Universität Zürich[UZH]

Institut für Computerlinguistik

Machine Translation of Film Subtitles from English to Spanish

Combining a Statistical System with Rule-based Grammar Checking

Masterarbeit der Philosophischen Fakultät der Universität Zürich

Referent: Prof. Dr. M. Volk

Verfasserin:
Jeanette Isele
Matrikelnummer 08-710-386
Gerenstrasse 12b
8602 Wangen

June 30, 2013

Abstract

In this project we combined a statistical machine translation system for the translation of film subtitles from English to Spanish with rule-based grammar checking. At first we trained the best possible statistical machine translation system with the available training data. The largest part of the training corpus consists of freely available amateur subtitles. A smaller part are professionally translated subtitles provided by subtitling companies. In the next step we developed, applied and evaluated the grammar checker.

We investigated if the combination of a statistical system with a rule-based grammar checker is reasonable and how we can improve the results. With the trained statistical machine translation system an application of the grammar checker would be recommendable, especially in order to correct disagreements between nouns, articles and adjectives. The precision of the grammar checker is very satisfying. With additional linguistic information, for example, syntactical information, we would probably be able to improve the grammar checker and include the correction of other kinds of errors. In addition, the evaluation showed that the improvement of the statistical machine translation system causes a significant decrease of the number of the considered errors. Furthermore, we elaborated various possibilities as to how the statistical machine translation system can be improved. Thus, one might examine, if the improvement of the system yields a significant decrease of the number of the errors. If this should be the case we have to question if the additional use of a grammar checker is still reasonable or if the number of the considered grammatical errors is too low.

Additionally, we compare the performance of the trained machine translation system with the state of the art performance in the SUMAT project for the automatic translation of film subtitles from English to Spanish. According to automatic evaluation scores, the system we trained in our project was slightly better than the system of the SUMAT project. This result shows that the use of freely available amateur subtitles for the training of a statistical machine translation system for the translation of professional subtitles is recommendable, even though their quality is not optimal.

Zusammenfassung

In diesem Projekt wurde ein statistisches maschinelles Übersetzungssystem für Filmuntertitel vom Englischen ins Spanische mit einem regelbasierten Grammatikkorrektursystem kombiniert. In einem ersten Schritt wurde mit den vorhandenen Trainingsdaten ein möglichst gutes statistisches maschinelles Übersetzungssystem trainiert. Darauf basierend wurde das Grammatikkorrektursystem entwickelt, angewandt und evaluiert.

Es wurde geprüft, ob die Kombination eines statistischen Systems mit einem regelbasierten Grammatikkorrektursystem sinnvoll ist und wie die Resultate noch verbessert werden könnten. Mit dem statistischen maschinellen Übersetzungssystem, das in diesem Projekt trainiert wurde, ist eine Anwendung des Grammatikkorrektursystems zu empfehlen, vor allem für die Korrektur von Fehlern, die die Kongruenz zwischen Substantiven, Artikeln und Adjektiven betreffen. Die Präzision des Grammatikkorrektursystems ist sehr zufriedenstellend. Mit zusätzlichen linguistischen Informationen, z.B. syntaktischen Informationen, könnte das Grammatikkorrektursystem noch verbessert und zusätzliche Arten von Fehlern könnten berücksichtigt werden. Die Evaluation zeigte auch, dass eine Verbesserung des statistischen maschinellen Übersetzungssystems einen signifikanten Abfall der betrachteten grammatischen Fehler zur Folge hat. Ausserdem konnten verschiedene Möglichkeiten herausgearbeitet werden, um das statistische maschinelle Übersetzungssystem zu verbessern. Es müsste also geprüft werden, ob die Anwendung des Grammatikkorrektursystems nach der Verbesserung des statistischen maschinellen Übersetzungssystems immer noch sinnvoll ist oder ob die Anzahl der beachteten Grammatikfehler zu gering ist.

Zusätzlich wurde die Performanz des trainierten maschinellen Übersetzungssystems mit dem aktuellen Forschungsstand für die automatische Übersetzung von Filmuntertiteln vom Englischen ins Spanische im SUMAT-Projekt verglichen. Gemäss automatischen Evaluationsmassen ist das beste Übersetzungssystems, welches in unserem Projekt trainiert wurde, etwas besser als dasjenige, aus dem SUMAT-Projekt. Dieses Resultat zeigt, dass die Verwendung frei verfügbarer Amateur-Untertitel für das Training eines statistischen maschinellen Übersetzungssystems für die Übersetzung professioneller Untertitel empfehlenswert ist, obwohl sich die Eigenschaften und Qualität von professionellen und Amateur-Untertiteln deutlich unterscheiden.

Acknowledgement

I would like to thank all the people who supported me with this project and paper.

I thank Martin Volk for the good supervision of my master's thesis and for his hints and advices which were important for the success of this project. I also would like to express my gratitude to Mark Fishel, Rico Sennrich and Simon Clematide who helped me to solve technical problems and answered my questions concerning the SUMAT project and the Moses tool. I want to thank the SUMAT research group of Vicomtech, especially Arantza del Pozo, Thierry Etchegoyhen and Volha Petukhova, who provided me with their test sets and important information about the state of the art in the SUMAT project. I also want to thank the VSI Group for the provision of their professionally translated subtitles for this project.

Many thanks to my father Roland Isele and my friends Mirjam Marti, Martina García and Jemeima Christen for proofreading the text and for supporting me with the revision.

Finally I want to thank my dear family and friends for their precious support, they kept me motivated and were always willing to lend me an ear when I had difficulties with my master's thesis. Thank you!

Contents

List of Tables

List of Acronyms

BLEU Bilingual Evaluation Understudy
EM algorithm Expectation Maximum Algorithm
MT Machine Translation
SMT Statistical Machine Translation
NER Named Entity Recognition
MERT Minimal Error Rate Training
METEOR Metric for Evaluation of Translation with Explicit ORdering
NLP Natural Language Processing
OCR Optical Character Recognition
OPUS Open Parallel corpUS
POS Part-Of-Speech
SUMAT SUbtitling by MAchine Translation
TER Translation Error Rate

1. Introduction

1.1. Motivation

Nowadays, large amounts of subtitles are produced. "Ideally subtitles are created for each language independently, but for efficiency reasons they are often translated from one source language to one or more target languages" (Volk et al., 2010, 53). Recent research projects, for example the SUMAT-project[1], have tested and evaluated the use of statistical machine translation systems for the translation of film subtitles. Since subtitles are generally short and syntactically simple, they are well suitable for statistical machine translation.

Other projects combine statistical and rule-based approaches or include additional linguistic information for the training of the statistical machine translation system (e.g. SUMAT[2] and Hardmeier 2008). Most of these projects did not achieve significant improvement of the translation quality.

Up until now, the translation quality of machine translation systems has been improved considerably. Nevertheless, the translations are rarely perfect and require a human revision. Among others, grammatical errors occur in the automatically translated texts.

1.2. Project and Research Questions

In this project, we combined a SMT system with a rule-based grammar checker. We trained and tuned the SMT system with Moses[3] (see section 6). The training corpus comprises two sub-corpora with parallel subtitles in English and Spanish: First, the OpenSubtitles corpus[4] (in total almost 33 million subtitles per language)

[1] http://www.sumat-project.eu/

[2] internal reports of 2012 and April 2013

[3] http://www.statmt.org/moses/

[4] http://opus.lingfil.uu.se/OpenSubtitles_v2.php

arranged and aligned by Jörg Tiedemann and second, almost 80'000 professionally translated subtitles per language provided by the VSI Group[5]. In the beginning we extracted a part of the corpus which we used as development and test set. With the available data, we then tried to train the best possible SMT system. We evaluated the trained systems with automatic evaluation scores and compared the results to the SUMAT project.

In a further step, we tried to improve the translated output with a rule-based grammar checker. We chose Python[6] as programming language for the grammar checker. The rules of our grammar checker are based on the part-of-speech tags and morphological analysis generated by Freeling[7]. Freeling is "an open source language analysis tool suite" which provides different NLP tools, such as part-of-speech tagging, morphological analysis and named entity recognition, for various languages (Center, 2012, 2). We also included the style and grammar checker LanguageTool[8] in order to improve the precision. LanguageTool is an open source spelling- and grammar-checker for more than 20 different languages. Finally, we evaluated manually the corrections made by our grammar checker.

The main research questions which shall be answered in this thesis are:

- Does the combination of a statistical system with a rule-based grammar-checker improve the quality of the translation?
- Is the combination of a statistical system with a rule-based grammar-checker recommendable in terms of the cost-benefit ratio?
- How can we improve our approach in order to improve the translation quality?

In order to answer the third question, we have to consider the improvement of the SMT system and the improvement of the grammar checker as well. Thus, we will also answer the following additional questions.

Questions concerning the improvement of the SMT-system:

- What should be considered in the composition of the training, development and test corpora for the training of the SMT-system?
- How can the training data be improved?
- Is the combination of amateur and professional subtitles recommendable?

[5] http://www.vsi.tv/
[6] http://www.python.org/
[7] http://nlp.lsi.upc.edu/freeling/
[8] http://www.languagetool.org/

Questions concerning the improvement and expansion of the grammar checker:

- Which kind of errors can be corrected with the grammar checker?
- or which error classes with low precision do the grammar checker rules have to be improved?
- Which possibilities do exist to extend the grammar checker?

Apart from answering these research questions, one of the main goals of this project is to investigate and evaluate this approach for future projects. We focus on a detailed description of all the steps of the project and mention open research questions.

1.3. Thesis Structure

In this project we trained a SMT system for film subtitles from English to Spanish and applied a grammar checker in order to correct some of the grammatical errors.

This paper contains nine sections. In this introduction, we present the setting of the project, the objectives and the research questions. These are followed by an introduction into statistical machine translation (see section 2). We explain the subtitling process and the characteristics of subtitles in section 3. This is important in order to understand certain characteristics of the corpora which influence the translation quality. Then we create an overview about statistical machine translation for subtitles (see section 4). In this context, we present the SUMAT project. In section 5 we describe and discuss the composition of the corpus, its characteristics and influence into the translation quality of the statistical machine translation system. In section 6 we explain the training of the translation system with Moses. This is followed by the evaluation and comparison of the qualities of the trained systems (see section 7). For the translation of the test sets, we use the best of our trained systems. In section 8 we present the self-created grammar checker. For each considered error class, we render a restricted error analysis, describe the development and structure of the rules and evaluate the corrected sentences. An overall evaluation of the grammar checker follows (see section 8.8.1). Additionally, we test if the grammar checker is text type specific (see section 8.8.2). The last section summarizes the results, answers the research questions and discusses open questions and future perspectives.

2. Statistical Machine Translation

In statistical machine translation systems, statistical methods are applied to produce translations from a source language into a target language. The translation model is the statistical model that proposes translation options and their probabilities. The translation model is trained with a parallel training corpus. Large amounts[9] of parallel and human-translated texts for the corresponding language pair are required (Carstensen et al., 2010, 647f.).

The parallel corpus must be sentence-aligned in order to know which sentence of the source text corresponds to which sentence of the target text. To achieve the sentence-alignment, the lengths (number of characters) of the sentences of the source and target language are compared and anchor points are used. Anchor points are tokens that are identical or at least similar in the source and the target language, such as numbers or proper names. Note that many-to-one and one-to-many alignments of sentences are possible as well (Carstensen et al., 2010, 648). The aligned sentences are used as input for the training of the translation model with Moses.

The translation model calculates the translation options and the corresponding probabilities for the input sentences. For the generation and calculation can be used the IBM models which apply the expectation maximization algorithm. The IBM models are based on generative modeling, which means that the generation of the translation is broken into smaller steps. In word-based models each word is translated independently and the corresponding word translation probabilities are used to calculate the translation probability of the whole sentence (Koehn, 2010, 86f.). In order to translate individual words and to calculate the corresponding translation probabilities, the words of the source and target sentences have to be aligned, and we need the lexical translation probability distribution (Koehn, 2010, 82f.). The difficulty is that the word alignment and the probability distribution interdepend. To calculate the probability distribution we need the word alignment, and to calculate the word alignment, we need the probability distribution. In other words, changes in the word alignment cause changes in the probability distribution and vice versa.

[9] at least about 1 mio. words

Therefore, we apply the expectation maximization algorithm (abbreviated EM algorithm). This means that in a first step, the model is initialized with uniform distributions and applied to the data. This is the expectation step (we expect the alignments) with the result that the words of the sentences in the source and target language are aligned. In the following maximization step, we calculate the lexical translation probabilities based on the expected alignments. Then, we estimate the alignments again, based on the calculated lexical translation probabilities. These steps are iterated various times until convergence (Koehn, 2010, 88). The calculated probabilities for each word pair are saved in so-called translation tables or t-tables. For the translation of each word we use the most probable aligned word of the target language. This approach is called maximum likelihood estimation (Koehn, 2010, 83). The translation system must also consider many-to-one and one-to-many alignments, differences in the word order in the source and target language and words without equivalents in the target language (Carstensen et al., 2010, 649).

Today, the majority of the translation models are phrase-based. In phrase-based models, words and tokens are not translated independently, instead, phrases (multi-word units, sequences of words) are translated and combined. The phrase pairs and their probabilities are saved in phrase translation tables with the corresponding probability (Koehn, 2010, 128-131). After the translation, the phrases have to be reordered because the order of phrases of the source and target language often show differences. Therefore a distance-based reordering model is used (Koehn, 2010, 127-129).

In addition to the translation model, a language model for the target language is created. For the training of the language model, we need a corpus in the target language. Obviously, we can use the texts in the target language of our training corpus. Optionally, we can complement this corpus with additional monolingual data. The language model contains information about typical word groups and word order of the target language (Carstensen et al., 2010, 650). In other words, it "measures how likely it is that a sequence of words would be uttered" by a speaker of the target language. We use the language model in order to guarantee the fluency and correct word order of the output and to improve the word choice. To achieve this, the language model takes each possible target sentence and calculates the probability that it is uttered (Koehn, 2010, 181). Language models are based on the calculation of probabilities of n-grams, which are the likelihood that certain words follow each other. The probability of n-grams can be calculated with the Markov chain (Koehn, 2010, 182). The Markov chain decomposes the n-gram (= word sequence) to calculate the probabilities. First the probability of the appearance of the first word of the sequence is calculated. Second, the Markov chain calculates

the probability that the second word is located after the first word of the sequence (conditioned probability), then the Markov chain calculates the probability that the third is located after the first two words of the sequence (conditioned probability) and so on. The probability of the complete n-gram is the product of all of these calculated probabilities (Koehn, 2010, 182):

$p(w_1, w_2, ..., w_n) = p(w_1) \cdot p(w_2|w_1)...p(w_n|w_1, w_2, ..., w_{n-1})$

It is impossible to consider all sizes of word sequences. Therefore, the Markov assumption is used to reduce the complexity. The Markov assumption claims that "only a limited number of previous words affect the probability of the next word" (Koehn, 2010, 182). It is common to use trigrams for the training of the language model, which means that up to two preceding words are considered (Koehn, 2010, 183).

The main goal for the translation of a text is to create and choose the most probable sentences of the target language for the translation of a text (Carstensen et al., 2010, 651):

$\widehat{T} = argmax_T \; p(T|S)$
(T = sentence in the target language, S = sentence in the source language)

With the Bayes' theorem this formula can be transferred into the following formula (Carstensen et al., 2010, 651):

$\widehat{T} = argmax_T \; p(T) \cdot p(S|T)$

*T*he language model calculates *P(T)* the the probability of the sentence being uttered in the target language. *P(S|T)* is the conditioned probability. Finally, the translation model chooses the sentence of the target language with the highest probability (T) (Carstensen et al., 2010, 651).

3. Subtitles and their Translation

3.1. Introduction into Subtitling and Subtitles

Subtitling belongs to audiovisual translation. "Audiovisual translation" stands for "different translation practices used in the audiovisual media – cinema, television, VHS – in which there is a transfer from a source to a target language, which involves some form of interaction with sound and images" (Díaz Cintas and Remael, 2007, 12). For example, it includes subtitling, dubbing, voice-over, narration and interpreting.

Subtitling is to present the dialogue and all the other considerable linguistic elements of a film, for example, inserts, graffiti, song texts, noises or letters, in a written form (Díaz Cintas, 2009b, 5). The subtitles are usually located at the bottom of the screen and are synchronized with the corresponding dialogue and image (Díaz Cintas and Remael, 2007, 8). Subtitles are an additional mode to carry the meaning of a scene to the viewer (Nagel et al., 2009, 52). They are integrated into the semiotic system and have to interact with the visual and audio mode (Díaz Cintas and Remael, 2007, 45). Elements or messages that are not expressed adequately by the subtitles can be compensated by another mode, for example, with gestures, voice, actions or movements of a person (Fong, 2009, 93f.). In other words, subtitles do not replace but complement the other, or at least some of the other semiotic modes (Díaz Cintas and Remael, 2007, 11f.). Subtitling is a multi-modal process. Thus, the creation of subtitles is very complex: the subtitler has to consider both the audio and the visual mode (Fong, 2009, 83). In order to guarantee the harmony of the different modes and channels, the duration of the dialogue should be more or less equivalent to the time the viewers need on average to read the subtitles (Díaz Cintas, 2009a, 46f.).

The function of the subtitles can vary. From a linguistic point of view, one distinguishes between intralingual, interlingual and bilingual subtitling. Intralingual subtitles are used for the deaf and the hard-of-hearing, for language learners, for karaoke effects or to represent dialectal speech in the standard language (Díaz Cin-

tas and Remael, 2007, 13f.). The speakers are sometimes indicated with different font colors so that deaf people can distinguish them without difficulties. Furthermore, paralinguistic information has to be included in the subtitles, for instance in order to indicate sound effects (Díaz Cintas and Remael, 2007, 14). In this project, we are interested in interlingual subtitles. They are produced through translation, adaption and conversion from a spoken dialogue in a source language into a written dialogue in a target language. This process is called "diagonal subtitling", because it results in a change of mode (spoken to written) and language (Díaz Cintas and Remael, 2007, 17). Bilingual subtitles are particularly used in bilingual areas or at film festivals. In the case of bilingual subtitles, one subtitle line is reserved for one language and the other one for the other language. This causes additional space constraints for the subtitles (Díaz Cintas and Remael, 2007, 18).

The subtitling process contains various challenges and difficulties, for example, the conversion of idiolects, sociolects or dialects, overlapping speech and the inclusion of overlapping noises and music into the subtitles (Díaz Cintas, 2009b, 4). Additionally, subtitlers have to consider the different perceptions of correctness and informality in the spoken and written language. For example, subtitles often contain milder vulgarisms and swearwords than the original spoken dialogue (Goldstein, 2009, 36) (Fong, 2009, 92). Furthermore, there are stylistic differences between written and spoken language. The written subtitles are often verbalized in a more formal way and with less colloquial expressions. A transcription of the dialogue would be inappropriate as subtitles (Nagel et al., 2009, 57f.). Moreover, it is sometimes difficult to include important gestures, for example a nod of the head, into the subtitles (Díaz Cintas and Remael, 2007, 52).

Given the mentioned difficulties, if the subtitler has enough time, he or she watches the entire film before subtitling and takes notes about problematic or difficult issues. He or she should consider five aspects in particular: elements of the dialogue with polysemious reading,"the gender and number of some nouns, pronouns and adjectives that are not marked in English", "the degree of familiarity shown among the characters", deictic units with referents on the screen and "exclamations with no fixed meaning" (Díaz Cintas and Remael, 2007, 31). Then the translator can start with the translation process. Nowadays subtitlers have a lot of programs and tools which support the subtitling process. They use word processing programs and a time code reader. They run some seconds of the film with the time codes, time the subtitles and put them to the right place on the screen (Fong, 2009, 10f.). Ideally the subtitler receives, beside the film, a script of the dialogue. Sometimes the subtitler even has to work without the film, only with the script of the dialogue (Díaz Cintas and Remael, 2007, 30). This can reduce the translation quality because the subti-

tler does not know the other channels and it might be difficult to understand the message of a dialogue (Goldstein, 2009, 67f). The next subtitling step is the revision and proofreading which is possibly done by another person. Finally, the simulation of the whole film allows the possibility to make some further amendments. Some of the described steps may be skipped due to time and cost constraints (Díaz Cintas and Remael, 2007, 33f.).

The discussed aspects will help us to understand and elaborate the characteristics of subtitles (see section 3.2), which we require in order to understand the advantages and disadvantages of subtitles for machine translation (see section 3.5).

3.2. Characteristics of Subtitles

Subtitles are heterogeneous. Although some general conventions exist, they depend on the individual guidelines of the companies, on the wishes of the clients and on the medium (Díaz Cintas and Remael, 2007, 23). Subtitles are used to express the message together with the visual channel and, if the subtitles are not for the deaf and hearing-aided, also with the acoustic channel. It is sometimes difficult to understand them without the context (Díaz Cintas, 2009a, 37).

Space and time constraints are typical for subtitles. Space constraints depend on the format of the subtitles, often two lines can be shown simultaneously on the screen (Díaz Cintas, 2009a, 21f.). The number of characters contained in a line depends on the workstation and the different subtitling companies, typically a line contains 32-41 characters (Díaz Cintas and Remael, 2007, 9). Some film festivals use subtitles up to 43 characters, whereas television stations generally use a maximum of 37 characters per line. This means that cinema subtitles present a major part of the original spoken dialogue in contrast to subtitles of television stations and DVDs. Obviously, different subtitle scripts exist for the same film depending on the media. Sometimes even various subtitle versions for the same media exist. For example, different TV stations have their own subtitle versions. A film of 90 minutes results in 900 subtitles in the cinema version, compared to 750 subtitles in the DVD version and 650 subtitles in the TV version. Sometimes the same subtitles are used for all media to avoid extra-costs, although this reduces their readability (Díaz Cintas and Remael, 2007, 24). Remember that a good timing of the subtitles is important, because the viewer must be able to read the whole subtitle and the length of the dialogue must correspond to the time the viewer needs to read the corresponding subtitle(s) (Díaz Cintas, 2009a, 21f.).

