Machine Translation

Considered as a sequence-to-sequence problem, MT aims at rendering a particular source language text into the desired target language without human involvement, and is regarded as a multilingual task within the classification of NLP tasks.

In both statistical and neural approaches to machine translation, the system parts from parallel corpora, in other words, parallel data corresponding to the source language and target language in question.

In the succeeding sections, both Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) will be further explained and illustrated.

3.2.1. Statistical Machine Translation

Despite the bunch of NMT projects undergone during the last wave of neural network research in the late 80s, which achieved similar results to the current NMT systems, the training data set sizes were not large enough to consider these models as robust MT systems (Koehn, 2020). This problem originated due to the lack of powerful computational resources and funding. Thus, during this period, known as AI winter, Phrase-Based SMT surged to fill this gap.

Statistical Machine Translation (SMT) systems analyze bilingual text corpora to generate a series of statistical models to find out the most probable translation for a particular phrase, i.e. a sequence of words with no linguistic motivation, given a statistical weight based on a series of grammars.

As proposed by Koehn, the SMT task can be split into multiple components, which are shown in Figure 3.2. We first deal with a so-called bilingual corpus on which some statistical processes occur, namely measuring the probability of certain words or phrases within the dataset. These statistical processes will also take place parallelly on the monolingual dataset. After this statistical analysis is finished, we obtain our translation and language model, which will be applied during the decoding process.

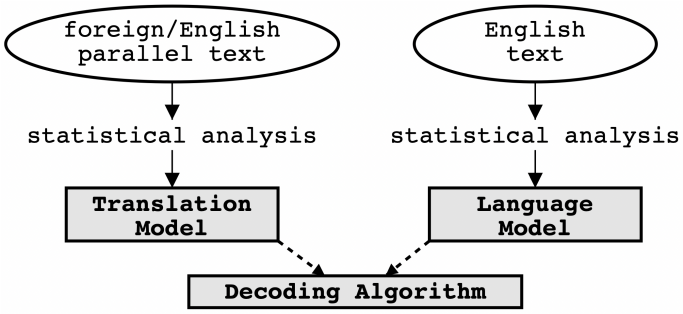

Figure 3.2.: Statistical Machine Translation Pipeline. Figure from Koehn (2009).

3.2.2. Neural Machine Translation

The renaissance of neural model research supposed a turning point for MT. After a period in which NMT could not compete against the already existing Phrase-Based SMT due to its computational complexity, the situation changed in 2015. At this point, the improved recurrent neural network encoder-decoder model (Sutskever et al., 2014; Cho et al., 2014) was combined with the attention mechanism (Bahdanau et al., 2014) (Bentivogli et al., 2016). Furthermore, Devlin's (2014) joint language and translation model showed great improvements with respect to the attempts at neural translation in the 90s, as well as the statistical approaches used at the time.

Similar to SMT, the objective of an NMT system is to predict the likelihood of a sequence of words and capture them within a language model, but applying artificial neural networks instead of statistical models. Usually, NMT systems are represented by an encoder-decoder model, as depicted in Figure 3.3. Both the encoder and decoder part of the mechanism are constituted by a Recurrent Neural Network (RNN) each, generally used for sequence-to-sequence problems. However, in recent years, the focus has been shifted to the Transformer architecture and the Attention mechanism. In the Encoder-Decoder model, the former encodes a given source sentence into a series of representations passed onto the context vector, which will be then decoded by the latter, delivering the output for the target language.

3.2.3. Evaluation Metrics

The quantitative evaluation of MT systems is paramount to gauge how good a system is performing in comparison to another, and make amendments accordingly, or to assess whether a change in the system has led to better results (Koehn, 2009).

Since there are many possible translations for one sentence and hence not a single right answer as it may be the case in fields such as ASR, MT outputs are challenging to assess. Essentially, for a MT system to be successful, its outputs should be as fluent and adequate as possible. Fluency involves grammatical correctness and idiomatic word choices (Koehn, 2009) so that it fits the pragmatic conventions of the target language, whereas adequacy

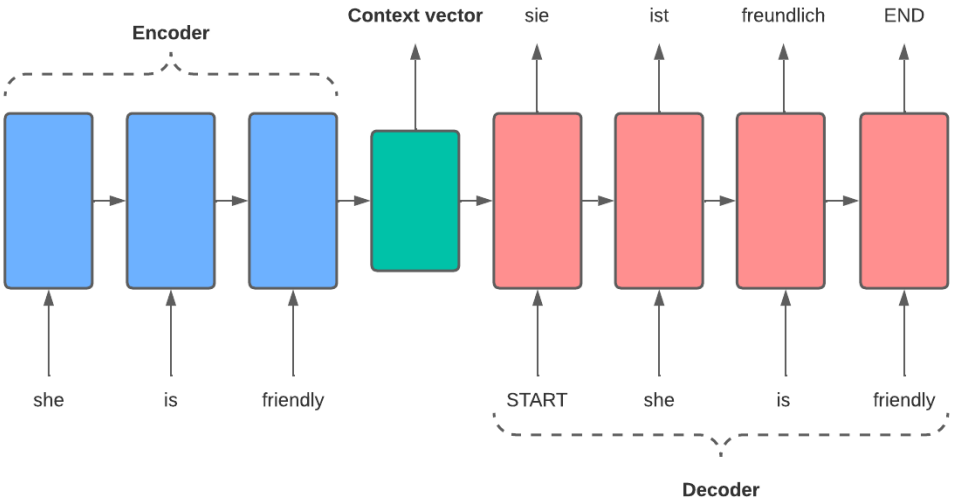

Figure 3.3.: Encoder-Decoder model for NMT.

is concerned with the conveyance of the meaning of the source language in the translated output.

Manual Evaluation

One method for evaluating a MT output is to classify whether it is correct or not with the help of bilingual evaluators or, if not available, monolingual evaluators and a reference translation.

Due to the broadness of correctness as a metric, fluency and adequacy are also part of the manual evaluation picture. When checking how fluent the rendition of the target text is, evaluators apply a scale ranging between 'incomprehensible' and 'fluent'. Table 3.1 elicits an evaluation tool with the definitions of adequacy and fluency, which human annotators can follow as a guide when assessing the translation quality.

This approach only involves the target-side text, removing the need for the evaluator to be proficient in both languages. In the case of adequacy evaluation, however, the evaluator is required to have good command of the source and target language since they rate a text based on whether all the meaning is retained in the translation.

	Adequacy	**Fluency**
5	all meaning	flawless target language
4	most meaning	good target language
3	much meaning	non-native target language
2	little meaning	disfluent target language
1	none	incomprehensible

Table 3.1.: Scheme for the manual evaluation of adequacy and fluency. Table adapted from Koehn (2009).

Automatic Evaluation

Overtly, the use of human evaluators to assess the quality of a particular text, either human-translated or machine-translated, carries a subjective component with it, which

results in uneven levels when it comes to intra-rater and inter-rater agreement [1]. This along the lack of standardised manual evaluation metrics has accelerated the search for more automatised methods.

As with any other NLP task, the automatic performance evaluation of a MT system relies mostly on the comparison with human-annotated golden data (Yvon, 2019). Some of the most popular automatic evaluation metrics are:

Bilingual Evaluation Understudy (BLEU) Introduced by Papineni et al. (2002), this automatic metric leverages how similar a hypothesised translation is to a reference translation. The more words and/or sentences the hypothesis shares with its reference, the higher the BLEU score.

Translation Edit Rate (TER) TER (Snover et al., 2006) appears as an alternative metric to BLEU. It is a character-based approach that measures the edit distance between a machine translated output and the reference translation, thus, it is commonly applied to assess post-editing effort. For Human-targeted Translation Edit Rate (HTER) post-editing of the system output is required (Snover et al., 2006).

Metric for Evaluation of Translation with Explicit Ordering (METEOR) This automatic metric proposed by Banerjee and Lavie (2005) combines precision and recall obtaining high correlation with human judgement. Since it allows multiple reference translations, the problem of flexibility in word matching is addressed. This allows the system to regard morphological variants and synonyms as feasible matches.

3.2.4. Quality Estimation

As NLP tasks become more sophisticated, so do their evaluation metrics. Unlike the previously mentioned automatic metrics, the following technique uses Machine Learning (ML) to assign quality scores to machine-translated segments. The dataset used for our experiments stem from a shared task on Quality Estimation (QE), which is exhaustively introduced in Section 5.1.

Automatic Quality Estimation (AQE) Traditional evaluation techniques such as BLEU or WER for machine translation have mostly relied on human-annotated data to leverage the quality of the outputs of a given MT. As opposed to the aforementioned techniques, AQE does not require any gold data to make qualitative predictions instead "[it estimates] the quality of a system's output for a given input, without any information about the expected output" (Specia et al., 2009). AQE can occur at different levels:

- Word-level: At this level, the metric predicts whether the word has been translated correctly or whether it is missing.
- Phrase-level: Once word-level AQE is carried out, phrase-level QE assesses the overall quality of translated phrases.
- Sentence-level: This approach makes use of TER or HTER scores to analyse that amount of words that have to be amended for the translation to be equal to its gold reference.
- Document-level: At document-level, AQE is designed to rate machine-translated text with regards to the type and severity of the errors that it may contain.

[1] Intra-rater agreement denotes the level of consistency in the evaluations of a particular rater, and inter-agreement, between raters.

Furthermore, before AQE gained momentum, it was a popular manual evaluation metric, alternative to the fluency and adequacy frameworks described before. The datasets involved in the experiments of this thesis includes AQE data on the word-level.

Manual Quality Estimation (MQE) On the other side of the spectrum, evaluators can also resort to identify issues in the target text. This is a language-dependent approach with which evaluators look for missing words, part of speech, incorrect word order, or added words, among others, in the machine-translated text.

3.3. Morphological Challenges in Natural Language Processing

As recurrently mentioned throughout the course of this thesis, morphology is a complex topic in itself. Each language belongs to a different morphological group, and even when pertaining to the same morphological typology, two languages can have slightly different morphological rules as it might be the case of Spanish and Portuguese. Despite both being fusional languages, their grammatical features vary dramatically. This means that there is no one-size-fits-all sort of solution for morphological issues in NLP. This is exacerbated by the variability of morphological rules and irregularities across langauges.

Inflectional morphology poses a challenge for NLP tasks since it makes instances of the same stem appear to be different words. For instance, the present form of the verb *to walk*, *walk*, would not be related to the past form of the same verb, *walked*. This increases sparcity and is a problem in information extraction and retrieval.

When it comes to morphology in MT, there are recurring issues in both translation sides of the spectrum (Burlot and Yvon, 2017):

- Source side: Morphological variation on the source leads to a higher appearance of Out-Of-Vocabulary (OOV) source tokens due to flexible word ordering (Bisazza and Federico, 2016), which increases the translation difficulty.
- Target side: Morphological variation on the target side forces the MT system to generate forms unseen during training.

In general, unseen word forms are difficult to generate due to data sparsity, and even with the recent developments in the NMT field, which include Byte Pair Encoding (BPE) splitting (Bollmann and Søgaard, 2021), there is still much work to be carried out. The paper by Klein and Tsarfaty (2020) presented in Chapter 2 offers an insight into word splitting in the context of morphologically rich languages. From their work, we can draw upon the conclusion that inferring linguistic context to such tokenization approaches can boost the performance of NLP systems.

On the other side of the picture, if sparsity is actually not the main issue, due to the appearance of all necessary tokens in the dataset, then the system has to decide what to output given the target-side context and the source-side input. As a matter of fact, when translating the sentence *the dog eats the apple.* into German, the system will have to determine the right conjugation, number and case for the words that are to be inflected. If the system works, it will output *der Hund isst den Apfel.*. This morphological interaction between source input and target output is to be examined in the subsequent chapters of this thesis.

4. Methodology

With the key concepts at hand, we now explain the methods that will serve us as a foundation to obtain the data in the required format for the addition of source-side linguistic features to the target side in Chapter 5. Furthermore, the different shared tasks explored by Bollmann and Søgaard in their paper will be briefly introduced along with the morphological and control features, which are indispensable for Chapter 6.

The steps that we present next are adopted from Bollmann and Søgaard's paper. The objective is to evaluate the role of morphology in the errors committed by an NLP system using a series of machine learning-based random forest classifiers. In the following, we will explain every step of the pipeline, from data crawling to the training of the system, including a brief introduction to the aforementioned classifier[1]. These procedures, implemented originally by Bollmann and Søgaard (2021), will be replicated in our experiments in Chapter 6. Figure 4.1 depicts the pipeline summarizing all steps followed to extract and analyze the data, including the step designed for the experimental part of this thesis, the source-side feature extraction.

4.1. Data Preparation

For the experimental setup of their paper, Bollmann and Søgaard (2021) gathered data from the four different shared tasks presented in Chapter 2. All the datasets used for these tasks include system outputs, as well as their gold annotations. It is noteworthy that the authors picked the system outputs from the MT shared task specifically for its gold annotations. These gold data provide us with word-level error labels for the MT outputs, which can be seen along with the morphological and control features of the word in question once the morphological features are extracted at a later step. All of these facts will be analyzed in-depth in Section 5.1. Furthermore, despite it being a shared task on MT, where both the source and target language depend on each other, the source-side and target-side datasets are analyzed separately in the original paper. Thus, we deem necessary to analyze the interaction between both languages. For this purpose, we extract the linguistic features from the source side and transfer them to the existing features on the target side. Specifically, we extract morphological and lexical features.

4.1.1. Data Crawling

Firstly, it is paramount to extract the data that will be used as a base for the system. This occurs by downloading the datasets from their respective websites. The repository for the paper has already some predefined shell files with the links to the websites for each of the shared tasks.

4.1.2. Extraction and Addition of Features

Once the datasets have been crawled from their corresponding websites, they need to be tagged morphological and lexically. UDPipe and UDLex models are commissioned to carry out this task. A more thorough description of these models in the Chapter 5 of this

[1] For an extended description of the data used for this application, consult https://github.com/coastalcph/eacl2021-morpherror

thesis. The output of these models will be a series of CoNLL files, where the last column contains a series of miscellaneous features, along with annotations of the errors found in the dataset. These miscellaneous features offer information about the lexicality (Section 5.2.2), length and frequency of a word, among others.

After tagging the datasets morphological and lexically, as well as having added string-based features to every token in the sentences, the relative importance of all the features extracted in the previous step will be gauged. At this stage, a series of control features that are not morphologically-motivated will be generated. The goal of incorporating both linguistic and non-linguistic features is to evaluate the predictiveness of errors of individual features, represented by Feature Importance scores.

4.2. Data Analysis

The pipeline of the methodology for the baseline system culminates when the classifiers are run along with the pre-processed data. The program outputs accuracy and F_1 scores for the classifiers. These are used as a measure of the performance of the classifiers at the time of predicting errors. Alongside these scores, there are FI scores which represent how predictive of error a particular feature is. The FI score for every feature category is calculated by retraining the classifiers without the features for that category.

Random Forest Classifier Classification is a supervised task for which labelled data is required to make predictions on new data. Having Decision Trees (Quinlan, 1986) as a foundation, Random Forest classifiers (Breiman, 2001) is an ensemble machine learning algorithm. The training data is fed to various decision trees simultaneously, which select features randomly for the splitting of the nodes. In the context of this thesis, random forest classifiers will provide us with the necessary metrics to evaluate the results of our experiments in Chapter 6. The results of classifiers are evaluated via stratified 5-fold-cross-validation.

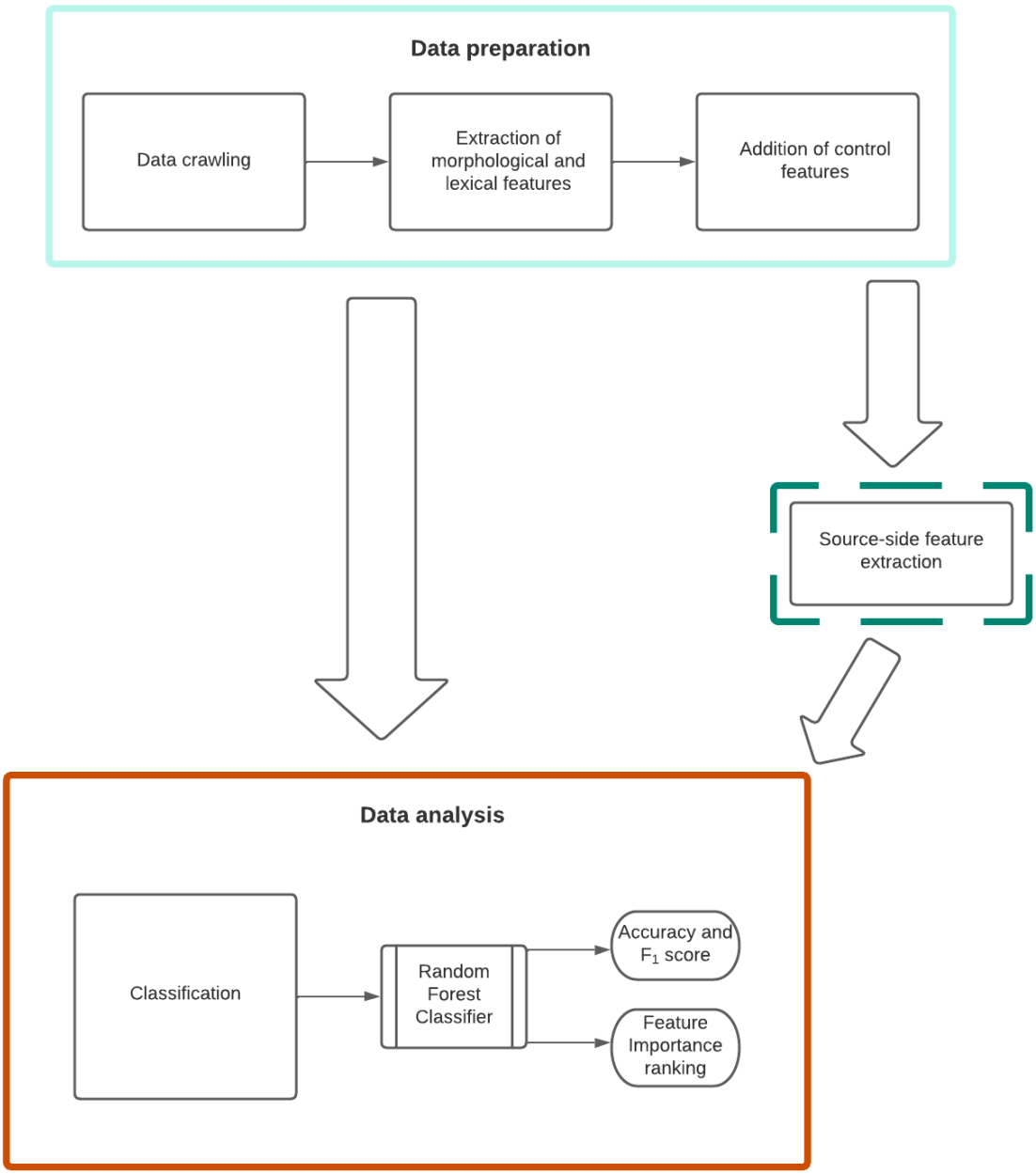

Figure 4.1.: Methodology pipeline. The source-side feature extraction is further explained in Section 5.5.1.

5. Data and Features

Having presented the methods for the experimental part of this thesis and before introducing such experiments, we will provide an insight into the data and packages used. Next, we will deep dive into a detailed insight of the MT dataset, as it is in the foreground of this thesis.

5.1. The WMT19 Dataset

In this thesis, we examine the influence of morphological source-side features in the output of a MT system, that is, the target side. Thus, we will be shifting our attention to the AQE shared task. For this purpose, Bollmann and Søgaard (2021) chose the data from the AQE shared task organized by the Workshop on Machine Translation in 2019 (WMT19). These datasets were translated using proprietary MT engines, and contain jargon from the IT-domain, including Amazon reviews for the English-German combination, and excerpts from the Microsoft Office software for the English-Russian combination. All these datasets are obtainable on the website[1] for the shared task in the English-German and English-Russian language pairs. The folders corresponding to the training and development data[2] for this shared task and language combinations contain the following set of files:

- file.hter: File containing the HTER (Snover et al., 2006) for every single sentence.
- file.mt: File storing the tokenized sentences resulting from the MT system assigned to the shared task.
- file.pe: File with tokenized post-edited target-side sentences from the mt file. They serve as gold data for the MT system.
- file.source_tags: File containing OK and BAD tags with respect to the source sentence. Tokens are tagged as OK if they were correctly translated, and BAD otherwise. Gaps, in other words, omissions, are not tagged.
- file.src: File storing tokenized source-side sentences.
- file.src-mt.alignments: Each line in this file aligns a word from the source sentence with the respective word in the target sentence. This file is used during the experimental part of this thesis to ensure a correct alignment between the source and target side.
- file.tags: File containing OK and BAD as output tags for each token. In the website WMT20 shared task on QE[3], it is stated that each gap between two words is tagged as BAD if one or more missing words should have been there, and OK otherwise. Note that number of tags for each target sentence is 2*N+1, where N is the number of tokens in the sentence.