In a subtitled film the text in the original language (speech) and the text in the target language (subtitles) co-exist. The viewer has the possibility to judge the translation immediately and to compare the source and target text. This has an impact on creating and translating subtitles: The subtitler tries to maintain words or expressions with phonetic and morphological similarities in the source and target language, in order that the viewer can recognize them in the spoken dialogue. For the same reason, the subtitler tries to avoid omissions of prominent words or expressions, otherwise the viewers would think that the subtitler forgot to include them into the subtitles (Díaz Cintas and Remael, 2007, 55f). At the same time, it is important that all reformulations are idiomatic, that they sound natural and that they do not contain many calques (Díaz Cintas and Remael, 2007, 150). However, it is sometimes difficult to satisfy these needs, because of the mentioned time and space constraints (Díaz Cintas and Remael, 2007, 9).

As we already mentioned, a subtitler must include all important information in the subtitles. Typically, condensation, reformulation, reduction omissions and ellipsis occur in subtitles, because of space and time constraints (Fong, 2009, 95). This means that subtitling is an adaption and not only a conversion from speech to written text (Díaz Cintas and Remael, 2007, 9). Reductions of the original dialogue are very common, whereas the original script is hardly ever expanded (Hillman, 2011, 385). Subtitlers often omit elements like "explications, vocatives, appellatives, proper names, circumstantials and modifiers" (Fong, 2009, 95). To decide if certain expressions or elements might be omitted, Cintas says "the question to be asked in case of doubt is: What requires more effort on the part of the viewer? A shorter subtitle with less information (quicker reading, more thinking)? Or a slightly longer subtitle with more information (slower reading, less thinking)?" (Díaz Cintas and Remael, 2007, 148). This shows another important characteristic of subtitles: their readability. For the viewer it should be as convenient as possible to read and understand the subtitles. This has an impact on the grammar and word order of the subtitles. The subtitler often rearranges the sentences of the spoken dialogue because the syntactic structure should be as simple and as common as possible. Line breaks within the subtitles have a significant impact on the readability of the subtitles. For example, if we have a sentence consisting of a main and a subordinate clause, the break should be between these two clauses. Also, if the sentence does not contain subordinate clauses, the breaks should be at grammatically and syntactically reasonable locations (Díaz Cintas, 2009a, 23f.). "Subtitled text should appear segmented at the highest syntactic nodes possible." (Díaz Cintas and Remael, 2007, 173). Thus, subtitles are as independent as possible from each other and frequently subtitles are clauses with a simple syntax. In other words: subtitles

should be structured "that they are semantically and syntactically self-contained" (Díaz Cintas and Remael, 2007, 172). Subtitles therefore divide long and complex sentences into various smaller ones.

In order to reformulate and condensate the spoken dialogue, subtitlers often apply the following modifications: simplifying verbal periphrases, generalizing enumerations, using a shorter near-synonym or equivalent expressions, using simple rather than compound tenses, changing word classes, short forms and contractions, changing negations or questions into affirmative sentences or assertions, indirect questions into direct questions, simplifying indicators of modality, turning direct speech into indirect speech, changing the subject of a sentence or phrase, manipulating of theme and rheme, turning long or compound sentences into simple sentences, active sentences into passive or vice versa, using pronouns (demonstrative, personal, possessive) and other deictics to replace nouns, or noun phrases and merging two or more phrases/sentences into one (Díaz Cintas and Remael, 2007, 151-162). "Most subtitles display a preference for conventional, neutral word order, and simple well-formed stereotypical sentences" (Díaz Cintas and Remael, 2007, 185).

Dialogues of films often contain colloquial speech which the subtitlers must transfer to a written form (Romero, 2011, 19). Subtitles present spoken language in a written form, thus containing characteristics of spoken as well as written language (Hillman, 2011, 386). The spoken dialogue is also based on a script which is a form of prefabricated spoken language and thus differs from spontaneous spoken language (Romero, 2011, 19f.). For example, the prefabricated dialogues of films only contain a limited number of colloquial words or expressions and are clearer and more coherent than spontaneous speech

The discussed characteristics show that subtitles use a special and unique style: "Grammar and lexical items tend to be simplified and cleaned up, whereas interactional features and intonation are only maintained in some extent" (Díaz Cintas and Remael, 2007, 63f.).

3.3. Fansubs

In the last years, the technology for audiovisual translation has developed very fast and nowadays, subtitling programs are accessible and affordable for individuals as well. This has facilitated fansubbing (also called amateur subtitling). Fansubs came up in the 1980s, when Europeans and Americans wanted to watch and understand Japanese animated films. Because few or no subtitles were available in foreign lan-

guages, the fans decided to create the subtitles by themselves and to make them available in the Internet for free. Today, fansubs are not restricted to Japanese films anymore but created for different kinds of film and language pairs (Díaz Cintas and Remael, 2007, 23-25). Fansubs often appear shortly after the premiere of a film, which often causes a competition between different fansubs for the same film. Because of this, producers of fansubs need to work as quickly as possible, which often results in a reduction of quality. In order to save time, the amateur subtitlers often use the intralingual subtitles for the deaf and hard of hearing instead of the original dialogue and translate them into their language. Some fansubs communities also pay attention to a good quality. The quality is of fansubs is thus heterogeneous (Goldstein, 2009, 32). Also note that a producer of fansubs has neither the equipment and material available nor the expertise of a professional subtitler (Díaz Cintas, 2009a, 49f.).

Fansubs often make use of different font colors in order to identify the speakers and introduce glosses and meta linguistic notes to explain certain film scenes, actions or elements. In contrast to professional subtitles, fansubs can contain more than two lines per subtitle, mostly however they consist of two lines. Additionally, the allowed number of characters in fansubs is higher than in professional subtitles (Díaz Cintas and Remael, 2007, 23-25).

3.4. The Translation of Subtitles

Díaz (2009b) states that "with an average 30% to 40% expansion rate from English into most other European languages, reduction is obviously the most important strategy in subtitling." (Díaz Cintas, 2009a, 26). This may cause non-verbatim translations. The indicated expansion rate is surprisingly high and cannot be confirmed with our corpora and translations from English to Spanish.

The actions and dialogues of films are embedded in cultural contexts. For the translation the subtitler has to consider that the cultural contexts of the source and target language differ. It is difficult to translate sentences, expressions or words referring to a cultural context (= culture-bound terms) that are only known in the source culture, but not in the target culture (Díaz Cintas and Remael, 2007, 45). To deal with these cultural differences, the subtitler has three possibilities: foreignization, neutralization and naturalization. In the case of foreignization, the translator tries to maintain all source-culture elements (e.g. personal relations, institutions or social customs) without changing them. The viewer has to adapt his or her point of the target culture of the film. In contrast, in the case of naturalization, the subtitler

tries to familiarize. This means that elements of the source culture (culture of the source language) are adapted or replaced by similar elements of the target culture (culture of the target language). In the case of neutralization the subtitler tries to avoid the identification with any culture. These approaches are often combined in subtitled films (Fong, 2009, 43).

Ambiguities are another challenge to the translation of subtitles. In translations from English to Spanish for example, the second form of address causes difficulties: the English pronoun *you* can be used in formal and informal situations, whereas in Spanish two different forms exist (*tú* as informal pronoun of address, *usted* as formal pronoun of address). The subtitler has to consider the context and decide for each situation, which Spanish address form would be adequate (Díaz Cintas and Remael, 2007, 189f.).

3.5. Consequences for Statistical Machine Translation

Compared to other types of texts, subtitles are suitable for the statistical machine translation. The space and time constraints (see section 3.2) have a positive and negative impact for the SMT of subtitles. The advantage is that each subtitle is semantically and syntactically as self-contained as possible. Thus, we can consider each subtitle as an independent sentence or at least as an independent clause. A subtitle-based input for the training of the translation and language model is possible. In addition, we discussed that subtitles are syntactically and grammatically as simple as possible, what facilitates the automatic translation. Furthermore, subtitles are short (generally up to 41 characters). This is also an advantage for statistical machine translation. The heterogeneity and especially different guidelines of the subtitling companies are a disadvantage for SMT, especially if the training corpus for the SMT system contains subtitles from different companies. The output of the SMT has to be adapted to the rules of the corresponding company again. Especially for fansubs (see section 3.3), the heterogeneity and the differences in quality are substantial.

The translation of culture-bounded terms (see section 3.4) and marked speech can be difficult as well, because a statistical machine translation system cannot choose between neutralization, foreignization and naturalization. Thus, translators or proofreaders must check manually if the SMT system translated the culture-bounded terms and marked speech adequately.

We discussed that subtitles contain characteristics from spoken as well as written

language (see section 3.2). This might not have a considerable impact for the statistical machine translation of subtitles as the system is trained with a corpus which already contains these characteristics. It might be tested in other projects if these characteristics augment the heterogeneity of the subtitles.

The consistent use of standard language and the correction of grammatical and phonetic errors reduce the heterogeneity of the subtitles. This probably has a positive impact on the translation quality.

The omissions and reductions from the dialogue to the subtitles (see section 3.2) do not have a considerable impact on the machine translation because the adaptions have already been made in the subtitles of the source language.

We have discussed the difficulty to understand the subtitles without the acoustic and visual mode. Manual evaluations of SMT-systems for subtitles can be problematic because without the acoustic and visual mode it may be difficult to decide if a translation is adequate or not.

Some of the subtitle characteristics are especially suitable for SMT whereas others cause difficulty for the automatic translation. A post correction and adaption of the produced translation remains indispensable when using a state of the art SMT system for automatic translation.

4. Statistical Machine Translation of Film Subtitles: SUMAT

In this chapter, we will present the state of the art in the development of statistical machine translation systems for film subtitles. We focus on translation systems for the language pair English-Spanish. Until now "no effective tools or services that can provide automatic subtitle MT services" (Petukhova et al., 2012, 21) have been available. For the training of a translation system, large amounts of high quality parallel subtitles are necessary. "Such data is available for some language pairs, e.g., in the OPUS OpenSubtitle corpus (Tiedemann, 2009), but being based on openly available subtitles with no quality checking, their usefulness has yet to be determined" (Petukhova et al., 2012, 21).

4.1. The SUMAT Project

SUMAT[10] stands for "Service for SUbtitling by MAchine Translation" and is a European project with the goal "to increase the efficiency and productivity of the European subtitle industry, while enhancing the quality of its results, thanks to the effective introduction of SMT technologies in the subtitle translation processes" (SUMAT, 2012, 3). Therefore, the partners which cooperate with the SUMAT project will provide an online translation service in order to semi-automate the translation of subtitles (Petukhova et al., 2012, 21). The partners that cooperate with SUMAT are Vicomtech, Athens Technology Center, Invision Ondertiteling, VoiceScript International, Applied Language Solutions, Deluxe Digital Studios, University Of Maribor and Titelbild Subtitling and Translation. Moreover, SUMAT subcontracted TextShuttle.

In total the SUMAT project covers nine European languages in total. The partners of the SUMAT project develop machine translation systems for 14 language pairs. "The targeted language pairs are: English-Dutch; English-French; English-German;

[10] www.sumat-project.eu

English-Portuguese; English-Spanish; English-Swedish and Serbian-Slovenian. The translation service will be working in both directions"(SUMAT, 2012, 3). The main project is divided into four steps:

1. Collection and preparation of the corpora
2. Development of the statistical machine translation systems
3. Development of an online platform to access the translation systems
4. Evaluation (user-based)

For the training corpus, at least 700'000 subtitles per language pair are needed. The partners of the SUMAT project decided to use the professionally translated subtitles from the project partners of the SUMAT project. Until 2012, the project collected 987'818 parallel subtitles for the language pair English-Spanish (see table1).

Language pair	**number of parallel subtitles**
English-Dutch	1'452'963
English-French	1'566'431
English-German	2'282'329
English-Portuguese	762'716
English-Spanish	987'818
English-Swedish	959'304
Serbian-Slovenian	215'097
Total	8'226'658

Table 1.: Corpora of the SUMAT project (2012) (SUMAT, 2012, 4)

For the training of the language model, the partners of the SUMAT project used additional monolingual subtitles of the target language. They plan to do experiments in which they include parallel subtitles from EuroparlTV[11], TED[12] and OpenSubtitles[13] for the training of the translation and language model (SUMAT, 2012, 5). The subtitles of the OpenSubtitle corpus were produced by amateurs and the quality is therefore not guaranteed (see section 5.2). So far the partners which cooperate with the SUMAT project have used the OpenSubtitle corpus only for language pairs for which not sufficient professional subtitles are available (Petukhova et al., 2012, 24).

All subtitles provided by the project partners were produced by "professionally trained translators and experienced freelancers" (Petukhova et al., 2012, 22) and

[11] http://www.europarltv.europa.eu/en/home.aspx

[12] http://www.ted.com/

[13] http://www.opensubtitles.org/en

the target language is always their mother tongue. After the translation, editors, proofreaders and reviewers revise the subtitles. The subtitles have to fullfill the following criteria:

- the subtitles are not longer than two lines
- the character number is adequate
- "all proper names have been consistently treated"
- the punctuation marks are correctly used
- the capitalization is correct
- the line breaks are appropriate
- all important and essential elements are translated
- no considerable changes of meaning between the source and target language
- no grammatical, syntactical and spelling mistakes (Petukhova et al., 2012, 22f.)

The partners of the SUMAT project use Moses (see section 6) for the training of the SMT system. For each language pair a different part of the corpus was used as development and test set. For the training of the language model, they used the IRST Language Modeling toolkit (see section 6) which is freely available (SUMAT, 2012, 5f). "The scores obtained on the subtitles training and test sets are quite promising having obtained BLEU scores above 20 (except for Slovenian-Serbian, Serbian-Slovenian and English-German), that can be translated into reasonable quality for the majority of the SUMAT systems" (SUMAT, 2012, 7). In the SUMAT project the SUMAT test set of 2012 (4000 lines per language) was used for the evaluation of the English-Spanish system (version 2012). The obtained BLEU score was 25.5%[14]. In our project (see section 8) we used the translation of the SUMAT test set to develop our grammar checker and to make a restricted error analysis. We did not find out the size of the used training corpus and how the test set was extracted. In 2013, the partners of the SUMAT project trained a new translation system for the language pair English-Spanish. They used different training, test and development sets than for the system of 2012. For the training of the translation model of 2013 they used 803'064 parallel subtitles and the development set contained 2'000 subtitles[15]. For the translation of their test set of 2013 (4000 lines) they obtained a BLEU score of

[14]internal report of 2012

[15]internal information

29.3% [16].

To improve the results, the SUMAT researchers tried to include linguistic features and annotations for the training of the models. They applied POS-tagging, lemmatization, dependency parsing, compound splitting, named entity recognition and phrase tables filling. They tested each of these linguistic features on one Romanic and one Germanic language and plan corresponding experiments with Slavic languages (SUMAT, 2012, 7). The first experiments for the language pair English-Spanish showed that the inclusion of linguistic features (Part-of-Speech Tags, Lemma on target side, syntactic information on target side and their combination) cause a reduction of the automatic evaluation scores in the system of 2012 as well as in the system of 2013[17]. Christian Hardmeier (2008) also made experiments with SMT translations of film subtitles in which he included additional linguistic information like part of speech tags and morphological information. He observed that the achieved improvement was small. "On the whole, none of the gains was large enough to justify the effort of producing the annotations" (Hardmeier, 2008, 78). Therefore, in our project, we train the statistical translation system without including any of these linguistic features.

[16]internal report of April 2013

[17]internal reports of 2012 and April 2013

5. Corpora

The corpus we used for the training of our SMT system consists of two parts: the OpenSubtitle corpus and the VSI subtitles. In this chapter we will discuss the composition and some characteristics of the corpus in order to understand the positive and negative impacts on the performance of our SMT systems and to elaborate possibilities for the improvement of the training corpus. We will discuss the characteristics of subtitles (see section 3.2) and fansubs (see section 3.3) and their expected impacts for SMT (see section 3.5).

Each part of our parallel corpus consists of two files, one with the English subtitles and the other with the aligned Spanish subtitles. For both corpora there is only a little meta-information. For example, we cannot identify the film boundaries nor do we have complete information about the included films and the original film language. Furthermore, there are no time codes or names of the subtitlers available.

One line in our corpus always refers to one subtitle, which is either one or two lines displayed simultaneously on the screen.

For the rest of this document, we always refer to only one line of our corpus when we talk about a subtitle.

5.1. Parallel Corpus of VSI Subtitles

The VSI[18] Group (Voice and Script International) is an international dubbing and subtitling company and provided a set of 77'781 parallel subtitles for English-Spanish for this project. In the following subsections we discuss the composition and characteristics of the VSI subtitle corpus.

[18] http://www.vsi.tv/

5.1.1. Composition and Preparation of the Parallel Corpus in SUMAT

The researchers of the SUMAT project also used the VSI subtitles as part of their training corpus. The VSI subtitles which we used were extracted from the complete SUMAT corpus after preprocessing. We describe the preparation of the SUMAT corpus in detail in order to understand the characteristics of the parallel subtitles from VSI.

The subtitling companies provided the subtitles in the original format and with the original names. The corresponding genre and domain were available as meta-information. Additionally, they distinguished between subtitles of scripted and unscripted films. Scripted means that the dialogue of the film is based on a prefabricated script, for example in series, dramas etc. Unscripted means that no prefabricated script of the spoken dialogue exists, as for example, in talk shows, reality shows or interviews[19].

Most of the subtitles were in a text-based (*.txt*) or binary format (*STL, 890, PAC, o32/s32/x32*). The provided subtitles in the text-based format were heterogeneous. Each subtitle contained the index and time code at the beginning. Some subtitles additionally contained the duration information or fields named *TIMEIN* and *TIMEOUT*. In other subtitles, the end time was not indicated at all, which means that important information was missing.

The group of files in the *o32/s32/x32* format "turned out to be too complicated to support without format specifications and was manually converted to .txt." (Fishel et al., 2012, 3) . For subtitles in other formats, the partners of the SUMAT project used corresponding format converters. The success rate of the conversion was 99.4%, which means that in the conversion step 0.6% of the data was lost. The binary formats additionally included formatting information, such as fonts, text positioning and coloring. The partners of the SUMAT project excluded this information for the training step because only the subtitle text can be used for the training of the translation system. Therefore, the researchers of the SUMAT project extracted and separately saved the additional formatting information in order to integrate them again after the translation process (Fishel et al., 2012, 3f.)The partners of the SUMAT project converted all files to Unicode (Fishel et al., 2012, 4). Finally, all files were formatted in Unicode.

In the next step, the partners of the SUMAT project applied the Lingua:Ident

[19]internal information from Mark Fishel

package to identify the language of the subtitles and to filter out all the subtitles in languages which were not within the scope of the project. This caused the exclusion of 0.8% of the converted subtitles.

Then, the partners of the SUMAT project had to align the corresponding language versions of the provided subtitle files. The researchers of the SUMAT project realized the automatic document-alignment, the mapping between subtitles of the same film in different languages, by using the document names and the meta information. In most of the cases the document name consists of an arbitrary name representing the film and a language acronym (e.g. *Movie_Title_en.txt* and *Movie_Title_fr.txt*). This allows the documents to be aligned with the same name and different language acronyms. The documents' meta information contains genre and domain. The documents can only be aligned if they are labeled with the same genre and domain. With these techniques the partners of the SUMAT project were able to align more than half of the documents. Subsequently, they compared the time codes to align the remaining documents: "two documents are considered parallel if at least 90% of the time codes correspond to each other" (Fishel et al., 2012, 5). Finally, the partners of the SUMAT project aligned the remaining documents manually. In some cases no corresponding counterpart was found. In total, the partners of the SUMAT project aligned 83% of all the files. For the language pair English-Spanish were generated 1641 document pairs (Fishel et al., 2012, 5f).

In the next step, the subtitles of these document pairs had to be aligned. The basic idea was to match the subtitles with the same time code within the aligned documents. A difficulty was that some companies adjust the time codes during the translation process. Therefore, the partners of the SUMAT project implemented an algorithm which accepts certain shifts of the time codes with the condition that this shift "stays almost constant" (Fishel et al., 2012, 6). This algorithm also allows many-to-one alignments. With this method 85% of the subtitles in document-aligned files can be aligned. The partners of the SUMAT project aligned 779'500 subtitles in total for the language pair English-Spanish (Fishel et al., 2012, 6).

The partners of the SUMAT project also prepared a sentence-based corpus for experiments with linguistic features to verify if they obtain better results with a sentence-based training corpus. The sentence-based corpus did not improve the translation quality, sometimes the translation quality was even reduced which therefore lead to the decision to work with the subtitle-based corpus.[20]

[20]internal information from Mark Fishel

5.1.2. Characteristics of the VSI Subtitle Corpus

In this section we discuss the characteristics of the VSI subtitle corpus. In total, this corpus contains 77'781 parallel subtitles. In the Spanish part we found 76'817 different subtitles. This means that only 1.24% of the subtitles are repetitions. In the English part we detected 76'724 (=98.64%) different subtitles, which means that only 1.36% are repetitions. Many of the repetitions in the VSI subtitles are short and common expressions or clauses which are typical for spoken dialogues (see ex. 5.1).

(5.1) *I don't know.*
in the world.
Okay.
It's interesting.
I couldn't believe it.

We also found some unexpected repetitions which contain proper names. (see ex. 5.2-5.5).

(5.2) *That is an amazing run by Toby Dawson!*

(5.3) *Joining me now is CEO Lars Rebien Sørensen.*

(5.4) *I name this ship British Reliance.*

(5.5) *so then we could call Ian and let him know his funds are available.*

We can explain the first two examples (ex. 5.2 and 5.3) as coincidental. The repetitions occur in different contexts and we can assume that they occur in different scenes of the same film or serie. In examples 5.4 and 5.5 the contexts (the adjacent lines in the subtitle file) are very similar, which means that the adjacent lines differ only in the punctuation, line breaks and detailedness (compare ex. 5.6 and 5.7). We can conclude that our set contains different subtitle versions for the same film (see section 3.2). The repetitions will most likely not have a considerable negative effect on our statistical machine translation system because of the small number of repetitions in relation to the big corpus.