Figure 5.1 provides us with an overview of the contents of these files for a particular sentence.

[1] https://www.statmt.org/wmt19/qe-task.html
[2] We only use the training data for our experiments because the amount of development data is not as representative, and they are not used to optimize the classifier.
[3] https://www.statmt.org/wmt20/quality-estimation-task.html

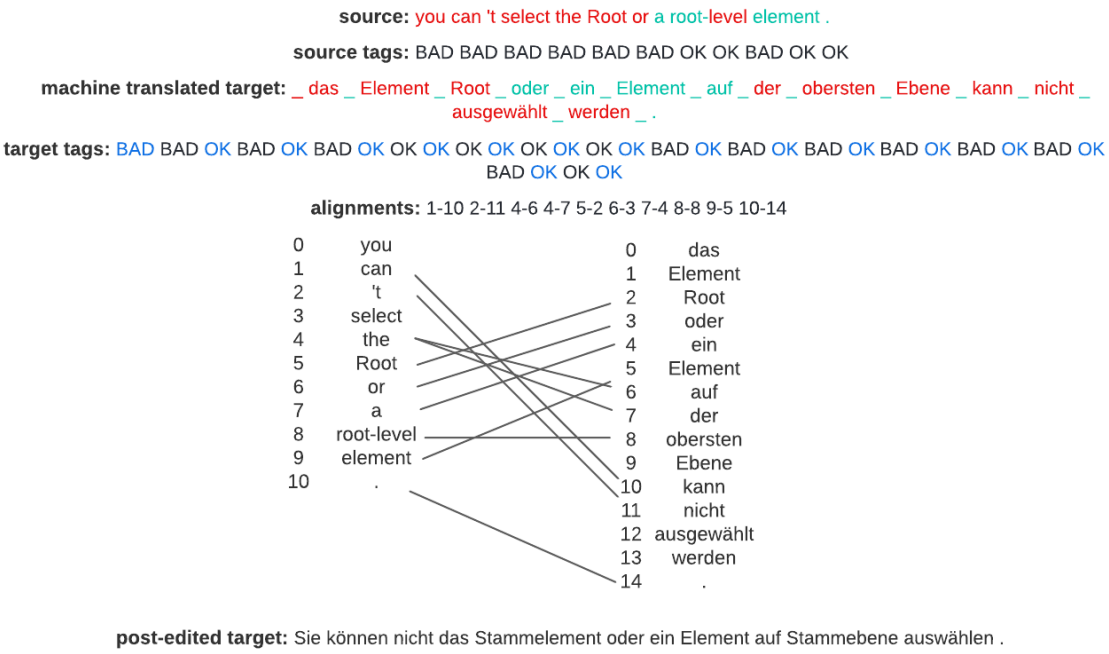

Figure 5.1.: Overview of the file contents for WMT19 dataset.

The source sentence is on the top followed by its corresponding tags for each token in the sentence on the bottom. The same applies to the target sentence. Target tags also include gap before tags, which indicate whether there has been omission in the output. Gap before tags are marked in blue and normal tags in black. If the output of the MT system was the expected one, the token is highlighted in green. Otherwise, the token is red. The human post-edited output corresponding to the reference translation is on the bottom of the figure. The gap before tags, however, are not employed for the experimental setup of this thesis as they describe omission errors, and it is unclear to which tokens these omissions should be ascribed to (Bollmann and Søgaard, 2021).

5.2. Morphological Features

Since morphological features play a crucial role in the experimental part of this thesis, in the following sections, we provide an insight into the morphological features as proposed by Bollmann and Søgaard, as well as the tools used to extract them.

It is noteworthy that the machine translated data from WMT19 are not completely correct, as we could see in the example in Figure 5.1, which make them not ideal for morphological analysis. The post-edited data would provide us with more successful outputs, but it is out of the objective of this thesis since we want to evaluate the outputs of an MT system.

5.2.1. Universal Dependencies

Universal Dependencies (UD) provides a framework of morphosyntactic rules that can be applied in a cross-lingual fashion, that is, grammatical notions are indicated by the manipulation of word forms or through dependency relations. In the context of UD, morphology has three levels of representations for a syntactic word[4]:

- Lemmata: Responsible for the semantics of the word.

[4]This information is available on the official website for Universal Dependencies: https://universaldependencies.org/u/overview/morphology.html

- Part of Speech (POS) tags: Representing the abstract lexical category of the word. It can vary depending on the role of the word within the syntax of a sentence, especially for words that can assume different POS tags such as *bank*, which can be both a verb and a noun.[5]
- Lexical and grammatical features corresponding to a particular word.

5.2.2. Lexical Features

Lexical features are responsible for providing information about the importance of a token within a dataset with syncretic attributes (Bollmann and Søgaard, 2021). Additionally, these features can also determine to what extent the POS tag of the token can be based on context. For instance, the English word *book* could either be a verb or a noun. The same applies to the ambiguity of the lexemes in a token, e.g., *ruling* is an inflection of the verb *to rule* or it could be a single free morpheme, *the ruling*. Table 5.1 summarize all possible lexical features for a particular dataset.

Feature name	Definition
+posambig	The POS tag has between 1 and 4 interpretations.
-posambig	The POS tag has one interpretation.
ˆposambig	The POS tag has more than 4 interpretations.
+lexambig	There are between 1 and 4 lemmata for the token.
-lexambig	There is 1 lemma for the token.
ˆlexambig	There are more than 4 lemmata for the token.
+syncretic	The token is representative for 1 to 4 morphological feature sets.
-syncretic	The token is representative for 1 morphological feature set.
ˆsyncretic	The token is representative for 4 morphological feature sets.

Table 5.1.: Summary of the different lexical features available. Adapted from Bollmann and Søgaard (2021).

In Figure 5.2, we can observe some examples for the lexical features presented in Table 5.1. For instance, the token *stall* is considered a *+posambig* POS tag since it can either be a verb or a noun. Something similar occurs with the token *handling*, which can adopt two different POS tags. Apart from this, it is also linked to the *+lexambig* property, which means that the word belongs to different lemmata. *handling* can either be inflected from the verb *to handle* or from the noun *handle*. As for syncretism, the token *work* can be relevant for different morphological classes. As a verb, it can either express imperativeness, as in *Work now!* or indicativeness *I work now*.

5.2.3. String-based Features

Finally, the above mentioned features are complemented by a series of string-based features, which compare the token with its lemma to compute how different they are from each other. These features show whether the lemma has been changed in the beginning of the lemma, *EDIT=PRE*, or at the end, *EDIT=SUF*. Likewise, the sequence could have been altered in the middle, and this is expressed by the *EDIT=IN* feature. On the contrary, *EDIT=FULL* means that there have not been any substitutions with respect to the lemma.

[5]Appendix A offers a table with different universal POS tags as a refresher for the interest reader.

Feature name	Definition
edit=edit_pre	Token has been modified in the beginning of the sequence with a prefix.
edit=edit_post	Token has been modified at the end of the sequence with a suffix.
edit=edit_in	Token has been modified in the middle of the sequence with an infix.

Table 5.2.: Summary of the different string-based features available.

Coming back to the example in the previous section, *handling* is labeled with the *edit=edit_post* property. *handling* stems from the lemma *handl-*, and *edit_post* essentially implies that the a suffix has been added to the lemma, in the case of *handling*, the inflectional morpheme *-ing*.

5.3. Control Features

Aside the morphological features, control features, which are not linguistically motivated, are put into place to assess the relative importance of the morphological features. A comprehensive listing of all morphological and control features is available in the Appendix B of this thesis. These features are not extracted using tools as it is the case for the morphological features, but applying hand-crafted features.

5.3.1. String Length Features

Every token in the dataset gets mapped to a single string length feature, which indicates the length of the token in question. This assignment occurs solely during the analysis of the CoNLL file by the random forest classifier, and the results are to be seen after running the analysis. These results will be illustrated in Chapter 6.

5.3.2. Token Frequency Bins

Similar to string length features, token frequency bin features denote how often a token occurs within the dataset. While the tokens in the target sentence in Figure 5.3 are rather uniform regarding this aspect, obtaining most of them either *freq=99* or *freq=rare*, the tokens in the source sentence in Figure 5.2 are more diverse, being linked to other frequency features.

5.4. Tools for Feature Extraction

The different morphological features analyzed in the work of Bollmann and Søgaard are extracted using various tools. Table 5.3 offers a better visualization of which tools are applied for the extraction of these features.

Morphological features	Tool
Universal Dependencies	UDPipe
Lexical features	UDLexicons or UniMorph
String-based features	EdLib

Table 5.3.: Summary of the different string-based features available.

UDPipe

Devised by the Institute of Formal and Applied Linguistics at the Charles University in Prag, UDPipe (Straka and Straková, 2017) offers a language-agnostic trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL files. Essentially, a CoNLL file is equivalent to a comma-separated values (CSV) file, which requires 10 columns [6].

UDPipe models are put in place as morphological taggers, and provide the words in the CoNLL files with their corresponding UD (Section 5.2). Since we are focusing on machine translation data, it is worth mentioning that the morphological annotation for these datasets might not be as perfect as that for monolingual NLP tasks.

UDLexicons and UniMorph

UDLexicons (Sagot, 2018) were designed to complement UD in the context of CoNLL. This tool provides coverage for a total of 38 languages with a collection of 53 morphological lexicons. If a language is not covered by UDLexicons, the script resorts to UniMorph (Kirov et al., 2018) [7].

UniMorph covers 167 languages, and has the goal of annotating the lexical meaning for an inflected word in any language supported by UniMorph. Depending on how often the token has appeared in the dataset, ambiguity or syncretism of a POS or a lexem can accompanied by different symbols.

Edlib

EdLib (Šošić and Šikić, 2017) is a Python library designed for the extraction of string based features. The string alignment is calculated by applying the Levenshtein algorithm integrated within the EdLib Python package.

[6] Each of the columns represent Id, Form, Lemma, UPosTag, XPosTag, Feats, Head, DepRel, Deps, and Misc, respectively.

[7] See Sagot (2018) for more information and explained difference with UniMorph.

```
# sent_id = 72

1  however  however  ADV  RB  _  _  _  _  freq=rare|tag=BAD
2  ,  ,  PUNCT  ,  _  _  _  _  freq=p99|tag=BAD
3  without  without  ADP  IN  _  _  _  _  freq=gt3|lex=+posambig,-lexambig,-syncretic|tag=BAD
4  error  error  NOUN  NN  Number=Sing  _  _  _  freq=gt3|lex=-lexambig,-posambig,-syncretic|tag=OK
5  handling  handle  NOUN  NN  Number=Sing  _  _  _  edit=edit_post|freq=p90|lex=+lexambig,+posambig,+syncretic|tag=OK
6  ,  ,  PUNCT  ,  _  _  _  _  freq=p99|tag=OK
7  an  a  DET  DT  Definite=Ind|PronType=Art  _  _  _  edit=edit_post|freq=p99|lex=+posambig,-lexambig,-syncretic|tag=BAD
8  application  application  NOUN  NN  Number=Sing  _  _  _  freq=p99|lex=-lexambig,-posambig,-syncretic|tag=BAD
9  may  may  AUX  MD  VerbForm=Fin  _  _  _  freq=gt3|lex=+posambig,+syncretic,-lexambig|tag=BAD
10  easily  easily  ADV  RB  _  _  _  _  freq=rare|lex=-lexambig,-posambig,-syncretic|tag=BAD
11  stall  stall  VERB  VB  VerbForm=Inf  _  _  _  freq=gt3|lex=+posambig,-lexambig,^syncretic|tag=BAD
12  or  or  CCONJ  CC  _  _  _  _  freq=p95|lex=+posambig,+syncretic,-lexambig|tag=BAD
13  frustrate  frustrate  VERB  VB  VerbForm=Inf  _  _  _  freq=rare|lex=+posambig,-lexambig,^syncretic|tag=BAD
14  the  the  DET  DT  Definite=Def|PronType=Art  _  _  _  freq=p99|lex=+posambig,-lexambig,-syncretic|tag=OK
15  user  user  NOUN  NN  Number=Sing  _  _  _  freq=p99|lex=-lexambig,-posambig,-syncretic|tag=BAD
16  if  if  SCONJ  IN  _  _  _  _  freq=gt3|lex=+posambig,-lexambig,-syncretic|tag=OK
17  something  something  PRON  NN  Number=Sing  _  _  _  freq=rare|lex=+posambig,-lexambig,-syncretic|tag=BAD
18  doesn  doesn  VERB  VBP  Mood=Ind|Tense=Pres|VerbForm=Fin  _  _  _  freq=rare|tag=BAD
19  't  't  PART  RB  _  _  _  _  freq=rare|tag=OK
20  work  work  VERB  VBP  Mood=Ind|Tense=Pres|VerbForm=Fin  _  _  _  freq=gt3|lex=+posambig,-lexambig,^syncretic|tag=BAD
21  as  as  ADP  IN  _  _  _  _  freq=p95|lex=+lexambig,+syncretic,^posambig|tag=OK
22  expected  expect  VERB  VBN  Tense=Past|VerbForm=Part  _  _  _  edit=edit_post|freq=rare|lex=+lexambig,+posambig,^syncretic|tag=OK
23  .  .  PUNCT  .  _  _  _  _  freq=p99|tag=OK
```

Figure 5.2.: Source sentence in CoNLL format.

```
# sent_id = 72

1  bei bei  ADP  APPR AdpType=Prep|Case=Dat _  _  _  freq=p99|gap_before_tag=OK|gap_tag=OK|tag=BAD
2  der der  DET  ART  Case=Dat|Gender=Fem|Number=Sing|PronType=Art _  _  _  freq=p99|gap_tag=OK|lex=+posambig,-lexambig,^syncretic|tag=BAD
3  Fehlerverarbeitung Fehlerverarbeitung NOUNNNGender=Fem|Number=Sing|Person=3  _  _  _  freq=rare|gap_tag=OK|tag=OK
4  kann können AUX  VMFIN  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|VerbType=Mod  _  _  _
   edit=edit_in,edit_post|freq=p99|gap_tag=BAD|lex=+syncretic,-lexambig,-posambig|tag=OK
5  eine  eine  DET  ART  Case=Nom|Gender=Fem|Number=Sing|PronType=Art  _  _  _  freq=p99|gap_tag=OK|lex=+syncretic,-lexambig,-posambig|tag=BAD
6  Anwendung Anwendung NOUNNNGender=Fem|Number=Sing|Person=3  _  _  _  freq=p99|gap_tag=OK|lex=+syncretic,-lexambig,-posambig|tag=BAD
7  den  den  DET  ART  Case=Acc|Gender=Masc|Number=Sing|PronType=Art  _  _  _  freq=p99|gap_tag=OK|lex=+posambig,+syncretic,-lexambig|tag=BAD
8  Benutzer  Benutzer  NOUNNNGender=Masc|Number=Sing|Person=3 _  _  _  freq=p99|gap_tag=OK|lex=-lexambig,-posambig,^syncretic|tag=BAD
9  jedochjedochADV  ADV  _  _  _  _  freq=p99|gap_tag=OK|tag=BAD
10 leicht  leicht  ADJ  ADJD Degree=Pos|Variant=Short  _  _  _  freq=p99|gap_tag=OK|tag=BAD
11 abfangen abfangen VERB VVINFVerbForm=Inf  _  _  _  freq=p90|gap_tag=OK|tag=BAD
12 oder  oder  CCONJ  KON  _  _  _  _  freq=p99|gap_tag=OK|tag=BAD
13 tftraut tftrauen  VERB VVFINMood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin _  _  _  edit=edit_post|freq=rare|gap_tag=OK|tag=BAD
14 ,  ,  PUNCT  $,  PunctType=Comm  _  _  _  freq=p99|gap_tag=OK|tag=OK
15 wenn  wenn  SCONJ  KOUS _  _  _  _  freq=p99|gap_tag=OK|tag=BAD
16 etwas etwas  PRON  PIS  Gender=Neut|Number=Sing|Person=3|PronType=Ind,Neg,Tot  _  _  _
   freq=p99|gap_tag=OK|lex=+posambig,-lexambig,^syncretic|tag=BAD
17 nicht  nicht  PART PTKNEG Polarity=Neg _  _  _  freq=p99|gap_tag=OK|lex=+posambig,-lexambig,-syncretic|tag=OK
18 wie  wie  CCONJ  KOKOM  ConjType=Comp  _  _  _  freq=p99|gap_tag=OK|lex=+posambig,-lexambig,-syncretic|tag=OK
19 erwartet  erwarten  ADJ  ADJD Degree=Pos|Variant=Short  _  _  _  edit=edit_post|freq=p99|gap_tag=OK|lex=-lexambig,-posambig,^syncretic|tag=OK
20 funktioniert  funktionieren VERB VVFINMood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin _  _  _
   edit=edit_post|freq=p99|gap_tag=OK|lex=+syncretic,-lexambig,-posambig|tag=BAD
21 .  .  PUNCT  $.  PunctType=Peri _  _  _  freq=p99|gap_tag=OK|tag=OK
```

Figure 5.3.: Target sentence in CoNLL format.

5.5. Source-side Feature Extraction

While Bollmann and Søgaard's paper offers a wide insight into the first three shared tasks (Section 2.1), we find that the MT shared task was not as explored with many questions to be answered. In the following sections, we will provide a more thorough analysis of how we process the WMT19 datasets by combining the morphological features from the source side with their corresponding aligned segments on the target side before they are used as an input for the random forest classifiers.

The central task at the stage of extracting the source-side features is to map every line in the target file with its corresponding source POS tags, morphological and lexical features so that we can analyze to which extent the source-side features interact with those in the target side. For this matter, pre-processing the datasets will account for a substantial part of this thesis [8].

Figures 5.2 and 5.3 show an example of how the CoNLL format looks like. The relevant features for our experiments are highlighted in different colors. From the source side, the features marked with the following colors are extracted:

- Blue: POS tags.
- Green: Lexical features.

These features are added to the target side for which the next features are particularly relevant for the error analysis:

- Blue: POS tags and morphological features for the POS tag in question.
- Orange: Frequency and string-based features.
- Green: Lexical features.

In Section 5.5.1, we present a sample target sentence of the dataset with the additional source-side features.

5.5.1. Data Pre-processing

Since the objective of this work is to evaluate the role of morphological features on the target output of a machine translation system, we want to match the source morphological features of a particular source word with its corresponding aligned word. Figure 5.4 shows a pipeline comprising the steps followed during the data pre-processing stage.

Extracting Alignments

Given the source and target files and their corresponding alignment files, we proceed to pre-process our data so that we can obtain the desired results when inserting the data into the classifier. At a first step, the alignments in the corresponding alignment file are extracted and stored in a dictionary where each key represents the sentence number and the values are tuples containing the alignments in the form of lists, where the digit to the left corresponds to the index of the source word and the digit on the right to that of the target word. Figure 5.5 illustrates this process for a given source and target sentence.

[8] The Python script designed for the source-side feature extraction is available in the `experiments_lin` folder in the CD.