(5.6) *In 604 years*
of recorded ship building in Sunderland on the River Wear,
no larger ship has been launched
than this 16,000-tonne motor tanker.
She's another addition to the fleet
which carries the products of the Persian oilfields.

I name this ship British Reliance.
May God bless her and all who sail in her.

(5.7) *In 604 years of recorded shipbuilding,*
no larger ship has been launched than this 16,000-tonne motor tanker.
She's an addition to the fleet,
which carries the products of the Persian oil fields across the world.
I name this ship British Reliance.
May God bless her and all who sail in her.

An English subtitle contains on average 44.5 characters per subtitle and a Spanish subtitle contains on average 44.7 characters. The number of characters of the English and Spanish subtitles is almost the same. In total, the English version contains 733'813 tokens, on average 9.43 tokens per line. Excluding the tokens which only consist of punctuation marks or special characters, we counted in total 641'607 tokens and on average 8.25 tokens per line (= subtitle). In total, the Spanish version contains 683'031 tokens and on average 8.78 tokens per line (= subtitle). Excluding the tokens which only consist of punctuation marks or special characters, we counted in total 593'087 tokens and on average 7.63 tokens per subtitle. This shows that the Spanish version on average contains fewer tokens per line than the English one.

In the following, we will analyze the subtitles with the highest and lowest number of tokens. The shortest subtitles in both languages only contain one token. The longest Spanish subtitle (concerning the number of tokens, see ex. 5.8) contains 26 tokens and the corresponding English sentence (see ex. 5.9) only contains 17 tokens. The longest English sentence (see ex. 5.10) even consists of 29 tokens, whereas the corresponding Spanish sentence (see ex. 5.11) only consists of 11 tokens.

(5.8) *De quien aspira a ser el hombre más rápido del mundo, pasamos al hombre apodado "el hombre sin piernas más rápido del mundo"*

(5.9) *They didn't play much on court, and she retire*[21] *due to a foot injury.*

(5.10) *And then saying, "What's stopping you getting on it?" Now most people, on it. What's stopping you getting on it."*

(5.11) *Hay que determinar por qué no todos demuestran ese compromiso.*

[21] error in the corpus

In the longest sentences of both languages we observe a remarkable difference of the number of tokens in the source and target language. In both cases it is not a verbatim translation. The sentences of the source and target language do not correspond to each other, neither in regard to the grammatical structure nor to the lexical elements. We can explain these observations either by an adaptation and reformulation done by the subtitle translator or by an alignment error. If we consider the adjacent subtitles of the longest English subtitle (see ex. 5.12) and compare them to the corresponding Spanish subtitles (see ex. 5.13), we observe that the adjacent subtitles are aligned perfectly. Therefore, we assume that the reason for this non-verbatim translation in this example is a reformulation done by the translator. If we consider the adjacent subtitles of the longest Spanish line, we can observe, that they do not correspond to each other. Furthermore, the proper names in the adjacent subtitles of the source and target language differ. The Spanish subtitles (see ex. 5.14) contain the proper names *Shanghái* and *Oscar Pistorius*, in contrast to the English subtitles (see ex. 5.15), which contain *Roger Federer* and *Basel*. Therefore, we assume that the reason for the non-verbatim translation in this example is an alignment error. Non-verbatim translations can have a negative impact for the machine translation system. The examples showed that the corpus still contains some noise data. A criterion to exclude sentences which are not aligned correctly or not translated literally would be the comparison of the number of tokens in the English and Spanish subtitle. If the number of tokens in the source and target language differs considerably, as in the examples we discussed, we can assume a non-literal translation or an alignment error.

(5.12) *The first thing is you've got to know where the bus is, where the bus is going.*
That's all about the strategy.
And then saying, "What 's stopping you getting on it?" Now most people on it . What 's stopping you getting on it."
Ultimately, the goal is you've got to be on the bus.
If you're not on the bus and you don't want to go where it's going,
then it isn't the right business for you.

(5.13) *Primero, hay que saber dónde está, adónde va,*
es decir, entender la estrategia.
Hay que determinar por qué no todos demuestran ese compromiso.
Todos los empleados deben subirse al autobús.
Los que no deseen hacerlo ni seguir su dirección,
no deberían trabajar con nosotros.

(5.14) *Gay marcaría 9,71, el récord estadounidense, insuficiente*
semanas más tarde, un tiempo de 9,69 en Shanghái, le demostró al mundo
De quien aspira a ser el hombre más rápido del mundo, pasamos al hombre
apodado" el hombre sin piernas más rápido del mundo"
A Oscar Pistorius le fueron amputadas las piernas
Luego de lograr el oro en los 200m. y el bronce en los 100 m.
la IAAF dictaminó que las prótesis que usa otorgan a Pistorius

(5.15) *It was a fairytale story as she won the US Open in her 3rd comeback.*
Roger Federer met his wife through tennis,
They didn't play much on court, and she retire due to a foot injury.
She had the best seat in the house since
They weren't rushed of upcoming nuptials,
but in 2009, they were married in his hometown of Basel.

We discussed that subtitles should be semantically and syntactically as self-contained and independent as possible and that, in an optimal case, each subtitle represents one sentence or at least an independent clause (see section 3.2). To get a notion of how many subtitles are full sentences, we counted all sentences which either end with a terminal punctuation mark and start with an uppercase letter, or which start and end with quotes. In the Spanish version we also had to consider the initial punctuation marks. In the English version 22'028 subtitles (28.32%) and in the Spanish version 21'020 subtitles (27.03%) are full sentences.

Oral feature:	**absolute frequency:**	**relative frequency** (in relation to the number of lines):
... (suspension points)	1223	1.57
hm/hmm	1	-
ahem	0	-
shh	0	-
ugh	0	-
yeah	256	0.33%
oh	95	0.12%
ah	2	-
uh	0	-

Table 2.: Interjections and suspension points in the English part of the VSI subtitles

We also discussed that subtitles have characteristics of spoken as well as written texts. To get a notion of the use of oral features in the corpus, we chose a random collection of oral features (interjections and suspension points) and investigated their frequencies in the corpus (see tables 2 and 3). This will be particularly interesting in comparison with their frequencies in the OpenSubtitle corpus. We used the lowercased corpus to count the frequencies and the relative frequencies are rounded on two right-of-comma positions. If the absolute frequency is low and the rounded relative frequency would is 0.00%, the relative frequency is not indicated. The result shows that only few of these interjections occur. The suspension points occur frequently in both languages.

Oral feature:	**absolute frequency:**	**relative frequency:**
... (suspension points)	958	1.23%
ay	0	-
eh	3	-
ah	3	-
bah	0	-
oh	4	-
uf	0	-
uy	0	-

Table 3.: Interjections and suspension points in the Spanish part of the VSI subtitles

Interestingly, we found some lines which are completely written in uppercase letters. Without watching the film it is difficult to understand what these uppercase letters stand for. They most likely represent written text which appears somewhere in the image (e.g. street names). It is important for SMT, that they occur in the source and target language and that they are translated literally. Examples (see ex. 5.16 and 5.17) show, that in the VSI subtitles uppercase subtitles are translated literally.

(5.16) English:
WHY GOLDMAN SACHS ASSET MANAGEMENT
FOR ABSOLUTE RETURN INVESTING?
Spanish:
¿POR QUÉ GOLDMAN SACHS ASSET MANAGEMENT
PARA INVERTIR EN RETORNO ABSOLUTO?

(5.17) English:
FOR FURTHER INFORMATION
ON THE GS US EQUITY ABSOLUTE RETURN PORTFOLIO,

OR VISIT www.gs.com / gsam.
Spanish:
PARA MÁS INFORMACIÓN
SOBRE EL GS US EQUITY ABSOLUTE RETURN PORTFOLIO,
O VISITE www.goldmansachsfunds.es

5.2. OpenSubtitle Corpus

OPUS[22] is an open multilingual parallel corpus which consists "of translated open source documents available on the Internet" (Tiedemann and Nygaard, 2004, 1). The available corpora were already preprocessed, so that they can be used directly for other applications, for instance, as training corpora for the training of an SMT system with Moses. Currently, the OPUS corpus consists of manuals of open-source software, texts from the European Union, film subtitles and biomedical data.

For the training of our SMT system we used the OpenSubtitle corpus, in addition to the VSI subtitles. The OpenSubtitle corpus is "a parallel corpus of movie subtitles", contained in OPUS (Tiedemann, 2009, 2). The corpus contains 32'947'747 sentences per language and was processed and composed by Jörg Tiedemann.

5.2.1. Composition of the OpenSubtitle Corpus

The website www.OpenSubtitles.org provides a large amount of freely available movie subtitles. These subtitles were created by fans and amateurs and therefore, high quality is not guaranteed. Consequently, the database contains a lot of noise data, thus preprocessing and filtering is advisable (Tiedemann, 2009, 2).

To create the corpus, Jörg Tiedemann (2009) used all the files which were available in the subviewer-format and converted all the files in the microDVD format to the subviewer-format. Then he converted the files to Unicode and applied "textcat", a language identification tool, to check if for each file the right language is indicated as meta information. If "textcat" confirmed the indicated language, the file was used for the corpus. With this technique he excluded a large amount of noise data. A disadvantage is that "textcat" is only trained for 46 languages and consequently, all the files in other languages were excluded (Tiedemann, 2009, 2).

[22] http://opus.lingfil.uu.se/

Another challenge was that often multiple subtitle uploads for the same movies exist. In these cases, Jörg Tiedemann and his team only included the most current version. "However, we include multiple copies of subtitles for the same movie if they correspond to different video files and have a corresponding subtitle file in a different language in the database" (Tiedemann, 2009, 3). In consequence, there are various identical or at least similar subtitle files in the OpenSubtitle corpus.

Then, Jörg Tiedemann converted all files which should be included in the corpus to xml. Additionally, he tokenized and split the subtitles at the sentence boundaries with regular expressions. Consequently, the corpus is sentence-based and not subtitle-based like the VSI corpus (Tiedemann, 2009, 3).

For the alignment Jörg Tiedemann used the timing information, which caused several problems. The corpus is sentence-based, but the time-slots are adjusted to subtitles. Hence the given time-slots do not really correspond to the sentences (one sentence = one line in the corpus). Therefore, Tiedemann adjusted the time-slots of the subtitles for sentences (Tiedemann, 2007, 4). In order to do this, he used a script which interpolated the time-slots and "used the nearest "time events" and calculated the time proportional to the strings in between"(Tiedemann, 2007, 4). Finally, the script verified the adjusted time information. If the script detected decreasing or identical time indications, it added 0.0001 seconds. This was done iteratively until the time indications were consistent. In the next step, each sentence of the source language was aligned with the sentence of the target language which shows the highest time overlap. An advantage is that with this approach, insertions, deletions and supplementary information are not aligned, which is correct (Tiedemann, 2007, 6).

A disadvantage of this approach is that "time information is unfortunately not very reliable. There are often slight differences in the timing that cause devastating errors when aligning purely based on this information" (Tiedemann, 2009, 4). Consequently, the subtitle files have to be synchronized before the alignment step. Jörg Tiedemann used the speed (duration of the subtitle) and the starting time of subtitles as indicators for timing differences. Generally the speed is continuous in the subtitles of the source language as well as in the subtitles of the target language. Therefore, the difference in the speed can be calculated with automatically detected anchor points. This enables the calculation of the different starting times. With this approach the time information of the target language was adjusted and they yielded better alignment-results than with length-based approaches (Tiedemann, 2007, 5f.).

5.2.2. Characteristics of the Parallel OpenSubtitle Corpus

In total, the corpus contains 32'947'747 sentences per language. In the Spanish part we counted 11'146'550 and in the English part, 9'136'575 different sentences. This means that in the Spanish part 66.17% and in the English part 72.27% of the sentences are repetitions. We cannot explain the high quantity of repetitions to be coincidence. We observed that not only short and common sentences, but also longer and rarer sentences are repeated. The reason for the high number of repetitions are different subtitle versions, respectively film versions, included in the corpus (see section5.2.1). In order to get a notion of how many versions of a film are included in the corpus, we tested it with two well-known films: *My Sister's Keeper* and *Into the Wild*. We counted five versions of *My Sister's Keeper* and nine versions of *Into the Wild*. This high quantity of repetitions can cause two types of problems for SMT. On one hand, it makes an extraction of a development and test set complicated, because the probability of the sentences of these sets appearing is high, which would manipulate the results. On the other hand, the repeated sentences will receive more weight than the sentences of films of which only one version is included. It must be tested if the repetitions really have a negative or manipulating impact on the translation quality.

Besides the exact repetition of sentences, passages and complete movies, we found sentences which are not identical but very similar. In most of these cases, we identified only slight typographical differences. Comparing the examples 5.18 and 5.19, we notice that the only difference are the missing quotes in the end of the third line in example 5.19. Because the corpus is sentence-based, there are no differences concerning the line breaks, as we observed it in the VSI-subtitles (see section 5.2.2). In a lot of cases in which we observe little typographic differences, the adjacent sentences are also identical or similar and therefore, we assume that they are from different subtitle versions of the same movie.

(5.18) *Chris was writing his story, and it had to be Chris who would tell it.*
"To call each thing by its right name.
"By its right name."

(5.19) *Chris was writing his story, and it had to be Chris who would tell it.*
"To call each thing by its right name.
"By its right name.

(5.20) *You just know it... ...through and through.*

(5.21) *You just know it. . . . . .through and through.*

Examples 5.20 and 5.21 also show typographic differences. In the second example (ex. 5.21), the suspension points are tokenized. Additionally, this example shows another interesting characteristic of the corpus: in both languages the suspension points are often doubled. The reason is probably that the subtitles are spread over two lines on the screen. In other words: In the original, the first line on the screen ends with suspension points and the second line starts with suspension points. For the parallel corpus, the two lines shown simultaneously on the screen are connected and therefore, the suspension points appear twice. Further projects might test if the double use of the suspension points has any effect (positive or negative) on the translation quality of the corpus.

The English version of the OpenSubtitle corpus contains 253'880'611 tokens and on average 7.71 tokens per sentence. If we only take the tokens which contain at least one letter or digit into consideration, we count a total of 203'496'702 tokens and an average of 6.2 tokens per sentence. In the Spanish version we count a total of 239'972'272 tokens and an average of 7.28 tokens per sentence. Excluding all tokens which only consist of punctuation marks or special characters, we count in total 179'724'765 tokens and on average 5.45 tokens per sentence. The number of tokens in the English and Spanish sentences is similar. On average, the English sentences contain about one token more. As in the VSI corpus (see section 5.1.2), the shortest sentences of both languages only contain one token. The longest English sentence consists of 819 tokens, which we found out to be a song text. The corresponding Spanish subtitle does not include the song text, instead, there is a link to a blog. The longest Spanish sentence contains 416 tokens and seems to be a senseless combination of words. Both examples show that the corpus contains some extremely long sentences which are not useful for our SMT systems, because they are not verbatim translations.

The English version contains per sentence on average 31.45 characters, whereas a Spanish sentence contains on average 33.87 characters. The number of characters of the English and Spanish sentences is almost the same, as we observed for the number of tokens already.

We calculated the frequencies for some oral features (selected interjections and suspension points) as we did for the VSI subtitles (see tables 4 and 5). We used the lowercased corpus again to count the frequencies. If an interjection or suspension points occur more than once in a line, we count them only once.

The results show that for both languages, the suspension points occur in 6-8% of the sentences. Regarding interjection, we note that in English *yeah* (0.54%), *oh* (0.28%) and *ah* (0.02%) occur often. In Spanish, the most frequent of the selected interjections are *ah*(0.08%), *eh*(0.17%), *ay* (0.05%) and *oh* (0.17%).

Oral feature:	**absolute frequency:**	**relative frequency:**
... (suspension points)	11986527	6.03%
hm/hmm	5561	0.02%
ahem	337	0.00%
shh	1991	0.01%
ugh	409	-
yeah	176984	0.54%
oh	93725	0.28%
ah	6228	0.02%
uh	64660	0.20%

Table 4.: Interjections and suspension points in the English part of the OpenSubtitle corpus

Oral feature:	**absolute frequency:**	**relative frequency:**
... (suspension points)	2587600	7.85%
ay	15210	0.05%
eh	56557	0.17%
ah	24847	0.08%
bah	258	0.00%
oh	55063	0.17%
uf	156	-%
uy	451	-%

Table 5.: Interjections and suspension points in the Spanish part of the OpenSubtitle corpus

Some of the sentences also contain information which does not appear in the spoken dialogue, for example, the description of certain actions, gestures or the name of a speaker (see ex. 5.22, 5.23, 5.24 and 5.25). Probably, this additional information in the subtitles occurs because the creators of amateur subtitles sometimes do not differentiate between interlingual subtitles for foreigners and intralingual subtitles for the deaf and hearing-aided. All of this information is written in uppercase letters and sometimes, this information is additionally indicated with brackets. When the non-dialogic information expresses the speaker name, a colon follows the name (see ex. 5.24). Considering the corresponding Spanish translations, we observe that

the additional information was not translated literally, but the information was included in the dialogue. In the first two examples (see ex. 5.22 and 5.23) in the English subtitles, dialogic and non-dialogic information is combined. In the corresponding Spanish subtitles, the non-dialogic information is omitted. In the other two examples (see ex. 5.24 and 5.25), the English subtitles only consist of non-dialogic information. In the corresponding Spanish subtitles this information was integrated into the dialogue. For example, the indication of *Melvin* as a speaker (ex. 5.24) in the English subtitles was replaced by a vocative in the Spanish subtitles. The described action *banging on door* in example 5.25 was replaced by a command in the Spanish subtitles (translated in English: *open the door!*). The integration into the dialogue as well as the omission cause a change of the content of the subtitles and the translations cease to be literal. This can have a negative impact on the performance of the SMT.

Furthermore, we found additional meta information about the film in some sentences, for example information such as the name of the company or person who synced the film (see ex. 5.26). We found out that they are translated literally and therefore, this should not have a negative impact on the translation quality of the SMT system.

(5.22) English: *I... (LAUGHlNG)*
Spanish: *¡Ah!*

(5.23) English: *(STAMMERlNG) I'il just tell you it is perfect.*
Spanish: *Tengo que decirte que es perfecto.*

(5.24) English: *- MELVlN:*
- Mama, Frank's on the phone! Spanish: *¡Melvin!*
¿Dónde está mamá? - ¿Qué? - ¡Que Frank está al teléfono! - ¡Eh, mamá! - ¿Cómo estás, Frankie?

(5.25) English: *(BANGlNG ON DOOR)*
Spanish: *¡Abran la puetra, joder!*

(5.26) English: *Subs synched by ShooCat Thx to javaopera for providing them*
Spanish: *Subs sincronizados por ShooCat Gracias a javaopera por ofrecerlos*

In the English as well as in the Spanish sentences, we found spelling mistakes which look like OCR errors (see ex. 5.27 - 5.35), also visible in the examples above). The mistakes are mainly confusions between letters which look similar to each other, for example *I* and *l*, *i* and *l* or *f* and *t*. These mistakes could have a negative

effect on the word-alignment and translation quality, as well as result in additional ambiguities (e.g. *if* and *it* are confused).

(5.27) *I said I'il fix it, didn't I?* (*I'il* instead of *I'll*)

(5.28) *- l think I can find a way out.* (*l* instead of *I*)

(5.29) *ln the glow how fair you shine!* (*ln* instead of *In*)

(5.30) *I had thought he would refuse... fhat he would quit the French Navy.* (*fhat* instead of *that*)

(5.31) *Hey, what'd you stop tor?* (*tor* instead of *for*)

(5.32) *Será el legado de ml Reino.* (*ml* instead of *mi*)

(5.33) *Intentó levantarme como al tipo de lslandia.* (*lslandia* instead of *Islandia*)

(5.34) *Tú nl ves lo que haces.* (*nl* instead of *ni*)

(5.35) *- Lo fengo fodo lisfo para fi - Lo fengo fodo listo para ti* (*fengo* instead of *tengo*, *fodo* instead of *todo*, *lisfo* instead of *listo*, *fi* instead of *ti*)

5.2.3. Characteristics of the Monolingual Data for Spanish

The OpenSubtitle corpus also contains monolingual subtitle data. In our project we use the Spanish monolingual part of the corpus to complement the training data of the language model. Our experiments have shown that not all sentences from the parallel corpus for the language pair English-Spanish are contained in the monolingual corpus for Spanish and vice versa. Therefore, we combined both corpora for the training of the language model. The monolingual corpus contains 84'961'910 sentences. In total, the corpus consists of 624'048'972 tokens and an average of 7.35 tokens per sentence. Excluding the tokens, which only consist of punctuation marks or special characters, we count in total 465'212'694 tokens and an average of 5.48 tokens per sentence. The average number of tokens per sentence is almost the same as in the Spanish part of the parallel OpenSubtitle corpus.

In total, the monolingual Spanish corpus contains 19'949'864 different lines, which means that 76.52% of the lines are repetitions. Consequently, we assume that the monolingual corpus also contains various subtitle versions for the same film.

5.3. Comparison of the two Parts of our Parallel Corpus

The corpus which we used for the training of our SMT system contains the VSI corpus and the OpenSubtitle corpus. In the last two sections (see section 5.1 and 5.2) we analyzed the characteristics of both corpora. In the present section we compare their characteristics and predict possible influences on the performance of the machine translation system.

First, we observe that the OpenSubtitle corpus contains 431 times as many lines as the VSI corpus. Therefore, the subtitles of VSI will have a lower impact on the translation quality of our system than the OpenSubtitles. Additionally, the quality of the two corpora differs. The VSI corpus contains only professional subtitles, whereas the Opensubtitle corpus consists of amateur subtitles. We have discussed that the OpenSubtitle corpus contains OCR-like errors which do not appear in the VSI subtitles. Furthermore, the use of non-dialogic information in the source and target language of the OpenSubtitle corpus is inconsistent. As a consequence, the OpenSubtitle corpus contains subtitle pairs which are not literal translations. In addition, we observed that the Opensubtitle corpus contains much more repetitions (in other words: various subtitle versions for the same film) than the VSI subtitles.

On average, the VSI subtitles contain more tokens than the subtitles of the OpenSubtitle corpus. A possible explanation is that the VSI subtitles are subtitle-based, whereas the OpenSubtitles are sentence-based. Further projects might investigate whether the combination of a sentence-based and a subtitle-based corpus has an impact on the translation quality. Another difference is that the VSI-Subtitles were already tokenized in the SUMAT project whereas the OpenSubtitles must still be tokenized.

We conclude that the two corpora are heterogeneous in different aspects, for example, the amount of subtitles, quality and length of the lines. The impact of this heterogeneity on the SMT-system might still be investigated with a systematic analysis.

5.4. Improvement Possibilities for the Corpora

The characteristics of the two parts of our corpus (described in section 5.1.2 and 5.2.2) showed that there are various aspects which can have a negative impact on the

translation quality. We will propose some possibilities of how to improve the corpora. Furthermore, we will include some of the proposed improvements in our project, further projects might analyze systematically if the improvement steps would improve the translation system.

5.4.1. Weighting of the Corpora

The impact of the high quality subtitles (VSI subtitles) on the translation quality is low due to the significantly smaller number of subtitles compared to the number of OpenSubtitles. Further projects might test if models with weighted corpora yield better translation qualities. (Sennrich, 2012) proposes and describes different possibilities in order to automatically calculate the weights for the combination of different corpora (in- and out-of-domain corpora). These approaches may be tested and applied for the combination of our subtitle corpora of heterogeneous quality.

5.4.2. Exclusion of Non-literal Translations

The discussed examples of both corpora lead to the assumption that long sentences are often not translated literally, what can have a negative impact on our SMT system. We discovered that the numbers of tokens of these sentences respectively subtitles in the source and target language differ considerably from each other. Additionally, we showed that on average the number of tokens in the English and Spanish sentences is almost the same. Therefore, we propose to test an approach in which we filter out long subtitles that have a considerable difference of the numbers of tokens in the source and target language. Moreover, this approach can be used for the error detection and confidence estimation in the translated output: If the difference of the number of tokens is considerable it is probable that the sentence or subtitle, has not been translated literally and that it has to be translated or corrected manually.

In this project, we filtered out lines (of both languages) which contain more than 35 tokens. Further projects might investigate if this number of maximal tokens can be improved.

Another source for non-literal translations is non-dialogic information (see section 5.2.2). Non-dialogic information is marked in the English text what enables the detection of non-dialogic information. It might be tested if an exclusion of all sentence-pairs which non-dialogic information improves the translation quality.

5.4.3. Exclusion of Repetitions

Another problem we detected was the high number of repetitions in the OpenSubtitle corpus, which occurred because different and similar subtitle versions for the film were in the corpus. To avoid (negative) impacts on the translation quality, we tried to filter out the repetitions. In a first experiment, we filtered out all repeated sentence-pairs of the English and Spanish corpus. With this approach we did not only filter out the non-coincidental repetitions of subtitles for the same movie, but also coincidental repetitions, for example, short and common expressions. The exclusion of non-coincidental repetitions caused a decline of the translation quality (see section 7.2). In order to ascertain that we only exclude identical sentences of different subtitle versions for the same movie, we applied the following criteria for the exclusion of repetitions. A sentence-pair is only filtered out if:

- the English sentence appears more than once in the English part of the corpus
- the Spanish translations of these repeated English sentences are also identical
- ten preceding and following sentences of the considered English sentence and its repetition are identical

The filtered results showed, that we did not exclude many of the repetitions of subtitle versions for the same film, because of little typographical differences (e.g. blanks) in one of the adjacent lines. Therefore, we decided to consider only 5 preceding and following sentences of the considered English sentence and its repetition. With this approach we reduced the corpus to 26'850'109 lines. Control samples showed that the corpus still contains a lot of repetitions of different versions for the same film.

In the monolingual Spanish data we did not filter out the repetitions.

5.4.4. The Use of Sentence- or Subtitle-based Corpora

The experiments of SUMAT showed that the use of a sentence-based corpus does not improve the quality of the SMT system[23]. The OpenSubtitle corpus is only available as sentence-based version. Therefore, further projects might investigate if it is more recommendable to mix sentence- and subtitle-based corpora or to use the sentence-based corpus of VSI subtitles in combination with the sentence-based version of the OpenSubtitle corpus.

[23]internal information from Mark Fishel

5.4.5. Correction of OCR-Errors

In the OpenSubtitle corpus we found OCR-like errors (see section 5.2.2). In a first attempt we used a spell checker to detect and correct spelling mistakes. Most of the automatic spell checkers only detect mistakes and propose correction possibilities but do not correct the errors automatically. We decided to use the spell checker of LanguageTool (see section 8.1.2) using the option to automatically apply the most probable correction possibility automatically. We tried to correct the examples discussed above (see ex. 5.22 - 5.35). Only in a few cases did the spelling work perfectly, for example LanguageTool corrected *BANGlNG* to *banging* (see ex 5.41)and *lslandia* to *Islandia* (see ex 5.43). Note that the correction process of LanguageTool changed the uppercase letters of *BANGlNG* to lowercase letters. In some other cases LanguageTool did not detect and correct the spelling mistakes at all. LanguageTool only uppercase the first letters when they occurred at the beginning of the sentence. For instance *l* was corrected to *L* instead of *I* (see ex 5.37) or *ln* was corrected to *Ln* instead of *In* (see ex 5.38). In the remaining cases the correction of LanguageTool was completely wrong. LanguageTool changes for instance *fhat* to *chat* instead of *that* (see ex 5.39), *fengo* to *fango* instead of *tengo* (see ex 5.45) and *fodo* to *godo* instead of *todo* (see ex 5.45). The results show that it is not appropriate to use the spell checker of the LanguageTool to correct the OCR-like errors in our corpus.

(5.36) *I said I'il fix it, didn't I?* was changed to *I said I'Al fix it, didn't I?* (*I'Al* instead of *I'll*)

(5.37) *l think I can find a way out.* was changed to *L think I can find a way out.* (*L* instead of I)

(5.38) *ln the glow how fair you shine!* was changed to *Ln the glow how fair you shine!* (*Ln* instead of In)

(5.39) *I had thought he would refuse... fhat he would quit the French Navy.* was changed to *I had thought he would refuse... chat he would quit the French Navy.* (*chat* instead of that)

(5.40) *Hey, what'd you stop tor?* was changed to *Hey, what'd you stop tor?* (*tor* instead of for)

(5.41) *(BANGlNG ON DOOR)* was changed to *(banging ON DOOR)* (correct)

(5.42) *Será el legado de ml Reino.* was changed to *Será el legado de mal Reino.* (*mal* instead of mi)

(5.43) *Intentó levantarme como al tipo de lslandia.* was changed to *Intentó*

levantarme como al tipo de Islandia. (correct)

(5.44) *Tú nl ves lo que haces.* was changed to *Tú na ves lo que haces.* (*na* instead of ni)

(5.45) *Lo fengo fodo lisfo para fi - Lo fengo fodo listo para ti* was changed to *Lo fango godo silfo para di - Lo fango godo listo para ti* (*fango* instead of tengo, *godo* instead of todo)

We decided to apply manually created rules to correct at least some of the frequent OCR errors. We applied these rules before the tokenization, because OCR errors can cause tokenization errors. A disadvantage is that a correction of OCR errors in the untokenized corpus is more complicated, because the words are sometimes combined with punctuation marks. Our rules only consider single words or combinations of words with quotes. We only created correction rules for a selection of common words. A manual analysis showed that sometimes *if* is used instead of it. Our rules do not correct such ambiguous errors, because we cannot decide automatically if *if* or *it* is meant.

We applied our correction rules after the elimination of repetitions. Table 6 shows which and how many words we corrected with our rules. In the table we do not distinguish between uppercase and lowercase words. The results show that our rules corrected much more words in the English part of the corpus than in the Spanish part. Possible reasons for this are our selection of common words for the correction rules and the development of more rules for the correction of English words than for the correction of Spanish words.

In a further experiment we tested the application of the following generalized rules for the correction of the English lines:

1. If *I* occurs at the end of a token and the letter before is lowercase, it is separated with a space because in most of the cases it is due to a tokenization error.

2. If *I* occurs in the middle of a word (if it is neither the first nor the last letter) between lowercase letters, it is changed to *l* because normally *I* does not appear in the middle of a word.

In total these generalized rules modified 61'132 words (see ex. 5.46-5.50). Example 5.50 shows, that the first rule causes some errors. In a test sample we observed that our rules never change correct words, but often our rules are able to correct wrong words.

English corrections:	**Number of corrections:**	**Spanish corrections:**	**Number corrections:**
ln changed to *in*	7'361	*lr* changed to *ir*	71
ls changed to *is*	9'987	*nl* changed to *ni*	100
lt changed to *it*	211	*ml* changed to *mi*	183
lf changed to *if*	17'777	*tl* changed to *ti*	1
'il changed to *'ll*	620'853	*sl* changed to *si*	49
'Il changed to *'ll*	14'054	*conmlgo* changed to *conmigo*	5
I've changed to *I've*	482	*contlgo* changed to *contigo*	8
I changed to *I*	1'240	*proplo* changed to *propio*	8
I'm changed to *I'm*	560	*propla* changed to *propia*	5
'ii changed to *'ll*	12	*ei* changed to *el*	360
fo changed to *to*	123	*dei* changed to *del*	201
tor changed to *for*	267	*ai* changed to *al*	117
foo changed to *too*	31	*eiia* changed to *ella*	13
nof changed to *not*	19	*eiias* changed to *ellas*	2
fhing changed to *thing*	14	*eiios* changed to *ellos*	4
fhat changed to *that*	13	*ia* changed to *la*	409
fhis changed to *this*	409	*ias* changed to *las*	113
fhese changed to *these*	3	*io* changed to *lo*	319
fhose changed to *those*	0	*ios* changed to *los*	186
whaf changed to *what*	11	*ie* changed to *le*	66
buf changed to *but*	9	*ies* changed to *les*	18
fhere changed to *there*	2		
fhe changed to *the*	116		
'f changed to *'t*	781		
total:	674'335	total:	2'594

Table 6.: Correction of OCR-errors

(5.46) *There are times whenI can only stand you as a human being.* was changed to *There are times when I can only stand you as a human being.* (correct, rule 1)

(5.47) *You're the kind of counsellorI like in my courtroom.* was changed to *You're the kind of counsellor I like in my courtroom.* (correct, rule 1)

(5.48) *FoIlow me.* was changed to*Follow me.* (correct, rule 2)

(5.49) *"Frodo was reaIly courageous, wasn't he, Dad?"* was changed to *"Frodo was really courageous, wasn't he, Dad?"* (correct, rule 2)

(5.50) *Did you know your girI belonged to those Japanese patriotic organizations?* was changed to *Did you know your gir I belonged to those Japanese patriotic organizations?* (incorrect: *gir I* instead of *girl*, rule 1)

For the correction of the Spanish part of the corpus we applied only one rule which corrected 92'798 words:

1. *I* is always changed to *l* if it does not occur in the initial position of a word

To the monolingual data we only applied the generalized rules.

6. Training of the Translation System with Moses

For the training of the SMT system and the preprocessing of the corpora we used the tools and scripts provided by Moses[24]. Moses is an open source implementation for SMT, available since 2005. Users can train SMT systems with their own parallel corpora. "The two main components in Moses are the training pipeline and the decoder" (Koehn, 2013, 12). The training pipeline consists of tools written in Perl and C++ which create a SMT system that takes plain text as input. The decoder is written in C++ and uses the created SMT system to translate a text from the source language into the target language (Koehn, 2013, 12).

Moses needs two sentence aligned files as input. One of the files contains the English text and the other file the corresponding Spanish translation. The user manual of Moses recommends the use of a sentence-based corpus, where each line should consist of one sentence. The input files should not contain empty lines or sentences which are longer than 100 tokens and the whole input text should be lowercase (Koehn, 2013, 169).

In this project we used the OpenSubtitle corpus and the VSI subtitles for the training. First we preprocessed both already aligned corpora. The OpenSubtitle corpus is sentence-aligned and the VSI corpus is subtitle-aligned. The VSI subtitles were tokenized in the SUMAT project already. We tokenized the OpenSubtitle corpus with the tokenizer provided by Moses. Then we applied the cleaning script of Moses. With this script we filtered out sentences which contain more than 35 tokens and empty lines. Finally, we lowercased the whole corpus with a script provided by Moses.

The training-script does the word alignment and creation of the lexical translation tables (see section 2) with Giza++[25] based on the IBM models, the extraction and scoring of the phrases (phrase-based model, see section2), the creation of the lexi-

[24] http://www.statmt.org/moses/

[25] http://www.statmt.org/moses/giza/GIZA++.html

calized reordering model and the building of the generation models and the configuration file (Koehn, 2013, 167). The language model was built with the IRSTLM[26] toolkit.

After the training of the SMT system, we improved the SMT system with minimum error rate training (MERT). Moses provides a MERT-script for this. Different models (e.g. language model, translation model, reordering model) are involved to find the optimal translation for an input sentence. In order to combine the probabilities, a discriminative linear model is used (Bertoldi et al., 2009, 1). The log probabilities of the different models are the features of this discriminative linear model and can be weighted differently. MERT finds a set of weights that will offer the best translation quality (Bertoldi et al., 2009, 2). For the calculation of these weights, the script uses a development set, which consists of a small part of the corpus which was not used for the training.

It must be noted that repetitions of this experiment may yield slightly different results. This is due to the fact that "Machine Translation training is non-convex. This means that there are multiple solutions and each time you run a full training job, you will get different results. In particular, one will see different results when running Giza++ (any flavour) and MERT."[27].

[26] http://hlt.fbk.eu/en/irstlm

[27] http://www.statmt.org/moses/?n=Moses.FAQ

7. Automatic Evaluation

In this chapter we describe the automatic evaluation of our trained translation systems. The first section (see section 7.1) gives a brief introduction into the functionality of the applied automatic evaluation scores and section 7.2 will show and discuss the performance scores we obtained for the trained SMT systems.

7.1. Automatic Evaluation Scores

An advantage of automatic evaluation scores is that they are cheaper than human evaluations and take less time. A main idea of most automatic evaluation scores is: "The closer a machine translation is to a professional human translation, the better it is." (Papineni et al., 2002, 311). For the automatic evaluation of a translation we need a metric to calculate the evaluation score and reference translations.

7.1.1. BLEU

BLEU stands for Bilingual Evaluation Understudy and is based on the word error rate metric (WER). The BLEU score is based on the comparison of n-grams (sequences of n tokens) in the evaluation translation and reference translation(s). The more n-grams of the translation and the reference translation(s) are the same, the better the translation and the higher the BLEU score (Papineni et al., 2002, 312). Each n-gram of the reference translation can be mapped only once to a corresponding n-gram of the evaluated translation, which means that the BLEU score requires a one to one relationship. For the comparison the modified n-gram precision is calculated. The modified n-gram precision is defined as the number of identical n-grams of the translation and the reference translation(s) divided by the total number of n-grams in the translation. The BLEU score is calculated from the combined modified n-gram precisions with the following formula (Papineni et al., 2002, 315) (Koehn, 2010, 226):
$BLEU\ score = brevity\ penalty\ *\ exp(\sum_{i=1}^{n} w_i log precision_i)$

(n = the maximum order of n-grams to be matched, typically 4)
(w = the weights for the different precisions, typically 1)

The logarithm of the modified precisions is used because "the modified n-gram precision decays roughly exponentially with n" (Papineni et al., 2002, 314). Weights are introduced to improve the combination of the modified n-gram precisions. Typically, the BLEU score is calculated using only the modified n-gram precisions for unigrams, bigrams, trigrams and fourgrams.

The formula for the BLEU score also considers the length of the translated sentence compared to the reference translation. In the optimal case, the length of the evaluated translation is the same as the length of the reference translation(s). In case the evaluated translation is longer than the reference translation, the length difference has a negative impact on the modified n-gram precisions, because not all n-grams of the evaluated translation can be mapped. In the reverse case, the length difference has no impact on the modified precision (Papineni et al., 2002, 314). Therefore, a multiplicative "brevity penalty factor" was introduced in the formula. "With this brevity penalty in place, a high-scoring candidate translation must now match the reference translations in length, in word choice, and in word order" (Papineni et al., 2002, 315). The brevity penalty factor is calculated with the formula $e^{1-lengthreference/lengthtranslation}$, except the factor is set 1 in case the sentence of the evaluated translation is longer than the reference translation (Callison-Burch and Osborne, 2006, 2).

7.1.2. METEOR

METEOR stands for Metric for Evaluation of Translation with Explicit Ordering. The higher the METEOR score, the better the translation. The calculation of the METEOR score is "based on explicit word-to-word matches between the translation and a given reference translation" (Lavie and Agarwal, 2007, 1), this means that METEOR only considers unigrams. METEOR executes word alignment between the strings of the evaluated translation and the reference translation(s) with different word-mapping modules. These modules also consider synonyms and apply stemming in order to improve the word alignment (Lavie and Agarwal, 2007, 1). First all possible matches of words are detected. Then METEOR selects the best matches considering the word alignment probability for the whole sentence. If there is more than one word-alignment choice for the sentence with the same probability, METEOR chooses the alignments for which the word order of the aligned words is most similar in the evaluated and reference translation (= least number of crossing

unigram mappings) (Lavie and Agarwal, 2007, 2).

After the word alignment, the METEOR score can be calculated. First, METEOR calculates the precision dividing the number of mapped unigrams by the total number of unigrams (=tokens) in the evaluated translation. In contrast to BLEU score, METEOR additionally calculates the recall. To do this, METEOR divides the number of mapped unigrams by the total number of unigrams (=tokens) in the reference translation. Then, we compute the METEOR score by using the "parameterized harmonic mean of P and R" (= F mean) (Lavie and Agarwal, 2007, 2).

Additionally, the METEOR score considers the word order: if the word order of the aligned words in the evaluated translation is similar to the word order of the aligned words in the reference translation, the METEOR score is better than if the word order is completely different. To achieve this, a penalty score is introduced which calculates the number of chunks divided by the number of aligned words. A chunk is the biggest group of adjacent words occurring in the evaluated translation as well as in the reference translation (Lavie and Agarwal, 2007, 2).

The METEOR score is calculated with the following formula (Lavie and Agarwal, 2007, 2):
*METEOR score = (1-penalty) * F_{mean}*
In case of more than one reference translation, the metric is calculated for each of these reference translations and the result with the highest probability is chosen.

7.1.3. TER

TER stands for Translation Edit Rate. "TER is defined as the minimum number of edits needed to change a hypothesis so that it exactly matches one of the references, normalized by the average length (number of tokens) of the references." (Snover et al., 2006, 3). The fewer edits, the lower the TER score is and the better the translation. Edits are either insertions, deletions or substitutions of words or shifts of token sequences. TER does not consider the shift distances and the number of tokens in the shifted sequences (Snover et al., 2006, 3).

7.1.4. Levenshtein Distance

The Levenshtein distance between two strings is the number of edits which are needed to change one string (e.g. sentence of the evaluated translation) into a given target string (e.g. the reference translation). Edits are defined as insertions, deletions

or substitutions of characters (Carstensen et al., 2010, 557f.).

To calculate the Levenshtein distances we used a script provided by Mark Fishel. This script calculates the average number of required edits (keystrokes) per sentence to change the evaluated translation into the reference translation. Moreover, it calculates the average Levenshtein distance for sentences of different length and it calculates the percentage of exact matching sentences. Furthermore, the script counts how many sentences can be changed to the corresponding sentence of the reference translation with fewer than five edits (= lev-5-distance).

7.2. Test Systems, Results and Comparisons

7.2.1. Procedure

First, we trained different systems and evaluated them automatically. The automatic evaluation scores enable the comparison of the performance of our different systems. We made experiments in which we changed the composition of the corpora, trying to improve their quality. For each experiment, we trained a SMT system which we describe, evaluate and compare in the following sections 7.2.2 - 7.2.7. For the evaluation, we calculated three automatic evaluation scores with the tool MultEval[28]: BLEU (see section 7.1.1), METEOR (see section 7.1.2) and TER (see section 7.1.3). If necessary we also calculated the Levenshtein distance. We used the test set of the SUMAT project (version 2012), which consists of 4000 lines and is subtitle-based. On average, it contains 10.83 tokens per line in the English version and 9.54 tokens per line in the Spanish version. Consequently, the lines of the SUMAT test set are on average about one token longer than in the VSI corpus and about two or three tokens longer than the lines in the OpenSubtitle corpus. The test set contains no repetitions and is already lowercased and tokenized. This means that for the evaluations, we had to compare the tokenized and lowercase translation of our SMT system with the reference translation.

Then we evaluated our best performing SMT system more in detail with two additional different test sets: the VSI test set and the OpenSubtitle test set. For the tokenization and lowercasing of these test sets we applied the scripts provided by MOSES.

Finally, we compared the results of our best performing system with the results of the SUMAT project for the language pair English-Spanish.

[28] https://github.com/jhclark/multeval

7.2.2. System 1

We trained our first system (see table 7) with the major part of the parallel OpenSubtitle corpus for English-Spanish provided by Jörg Tiedemann (2009). Because we planned to use additional test sets, we excluded most of the subtitle versions of two films *My Sister's Keeper* and *Into the Wild*. After the exclusion of the subtitle versions for these two films, the training set still contained 32'937'896 subtitles (original size of the corpus: 32'947'747 subtitles).

As we discussed in section 5.2.2, the different subtitle versions of a film are not always exactly the same and therefore we did not manage to exclude all subtitle versions of these two films. Therefore, most of the lines of the test sets also exist in the training set and this would result in good but not meaningful evaluation scores. As a result, we could not use these extracted films as test sets.