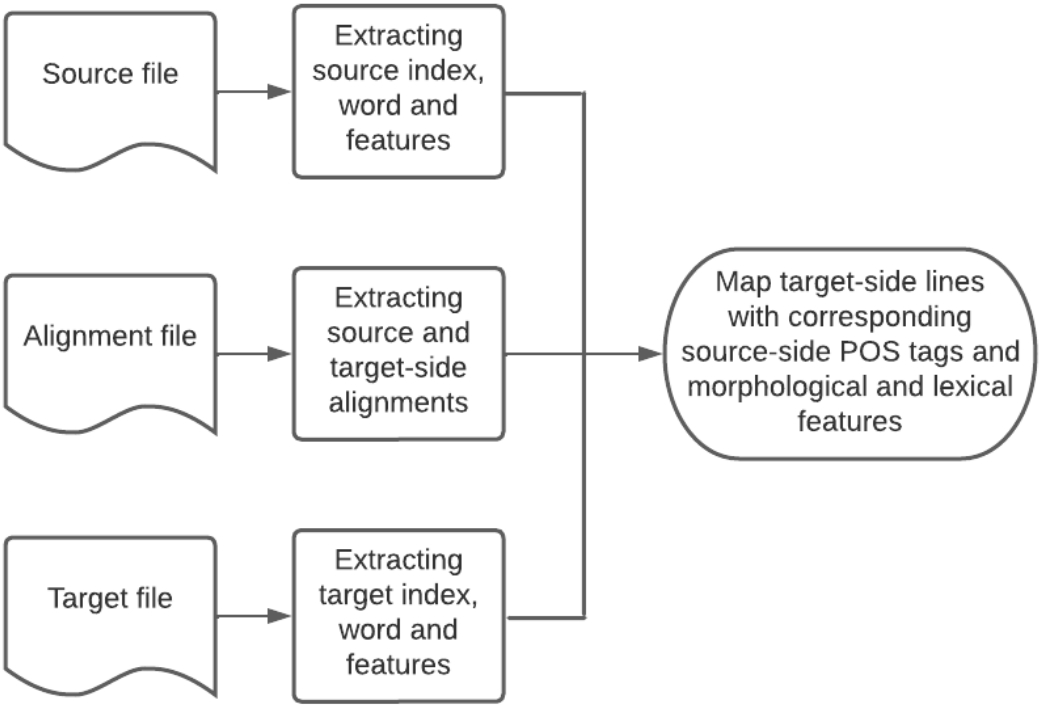

Figure 5.4.: Feature extraction pipeline

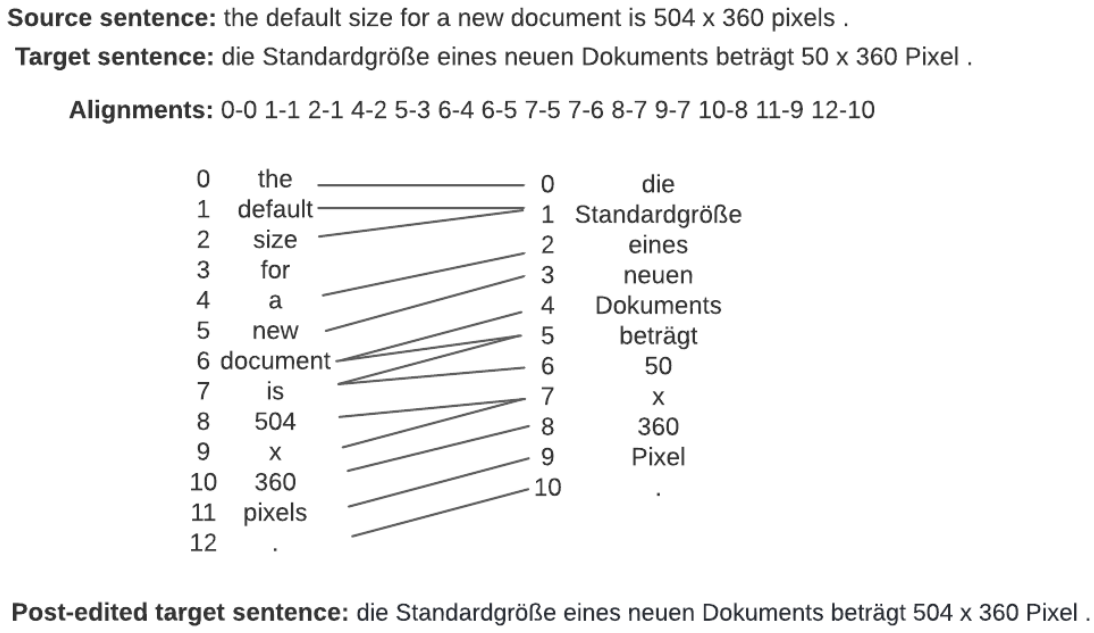

Figure 5.5.: Overview into alignment pipeline.

Extracting Source-side and Target-side Words and Features

The second step of the pre-processing is to extract the words, along with their morphological and lexical features for every source and target sentence in the CoNLL files as presented in Section 5.

Aligning Source-side and Target-side Words and Features

After extracting the source-side and target-side information for each sentence from the CoNLL files, we proceed to align the source-side words and features with the corresponding target words and features based on the aforementioned alignment dictionary. In some cases, there are multi-alignments, that is, a source word is mapped to several target words, and vice versa. In other cases, there are words with no alignments. We handle the prior by pairing the word with all its possible combinations as it is the case of *of the*, which is treated as the genitive *des* in German.

Figure 5.6 illustrates every possible mapping in a decision tree fashion, while Table 5.4 exemplifies these possibilities with output samples for the German sentence *in früheren Versionen wurde ein Computerprogramm als eine Abfolge von Schritten oder Anweisungen definiert , die der Computer ausführt*, which is aligned to the English *earlier we defined a computer program as a series of steps or instructions that the computer performs ..* If available, the target word is mapped to all its corresponding source-side POS tag and lexical features. Otherwise, the target word is matched with an unknown token, i.e. *UNK*. Adding unknown tokens to those words with no aligned words is important for us to obtain a new dataset with sentences that have the same length as those in the original dataset.

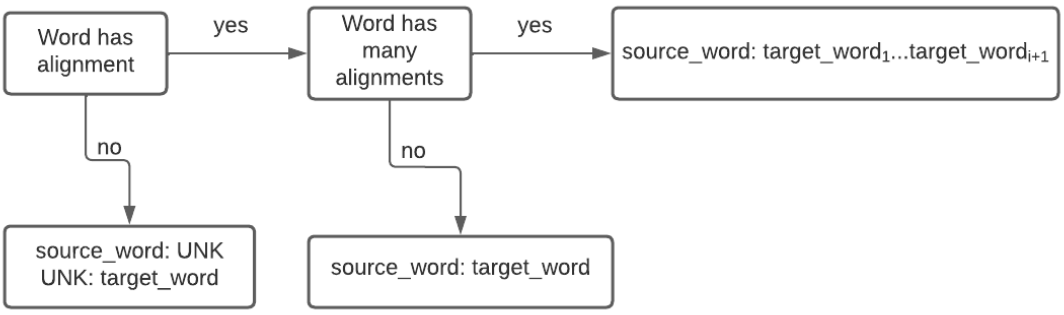

Figure 5.6.: Overview into different alignment types.

Alignment type	Target Word	Source-side Features
1-to-1 alignment	Versionen	en_tag=ADV en_lex=+lexambig,+posambig,+syncretic
Frequent multi-alignment	von	en_tag=NOUN-ADP en_lex=-lexambig,-posambig,-syncretic
Infrequent multi-alignment	früheren	en_tag=MULTI en_lex=+lexambig,+posambig,+syncretic
Unknown alignment	wurde	en_tag=UNK

Table 5.4.: Sample outputs for different alignment from the Python script designed for source-side feature extraction.

In the case of multi-alignments, we take the corresponding word and map it to all its possible alignments. This is crucial to guarantee the accuracy of the results in the experiment. Furthermore, there is no word being aligned to more than four words simultaneously. If there is a multi-alignment occurring less frequently in a dataset, it is simplified to a *MULTI* token to reduce noise during the classification step of the ex-

periment. This threshold varies from language to language. For instance, the threshold for the Russian dataset is set to 99, whereas for the German, this number ascends to 400, the reason being that there are 598 different multi-alignment combinations for the Russian training dataset, and 830 for the German training dataset. Tables 5.5 and 5.6 indicate the most common multi-alignment combinations for both language combinations. For tags appearing more frequently, we concatenate them by a '-'. Lexical features are also extracted. In the case of multi-alignments, we take the lexical feature set of the tag containing the most '+' or '^' symbols, as these indicate that a POS tag or lemma is especially relevant, as opposed to the '-' symbol. For unknown alignments, we do not add any type of lexical information, preserving only the target-side lexical features, if available.

Sample source words	**Multi-alignment**	**Sample target word**	**Target alignment**
Direction lines	en_tag=NOUN-NOUN	Grifflinien	tag=NOUN
of the	en_tag=ADP-DET	der	tag=DET
view text	en_tag=VERB-NOUN	anzuzeigen	tag=VERB
same document	en_tag=ADJ-NOUN	gleichen	tag=ADJ
the QuarkXPress	en_tag=DET-NOUN	den	tag=DET

Table 5.5.: Top 5 multi-alignment combinations for English-German mixed dataset with examples.

Sample source words	**Multi-alignment**	**Sample target word**	**Target alignment**
server connection	en_tag=NOUN-NOUN	серверу	tag=NOUN
will make	en_tag=AUX-VERB	сделает	tag=VERB
the list	en_tag=DET-NOUN	списка	tag=NOUN
to the	en_tag=ADP-DET	к	tag=ADP
invited to	en_tag=VERB-ADP	приглашен	tag=VERB

Table 5.6.: Top 5 multi-alignment combinations for English-Russian mixed dataset with examples.

Table 5.7 displays statistics for the ratio of multi-alignments and unknown alignments in the different datasets used for the experimental part of this thesis.

Dataset	**Language pair**	**Unknown tokens**	**Multi-aligned tokens**
Train	**ENG-DEU**	9.64%	10.94%
	ENG-RUS	3.55%	9.30%

Table 5.7.: Statistics for unknown and multi-aligned tokens in the datasets.

To find what representation would lead to better results in the experimental setup of this thesis, we looked at other alternatives to represent source-side features on the target side. Table 5.8 offers one of these alternatives. The results for the datasets containing this sort of representation obtained similar results to those of the datasets with the chosen representation, and are not further analyzed in Chapter 6. These datasets are available in the CD attached to this work.

Alignment type	Target Word	Source-side Features
1-to-1 alignment	Versionen	en_tag=ADV en_lex=+lexambig,+posambig,+syncretic
Frequent multi-alignment	von	multi_tag=NOUN-ADP multi_lex=-lexambig,-posambig,-syncretic
Infrequent multi-alignment	früheren	multi_tag=ADV-PRON multi_lex=+lexambig,+posambig,+syncretic
Unknown alignment	wurde	en_tag=UNK

Table 5.8.: Alternative for source-side feature representation in target data.