After running the cleanup-script provided by Moses, we trained the translation system with 32'766'356 parallel subtitles. We trained the language model with the Spanish part of the bilingual parallel corpus, which means 32'947'747 subtitles. For the corpus which we used to train the language model with, we decided not to run the cleaning-script, because it is not necessary to cut out long sentences.

The calculated BLEU Score for this system is 22%, the METEOR score is 45.7% and TER is 62.7%.

	System1
# of subtitles: (= number of subtitles per language in total)	32'927'896 (OpenSubtitles)
# of subtitles: (= number of subtitles per language to train the translation model)	32'766'356 (OpenSubtitles)
# of subtitles: (= number of subtitles in the target language to train the language model)	32'927'896 (OpenSubtitles)
# of subtitles: (= number of subtitles per language in the development set)	0
Test set of SUMAT (2012): **BLEU:** **METEOR:** **TER:**	4'000 lines 22% 45.7% 62.7%

Table 7.: System 1: Training with the parallel OpenSubtitle corpus

7.2.3. System 2 and System 3

	System 1	System 2
# of subtitles (total):	32'927'896 (Opensub.)	33'010'050 (Opensub.+VSI)
# of subtitles to train the TM:	32'766'356 (Opensub.)	32'853'252 (OpenSub.+VSI)
# of subtitles to train the LM:	32'927'896 (Opensub.)	33'010'050 (OpenSub.+VSI)
# of subtitles in the dev. set:	0	0
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	22%	24.3%
METEOR:	45.7%	48.2%
TER:	62.7%	59.8%

Table 8.: Comparison of system 1 (only OpenSubtitles as training data) and system 2 (OpenSubtitles and VSI subtitles as training data)

For the training of the second model (see table 8) we complemented the training set of system 1 with the VSI corpus. Accordingly, we used 33'010'050 subtitles for the training of the language model and 32'853'252 parallel subtitles for the training of the translation model.

We evaluated the translation quality with the SUMAT test set (version 2012). The BLEU score is 24.3% which shows that the addition of the VSI subtitles caused an increase of the BLEU score of 2.3 percentage points. We also got an increased METEOR score (48.2%) and TER score (59.8%) (see table 8).

	System 2	System 3
# of subtitles (total):	33'010'050 (Opensub.+VSI)	33'010'050 (Opensub.+VSI)
# of subtitles to train the TM:	32'853'252 (OpenSub.+VSI)	32'853'252 (OpenSub.+VSI)
# of subtitles to train the LM:	33'010'050 (OpenSub.+VSI)	33'010'050 (OpenSub.+VSI)
# of subtitles in the dev. set:	0	1'409 (OpenSub)
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	24.3%	25.9%
METEOR:	48.2%	48.3%
TER:	59.8%	57.1%

Table 9.: Comparison of system 2 (not tuned) and system 3 (tuned). Both systems were trained with the same data (OpenSubtitles and VSI subtitles)

We decided to tune this system with the MERT script provided by Moses (see section 6 and table 9). For this we used a small development set containing 1'409 subtitles in total (1'234 different subtitles) as input for MERT. This tuning improved the BLEU Score to 25.9%. Also, the TER score (57.1%) shows a considerable improvement. In

the METEOR score (48.3%) we only observe a small improvement of 0.1 percentage points.

The results show (see section 7.2.3) that the combination of the Opensubtitle corpus with the VSI corpus yields better results compared to the system that only uses the Opensubtitle corpus. Therefore, we used this combination for our further experiments.

7.2.4. System 4

For this experiment we used the cleaned training corpus of system 2. We excluded all repetitions (the identical subtitles of the source language which have identical translations in the target language) of the OpenSubtitles. This reduced the number of parallel subtitles in the training corpus to 19'644'742. With this approach we tried to avoid that some subtitles carry more weight purely because the corpus contains various subtitle versions for the same film (see section 5.2.2). We did not exclude the repetitions of the VSI corpus, because of the small number of repetitions in the VSI corpus (see section 5.1.2). For the training of the language model we used the Spanish part of the cleaned training corpus.

Compared to system 2 we observe a considerable decrease of all the evaluation scores (see table 10). We conclude that the exclusion of all repetitions in the OpenSubtitle corpus has a negative impact on the translation quality. Therefore we decided not to refine or tune this system.

	System 2	**System 4**
# of subtitles to train the TM:	32'853'252 (OpenSub.+VSI)	19'644'742 (OpenSub.+VSI)
# of subtitles to train the LM:	33'010'050 (OpenSub.+VSI)	19'644'742 (OpenSub.+VSI)
# of subtitles in the dev. set:	0	0
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	24.3%	21.8%
METEOR:	48.2%	46.1%
TER:	59.8%	63.3%

Table 10.: Comparison of system 2 (training corpus includes repetitions, not tuned) and system 4 (training corpus does not include repetitions, not tuned)

7.2.5. Systems 5 and 6

For the training of system 5 (see table 11) we improved the selection of the development and test set. We formed a test set by extracting a complete film (545 lines) from the complete OpenSubtitle corpus. We selected this film with 545 lines, because we had to choose a film which does not occur more than once in the corpus. To create the VSI test set we extracted 4'001 lines from the VSI corpus. We also extracted 15'001 lines from the VSI corpus and used them as a development set. We decided to only use VSI subtitles for the development set because they guarantee a high quality.

Moreover, we made several adaptions in the corpus. First, we excluded all repetitions of the OpenSubtitle corpus which have an identical Spanish translations and appear in the same context (5 identical preceding and following lines, see section 5.4). This reduced the number of subtitles in the OpenSubtitle corpus to 26'850'109. In total the corpus we used to train the translation model contained 26'787'230 subtitles after the cleanup.

	System 2	**System 5**
# of subtitles (total):	33'010'050 (Opensub.+VSI)	26'908'887 (Opensub.+VSI)
# of subtitles to train the TM:	32'853'252 (OpenSub.+VSI)	26'787'230 (OpenSub.+VSI)
# of subtitles to train the LM:	33'010'050 (OpenSub.+VSI)	111'870'797 (OpenSub.+VSI+monoling.)
# of subtitles in the dev. set:	0	0
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	24.3%	25.0%
METEOR:	48.2%	48.8
TER:	59.8%	59.1%

Table 11.: Comparison of system 2 (not tuned) and system 5 (not tuned, additional monolingual data, improvement of the extraction of repetitions, correction of OCR errors)

Secondly, we added the Spanish monolingual data (84'961'910 sentences of the OpenSubtitle corpus) to the corpus for the training of the language model. This means that the corpus for the training of the language model consists of three parts: the VSI-Subtitles, the Spanish part of the parallel Opensubtitle corpus (without repetitions which occur in the same context) and the monolingual Spanish OpenSubtitle corpus (still containing all repetitions). In a third step we tried to correct some of the OCR-like errors in the OpenSubtitle corpus (see section 5.4). We applied a script with rules to correct selected tokens and two generalized rules on the English

part of the OpenSubtitle corpus. On the Spanish part we applied the rules which correct specific tokens, but no generalized rules. We corrected no OCR errors in the Spanish monolingual data of the OpenSubtitle corpus.

The results (see table 11) show that these adaptions improve the translation quality. All evaluation scores increased. Therefore, we decided to tune this system by applying the MERT-script (see section 6). The development set which we used for the tuning contained 15'001 parallel VSI subtitles. The tuning caused a further improvement of all the evaluation scores, for example, the BLEU score increased 1.7 percentage points (see table 12).

	System 5	**System 6**
# of subtitles (total):	26'908'887 (Opensub.+VSI)	26'908'887 (Opensub.+VSI)
# of subtitles to train the TM:	26'787'230 (OpenSub.+VSI)	26'787'230 (OpenSub.+VSI)
# of subtitles to train the LM:	111'870'797 (OpenSub.+VSI+monoling.)	111'870'797 (OpenSub.+VSI+monoling.)
# of subtitles in the dev. set:	0	15'001 (VSI)
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	25.0%	26.7%
METEOR:	48.8%	49.8%
TER:	59.1%	56.1%

Table 12.: Comparison of system 5 (not tuned) and system 6 (tuned). Both systems were trained with the same corpus.

7.2.6. Systems 7 and 8

In the next experiment we improved the correction of OCR-errors. We extended our script with a generalized rule to correct OCR errors in the Spanish part of the training corpus (see section 5.4). We also applied this rule to the monolingual Spanish data.

The BLEU score and TER score of system 7 (see table 13) are equal as in system 5. The METEOR score of system 7 is even 0.1 percentage points worse than in system 5. The difference of this METEOR score is not significant and we can possibly explain it with the fact that training with Moses is non-convex (see section 6). We conclude that the correction of OCR-like errors with the generalized rule for the Spanish part does not have an effect on the automatic evaluation scores.

We created a new system 8 by tuning system 7. We tuned system 7 with the same development set as used for system 6 (see table 12) to check if the evaluation scores

of the tuned systems 6 and 8 show a significant difference. For system 8 (see table 14) we observe almost the same evaluation scores as for system 6. The BLEU score and METEOR score of system 8 are 0.1 percentage points higher, whereas the TER score is 0.1 percentage points worse.

	System 5	**System 7**
# of subtitles (total):	26'908'887 (Opensub.+VSI)	26'908'887 (Opensub.+VSI)
# of subtitles to train the TM:	26'787'230 (OpenSub.+VSI)	26'787'230 (OpenSub.+VSI)
# of subtitles to train the LM:	111'870'797 (OpenSub.+VSI+monolling.)	111'870'797 (OpenSub.+VSI+monoling.)
# of subtitles in the dev. set:	0	0
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	25.0%	25.0%
METEOR:	48.8%	48.7%
TER:	59.1%	59.1%

Table 13.: Comparison of system 5 (not tuned) and system 7 (not tuned). In system 7 additional OCR errors were corrected.

	System 6	**System 8**
# of subtitles (total):	26'908'887 (Opensub.+VSI)	26'908'887 (Opensub.+VSI)
# of subtitles to train the TM:	26'787'230 (OpenSub.+VSI)	26'787'230 (OpenSub.+VSI)
# of subtitles to train the LM:	111'870'797 (OpenSub.+VSI+monoling.)	111'870'797 (OpenSub.+VSI+monoling.)
# of subtitles in the dev. set:	0	15'001 (VSI)
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	26.7%	26.8%
METEOR:	49.8%	49.9%
TER:	56.1%	56.2%

Table 14.: Comparison of system 6 (tuned) and system 8 (tuned, improved correction of OCR errors)

An automatic comparison showed that the systems 6 and 8 translated 1'082 of the 4'000 sentences differently. We decided to compare 50 differently translated sentences manually. This manual evaluation (see table 15) should help to decide which system works better. We decided for each analyzed sentence which system translated better. Translations for which we were not able to decide the category are counted in the category *impossible to decide*. In most of the cases we had to consider the context, to decide which translation is better. Although this evaluation is based on a small number of translated sentences and contains subjective decisions,

it will give a perception of the performance of the two systems.

total:	50
System 6 translates better:	15
System 8 translates better:	11
Impossible to decide:	24

Table 15.: Manual comparison of the sentences that were translated differently by system 6 and 8

For 24 of 50 sentences we were not able to decide which translation is better (see ex. 7.3). Either both sentences were completely wrong or they showed lexical variations which are both possible. This shows that the quality of the translation is difficult to judge without any defined criteria and error classes to base the decision on. For 15 (30%) of the evaluated sentences system 6 produces a better translation, whereas for 11 (22%) sentences system 8 performs better (see ex. 7.1 and 7.2). In total, system 6 outperforms system 8 for more sentences than vice versa.

(7.1) **English subtitle:**
it was a massive undertaking .
System 6:
fue una tarea ***enorme*** *.*
System 8:
fue una tarea ***masiva*** *.*
Comparison:
System 6 performs better.

(7.2) **English subtitle:**
there was no such thing as multiple cameras , being synched together .
System 6:
no ***hay*** *nada como varias cámaras , están sincronizados juntos .*
System 8:
no ***había*** *tal cosa como varias cámaras , están sincronizados juntos .*
Comparison:
System 8 performs better.

(7.3) **English sentence:**
i 'm sure . when we did it , we weren 't even synched to film ,
System 6: *estoy seguro . cuando lo hicimos ,* ***no*** *estábamos sincronizados para filmar ,*
System 8:

*estoy seguro . cuando lo hicimos , **ni siquiera** estábamos sincronizados para filmar ,*

Comparison:

Impossible to decide which system performs better.

For the comparison of system 6 and 8 we additionally used the Levenshtein distance and the percentage of exact matches (see section 7.1.4 and table 16). The results show that system 6 achieves 3.65% exact matches, which is slightly more than the percentage of exact matches for system 8 (3.48%). System 6 also yields slightly more lev-5-matches than system 8. In contrast, the average Levenshtein distance for system 8 (22.16) is slightly better than for system 6 (22.17). The script also calculated the Levenshtein distance for sentences of different lengths (see table 17). Generally, we observe that the longer the sentences are the more difficult is the translation. For the sentences of all lengths system 6 yields on average slightly more absolute matches than system 8. For the lev-5-matches we observe the same. Only for sentences with 4-6 tokens system 8 yields on average more lev-5-matches. Calculating the average Levenshtein distance, system 8 outperforms system 6, except for sentences containing 1-3 tokens. The observed differences between the absolute matches, lev-5-matches and the average Levenshtein distances are little and not significant.

	System 6	System 8	System 9
Total absolute matches:	3.65%	3.48%	3.60%
Total Lev-5 matches:	10.35%	10.10%	9.85%
Total average lev dist:	22.17	22.16	22.14

Table 16.: Levenshtein distance and exact matches of system 6 and 8

tokens:	**1-3**	**4-6**	**7-9**	**10-12**	**13-15**	**16-18**	**19+**
Absolute matches (syst.6):	24.44%	9.08%	3.89%	0.79%	0.49%	0.00%	0.00%
Lev-5 matches (syst.6):	45.56%	22.35%	12.18%	4.28%	2.63%	0.52%	0.00%
Avg lev. dist. (syst.6):	7.53	12.83	18.55	25.42	30.81	35.75	40.93
Absolute matches (syst.8):	22.47%	9.03%	3.80%	0.60%	0.49%	0.00%	0.00%
Lev-5 matches (syst.8):	46.07%	22.28%	11.83%	4.28%	2.30%	0.51%	0.00%
Avg lev. dist. (syst.8):	7.65	12.73	18.48	25.32	30.64	35.75	39.73
Absolute matches (syst.9):	23.60%	8.97%	3.89%	0.61%	0.51%	0.00%	0.00%
Lev-5 matches (syst.9):	46.07%	21.52%	11.19%	3.86%	2.37%	0.54%	0.00%
Avg lev. dist (syst.9):	8.03%	12.83	18.58	25.66	30.94	35.87	40.00

Table 17.: System 6 and 8: Levenshtein distance, sentences of different length

The manual and automatic evaluation of systems 6 and 8 showed that the correction of the OCR errors in the Spanish part did not significantly improve the translation quality. We can think of three possible reasons: First, the comparison of two systems with small performance differences cannot provide significant results. The training with Moses is non-convex and, therefore, we get small performance variations which cannot be assigned to changes in the training corpora. Second, our rules possibly do not correct enough OCR-errors to result in a significant improvement of the translation quality. Third, the number of OCR errors in the training corpus is perhaps too small to negatively impact the translation quality. Further projects must investigate these possible explanations. Additionally, we might test if OCR errors are indicators for subtitles of bad quality and if we should exclude these sentences from the corpus.

Although the performance of both systems is almost the same, we decided to continue with the training corpora of system 8.

7.2.7. System 9

	System 8	**System 9**
# of subtitles (total):	26'908'887 (Opensub.+VSI)	26'908'887 (Opensub.+VSI)
# of subtitles to train the TM:	26'787'230 (OpenSub.+VSI)	26'787'230 (OpenSub.+VSI)
# of subtitles to train the LM:	111'870'797 (OpenSub.+VSI+monoling.)	111'870'797 (OpenSub.+VSI+monoling.)
# of subtitles in the dev. set:	15'001	15'001 (VSI)
Test set of SUMAT (2012)	4'000 lines	4'000 lines
BLEU:	26.8%	26.8%
METEOR:	49.9%	49.5%
TER:	56.2%	56.0%

Table 18.: Comparison of system 8 (tuned) and system 9 (tuned, all special characters are escaped)

For system 9 (see table 18) we improved the training corpus by escaping special characters of the VSI subtitles. We observed an inconsistency between the special characters in the VSI and OpenSubtitle corpus. The Moses tokenizer escapes special characters, for example, it replaces *'ll* with *apos;ll*. However, the VSI subtitles were not escaped, because the SUMAT project which tokenized the VSI subtitles works with non-escaped special characters. We decided to escape all special characters of our training corpus, development set and test sets. Apart from this modification we used the same corpora as for the systems 7 and 8. After the training, we tuned

the system with MERT. The BLEU score (see table 18) is exactly the same as for system 8, the TER score is slightly better, whereas the METEOR score is slightly worse. We observe no significant differences in the evaluation scores.

The number of exact matches of system 9 compared to system 8 is slightly higher (see tables 16 and 17), whereas the percentage of Lev-5 matches is slightly lower. The average Levenshtein distance is almost the same in both systems. If we consider the number of absolute matches and the average Levenshtein distance for sentences of different lengths, we only observe small differences which are not significant.

We see that the automatic evaluation scores show no significant differences between systems 6, 8 and 9. We decided to use system 9 for the test set translations, which are used for the improvement and evaluation of the grammar checker (see section 8).

7.3. The Final Translation System

In this section we evaluate our final Translation System (System 9) with additional test sets (see section 7.3.1). Furthermore, we compare our system and its performance with the system of SUMAT for English-Spanish (see section 7.3.2).

7.3.1. Evaluation with the VSI and OpenSubtitle Test Set

We already evaluated system 9 with the SUMAT test set (version 2012) which contains 4000 lines (see tables 12, 16 and 17). Within the SUMAT test set no repetitions occur. 72 of the English sentences of the test set also occur in the training corpus and none of them occurs in the development set.

For the evaluation of our final system (= system 9), we additionally used the VSI test set and the OpenSubtitle test set.

7.3.1.1. Evaluation with the OpenSubtitle Test Set

The OpenSubtitle test set contains 545 lines. Within the test set no repetitions occur. None of the test set sentences occurs in the training corpus or development set. We corrected the OCR-errors in the test set manually. On average, the English part of the test set contains 6.77 tokens per line. These are about 4 tokens less per line than in the SUMAT test set (version 2012).

	SUMAT test set	OpenSubtitle test set	VSI test set
# of lines:	4'000	545	4'001
BLEU:	26.8%	35.5%	25.4%
METEOR:	49.5%	57.6%	47.6%
TER:	56.0%	45.9%	58.8%
Absolute matches:	3.60%	17.80%	3.75%
Lev-5 matches:	9.85%	34.86%	9.80%
Avg lev. dist.	22.14	12.31	21.33

Table 19.: Evaluation of system 9 with the SUMAT test set, OpenSubtitle test set and VSI test set

System 9 yields a BLEU score of 35.5% for this test set (see table 19), which is is 8.7 percentage points higher than the BLEU score of the SUMAT test set (version 2012). Also the TER score and METEOR score show considerably more satisfying results for the OpenSubtitle test set than for the SUMAT test set (version 2012). The Levenshtein distances for sentences of different lengths show that the performance of the system is better for short sentences than for long sentences (see tables 19 and 20). On average, the sentences of the OpenSubtitle test set are shorter than in the SUMAT test set. This is probably one of the reasons why the evaluation scores for this test set are considerably better than for the SUMAT test set (version 2012). Another possible reason is the composition of the corpus. In section 5 we discussed that the major part of our corpus are OpenSubtitles. Therefore, we assume that the OpenSubtitle test set is more similar to the training corpus than the SUMAT test set which may result in a higher translation quality. Furthermore, the OpenSubtitle test set contains a lot of dialogue with common expressions. As some of these common expressions exist already in the training corpus, it is more probable that our system translates them correctly.

These reasons also explain why we can find considerably more exact matches in the OpenSubtitle test set than in in the SUMAT test set (version 2012). In total, our final system translated 17.80% of the sentences of the OpenSubtitle test correctly (see table 19). Moreover, we found in the OpenSubtitle test set 31.26% more lev-5-matches than in the SUMAT test set. Also, the average Levenshtein distance is considerably better than in the SUMAT test set of 2012.

number of tokens:	1-3	4-6	7-9	10-12	13-15	16-18	19+
Absolute matches (SUMAT):	23.60%	8.97%	3.89%	0.61%	0.51%	0.00%	0.00%
Lev-5 matches (SUMAT):	46.07%	21.52%	11.19%	3.86%	2.37%	0.54%	0.00%
Avg lev. dist (SUMAT):	8.03%	12.83	18.58	25.66	30.94	35.87	40.00
Absolute (OpenSub.):	38.26%	16.30%	12.20%	2.22%	0.00%	0.00%	0.00%
Lev-5 (OpenSub.):	60.83%	37.44%	23.58%	11.11%	5.00%	0.00%	0.00%
Avg lev. dist (OpenSub.):	5.3	10.31	13.78	22.00	26.15	31.86	41.12
Absolute (VSI):	16.22%	10.99%	2.34%	0.94%	0.56%	0.00%	0.00%
Lev-5 (VSI):	36.49%	23.58%	7.54%	4.12%	2.78%	0.00%	0.00%
Avg lev. dist (VSI):	8.41	13.15	20.22	25.78	30.62	35.27	39.00

Table 20.: Levenshtein distance for sentences of different lengths (test sets: SUMAT 2012, OpenSubtitle, VSI)

7.3.1.2. Evaluation with the VSI Test Set

The VSI test set contains 4'001 VSI subtitles. There are no repeated lines within the test set. 96 of the subtitles of the test set also exist in the training set and 4 subtitles of the test set occur in the development set as well. In total, we counted 38'877 tokens in the test set, which means 9.72 tokens per line on average. These are about 3 tokens more than the OpenSubtitle test set and approximately 1 token less per line than the SUMAT test set (version 2012).