Combining Source-side Features with Target Features

The dataset for our experiments will contain the target sentences with their aligned source-side morphological features, if available, as well as their original morphological and control features. Figure 5.7 holds a sample sentence displaying source-side features on the target side, which corresponds to the sentences in Figures 5.2 and 5.3. The features marked in blue correspond to the original target sentence, whereas the features marked in green represent the transferred linguistic source-side features. In contrast to the target features, the source-side features do not regard context, meaning that the attributes are token-dependent.

```
# sent_id = 72

1  bei  bei  ADP  APPR  AdpType=Prep|Case=Dat  _  _  _  en_tag=PUNCT|freq=p99|gap_before_tag=OK|gap_tag=OK|tag=BAD
2  der  der  DET  ART  Case=Dat|Gender=Fem|Number=Sing|PronType=Art  _  _  _  en_tag=UNK|freq=p99|gap_tag=OK|lex=+posambig,-lexambig,^syncretic|tag=BAD
3  Fehlerverarbeitung  Fehlerverarbeitung  NOUN  NN  Gender=Fem|Number=Sing|Person=3  _  _  _  en_tag=MULTI|en_lex=+posambig,-lexambig,-syncretic|freq=rare|gap_tag=OK|tag=OK
4  kann  können  AUX  VMFIN  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|VerbType=Mod  _  _  _  en_tag=AUX|en_lex=+posambig,+syncretic,-lexambig|edit=edit_in,edit_post|freq=p99|gap_tag=BAD|lex=+syncretic,-lexambig,-posambig|tag=OK
5  eine  eine  DET  ART  Case=Nom|Gender=Fem|Number=Sing|PronType=Art  _  _  _  en_tag=DET|en_lex=+posambig,-lexambig,-syncretic|freq=p99|gap_tag=OK|lex=+syncretic,-lexambig,-posambig|tag=BAD
6  Anwendung  Anwendung  NOUN  NN  Gender=Fem|Number=Sing|Person=3  _  _  _  en_tag=NOUN|en_lex=-lexambig,-posambig,-syncretic|freq=p99|gap_tag=OK|lex=+syncretic,-lexambig,-posambig|tag=BAD
7  den  den  DET  ART  Case=Acc|Gender=Masc|Number=Sing|PronType=Art  _  _  _  en_tag=UNK|freq=p99|gap_tag=OK|lex=+posambig,+syncretic,-lexambig|tag=BAD
8  Benutzer  Benutzer  NOUN  NN  Gender=Masc|Number=Sing|Person=3  _  _  _  en_tag=NOUN|en_lex=-lexambig,-posambig,-syncretic|freq=p99|gap_tag=OK|lex=-lexambig,-posambig,^syncretic|tag=BAD
9  jedoch  jedoch  ADV  ADV  _  _  _  _  en_tag=ADV|freq=p99|gap_tag=OK|tag=BAD
10 leicht  leicht  ADJ  ADJD  Degree=Pos|Variant=Short  _  _  _  en_tag=MULTI|en_lex=+posambig,-lexambig,^syncretic|freq=p99|gap_tag=OK|tag=BAD
11 abfangen  abfangen  VERB  VVINF  VerbForm=Inf  _  _  _  en_tag=CCONJ|en_lex=+posambig,+syncretic,-lexambig|freq=p90|gap_tag=OK|tag=BAD
12 oder  oder  CCONJ  KON  _  _  _  _  en_tag=MULTI|en_lex=+posambig,-lexambig,^syncretic|freq=p99|gap_tag=OK|tag=BAD
13 tftraut  tftrauen  VERB  VVFIN  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin  _  _  _  en_tag=DET|en_lex=+posambig,-lexambig,-syncretic|edit=edit_post|freq=rare|gap_tag=OK|tag=BAD
14 ,  ,  PUNCT  $,  PunctType=Comm  _  _  _  en_tag=UNK|freq=p99|gap_tag=OK|tag=OK
15 wenn  wenn  SCONJ  KOUS  _  _  _  _  en_tag=SCONJ|en_lex=+posambig,-lexambig,-syncretic|freq=p99|gap_tag=OK|tag=BAD
16 etwas  etwas  PRON  PIS  Gender=Neut|Number=Sing|Person=3|PronType=Ind,Neg,Tot  _  _  _  en_tag=PRON|en_lex=+posambig,-lexambig,-syncretic|freq=p99|gap_tag=OK|lex=+posambig,-lexambig,^syncretic|tag=BAD
17 nicht  nicht  PART  PTKNEG  Polarity=Neg  _  _  _  en_tag=PART|freq=p99|gap_tag=OK|lex=+posambig,-lexambig,-syncretic|tag=OK
18 wie  wie  CCONJ  KOKOM  ConjType=Comp  _  _  _  en_tag=ADP|en_lex=+lexambig,+syncretic,^posambig|freq=p99|gap_tag=OK|lex=+posambig,-lexambig,-syncretic|tag=OK
19 erwartet  erwarten  ADJ  ADJD  Degree=Pos|Variant=Short  _  _  _  en_tag=VERB|en_lex=+lexambig,+posambig,^syncretic|edit=edit_post|freq=p99|gap_tag=OK|lex=-lexambig,-posambig,^syncretic|tag=OK
20 funktioniert  funktionieren  VERB  VVFINMood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin  _  _  _  en_tag=VERB|en_lex=+posambig,-lexambig,^syncretic|edit=edit_post|freq=p99|gap_tag=OK|lex=+syncretic,-lexambig,-posambig|tag=BAD
21 .  .  PUNCT  $.  PunctType=Peri  _  _  _  en_tag=PUNCT|freq=p99|gap_tag=OK|tag=OK
```

Figure 5.7.: Target sentence in CoNLL format with additional linguistic source-side features.

6. Experiments and Results

This chapter surveys the results of our approach towards the interaction of source-side linguistic features with those of the target side. In the following sections, we will first run the classifier on the source-side and target-data separately, and subsequently we will proceed to the error classification with the newly created datasets resulting from the data pre-processing procedures in Section 5.5.1.

After running the classifiers, we offer an overview of the metrics obtained for the single-dataset experiment before moving on to our cross-dataset method. This will be followed by the evaluation of the impact of adding linguistic source-side features to the target side with respect to the results of the baseline system, analyzing thereafter language-specific peculiarities in Chapter 7.

6.1. Length of the Datasets

As for the contents of the training datasets, the English-German dataset contains a total of 13442 source and target sentences with an average length of 17.46 words and 19.01 words for the English sentences and German sentences, respectively. The alignment file comprises this way 13442 lines, each corresponding to the alignments for every sentence in the source and target side. The English-Russian files comprises significantly more data amounting to a total of 15089 sentences for both the source and target side with an average sentence length of 9.84 and 9.40 words for the English and Russian side, respectively. Table 6.1 summarizes this information.

Language pair	**Sentences**	**Average SS length**	**Average TS length**
English-German	13442 sentences	17.46 words	19.01 words
English-Russian	15089 sentences	9.84 words	9.40 words

Table 6.1.: Length and average sentence length for the training datasets and language pairs obtainable at the WMT19 shared task website. SS is for source sentence, and TS is for target sentence.

6.2. Error Classification

After having followed all steps presented in the Chapter 4, we proceed to run the `analyze.py` script by Bollmann and Søgaard (2021) on all datasets used for our experiments, along with the following flags to compute FI scores for morphological and control features:

`-n`: This flag generates non-morphological features as control features.

`-I`: This flag avoids creating features for preceding or following tokens, which means that the immediate context is not taken into account.

`-L`: This flag avoids local feature combinations for the same token.

`-w 0` (default: 4): This flag indicates the amount of context tokens to be included for wide context features. In the experiment carried by Bollmann and Søgaard, no context tokens were taken account since this flag is followed by 0.

`--method drop-category-upos`: This method evaluates the importance of the features appearing in the dataset preserving individual POS features.

Furthermore, as in the original experiment by Bollmann and Søgaard, we ran the classifier a second time to analyze the FI scores of control features on their own. In order to do this, we must add the following flag to the flags presented above:

`-M`: This flag avoids the creation of morphological features for the dataset in question.

Firstly, to obtain a better overview of whether our future experiments actually make a difference, we run the classifier using the original CoNLL files for every single language. Later, we present the results for the dataset crafted following the steps in Section 5.5. It should be mentioned that for the classifier to detect the source-side features added to the target-side sentences, we needed to modify the `features.py` script by adding the newly incorporated features *en_tag=* and *en_lex=*. This script detects all the different features within the datasets, and passes them onto the classifier for them to be ranked.

6.3. Evaluating Classifier Performance

To assess how performant the classifiers are when run with the different datasets experiments, we compare their accuracy metrics for each task. Classification accuracy indicates the percentage of the predictions that are correct. The accuracy results and F_1 scores for the random forest classifiers are displayed in Tables 6.2 and 6.3.

The F_1 scores ascertain how well the classifiers learned the task at hand Bollmann and Søgaard (2021). It should be noted that the classifier is trained to detect the errors in state-of-the-art systems, which is not common in other NLP classification tasks. This means that these systems are optimized to fix easily-detectable errors, making the classification task complicated. Furthermore, it should not be assumed that this task is well learnable by merely regarding morphological features (Bollmann and Søgaard, 2021).

The accuracy and F_1 values for the classifier run on the data with both morphological and control features (Table 6.2) vary across languages, but with stronger results for the mixed datasets in both the English-German and English-Russian language pairs. As for individual languages, we observe that the metrics for the target language in the English-German combination is lower than that of Russian in the English-Russian language pair.

Language	**Metrics**	
	Accuracy	**F_1**
English (ENG-DEU)	62%	28%
German (ENG-DEU)	58%	30%
Mixed (ENG-DEU)	76%	41%
English (ENG-RUS)	54%	24%
Russian (ENG-RUS)	70%	32%
Mixed (ENG-RUS)	80%	44%

Table 6.2.: Classifier accuracy and F_1 scores for the different training datasets and language pairs for morphological and control features.

As for the experiment where the classifier is run only with control features (Table 6.3), the metrics are generally lower, with the exception of the German data in the English-German language pair, which keep the same values as in the previous experiment.

Language	Metrics	
	Accuracy	**F_1**
English (ENG-DEU)	47%	25%
German (ENG-DEU)	58%	30%
Mixed (ENG-DEU)	57%	25%
English (ENG-RUS)	40%	21%
Russian (ENG-RUS)	56%	23%
Mixed (ENG-RUS)	56%	23%

Table 6.3.: Classifier accuracy and F_1 scores for the different training datasets and language pairs for control features.

6.4. Evaluating Feature Importance

In the context of FI, we observe some differences in the outputs provided by the classifier for the different language pairs, that is, when running the classifier on each CoNLL file individually. Next, we present the results for each of the original CoNLL files, which will be succeeded by the results of the mixed datasets for both the English-German and English-Russian language pairs.

6.4.1. English - German Training Dataset

Frequency and lexical features are on the top of the feature importance list for the English data in this language pair. For the German, morphological features indicating adposition types, e.g. preposition or circumposition, lead the ranking followed by the perfective verb aspect, case, and adjective degree. It should be mentioned that after running the classifier all features obtained the same score if they belonged to the same feature category. Therefore, we have summarized all feature scores by feature category.

These differences in the results of the different languages are thinkable due to the extent of their morphological complexity, with English relying more on syntactic changes to express phenomena that in other languages are manifested through morphology as it is the case with German. Tables 6.4 and 6.5 rank the different feature categories for the morphological and control features for the English and German CoNLL files, respectively, by their Feature Importance (FI).

Category	FI
Frequency	0.016
Lexical features	0.012
Length	0.008
Morphological features	0.005
POS tags	0.004
String based features	0.001

Table 6.4.: Morphological and control features ordered by their FI for the ENG side in the ENG-DEU training dataset.

Category	FI
Morphological features	0.014
Frequency	0.012
Lexical features	0.009
Length	0.007
String-based features	0.004
POS tag	0.0007

Table 6.5.: Morphological and control features ordered by their FI for the DEU side in the ENG-DEU training dataset.

Since morphological features are on top of the FI ranking for the German side, Table

6.6 holds the features which are top-ranked in the classifier output.