The evaluation scores for the translation of the VSI test set are the worst of all test sets (see table 19). We can explain the difference compared to the OpenSubtitle test set with the fact that the largest part of the training corpus are OpenSubtitles which means that the OpenSubtitle test set is more similar to the training data than the VSI test set. Compared to the results for the SUMAT test set, the BLEU score (see table 19) of the VSI test set is 1.4 percentage points worse, the METEOR score 1.9 percentage points worse and the TER score 2.8% worse.

The Levenshtein distance shows similar results (see table 19 and 20): the differences between the evaluation scores for the OpenSubtitle test set and the VSI subtitles is considerable, in contrast to the results for the SUMAT and VSI test set, where the evaluation scores are similar. The number of exact matches (the sentence of the translation and reference translation is the same) is slightly worse in the VSI test set, whereas the total average of the Levenshtein distance is slightly better than in the SUMAT test set.

We found significant differences between the test set of the OpenSubtitle corpus and the test sets which consist of SUMAT subtitles. This confirms the heterogeneity of the two parts of our corpus, which we discussed in section 5.3.

7.3.2. Comparison with the Results of the SUMAT Project

In this section we compare the performance of our final system (= system 9) with the performance of the systems of the SUMAT project (versions 2012 and 2013). For the comparison of the performance, we used the BLEU score, the METEOR score and the TER score.

The results show (see table 21) for our final system (= system 9) that the BLEU and METEOR score are slightly better than the corresponding scores for the SUMAT[29] system of 2012. In contrast, the TER score shows a slightly better result for the SUMAT system.

	System 9	**SUMAT system (version 2012)**
# of subtitles (total):	26'908'887 (Opensub.+VSI)	???
# of subtitles to train the TM:	26'787'230 (OpenSub.+VSI)	???
# of subtitles to train the LM:	111'870'797 (OpenSub.+VSI+mono.)	???
# of subtitles in the dev. set:	15'001 (VSI)	???
Test set of SUMAT (2012)	4'000	4'000
BLEU:	26.8%	25.5%
METEOR:	49.5%	47.7%
TER:	56.0%	54.6%

Table 21.: Comparison of system 9 (system trained in this project) and the SUMAT system (version 2013)

We also compared our final system with the SUMAT system of 2013[30]. For the training of the SUMAT system (version 2013), the SUMAT project used 803'064 parallel subtitles. Their development set consisted of 2'000 subtitles. For the performance evaluation of our final system (= system 9), we used the same test set as the SUMAT project for the evaluation of their system of 2013. Note that for the evaluation, they used another test set than for the evaluation of the system of 2012. The results (see table 22) show, that the BLEU score of our system final is 1.3%

[29]The scores for the SUMAT system are extracted from an internal report of SUMAT of 2012. We did not find out how much training data the SUMAT project used.

[30]the scores for the SUMAT system are extracted from an internal report of SUMAT of April 2013

better than for the SUMAT system of 2013. The METEOR score of both systems is similar, but our final system achieves a slightly better score (0.4% higher) than the SUMAT system of 2013. The TER score of our final system is 1.7% better than for the translation of the SUMAT system of 2013.

	System 9	**SUMAT system (version 2012)**
# of subtitles (total):	26'908'887 (Opensub.+VSI)	???
# of subtitles to train the TM:	26'787'230 (OpenSub.+VSI)	803'064
# of subtitles to train the LM:	111'870'797 (OpenSub.+VSI+mono.)	???
# of subtitles in the dev. set:	15'001 (VSI)	2'000
Test set of SUMAT (2012)	4'000	4'000
BLEU:	30.6%	29.3%
METEOR:	51.5%	51.1%
TER:	52.6%	54.3%

Table 22.: Comparison of system 9 (system trained in this project) and the SUMAT system (version 2013)

The comparison shows that the performance of our final system is slightly better than the performance of the SUMAT systems. For our final system we used considerably more training data than the SUMAT project did. We conclude that an inclusion of amateur subtitles can be useful to improve the performance of the translation system, although the amateur subtitles do not guarantee a high quality.

8. Grammar Checking

In this chapter we will correct the translation produced by the SMT system (see section 7) with a rule-based grammar checker. We need linguistic information for the formulation of the grammar rules. Freeling (see section 8.1.1) provides the part-of-speech tags and the morphological analysis of the Spanish sentences.

For each of our error classes (see section 8.2) we make a restricted error analysis based on the SUMAT test set. Then we use the analyzed errors for the development of the rules. Some of the developed rules compare the errors detected by our grammar checker with the errors found by LanguageTool. Based on these rules, we design a grammar checker program and then the grammar checker corrects the translated test sets. Finally, we evaluate the corrections and discuss improvements such as DVD manuals (see section 8.8.2).

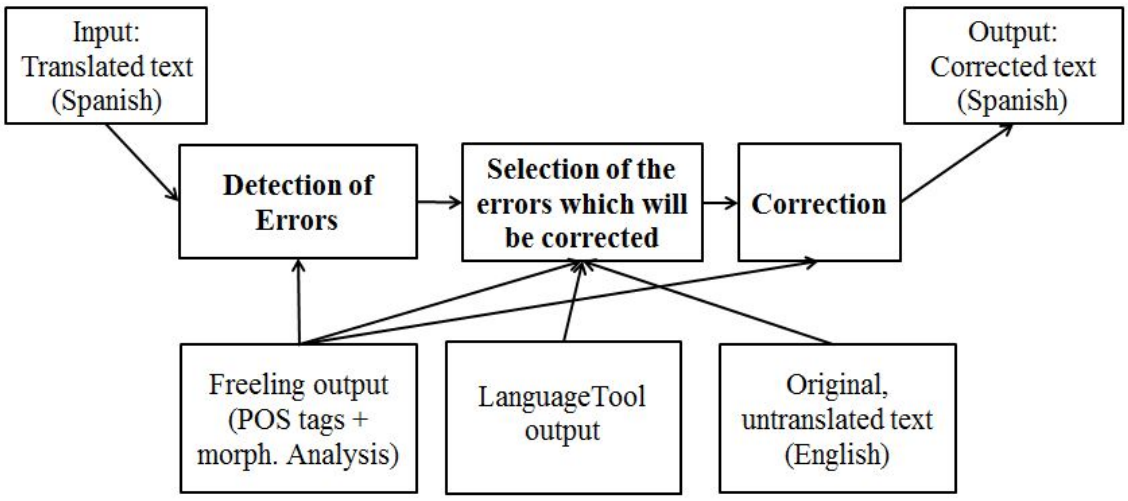

Figure 1: Program flow of our grammar checker

Our grammar checker program consists of three steps (see figure 1): In a first step, the grammar checker detects possible grammatical errors with the part-of-speech tags and the morphological analysis of the Freeling output. Then the grammar checker selects which of the detected errors will be corrected. To decide which errors should be corrected, the grammar checker uses different sources: the detected errors of LanguageTool, the English original text and/or the morphological analysis

of the Freeling output. Finally the grammar checker corrects the selected errors, for this the grammar checker uses the developed rules which require the morphological analysis of the Freeling output.

8.1. Tools to Provide Linguistic Information and to Propose Errors

This section describes the application of Freeling and LanguageTool.

8.1.1. Freeling

Freeling[31] is "an open source language analysis tool suite" (Center, 2012, 2) for the languages Spanish, English, Galician, Italian, Portuguese, Russian, Asturian, Catalan and Welsh. Freeling is maintained by the TALP Research Center of the "Universitat Politècnica de Catalunya" and provides among other things tools for POS-tagging, morphological analysis (also number, quantity and date detection, multiword detection) and named entity recognition (Center, 2012, 2).

In this project we use Freeling for the morphological analysis. Three input formats are available: plain text, tokenized text (one token per line) or split text (one token per line whereby a new line indicates a new sentences). The input has to be detokenized and recased in order to ensure a correct morphological analysis. For example, the uppercase letters at the beginning of a word are used as an indicator for proper names. We did the recasing and tokenisation with Perl-scripts provided by Moses.

For all input formats, Freeling applies its own tokenization and sentence-splitting. Because of this, the line breaks in the input and output of Freeling differ considerably and a mapping of the sentences of the output with the corresponding sentences of the input is difficult. Without this mapping, a correction is impossible. We applied Freeling using the option to consider a line break in the input file as a start of a new sentence. The result contained too many line breaks because Freeling introduces additional line breaks when enabling this option. Therefore, we tested another approach. We mapped the Freeling output with the corresponding sentence of the input file by counting the tokens. However, this approach did not work because Freeling applies its own tokenization. For example, the token *del* is separated into

[31] http://nlp.lsi.upc.edu/freeling/index.php?option=com_content&task=view&id=13&Itemid=42

the two tokens *de* and *el*. Thus, we introduced the line number at the beginning of each line before the analysis with Freeling, and the mapping succeeded.

The result of Freeling contains the original token, the corresponding lemma, the different possible morphological analyses and their probabilities.

The Freeling output format is:
⟨*line·number*⟩⟨*token*⟩⟨*lemma*⟩⟨*morph.·analysis*⟩⟨*prob.*⟩⟨*alt.·morph.·analysis*⟩⟨*prob.*⟩...

The examples 8.1-8.3 show some lines of the Freeling output[32]

(8.1) *3994 canciones canción NCFP000 1*

(8.2) *471 estaba estar VAII1S0 0.5 estar VAII3S0 0.5*

(8.3) *456 medio medio AQ0MS0 0.314286 medio NCMS000 0.262338 medio RG 0.262338 medio PI0MS000 0.158442 mediar VMIP1S0 0.0025974*

Freeling uses the EAGLES tag set for the morphological annotation. The morphological information is represented by a combination of uppercase alphabetic characters. The first character indicates the part-of-speech. The meaning of the other alphabetic characters depends on the part-of-speech tag. In ex. 8.1 *N* indicates that it is a noun; *C* shows that it is a common noun; *F* represents the gender (feminine), and *P* indicates the number. Freeling sets *0* if a value cannot be derived. This may have two reasons: Either this criterion was not analyzed by Freeling (but forms part of the EAGLES tag set) or a special characteristic was not identified. For example (see ex. 8.1), the next two *0* at the positions 5 and 6 show that we did not apply semantical classification (not analyzed). The last *0* shows that it is neither a diminutive nor an augmentative (no special degree was identified). In example 8.2, two different morphological analyses are possible. The verb form can either refer to the first or to the third person. Both analyses have a probability of 50%. Example 8.3 shows even five different possibilities of analysis. The possibilities are always sorted according to their probability.

8.1.2. LanguageTool

LanguageTool[33] is an open source spelling- and grammar-checker for more than 20 different languages. For Spanish 103 rules for the monolingual mode and one rule for the bilingual mode are registered. These rules are not sufficient to detect all

[32] The input string was *canciones estaba medio*. Freeling applies sentence-splitting and tokenizes the input string. Afterwards Freeling analyzes each token

[33] http://www.languagetool.org/

grammatical errors in Spanish texts.

Many different options are available in LanguageTool, for example: language detection, detection of errors with the existing rules, suggestions for the correction of errors, automatic correction of errors, indication of the mother tongue to find false friends and checking of the text in a bilingual mode. In addition, the users can create their own rules[34] and enable or disable certain rules.

In this project we used LanguageTool to confirm some of the grammatical errors our grammar checker detected. We selected the monolingual mode because only one rule is available for the bilingual mode. LanguageTool provides an option for the automatic correction of the detected errors. Corrections can only be made if the rules provide correction suggestions (see ex. 8.6). Especially for grammar errors, such suggestions are rare. Therefore, we did not select this option. We applied LanguageTool on the recased and retokenised translation of the SUMAT test set (version 2012) without selecting any additional options.

(8.4) *54.) Line 33, column 18, Rule ID: DET_NOM_PLUR[3]*
Message: Posible falta de concordancia de número entre «las» y «película».
así se agotaron las película al mismo tiempo.

(8.5) *5523.) Line 3808, column 33, Rule ID: EL_NOM_MASC[4]*
Message: Posible falta de concordancia de género entre «el» y «omnipotencia».
si me pregunta, ¿debo entender el omnipotencia divina

(8.6) *5529.) Line 3812, column 53, Rule ID: MI_FINAL1[8]*
Message: El pronombre personal 'mí' lleva tilde.
Suggestion: mí
...avoured constantemente expresar en la música mi tormento

LanguageTool found 3'868 errors in our data set (see ex. 8.4-8.6). From this result we extracted 238 grammatical errors, because the output contains, beside the grammatical errors, other errors as well. Our grammar checker then checks each error against the errors found with LanguageTool in order to confirm the detected errors.

[34] We decided to develop our own grammar checker instead of creating own rules with Language Tool, because in this way we had the possibility to include external sources such as the English original text or the alternative morphological analyses of the Freeling output.

8.2. Goals of the Grammar Checker

The focus of this grammar checker lies on precision. This means that we pay attention to the prevention of creating new errors. In other words, it should be possible to use the grammar checker without the risk to produce new errors. The precision of the grammar checker depends on the detection and correction. The goal of this project is a precision above 75% for each error class that is included in the grammar checker.

First, we concentrate on the correction of agreement errors. We study disagreements between nouns, determiners and adjectives (see sections 8.3-8.5). Disagreements with verbs are versatile and, therefore, we only take into account specific cases of disagreements between verbs and subjects (see section 8.6). Additionally, we consider word combinations in which prepositions require an infinitive if the following token is a verb (see section 8.7).

8.3. Disagreements between Determiners and Nouns

8.3.1. Restricted Error Analysis

Disagreements between determiners and nouns are the first error class that the grammar checker examines. Determiners and nouns can morpho-syntactically disagree in gender and number.

Spanish distinguishes two genders: masculine and feminine. The Freeling coding of the morphological analysis is: M for masculine, F for feminine and C if the masculine and feminine form of a word is identical. Freeling only codes C if the change of gender does not result in a change of meaning. If a word can be masculine as well as feminine and these two forms have different meanings, Freeling selects the most probable gender. For example *final* is always tagged as feminine. Concerning the gender two cases of disagreement can occur: Either the determiner is feminine and the noun masculine or vice versa.

Spanish distinguishes two numbers: singular and plural. Therefore, two cases of disagreements can be found: either the article is singular and the noun plural or vice versa. English distinguishes also between singular and plural. The regular plural markers of both languages are identical (*-s* or *-es*). So Freeling assigns the correct number to untranslated English words which means that we can also correct the number of untranslated words.

For the restricted error analysis[35] our script searches all disagreements between determiners and nouns in the output of Freeling. The input for Freeling was the SUMAT test set (version 2012) translated with the SUMAT system (version 2012). The script selects cases in which the noun follows the determiner directly. Finally we count and classify the detected disagreements manually.

Disagreement:	**Detected:**	**True Positives:**	**False Positives:**	**Impossible to decide:**
Det. (m) + noun (f):	64	21	42	1
Det. (f) + noun (m):	25	13	3	9
Det. (sg) + noun (pl):	26	23	2	1
Det. (pl) + noun (sg):	15	6	3	6
Total:	130	64	50	16

Table 23.: Restricted Error Analysis: Disagreements between determiners and nouns

For the error classification we defined the categories "true positives", "false positives" and "impossible to decide". True positives are cases in which the detected disagreements are real errors and must be corrected. False positives are cases in which the manual evaluation shows no disagreement and must not be corrected. "Impossible to decide" is a category for the cases which are neither true positives nor false positives. In these cases the translations are often completely wrong or the syntax structure is unclear. In the same category we find cases with untranslated nouns that do not exist in the target language[37](RAE, 2011) and for which it is impossible to determine the gender. For the development of the rules only true positives and false positives were considered.

In total, the script detected 130 disagreements (see table 23). In 64 of these disagreements the determiner is masculine and the noun feminine. 21 of the 64 disagreements are true positives and must be corrected (see ex. 8.7). One of the 64 disagreements we classify as "impossible to decide". It is impossible to decide if *llama* is a verb and *los* a pronoun or if *llama* is a noun and *los* the corresponding article (see ex. 8.8). 42 of the 64 disagreements are false positives. Some false positive appear because of ambiguous nouns that are used in feminine and masculine forms (with a change of meaning). A typical example is the noun *final* (see ex. 8.9). *Final* can occur as masculine (en: *the end*) as well as feminine (en: *the finale*). Other false positives

[35] We choose the term *restricted error analysis* because we only consider the disagreements which we can detect automatically with our script[36] and the morphological analysis of the Freeling output. For a complete error analysis for disagreements between determiners and nouns we would have to search the disagreements manually and this would be time-consuming.

[37] "untranslated" stands for all words that do not appear in the Spanish reference dictionary of the *Real Academía Española*: http://www.rae.es/rae.html

occur because of singular feminine nouns which begin with a stressed *a*. In these cases the preceding article has to be masculine, because of phonetic reasons (see ex. 8.9 (Raya, 2008, 30). For example, *alma* is a feminine noun beginning with a stressed *a* and thus the masculine article *un* is used[38].

(8.7) *y comprobado ambos su edad y* ***un melancolía*** *bordering en amargura .* (masculine article *un* instead of the feminine article *una*)

(8.8) *y dave solo para* ***los llama*** *, lib .*

(8.9) *pero hay algo* ***al final*** *, que pongamos en el final .* (correct)

Our script found 25 cases in which a feminine determiner precedes a masculine noun. A manual evaluation of these cases showed that 12 of them are true positives (see es. 8.10). Only 3 of them are false positives. All of the false positives are incorporated loanwords. For example *demo*[39], an abbreviation of the English noun *demonstration*, was incorporated in the Spanish vocabulary and is used with feminine determiners (see ex. 8.11). 9 other cases are classified as "impossible to decide" because some English words were copied untranslated into the Spanish text (see ex. 8.12) and the gender cannot be determined.

(8.10) ***las compañeros*** *constante desde su infancia ,* (feminine article *las* instead of the masculine article *los*)

(8.11) *como una droga fue un proceso muy sencillo porque* ***la demo*** *terminó .* (correct)

(8.12) *esto es antes de la carretera y pasar todas* ***las dustbins*** (*dustbin* is an English noun)

We found 26 disagreements in which the determiner is singular and the noun plural. 23 of the 26 disagreements are true positives and should be corrected. 15 of the true positives contain Spanish nouns (ex. 8.13) and 9 of them contain untranslated English nouns (ex. 8.14). 2 of the 26 detected determiner-noun disagreements are false positives. In example 8.15 the verb *haces* is tagged wrongly (as a noun) and the pronoun *lo* is also tagged wrongly (as a determiner).

(8.13) *busca , irónicamente , que los rusos están vendiendo* ***su grabaciones*** *en el oeste .* (*su* instead of plural *sus*)

(8.14) *" no peor que su predecessors .* (*su* instead of *sus*)

[38] http://lema.rae.es/dpd/?key=agua

[39] http://lema.rae.es/drae/?val=demo

(8.15) *si simplemente ... no sé cómo lo haces . jugar , a ver lo que haces .* (should not be corrected)

We found 15 cases of disagreement in which the determiner is plural and the noun singular. 6 of the 15 cases are true positives and should be corrected (ex. 8.16). 2 of the true positives contain an untranslated English noun (ex. 8.17). 3 of the 15 detected disagreements are false positives. They occur because of untranslated English words, wrong part-of-speech tags or more complex nominal phrases. In example 8.18 the nominal phrase does consist of a determiner and two nouns that are combined with a copula, instead of one noun only. The determiner has to be plural, because it refers to both nouns. Our search script finds these cases by mistake, because only the first noun is considered.

(8.16) *a* ***los espectador*** *reflexionar sobre algo* ("a los" instead of "al")

(8.17) *"* ***nuestros symphony !*** ("nuestros" instead "nuestro")

(8.18) *vale , claro . el premio al mejor álbum del año va a* ***ambos artista y productor*** *.*

These results of the restricted error analysis show that disagreements between determiners and nouns occur in the translations of SMT systems. We outlined that the morphological analysis of Freeling is not always a sufficient criterion to detect disagreement errors between determiners and nouns. Especially loanwords and untranslated English words cause difficulties. We also discovered that we must develop different rules for the detection and correction of disagreements of gender and number. Untranslated English nouns cause difficulties for the detection and correction of disagreements of gender, which is not the case for disagreements of number.

8.3.2. Development of the Rules

We used the true and false positives of the restricted error analysis to develop the rules. The first part of the grammar checker consists of the script that we used for the restricted error analysis. The program has two major functions: The first function is to filter the detected disagreements in order to exclude the false positives; the second function is to correct the disagreements.

There are different types of determiners: definite articles, demonstratives, possessives, interrogatives, exclamatives and indefinite determiners. One of the program's main ideas was to make the rules for the different types as independent as possible in order to allow adaptions, exclusions and evaluations for each type. Our experi-

ments showed that the precision with which disagreements of number were detected was not improved by additional filter rules. Therefore, we decided to correct almost all the disagreements of number we detected in the restricted error analysis. We realized that the plural forms *ambos* and *ambas* (see ex. 8.18) do not have a corresponding singular form. We introduced a rule that prohibits corrections of the number of *ambos* and *ambas*.

We added additional filter rules to exclude false positives from the disagreements of gender we detected in the error analysis. We applied the following rules on the set of errors:

- if a disagreement is found by LanguageTool and by our script, we correct the disagreement
- if a noun (that disagrees with the determiner) can occur with both genders, we do not correct the disagreement
- if a noun (that disagrees with the determiner) occurs in both the target language and the corresponding sentence of the source language, we assume that it is an untranslated English word and do not correct the disagreement
- if a singular feminine noun (that disagrees with the determiner) begins with the grapheme *a*, *á* or *ha* (= morpheme *a*), we do not correct the disagreement

We correct the gender or number always on the determiner. The correction of determiners has two advantages. First, it is a closed group of words and second the inflection of the determiners is almost regular. We simplified and generalized some of the grammar rules to reduce the complexity of the program. In the following, we list the summarized rules the grammar checker applies to correct the number and gender of the determiners. The sequence of rules in the program is important. Specialized rules (e.g. for exceptions) are applied before generalized rules.