Feature rank	Feature subcategory
1	Adposition: Circumposition, Postposition and Preposition
2	Verbal aspect: Perfective
3	Case: Accusative, Dative and Genitive
4	Comparative conjunction type
5	Adjectival degree: Comparative and Positive

Table 6.6.: Top 10 morphological features grouped by their subcategories for DEU data in ENG-DEU dataset ordered by their rank.

The ranking for the top 6 control features presented in Tables 6.7 and 6.8 indicates that frequency and especially length features are slightly more relevant for classification on the German side. Both tables share the same features, but for the last feature, which is *freq=p98* for the English data and *freq=p95* for the German data.

Category	FI
len10+	0.005
len4-6	0.005
len7-9	0.005
freq=gt3	0.002
freq=p90	0.002
freq=p98	0.002

Table 6.7.: Top 6 control features ordered by their FI for the ENG side in the ENG-DEU training dataset.

Category	FI
len10+	0.025
len4-6	0.025
len7-9	0.025
freq=gt3	0.006
freq=p90	0.006
freq=p95	0.006

Table 6.8.: Top 6 control features ordered by their FI for the DEU side in the ENG-DEU training dataset.

6.4.2. English - Russian Training Dataset

For the English-Russian dataset, features concerning animacy, aspect, and case are most predictive of error for the Russian data. As for the English data, length features also dominate the top 10, but unlike in the English - German combination, lexical features also seem to be rather prominent. Tables 6.9 and 6.10 illustrate the FI for the feature categories of both morphological and control features English and Russian CoNLL files. Again, as it was the case for the English-German language pair, features belonging to the same feature categories have the same scores, and the tables below show the summary for the scores for each of the feature categories.

The trend observed with the previous language pair is preserved here with regards to control features. Control features do not seem to be as prominent for the source side as they are for the target side. Interestingly, unlike for the English and German data in the English-German combination and English data in the current combination, for which length features presented higher results, frequency features on the Russian data are more present than length features.

As we did with German dataset in English-German combination, in Table 6.13 we provide the top-ranked Russian morphological features.

Category	FI
Morphological features	0.006
Lexical features	0.0002
POS tags	0.005
Length	0.022
Frequency	0.033
String based features	0.004

Table 6.9.: Morphological and control features ordered by their FI for the ENG side in the ENG-RUS training dataset.

Category	FI
Morphological features	0.050
Frequency	0.036
POS tags	0.012
Lexical features	0.024
Length	0.018
String based features	0.006

Table 6.10.: Morphological and control features ordered by their FI for the RUS side in the ENG-RUS training dataset.

Category	FI
len10+	0.001
len4-6	0.001
len7-9	0.001
freq=rare	0.001
freq=gt3	0.001
freq=p90	0.001

Table 6.11.: Top 6 control features ordered by their FI for the ENG side in the ENG-RUS training dataset.

Category	FI
freq=gt3	0.014
freq=p90	0.014
freq=rare	0.014
len10+	0.001
len4-6	0.001
len7-9	0.001

Table 6.12.: Top 6 control features ordered by their FI for the RUS side in the ENG-RUS training dataset.

6.4.3. Mixed Training Datasets

After having run the classifier with the original CoNLL files and with their results at hand, we proceed to assess the files containing target-side features with their corresponding source-side linguistic features.

German Training Dataset with English Morphological and Lexical Features

When we mix the English features with the German ones, English POS tags such as adjectives, adpositions and adverbs have a prominent presence in the ranking. We speculate that this is the case because these POS tags are difficult to translate into German.

For adjectives, the system has to grant the stem of the adjective with information regarding case, gender and declension. For adpositions, it might be the case that oftentimes the genitive is used instead of a preposition. As for adverbs, English presents a more regular structure when forming adverbs. To form regular adverbs, we need to append *-ly* to the adjective. In German, however, the adverbial repository is richer.

After the English POS tags, the classification is continued by the English lexical features that used to be on the top of the results when running the classifier on the English CoNLL file. These two sets of features are continued by the features that used to be on top of the results for the German CoNLL file in the previous experiment. Table 6.14 comprises the top 8 features for this experiment. As opposed to the experiments in the previous sections, in this case, different morphological features obtain different scores, hence, we put these features independently in 6.14. In terms of control features, the metrics for the mixed English-German dataset (Table 6.15) are similar to those of the German dataset when run with the classifier individually.

Feature rank	Feature
1	Animacy: Animacy and Inanimacy
2	Verbal aspect: Imperfective and Perfective
3	Case: Dative, Genitive, Instrumental, Locative and Nominative

Table 6.13.: Top 10 morphological features grouped by subcategories for RUS data in ENG-RUS dataset ordered by their rank.

Language	Category	FI
English	POS tags	0.064
English	Lexical	0.050
German	Frequency	0.023
German	Length	0.019
German	String-based	0.011
German	Gender	0.010
German	Case	0.009
German	Number and pronoun type	0.001

Table 6.14.: Morphological and control features ordered by their FI for German data with English-side features in ENG-DEU dataset.

Category	FI
len10+	0.026
len4-6	0.026
len7-9	0.026
freq=gt3	0.006
freq=p90	0.006
freq=p95	0.006

Table 6.15.: Top 6 control features ordered by their FI for German data with English-side features in ENG-DEU dataset.

Russian training dataset with English morphological features

When mapping the English side to the Russian side, there is a difference with respect to the same experiment carried out on the English-German dataset: the target features are very prominent in the ranking. These dissimilarities can be contemplated on Table 6.16. For this mixed dataset, the results for the control features are identical to those of the Russian experiment in the previous section. This is shown on Table 6.17.

Language	Category	FI
English	Lexical	0.072
English	POS tags	0.062
Russian	Frequency	0.056
Russian	Length	0.030
Russian	Case	0.027
Russian	Gender	0.017
Russian	String-based	0.009

Table 6.16.: Morphological and control features ordered by their FI for Russian data with English-side features in ENG-RUS dataset.

Category	FI
freq=gt3	0.014
freq=p90	0.014
freq=rare	0.014
len10+	0.001
len4-6	0.001
len7-9	0.001

Table 6.17.: Top 6 control features ordered by their FI for Russian data with English-side features in ENG-RUS dataset.

7. Discussion

Matching linguistic source-side features with the target side has enabled us to analyze the interaction between the morphology of the source and the target language. The following sections will discuss the results obtained in Chapter 6, as well as some of the limitations encountered in the datasets used for the experiments.

7.1. Effect of Adding Linguistic Source-side Features to the Target Side

Compared to the results delivered by the random forest classifier when run with the individual CoNLL files, the accuracy and F_1 scores for the mixed datasets are relatively higher. This means that the classifier performs better when combining linguistic source-side information with the target-side files.

After running the classifier with the individual CoNLL files, features pertaining to the target side generally present higher FI scores across feature categories than those on the source side, especially when it comes to linguistic features. To this end, one might expect that, when mixing linguistic source-side features with those on the target side, the target-side features would still rank higher than those in the source side. However, our results suggest otherwise for the mixed datasets. For the mixed dataset English-German dataset, the English POS tags and lexical features are on the top of the list with a difference 85% to the next feature on the list, the German frequency features. For the mixed English-Russian dataset, the English POS tags and lexical features lead the ranking, which is rather surprising given the prominence of the Russian morphological features when running the classifier with the individual Russian CoNLL file. Having said this, we do not discard any unaccounted-for side effect that favors source-side features over target-side features in the mixed datasets.

7.2. Limitations of Machine Translation Datasets

The data used for the experimental part of this thesis has potential limitations. Translating between two languages is an inherently hard task, but more so when the data to be translated belongs to a specific domain as it is the case of ours, which belong to the IT domain. This not only presents a hurdle for the MT system, but also for the tools applied in the linguistic annotation of the datasets. Furthermore, many errors in the datasets of the WMT19 shared task are not morphological, but either lexical or syntactical as can be observed in Figure 7.1.

A further issue is that for this MT shared task, there are nearly not as many languages available as for the other shared tasks in the Bollmann and Søgaard (2021) paper. Despite having access to datasets in the English-German and English-Russian language pairs, with German and Russian being reasonable representatives for morphologically rich languages, we believe that our results could have been strengthened by having further language pairs involving a morphologically poor and a morphologically rich language simultaneously. We also carried out a further experiment with an English-Chinese dataset found on the web-

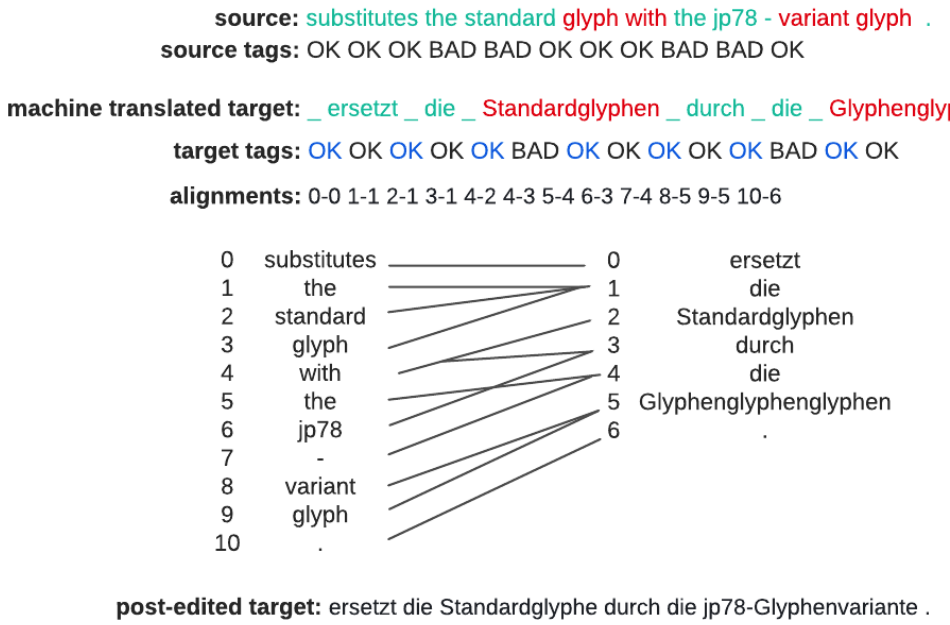

Figure 7.1.: Sentences containing lexical errors.

site for the WMT shared task in 2020[1]. The results were not as promising as those for the English-German and English-Russian datasets, which is explicable due to both of the languages having poor morphologies and the rather unexisting linguistic annotation for the Chinese side.

In terms of error analysis, the concept of 'error' itself is not as well-defined for MT tasks as it might be for other NLP tasks such as tagging or parsing. In the latter examples, context is not accounted for, and there is usually just one possible solution. For instance, given a verb, the parser discerns whether the verb is in the past tense or not. This is not the case for MT, for which a particular word can have many different interpretations depending on context.

[1]This dataset is available at: https://www.statmt.org/wmt20/quality-estimation-task.html. The results have also been included in the CD under `experiments_lin/en_zh`.

8. Conclusion

In the past chapters, we evaluated the interaction between the morphologies of a morphologically poor language, English, and two morphologically rich languages, German and Russian. We extended the work of Bollmann and Søgaard (2021) on error analysis and the role of morphology by further analyzing the datasets of a shared task, which was not as explored in the paper, the MT task on QE. The experiments for this shared task on the paper by Bollmann and Søgaard (2021) consisted on assessing the data for the source and target language independently, which were annotated with their corresponding linguistic, control features, and error tags. In MT, and translation in general, the source and target language coexist, and it is paramount to regard how much the source language is impacting the target language, and vice versa. Thus, in this thesis, we go a step forward by combining the linguistic features from the source side with those in the target side. For this purpose, we adopted the pipeline designed by Bollmann and Søgaard (2021) for pre-processing and analysis. This pipeline is augmented by a series of language-agnostic scripts, written for this thesis, and conceived for the extraction of linguistic source-side features and their transfer to the target side.

The results for our combined approach to the datasets are promising. Despite the target language features presenting higher FI scores than those in the source language, this situation changes when the random forest classifier is run with the mixed datasets. In this case, English POS tags and lexical features score higher than any other feature category pertaining to the target side. This is complemented by higher accuracy and F_1 scores for the classifiers when run with the combined variants. This suggests that the source side is indeed important for error classification in the context of MT systems.

8.1. Future Work

The original paper by Bollmann and Søgaard (2021) puts a great focus on the Dependency Parsing and SEM shared tasks due to them dealing with many more languages than the MT shared task. Therefore, running these experiments with a more representative amount of datasets, ideally containing a mixture between a morphologically poor and a morphologically rich language, could give us a better insight into morphological errors.

Due to time constraints, we did not assess further source-side feature combinations. For instance, by adding linguistic and non-linguistic source-side features simultaneously, we could check whether length and frequency features present other values than those for the current mixed datasets in Tables 6.15 and 6.17, which are almost identical to those of the target side. Furthermore, it would also be interesting to integrate the complete set source-side features into the target side to evaluate the classifier's performance and whether this results in a different ranking of the individual features.

A possible continuation based on the contributions of this thesis would be creating an automated method to classify the errors on the post-edited files from the WMT19 shared task, i.e. adding whether the error committed has either been an inflectional or derivational error to the OK/BAD labels. Likewise, having more sophisticated methods for the systematic evaluation of different feature combinations would be beneficial to further analyze potential errors and research directions.

Bibliography

Marios Andreou. *Lexemes*. 03 2019. ISBN 9780199772810. doi: 10.1093/obo/9780199772810-0232.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.618. URL https://aclanthology.org/2020.emnlp-main.618.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2014. URL https://arxiv.org/abs/1409.0473.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Luisa Bentivogli, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1025. URL https://aclanthology.org/D16-1025.

Arianna Bisazza and Marcello Federico. Surveys: A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational Linguistics*, 42(2):163–205, June 2016. doi: 10.1162/COLI_a_00245. URL https://aclanthology.org/J16-2001.

Arianna Bisazza and Clara Tump. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1313. URL https://aclanthology.org/D18-1313.

Marcel Bollmann and Anders Søgaard. Error analysis and the role of morphology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1887–1900, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.162. URL https://aclanthology.org/2021.eacl-main.162.

Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.

Franck Burlot and François Yvon. Evaluating the morphological competence of machine translation systems. In *Proceedings of the Second Conference on Machine Translation*, pages 43–55, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4705. URL https://aclanthology.org/W17-4705.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations

using rnn encoder-decoder for statistical machine translation, 2014. URL https://arxiv.org/abs/1406.1078.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29 (2):127–161, 2015. doi: 10.1007/s10590-015-9169-0.

Jacob Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. Fast and robust neural network joint models for statistical machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1370–1380, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1129. URL https://aclanthology.org/P14-1129.

Heidi Dulay, Marina Burt, and Stephen Krashen. *Language two*. Oxford University Press, 1983.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5401. URL https://aclanthology.org/W19-5401.

Johann Wolfgang von Goethe. *Versuch die metamorphose der Pflanzen zu Erklären*. 1790.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL https://aclanthology.org/W09-1201.

M. Haspelmath, M.S. Dryer, D. Gil, and B. Comrie. *The World Atlas of Language Structures*. Number v. 1 in Oxford linguistics. OUP Oxford, 2005. ISBN 9780199255917. URL https://books.google.de/books?id=amJNAp8LLREC.

Carl James. *Errors in language learning and use: Exploring error analysis*. Routledge, 1998.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1293.

Stav Klein and Reut Tsarfaty. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sigmorphon-1.24. URL https://aclanthology.org/2020.sigmorphon-1.24.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2009. doi: 10.1017/CBO9780511815829.

Philipp Koehn. *Neural Machine Translation*. Cambridge University Press, 2020. doi: 10.1017/9781108608480.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490, 2012. URL http://arxiv.org/abs/1201.0490.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, mar 1986. doi: 10.1007/bf00116251. URL https://doi.org/10.1007%2Fbf00116251.

Carlos Ramisch, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-4925.

Sebastian Ruder. Why You Should Do NLP Beyond English. http://ruder.io/nlp-beyond-english, 2020. Accessed: 20-06-2022.

Benoît Sagot. A multilingual collection of CoNLL-U-compatible morphological lexicons. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1292.

August Schleicher. *Zur Morphologie der Sprache*. Eggers, 1859. URL https://books.google.de/books?id=kfh3QwAACAAJ.

Abhishek Sharma. Part-of-speech(pos) tag: Dependency parsing: Constituency parsing. https://www.analyticsvidhya.com/blog/2020/07/part-of-speechpos-tagging-dependency-parsing-and-constituency-parsing-\in-nlp/, Jul 2020.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

Lucia Specia, Marco Turchi, Nicola Cancedda, Nello Cristianini, and Marc Dymetman. Estimating the sentence-level quality of machine translation systems. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*,

Barcelona, Spain, May 14–15 2009. European Association for Machine Translation. URL https://aclanthology.org/2009.eamt-1.5.

Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3009. URL https://aclanthology.org/K17-3009.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014. URL https://arxiv.org/abs/1409.3215.

Peratham Wiriyathammabhum, Douglas Summers-Stay, Cornelia Fermüller, and Yiannis Aloimonos. Computer vision and natural language processing: Recent approaches in multimedia and robotics. *ACM Comput. Surv.*, 49(4), dec 2016. ISSN 0360-0300. doi: 10.1145/3009906. URL https://doi.org/10.1145/3009906.

François Yvon. Quality Estimation for Machine Translation. *Computational Linguistics*, 45(2):391–394, 06 2019. ISSN 0891-2017. doi: 10.1162/coli_r_00352. URL https://doi.org/10.1162/coli_r_00352.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/K18-2001. URL https://aclanthology.org/K18-2001.

Çagri Çöltekin and Taraka Rama. Exploiting universal dependencies treebanks for measuring morphosyntactic complexity. 2018.

Martin Šošić and Mile Šikić. Edlib: a C/C++ library for fast, exact sequence alignment using edit distance. *Bioinformatics*, 33(9):1394–1395, 01 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw753. URL https://doi.org/10.1093/bioinformatics/btw753.

A. Universal POS Tags

Tag	Description
ADJ	Adjective
ADV	Adposition
ADP	Adverb
AUX	Auxiliary
CCONJ	Coordinating Conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PRON	Pronoun
PROPN	Proper Noun
PUNCT	Punctuation
SCONJ	Subordinating Conjunction
SYM	Symbol
VERB	Verb
X	Other

Figure A.1.: Universal POS tags. Figure from Sharma (2020).

B. Morphological and Control Features

Feature	**Definition**	
Morphological features		
U:POS={VALUE}	universal part-of-speech tag, e.g. U:POS=VERB	
U:{FEAT}={VALUE}	universal feature according to the UD specification, e.g. U:TENSE=PAST	
AMBIG$_{\text{POS}}$=NO	$\lvert P_t\rvert = 1$	where P_t is the set of all observed universal POS tags for t
AMBIG$_{\text{POS}}$=YES	$1 < \lvert P_t\rvert < 5$	
AMBIG$_{\text{POS}}$=HIGH	$\lvert P_t\rvert \geq 5$	
AMBIG$_{\text{LEX}}$=NO	$\lvert L_t\rvert = 1$	where L_t is the set of all observed lemmata for t
AMBIG$_{\text{LEX}}$=YES	$\lvert L_t\rvert > 1$	
SYNCRETIC=NO	$\lvert M_t\rvert = 1$	where M_t is the set of all observed morphological feature combinations for t
SYNCRETIC=YES	$1 < \lvert M_t\rvert < 5$	
SYNCRETIC=HIGH	$\lvert M_t\rvert \geq 5$	
EDIT=PRE	$x_0 \neq$ MATCH	where $[x_0, \ldots, x_n]$ is the sequence of edit alignments between t and l, $x_i \in \{\text{MATCH}, \text{MISMATCH}, \text{GAP}\}$
EDIT=SUF	$x_n \neq$ MATCH	
EDIT=IN	$\exists i, j, k : i < j < k$ $\wedge x_i =$ MATCH $\wedge x_j \neq$ MATCH $\wedge x_k =$ MATCH	
EDIT=FULL	$\forall i : x_i \neq$ MATCH	
Control features		
LEN=1-3	$1 \leq \lvert t\rvert \leq 3$	where $\lvert t\rvert$ is the string length of t
LEN=4-6	$4 \leq \lvert t\rvert \leq 6$	
LEN=7-9	$7 \leq \lvert t\rvert \leq 9$	
LEN=10+	$\lvert t\rvert \geq 10$	
FREQ=99	$P_{99} \leq f(t)$	where $f(t)$ is the absolute frequency count of t and P_n is the n-th percentile of the frequency distribution
FREQ=98	$P_{98} \leq f(t) < P_{99}$	
FREQ=95	$P_{95} \leq f(t) < P_{98}$	
FREQ=90	$P_{90} \leq f(t) < P_{95}$	
FREQ=UNCOMMON	$4 \leq f(t) < P_{90}$	
FREQ=RARE	$f(t) < 4$	

Figure B.1.: Morphological and control features. Figure from Bollmann and Søgaard (2021)