Changes from feminine to masculine:

- *la* is changed to *el* (specialized rule)
- *de la* is changed to *del* (specialized rule)
- *a la* is changed to *al* (specialized rule)
- *una* is changed to *un* (specialized rule)
- *aquella* is changed to *aquel* (specialized rule)
- *-una* is changed to *-ún* (generalized rule)

- *-a* is changed to *-e* (generalized rule) [40]
- *-a* is changed to *-o* (generalized rule)
- *-as* is changed to *-os* (generalized rule)

Changes from masculine to feminine:

- *el* is changed to *la* (specialized rule)
- *del* is changed to *de la* (specialized rule)
- *al* is changed to *a la* (specialized rule)
- *un* is changed to *una* (specialized rule)
- *aquel* is changed to *aquella* (specialized rule)
- *-ún* is changed to *-una* (generalized rule)
- *-e* is changed to *-a* (generalized rule) [35]
- *-o* is changed to *-a* (generalized rule)
- *-os* is changed to *-as* (generalized rule)

Changes from singular to plural:

- *el* is changed to *los* (specialized rule)
- *al* is changed to *a los* (specialized rule)
- *del* is changed to *de los* (specialized rule)
- *la* is changed to *las* (specialized rule)
- *aquel* is changed to *aquellos* (specialized rule)
- *ese* is changed to *esos* (specialized rule)
- *este* is changed to *estos* (specialized rule)
- *un* is changed to *unos* (specialized rule)
- *-ún* is changed to *unos* (generalized rule)
- *-o* is changed to *-os* (generalized rule)
- *-a* is changed to *-as* (generalized rule)

[40] Obviously determiners ending with *-e* exist in masculine as well as in feminine. Because Freeling tags these forms with *C* (common), no disagreements for determiners ending with *-e* are detected

- *-e* is changed to *-es* (generalized rule)

Changes from plural to singular:

- *a los* is changed to *al* (specialized rule)
- *de los* is changed to *del* (specialized rule)
- *los* is changed to *el* (specialized rule)
- *las* is changed to *la* (specialized rule)
- *aquellos* is changed to *aquel* (specialized rule)
- *esos* is changed to *ese* (specialized rule)
- *estos* is changed to *este* (specialized rule)
- *unos* is changed to *un* (specialized rule)
- *-unos* is changed to *ún* (generalized rule)
- *-os* is changed to *-o* (generalized rule)
- *-as* is changed to *-a* (generalized rule)
- *-es* is changed to *-e* (generalized rule)
- the number of *ambos* and *ambas* is never changed (because it does not exist in singular)

8.3.3. Evaluation

8.3.3.1. Evaluation Method

For the first evaluation step we used the test set of SUMAT (version 2012, translated by the SUMAT system). We developed the rules of the grammar checker based on the errors of the same test set, we already described in section 8.3.2. This means that the result of the evaluation of this test set is not representative. Therefore we evaluated additional test sets. We translated the SUMAT test set, the OpenSubtitle test set and the VSI test set with system 9 (see section7.2.7) and used their translations as our test sets for the evaluation. Still we do not know if all possible instances of disagreements between determiners and nouns appear in the test sets. In order to show which disagreement constructions the grammar checker is able to correct and which not, we evaluated the correction of artificially created Test Suites.

We manually evaluated all sentences of the test sets the grammar checker had corrected and we classified the corrections into three categories. The first category contains true positives (see ex. 8.19). True positives are cases where the correction improves the sentence. Note that improvement does not mean that the sentence has to be perfect after the correction (see ex. 8.20). The second category are false positives, which means that the corrected sentence is worse (see ex. 8.21). The third category is comprised of cases for which it is impossible to decide if the sentences are better or worse after the correction (see ex. 8.22).

(8.19) **Input:** *entonces supongo que nadie llegué* ***a abordando*** *este asunto*
Output: *entonces supongo que nadie llegué* ***a abordar*** *este asunto*
Explanation: The verb form after *llegar a* must be an infinitive. The grammar checker replaces the gerund with an infinitive and therefore we classify this correction as "true positive"

(8.20) **Input:** *me llenó de un* ***raro y genuino felicidad*** *.*
Output: *me llenó de un* ***raro y genuina felicidad*** *.*
Explanation: Both adjectives may be changed to the feminine form. However the grammar checker only corrects one of them, we classify this sentence as true positive, because the output sentence is better than the input sentence.

(8.21) **Input:** *y oí* ***la demo*** *,*
Output: *y oí* ***el demo*** *,*
Explanation: The feminine article which is used in the input sentence is correct. The grammar checker replaces this correct feminine article with an incorrect masculine article. Therefore we classify this correction as "false positive".

(8.22) **Input:** *en* ***la primera andrógino*** *siendo ...*
Output: ***en el primero andrógino*** *siendo ...*
Explanation: The substantive is missing, therefore we cannot decide which is the correct gender of the adjectives and article. Thus, we classified this correction as "impossible to decide".

After the classification, we calculate the precision from the number of true and false positives. The recall cannot be calculated because we do not have the number of false negatives. A manual analysis of each sentence of the test set is needed to get the number of false negatives, which is very time consuming and complex. This would exceed the scope of this project, especially because the focus of the grammar checker lies on high precision. The goal is to achieve a total precision of more than 75%.

This is an arbitrary decision which we made to have a consistent methodological approach. User studies might elaborate how high the precision of the grammar checker must be at least that the users declare the grammar checker as useful and helpful.

8.3.3.2. Results for the SUMAT Test Set Translated with the SUMAT System

The results show that the grammar checker corrected 77 disagreements between determiners and nouns in the SUMAT test set (version 2012, translated with the SUMAT system). In total, we classified 60 of the 77 corrections as true positives, 8 as false positives and for 9 corrections it was impossible to decide. The precision is 88.24% (see table 24), which is a satisfying result. We can observe that the majority of corrections are changes from masculine to feminine (20 of 77) or from singular to plural (23 of 77). For these kinds of corrections the precision is high (90.91 and 100%). The precision for changes from plural to singular and from feminine to masculine is lower (78.57% and 71.43%), but still satisfying.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + noun (f):	20	2	2	90.91%
Det. (f) + noun (m):	11	3	2	78.57%
Det. (sg) + noun (pl):	23	0	1	100%
Det. (pl) + noun (sg):	15	6	3	71.43%
Total:	60	8	9	88.24%

Table 24.: Evaluation of the SUMAT test set of 2012 translated by the SUMAT system: Disagreements between determiners and nouns

8.3.3.3. Results for the SUMAT Test Set Translated with System 9

The grammar checker corrected 26 disagreements in the translation of the SUMAT test set with our system 9 (see section 7.2.7). We have seen that system 9 works better than the SUMAT system (see section 7.3.2). Compared to the previous evaluation (see section 8.3.3.2), we found roughly two third less corrections. This means that the better performance of the SMT system, the lower the number of disagreements between determiners and nouns.

In total, we classified 21 of the 26 corrections as true positives, 4 as false positives and for 1 correction it was impossible to decide. The precision is 84% (see table 25).

This result is almost the same as in the previous evaluation (see section 8.3.3.2). We observe that the changes from masculine to feminine and from singular to plural show the best precisions (100%) (see table 25). This observation agrees with our previous evaluation (see table 24).

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + noun (f):	5	0	1	100%
Det. (f) + noun (m):	4	1	0	80%
Det. (sg) + noun (pl):	8	0	0	100%
Det. (pl) + noun (sg):	4	3	0	75%
Total:	21	4	1	84%

Table 25.: Evaluation of the SUMAT test set of 2012 translated by system 9: Disagreements between determiners and nouns

8.3.3.4. Results for the VSI Test Set Translated with System 9

In the VSI test set (4001 lines) the grammar checker corrected in total 22 disagreements. The number of corrections is almost the same as for the SUMAT system which was also translated with system 9.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + noun (f):	2	1	0	66.67%
Det. (f) + noun (m):	3	0	0	100%
Det. (sg) + noun (pl):	10	2	0	83.33%
Det. (pl) + noun (sg):	1	2	1	33.33
Total:	16	5	1	76.19%

Table 26.: Evaluation of the VSI test set translated by system 9: Disagreements between determiners and nouns

In total we classified 16 of the 22 corrections as true positives, 5 as false positives and for 1 correction it was impossible to decide (see table 26). The precision is 76.19%. This result is worse than the precisions in the previous evaluations (see sections 8.3.3.2 and 8.3.3.3), but the precision is still above our limit of 75% (see section 8.3.3.1). We realize that the precision for changes from plural to singular (33.33%) is low. This result as such is not meaningful, because the number of corrections for this kind of changes is low and we cannot derive any conclusions. However, we see that in all previous evaluations the precision of the change from plural to singular is the lowest of all the possible changes. Further projects might test, if it is possible

to improve the grammar checker for the changes from plural to singular, possibly with the inclusion of additional tools or additional linguistic information.

8.3.3.5. Results of the OpenSubtitle Test Set Translated with System 9

In the OpenSubtitle test set the grammar checker detected no disagreement between determiners and nouns. We can explain this with the shortness of the test set (545 lines) and with the fact that the sentences are shorter than the sentences in the other test sets. The probability of disagreements in short sentences is lower than in longer sentences, because less tokens and n-grams (bigrams and trigrams) are combined.

8.3.3.6. Comparison of the Results of the Test Sets

In all test sets together, the grammar checker corrected 126 disagreements. We classified 98 of these corrections as true positives. 17 corrections are false positives and for 11 corrections it is impossible to decide if the correction yields an improvement or not. The total precision is 85.22%. This is above 75% and, therefore, we recommend the application of the grammar checker for the correction of disagreements between determiners and nouns. Furthermore, we observed that the precision of the changes from plural to singular is the worst of all, hence these rules might be improved. This leads to the hypothesis that we must apply different rules for the selection of cases in which plural forms are changed to singular than for the selection of the reverse cases. This hypothesis and the calculated precisions must be proven with additional and bigger test sets in further experiments.

In the test sets translated by system 9, the grammar checker makes only few corrections of disagreements between determiners and nouns. In addition, the grammar checker sometimes corrects the number of determiners preceding untranslated English nouns. In the post-editing process the translator or proofreader has to translate the noun and adapts the determiner if necessary. Therefore, the benefit of the correction of the number of determiners preceding untranslated English nouns is questionable. A cost-benefit analysis can help to clarify this question and also to decide if our grammar checker yields any benefits.

8.3.3.7. Discussion of the False Positives

Wrong part-of-speech tags cause false positives. Sometimes Freeling tagged pronouns (e.g. *las* in ex. 8.23) as determiners and verbs as nouns (e.g. *arreglo* in ex.

8.23). To avoid the correction of these false positives, we added a rule to the grammar checker. We prevented the correction of determiners and nouns that can also occur with other part-of-speech tags. This experiment resulted in a meaningful decrease of the recall.

Wrong analyses of loanwords also cause false positives. In example 8.24 *graffiti* is not recognized as a plural form. Loanwords are often identical to the corresponding English word. To avoid the correction of these false positives, we added a rule to the grammar checker. We prevented the correction of the number of determiners preceding nouns that also appear in the corresponding line of the English source text. This experiment resulted in a meaningful decrease of the recall compared to the experiments without these rules.

Because both adaptions cause a meaningful decrease of the recall, we did not include them in the grammar checker.

(8.23) *y creo que ahora me* ***las arreglo*** *para expresar esa comprensión*
was changed to
y creo que ahora me ***los arreglo*** *para expresar esa comprensión*

(8.24) *y debo decir , no* ***tantos graffiti*** *en la pared .*
was changed to
y debo decir , no ***tanto graffiti*** *en la pared .*

8.3.3.8. Results of the Test Suites

In addition to the test sets, we applied the grammar checker on the Test Suite file and evaluated the corrections. Test Suite files contain sentences or phrases that are either constructed artificially or extracted from a corpus (see appendix C). They are used to evaluate MT systems in order to check if certain linguistic aspects can be translated correctly or not. It is important that the Test Suites contain only the linguistic aspects to be tested and no other characteristics that complicate the translation. Apart from the tested phenomena, the sentences or phrases of the Test Suites should be syntactically, grammatically and lexically as simple as possible (Eckel, 1998, 37). In this project we did not use the Test Suites for the evaluation of an MT system but for the evaluation of the grammar checker.

The Test Suites contain a selection of correct and incorrect sentences. In the correct sentences the determiner and noun agree, whereas in the incorrect sentences they disagree. The evaluation with the Test Suites (see appendix C) showed that the grammar checker is able to correct disagreements that contain all types of determin-

ers. The grammar checker never did corrections in sentences without disagreements. We observe that the grammar checker manages to correct determiner-noun combinations in which both, gender and number, disagree. The grammar checker is also able to correct some special cases, for example, the correction of determiners combined with a preposition (*del* and *al*) succeeds. Additionally, the grammar checker allows the combination of masculine determiners with feminine nouns, if the noun is singular and starts with *a*.

The rule that allows masculine determiners to precede singular feminine nouns which begin with *a* sometimes causes false negatives. This happens because the rule is simplified. It does not include the condition that the first syllable of the noun has to be stressed if a masculine determiner is combined with a singular feminine noun. We did not include this condition in the grammar checker, because we cannot automatically differentiate between stressed and unstressed syllables. The grammar checker assumes that all first syllables of singular feminine nouns beginning with *a* following a masculine determiner are stressed and makes no correction. In example 8.25, the grammar checker does not replace the masculine article before *alma* what is correct. Choosing the reverse assumption would result in a decrease of the precision. Disagreements between determiners and masculine nouns beginning with *a* are always corrected. In example 8.26), the wrong feminine article is replaced with the correct masculine article, therefore we classify this correction as "true positive".

Wrong part of speech tags also cause false negatives. In example 8.27 *vario* is tagged as a noun. As a result, the grammar checker does not correct the existing disagreement of number. This shows that the main problem of the grammar checker are wrong part-of-speech tags and not the morphological analysis. Another problem of the grammar checker is the different casing in the input files. The translation used as input for the grammar checker is lowercase. Instead of this, we used a recase version for the analysis with Freeling. The rules do not manage the mapping between uppercase determiners (e.g. at the beginning of a sentence) of the Freeling output and lowercase determiners of the input.

In addition, the corrections in the Test Suites show that, when two determiners precede the noun, the grammar checker only corrects the determiner which precedes the noun directly (see ex. 8.28).

(8.25) **Input:** *Tiene* ***una alma*** *buena.* - **Output:** *Tiene* ***una alma*** *buena.* (*una* instead of *un*).

(8.26) **Input:** *Tiene* ***una amigo bueno.*** - **Output:** *Tiene* ***un amigo*** *bueno.* (correct)

(8.27) **Input:** *Tengo que escribir* ***vario textos****.* - **Output:** *Tengo que escribir* ***vario textos****.* (*vario* instead of *varios*)

(8.28) **Input:** *Lo he dicho a* ***todo mi hermanos****.* - **Output:** *Lo he dicho a* ***todo mis hermanos****.* (*todo* instead of *todos*)

8.4. Disagreements between Adjectives and Nouns

8.4.1. Restricted Error Analysis

The grammar checker also corrects disagreements between adjectives and nouns. We made a restricted error analysis. Using the same categories and method as described in section 8.3.3.1, we counted and classified the disagreements between adjectives and nouns. The disagreements of the SUMAT test set (version 2012) translated by the SUMAT system (version 12) served us again as input for the restricted error analysis.

In Spanish the adjective follows the noun in most of the cases. Some adjectives can or even must precede the noun. Sometimes a change of the position of the adjective implies a change of meaning. In the error analysis we distinguish between cases in which the adjective precedes the noun and cases in which the adjective follows the noun (Raya, 2008, 26). The grammar checker examines again disagreements of number and gender.

First, we discuss the cases in which the adjective precedes the noun. In the morphological analysis of Freeling we found 100 disagreements in which this is the case (see table 29). We analyzed them manually. 72 cases are true positives and must be corrected. We ascribed 15 cases in the category "impossible to decide". The reasons for these cases are confusing sentence structures or untranslated English words.

Disagreement:	Detected:	True Positives:	False Positives:	Impossible to decide:
Adj. (m) + noun (f):	34	29	3	2
Adj. (f) + noun (m):	26	9	7	10
Adj. (sg) + noun (pl):	39	30	6	3
Adj. (pl) + noun (sg):	1	1	0	0
Total:	100	69	16	15

Table 27.: Restricted error analysis: Disagreements between adjectives and nouns (the adjective precedes the noun)

16 cases are false positives, which the grammar checker should not correct. Some of the false positives are named entities consisting of two tokens, which were not recognized by Freeling. In these cases one of the tokens is tagged as an adjective and the other token as a noun. For example, the named entities *nueva york* (see ex. 8.29 or *shea stadium* (see ex. 8.30) are not recognized and, therefore, the grammar checker detects a disagreement between the two tokens of the named entity, although the combination of these two words is completely correct. Probably, the named entity is not recognized because of a recasing error. Most of the other false positives occur because of wrong part-of-speech tags. In most of these cases Freeling confuses adjectives and nouns. In example 8.31, *portuguesa* is tagged as a noun instead of an adjective. In fact *baile* is the noun and the two following words are the corresponding adjectives.

The number of detected disagreements is almost the same for all possible sorts of disagreements, except for the disagreements in which the adjective is plural and the noun singular. For this kind of disagreement we found only one case.

(8.29) *en* ***nueva york*** *1 de agosto de 1971 .*

(8.30) *cuando los beatles ellos tocaron juntos como grupo en el* ***shea stadium*** *.*

(8.31) *como un* ***baile folclórico portuguesa*** *conocida como la folia .*

Disagreement:	Detected:	True Positives:	False Positives:	Impossible to decide:
Noun (f) + adj. (m):	21	12	7	2
Noun (m) + adj. (f):	7	6	0	1
Noun (sg) + adj. (pl):	12	4	6	2
Noun (pl)+ adj. (sg):	15	12	0	3
Total:	55	34	13	8

Table 28.: Restricted error analysis: Disagreements between adjectives and nouns (the adjective follows the nouns)

Secondly, we consider the cases in which the adjective follows the noun. We detected 54 disagreements (see table 28) with our script. Although typically the adjective follows the noun, we found with our restricted error analysis more disagreements in which the adjective precedes the noun. This may be due to two reasons: either combinations in which the adjectives precede the nouns are more error-prone or the position of the adjective relative to the noun often is wrong in the translation. A manual analysis of the detected disagreements speaks for the second reason.

We manually classified 39 disagreements as true positives, 9 as false positives and

7 disagreements as "impossible to decide" (see table 28). Particularly, we classified combinations containing the word form *juntos* or *juntas* as false positives, because these adjectives often do not refer to the noun that follows or precedes it directly. In example 8.32, we detected a disagreement between *tiempo* (*time*) and *juntos* (*together*), but in fact *juntos* refers to *todos* and therefore the agreement is completely correct. Other false positives occur as a consequence of untranslated elements (e.g. *stack-heel* in ex. 8.33) or named entities that were not recognized (e.g. *rachmaninov* in ex. 8.34).

(8.32) *todos tenían un buen tiempo* ***juntos*** *, no había egos .*

(8.33) *así que tottered brevemente en mi* ***stack-heel botas*** *y dijo :*

(8.34) *pero dado de* ***rachmaninov atractivo*** *para muchos scherzando cosas ,*

We discovered that the morphological analysis of Freeling is not a sufficient criterion to detect disagreement errors between adjectives and nouns. Particularly unrecognized named entities and untranslated English words cause challenges for the detection of disagreements.

8.4.2. Development of the Rules

The program structure for the detection and correction of disagreements between adjectives and nouns is almost the same as the one used for the detection and correction of disagreements between determiners and nouns (see section 8.3.2). We adjusted some of the detection rules in order to improve the precision.

The EAGLES tag set distinguishes ordinal and qualifying adjectives. Ordinal adjectives are ordinal numbers which are used as adjectives; the group of qualifying adjectives subsumes all the remaining adjectives. Tests showed that we can apply identical rules for the detection of disagreements between ordinal and qualifying adjectives.

For each detected disagreement the grammar checker tests first, if this error was also found by LanguageTool. If this is the case, the disagreement is corrected. If this is not the case and if it is, furthermore, a disagreement of gender, the grammar checker tests whether the noun can also occur with the other gender, the grammar checker does not correct the disagreement. For all the other cases the grammar checker tests, if the noun occurs in the corresponding line of the English source text. If no, the grammar checker does not correct the disagreements. If yes, the grammar checker corrects them. By means of this approach we excluded disagreements with

untranslated English words and loanwords. This rule was adapted for the disagreements between singular and plural: neither the adjective nor the noun is allowed to occur in the English source text.

The grammar checker always makes the correction on the adjectives. For the correction we applied the following rules:

Changes from feminine to masculine:

- if the adjective precedes: *buena* is changed to *buen* (specialized rule)
- if the adjective precedes: *mala* is changed to *mal* (specialized rule)
- if the adjective precedes: *primera* is changed to *primer* (specialized rule)
- if the adjective precedes: *tercera* is changed to *tercer* (specialized rule)
- the ending -*a* is changed to -*o* (generalized rule)
- the ending -*as* is changed to -*os* (generalized rule)

Changes from masculine to feminine:

- if the adjective precedes: *buen* is changed to *buena* (specialized rule)
- if the adjective precedes: *mal* is changed to *mala* (specialized rule)
- if the adjective precedes: *primer* is changed to *primera* (specialized rule)
- if the adjective precedes: *tercer* is changed to *tercera* (specialized rule)
- the endings -*e* and -*o* are changed to -*a* (generalized rule)
- the ending -*os* is changed to -*as* (generalized rule)

Changes from singular to plural:

- *feliz* is changed to *felices*
- if the adjective precedes: *gran* is changed to *grandes*
- if the adjective precedes: *buen* is changed to *buenos* (specialized rules)
- if the adjective precedes: *mal* is changed to *malos* (specialized ruled)
- if the adjective precedes: *primer* is changed to *primeros* (specialized rule)
- if the adjective precedes: *tercer* is changed to *terceros* (specialized rule)
- if the adjective ends in -*o*, *a* or *e*, *s* is added (generalized rule)

- if the adjective ends not in *-o*, *a* or *e*, *es* is added (generalized rule)

Changes from plural to singular

- *felices* is changed to *feliz*
- if the adjective precedes: *grandes* is changed to *gran* (specialized rule)
- if the adjective precedes: *buenos* is changed to *buen* (specialized rules)(footnote)
- if the adjective precedes: *malos* is changed to *mal* (specialized ruled)(footnote)
- if the adjective precedes: *primeros* is changed to *primer* (specialized rule)(footnote)
- if the adjective precedes: *terceros* is changed to *tercer* (specialized rule)(footnote)
- the last letter *-s* is omitted if the adjective ends in *-os* or *-as* (generalized rule)
- the last letter *-s* is omitted if the adjective ends in *-tes* (generalized rule)
- the last two letters *-es* are omitted if the adjective ends in *-es*, but not *-tes* (generalized rule) (footnote: simplified rule)
- the last letter *-s* is omitted with all the other adjectives ending in *-s* (generalized rule)

8.4.3. Evaluation

For the evaluation we used the same test sets and the same approach as for the disagreements between articles and nouns (see section 8.3.3). We analyze the corrected sentences manually to decide if the correction yields an improvement or a worsening, or to see that a decision cannot be made.

8.4.3.1. Results for the SUMAT Test Set Translated with the SUMAT System

First, we evaluated the corrections in the SUMAT test set (version 2012) which we translated with the SUMAT system (version 2012). This is the same test set we used for the restricted error analysis and as our basis for the development of the program (see section 8.4.1). In total, the grammar checker corrected 106 disagreements between adjectives and nouns (see table 29). 86 of the 106 corrections improve the sentences, whereas 10 corrections resulted in an even more incorrect sentence. 10 corrections fall in the category "impossible to decide". A lot of these cases are combinations in which the adjective is positioned between two nouns and

it is unclear to which noun the adjective belongs.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Adj. (m) + noun (f):	29	3	0	89.29%
Adj. (f) + noun (m):	9	2	2	81.82%
Adj. (sg) + noun (pl):	19	2	3	86.36%
Adj. (pl) + noun (sg):	1	0	0	100
Noun (f) + adj. (m):	9	3	1	75%
Noun (m) + adj. (f):	5	0	0	100%
Noun (pl) + adj. (sg):	11	0	2	100%
Noun (sg) + adj. (pl):	3	0	2	100
Total:	86	10	10	89.58%

Table 29.: Evaluation SUMAT test set of 2012 translated by the SUMAT system: Disagreements between adjectives and nouns

In total, the precision is 89.58% (see table 29), which is considerably above our goal of 75%. We also see that the precision for the different kinds of agreement is always at least 75%.

8.4.3.2. Results for the SUMAT Test Set Translated with System 9

In this part of the evaluation we translated the SUMAT test set with system 9, applied the grammar checker and classified the corrections. We used again the categories “true positives”, “false positives” and “impossible to decide” (see section 8.3.3.1).

In total, the grammar checker corrected 61 sentences (see table 30). Compared to the results of the translation with the SUMAT system (see table 29), we observe a clear reduction of the number of corrections. This endorses the hypothesis that an improvement of the system reduces the number of disagreements.

We classified 52 of the 61 corrections as “true positives”, 4 as “false positives” and 5 as “impossible to decide”. The total precision is 92.86% (see table 30) which is considerably above 75%. As in the previous part of the evaluation (see section 8.4.3.1), we see that the precision for the different kinds of agreement is always considerably above 75%.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Adj. (m) + noun (f):	12	1	1	92.31%
Adj. (f) + noun (m):	10	1	0	90.91%
Adj. (sg) + noun (pl):	12	2	2	85.71%
Adj. (pl) + noun (sg):	0	0	0	-
Noun (f) + adj. (m):	5	0	0	100%
Noun (m) + adj. (f):	4	0	2	100%
Noun (pl) + adj. (sg):	8	0	0	100%
Noun (sg) + adj. (pl):	1	0	0	100
Total:	52	4	5	92.86%

Table 30.: Evaluation SUMAT test set of 2012 translated by system 9: Disagreements between adjectives and nouns

8.4.3.3. Results for the VSI Test Set Translated with System 9

In this part of the evaluation we translated the VSI test set with system 9 (see table 18), applied the grammar checker and classified the corrections. We used the same categories as in the other evaluations (see section 8.3.3.1).

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Adj. (m) + noun (f):	8	0	5	100%
Adj. (f) + noun (m):	7	2	3	77.78%
Adj. (sg) + noun (pl):	21	3	1	87.5%
Adj. (pl) + noun (sg):	1	2	3	33.3%
Noun (f) + adj. (m):	7	1	0	100%
Noun (m) + adj. (f):	2	2	1	100%
Noun (pl) + adj. (sg):	6	0	0	100%
Noun (sg) + adj. (pl):	1	0	0	100
Total:	53	10	14	84.21%

Table 31.: Evaluation VSI test set of 2012 translated by system 9: Disagreements between adjectives and nouns

In total, the grammar checker corrected 77 disagreements. We classified 53 of the 77 corrections as “true positives”, 10 as “false positives” and 14 as “impossible to decide” (see table 32). The number of true positives is almost the same as in the SUMAT test set translated with system 9 (see section 8.4.3.2). The precision (84.21%) is lower than for the SUMAT test set, but still considerably above 75%.

8.4.3.4. Results for the OpenSubtitle Test Set Translated with System 9

The OpenSubtitle test set (545 lines) is smaller than the other evaluated test sets (approximately 4000 lines). The grammar checker made only 6 corrections in the OpenSubtitle test set. Possible reasons are the shortness of the test set and the considerably better translation quality of this test set compared to the other test sets (see section 7.2.7).

The precision is 80%, which is above 75% (see table 32). The calculated precision is not very meaningful, because the number of corrections is low and therefore the result is not meaningful.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Adj. (m) + noun (f):	2	0	0	100%
Adj. (f) + noun (m):	0	0	1	-
Adj. (sg) + noun (pl):	1	0	0	100%
Adj. (pl) + noun (sg):	0	0	0	-
Noun (f) + adj. (m):	1	1	0	50%
Noun (m) + adj. (f):	0	0	0	-
Noun (pl) + adj. (sg):	0	0	0	-
Noun (sg) + adj. (pl):	0	0	0	-
Total:	4	1	1	80%

Table 32.: Evaluation OpenSubtitle test set of 2012 translated by system 9: Disagreements between adjectives and nouns

8.4.3.5. Comparison of the Results of the Test Sets

The developed grammar checker corrects disagreements between adjectives and nouns with a high precision. For all the test sets the precision is considerably above 75%. Therefore, we included the corrections of disagreements between adjectives and nouns in the final version of the grammar checker.

The evaluation yields additional findings. The number of corrections in translations of different translation systems differs considerably. Furthermore, the number of corrections differs between test sets containing amateur subtitles and others containing professional subtitles.

8.4.3.6. Discussion of the False Positives

The main reasons for the false positives are wrong part-of-speech tags (assigned by Freeling) and named entities that were not recognized. The grammar checker never corrects disagreements containing named entities, except if the named entities are not identical in English and Spanish (see section 8.4.2). Example 8.35 shows the problem of named entities that are not recognized by Freeling and are not identical in English and Spanish. The named entity *new zealand* is correctly translated with *nueva zelanda*. Freeling does not recognize *nueva zelanda* as a named entity because of a recasing error. Thus *zelanda* is incorrectly tagged as an adjective and the grammar checker detects erroneously a disagreement with the following noun (*compañero*).

We identified the wrong tagging of the token *solo* as an additional source for false positives. *Solo* can occur as adverb or adjective. Sometimes Freeling tags *solo* incorrectly as an adjective instead of an adverb. Thus, the grammar checker detects wrongly a disagreement, which causes a false positive (see ex. 8.36).

(8.35) Input: ***nueva zelanda*** *compañero , marc jacobs , él pasó por su calor .*
Outfput: ***nueva zelando*** *compañero , marc jacobs , él pasó por su calor .*

(8.36) Input: *y los dos bateristas son sólo truenos , ringo y keltner son* ***solo truenos*** *.*
Output: *y los dos bateristas son sólo truenos , ringo y keltner son* ***solos truenos*** *.*

8.4.3.7. Discussion of the Errors in the Corrections

In two sentences the grammar checker changes the adjective *múltiples* to an incorrect singular word form (*múltipl*). This is a result of the simplification and generalization of the grammar rules included in the grammar checker. In cases in which a plural adjective ending with *-es* must be changed to singular, the grammar checker deletes the ending *-es*, if no vowel or *-t-* precedes the ending. According to this rule, the grammar checker corrects the plural adjective *múltiples* to *múltipl* instead of *múltiple* (see ex. 8.37). To avoid this mistake, we changed this rule in the following way: if *-l-* precedes the ending, the grammar checker reduces the ending *-es* to *-e*. This created new errors: for example, the plural adjective *fragiles* would be corrected to the singular form *fragile* instead of *frágil*. Therefore, we removed the alternation of this rule. We classified both cases in the category "impossible to decide", because the sentence contains different errors before and after the correction. Before the

correction the noun and the adjective disagreed, after the correction they agreed, but the form of the adjective was incorrect.

(8.37) Input: *candide thovex , una leyenda viviente y* ***múltiples*** *ganador de los x games ,*
Output: *candide thovex , una leyenda viviente y* ***múltipl*** *ganador de los x games ,*

8.4.3.8. Results of the Test Suites

We created Test Suites for a more systematic evaluation of certain linguistic phenomena, as we did for the disagreements between determiners and nouns. We decided that the Test Suites consists only of phrases, because phrases suffice to show if the correction succeeds. We need fewer phrases than sentences for the disagreements between determiners and nouns, because only two types of adjectives (qualifying and ordinal adjectives) exist.

The evaluation with our Test Suites confirms that, in general, the correction of disagreements between adjectives and nouns succeeds (see appendix annex:Tables). In cases in which the gender as well as the number are wrong, only the gender is corrected (see ex. 8.38). This is not acceptable and must be solved. We improved the grammar checker to enable corrections of gender and number in the same sentence. We used the output of the correction of gender as input to correct the number. In other words, if after the correction of gender a disagreement of number is detected, it will be corrected. This change improved the recall for the correction of disagreements of number.

With the evaluation of the modifications in these Test Suites we identified the following mistake: the grammar checker does not consider changes of accents. For example, the grammar checker changes the singular adjective *débil* to the plural form *débiles* instead of *debiles* (see ex. 8.39). To avoid such mistakes, phonetic and phonologic rules might be included in the grammar checker. Another solution would be the use of an existing spell checker. This must be investigated in future projects.

The results of the Test Suites also show that the grammar checker only corrects the adjective that precedes or follows the noun directly. Examples exist (see ex. 8.40) in which two adjectives combined with a conjunction belong to the same noun. Future releases of the grammar checker might be able to correct disagreements with patterns of this type.

(8.38) Input: *las casas* ***antiguo***

Output: *las casas* ***antigua*** (*antigua* instead of *antiguas*)

(8.39) Input: *las personas* ***débil***
Output: *las personas* ***débiles*** (*débiles* instead of *debiles*)

(8.40) Input: *la persona* ***amables*** *y* ***fuertes***
Output: *la persona* ***amable*** *y* ***fuertes*** (*fuertes* instead of *fuerte*)

8.5. Disagreements between Determiners and Adjectives

In this section, we discuss the detection and correction of disagreements between determiners and adjectives. In sentences and clauses, the determiners and adjectives always have to agree with the noun, therefore it would not be necessary to check the agreements between determiners and adjectives. Our grammar checker only considers disagreements between adjacent words, because we did not include a syntactical analysis and therefore it is unclear which words belong together if they are not adjacent. As a consequence the grammar checker does not detect disagreements between determiners and nouns, if an adjective is interposed. Therefore we decided to detect and correct also disagreements between determiners and adjectives to improve the recall. If further projects include syntactical information and improve the detection of disagreements between determiners and nouns, the disagreements between determiners and adjectives will not have to be considered anymore.

8.5.1. Restricted Error Analysis

We make a restricted error analysis for disagreements between determiners and adjectives. We use the same method as for the disagreements between determiners and nouns and between adjectives and nouns, and we also use the SUMAT test set (version 2012) translated by with SUMAT system (version 2012).

The grammar checker only has to consider cases in which the adjective precedes the noun. The reason is that if the adjective follows the noun, the rules of the grammar checker already ensure that the noun agrees with the following adjective and the preceding determiner. This means that also the determiner and the adjective agree.

Disagreement:	Detected:	True Positives:	False Positives:	Impossible to decide:
Det. (m) + adj. (f):	18	16	1	1
Det. (f) + adj. (m):	5	3	1	1
Det. (sg) + adj. (pl):	11	9	0	2
Det. (pl)+ adj. (sg):	10	4	1	5
Total:	44	32	3	9

Table 33.: Restricted error analysis: Disagreements between determiners and adjectives

In total, our script for the error analysis found 44 disagreements between determiners and adjectives (see table 33). 32 of the 44 cases are real disagreements and should be corrected. 3 of the 44 detected disagreements are false positives and in 9 cases it is impossible to decide. These results show that if we correct all detected disagreements, the precision would be 91.43%, which is already above the required 75%.

We classified the following sentences as false positives:

(8.41) *nadezhda von meck la viuda de* ***una promotoro exprés*** (*promotoro* is wrongly tagged as an adjective) -

(8.42) *cuando los beatles ellos tocaron juntos como grupo* en **el shea stadium** . (*shea stadium* is not recognized as a named entity)

(8.43) *porque tuvimos anuncio libs y* ***esos bv piezas*** *que fui allí ,* (*bv* is an untranslated abbreviation)

In example 8.41 the grammar checker detected erroneously a disagreement between the determiner and adjective, because the grammar checker changed *promotora* (feminine) erroneously into *promotoro* (masculine) in a previous part of the program. The grammar checker applied this change because of wrong part-of-speech tags assigned by Freeling. Freeling tagged *exprés* as a masculine noun and *promotora* as a feminine adjective. This means that without changing anything, the agreement in phrase (*una promotora exprés*) is correct.

In example 8.42, the uncorrected phrase *en el shea stadium* is grammatically correct even though the grammar checker detected a disagreement between the adjective and determiner. The reason for this is that Freeling did not recognize the masculine named entity *shea stadium* because of a recasing error and Freeling tagged *shea* as a feminine adjective.

The detected disagreement in example 8.43 contains *bv*, an untranslated English ab-

breviation for which the gender and number is unclear. Freeling tags *bv* as a singular adjective. Therefore, the grammar checker detects erroneously a disagreement between *bv* and the plural determiner. If we consider the plural feminine noun *piezas*, we see that the use of a plural determiner is correct and in fact no disagreement exists.

8.5.2. Development of the Rules

This section describes the development of the rules for the detection and correction of disagreements. The restricted error analysis showed that if the grammar checker corrects all detected disagreements, the precision would be 91.43%, which is considerably above 75%. Therefore, not many rules are required to distinguish between true and false positives. The only rule we use to distinguish between true and false positives in the final version of the grammar checker is that the adjective is not allowed to occur in the corresponding line of the English source text. With this rule we exclude untranslated English words and not recognized named entities that are identical in English and Spanish. We did not include LanguageTool, because no rules exist in LanguageTool to detect disagreements between determiners and adjectives.

The grammar checker always makes the corrections on the determiner. The determiner has to agree with the noun and the previous section of the grammar checker ensures that the adjective agrees with the noun. If we ensure in this part of the grammar checker that the determiner is adjusted to the agreement of the adjective, we automatically ensure the agreement between the determiner and noun. For the corrections of the determiner we used the same rules as listed in section 8.3.2.

8.5.3. Evaluation and Improvement

The evaluation procedure is the same as as for the disagreements between determiners and noun and disagreements between adjectives and nouns (see section 8.3.3.1).

8.5.3.1. Results for the SUMAT Test Set Translated with the SUMAT System

First, we evaluated the SUMAT test set (version 2012) that we translated with the SUMAT system (version 2012). In total, the grammar checker corrected 39 disagreements. 32 of the 39 corrections are true positives and 1 correction is a false

positive. We classified 6 corrections in the category "it is impossible to decide". The precision is 96.97%, which is considerably above 75% (see table 8.5). We observe that the number of disagreements between determiners and adjectives is lower than the number of disagreements between adjectives and nouns, and determiners and nouns (see sections 8.3.3.2 and 8.4.3.1).

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + adj. (f):	16	0	1	100%
Det. (f) + adj. (m):	3	1	1	75%
Det. (sg) + adj. (pl):	9	0	2	100%
Det. (pl) + adj. (sg):	4	0	2	100
Total:	32	1	6	96.97%

Table 34.: Evaluation SUMAT test set of 2012 translated by the SUMAT system: Disagreements between determiners and adjectives

8.5.3.2. Results for the SUMAT Test Set Translated with System 9

In the translation of the SUMAT test set (version 2012) with system 9 the grammar checker corrected fewer disagreements than in the translation with the SUMAT system (version 2012). In total, the grammar checker made 19 corrections (see table 35). 15 of the 19 corrections are true positives, we classified 4 corrections in the category "impossible to decide". No false positives occur, therefore, the precision is 100%.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + adj. (f):	5	0	0	100%
Det. (f) + adj. (m):	3	0	2	75%
Det. (sg) + adj. (pl):	7	0	1	100%
Det. (pl) + adj. (sg):	0	0	1	-
Total:	15	0	4	100%

Table 35.: Evaluation SUMAT test set of 2012 translated by system 9: Disagreements between determiners and adjectives

8.5.3.3. Results for the VSI Test Set Translated with System 9

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + adj. (f):	2	0	1	100%
Det. (f) + adj. (m):	1	0	0	100%
Det. (sg) + adj. (pl):	8	0	1	-%
Det. (pl) + adj. (sg):	3	1	2	75
Total:	12	1	4	92.31%

Table 36.: Evaluation VSI test set translated by system 9: Disagreements between determiners and adjectives

In the VSI test set the grammar checker made 17 corrections which is almost the same as in the SUMAT test set translated with system 9. 12 of the 17 corrections are true positives and 1 is a false positives. We classified 4 corrections in the category "impossible to decide". The precision is 92.31% (see table 36), which is considerably above 75%.

8.5.3.4. Results for the OpenSubtitle Test Set Translated with System 9

In the OpenSubtitle test set the grammar checker corrected only 2 disagreements between determiners and adjectives (see table 37). For one of the corrected disagreements it is impossible to decide if the correction causes an improvement or worsening of the sentence. The other corrected disagreement is a true positive, therefore, the precision is 100%. This result is not meaningful because of the low number of corrections.

Disagreement:	True Positives:	False Positives:	Impossible to decide:	Precision:
Det. (m) + adj. (f):	1	0	0	100%
Det. (f) + adj. (m):	0	0	1	-%
Det. (sg) + adj. (pl):	0	0	0	-%
Det. (pl) + adj. (sg):	0	0	0	-
Total:	1	0	1	100%

Table 37.: Evaluation OpenSubtitle test set translated by system 9: Disagreements between determiners and adjectives

8.5.3.5. Comparison of the Results of the Test Sets

For all evaluated test sets, the precision is above 90%, which considerably exceeds our limit. Therefore, we decided to include our developed rules for the correction of disagreements between determiners and adjectives in our final version of the grammar checker.

We observed that the number of corrections of disagreements between determiners and adjectives is lower than the number of corrections of disagreements between determiners and nouns and between adjectives and nouns. Additionally, this evaluation confirmed our conclusion that the number of disagreements depends on the quality of the translation systems, as well as on the characteristics of the subtitles (VSI subtitles or OpenSubtitles).

8.5.3.6. Discussion of the False Positives

In total, we classified only two corrections as false positives. In the following, we discuss the false positive of example 8.44. This false positive occurred because of a disagreement between the adjective and the noun that the grammar checker did not correct. The detection of this disagreement failed because of a wrong translation of a compound (see ex.8.44). The translation system did not find the correct Spanish translation for the compound *snow park*. The system translated *snow* as *nieve* and used *parks* untranslated. The order of the compound components in the resulting Spanish translation is identical to the order of the compound components in the English source text. This order of compound components is wrong in Spanish, because in Spanish, the inflected part must precede the non-inflected part. The grammar checker assumes a correct order of the compound components and checks the agreements between the adjective and the first compound component. Therefore, the grammar checker did not detect the disagreement between the adjective and the second compound component *parks*. Consequently, the grammar checker corrects the disagreement between the determiner and adjective. As a result of this, the determiner agrees with the adjective, but the determiner and adjective do not agree with the compound noun.

(8.44) Input:*en val senales y kronplatz , **2 de los mejor italiano nieve parks** .*
Output: *en val senales y kronplatz , **2 del mejor italiano nieve parks** .*

8.5.3.7. Results of the Test Suites

We used Test Suites again for a more systematic evaluation of the correction of disagreements between determiners and adjectives in selected linguistic phrases. The results of the Test Suites for the correction of disagreements between determiners and nouns showed (see section 8.3.3.8), that the grammar checker corrections succeed for the different kinds of determiners. Therefore, it is not necessary to test the correction for all determiners again. Thus, we restrict the Test Suites to a selection of a few determiners, which occur in different morpho-syntactic constructions with adjectives. Hence, the number of phrases in the Test Suites is lower than for the evaluation of corrections of disagreements between determiners and nouns.

The translations of these Test Suites show that the correction of disagreements between articles and adjectives succeeds, except for a few cases in which the adjectives have an identical form in masculine and feminine (see appendix annex:Tables). For these cases, Freeling cannot decide if the gender is masculine or feminine and sets a *C* (common) in the morphological analysis. Consequently, the grammar checker cannot detect a disagreement of gender between determiners and such adjectives. This reduces the recall. Examples 8.45 and 8.46 show that *grandes* can be masculine as well as feminine. In both cases the gender of the determiner is wrong, but the grammar checker detects no disagreements.

(8.45) Input: ***las*** *grandes edificios*
Output: ***las*** *grandes edificios* (*las*(f) instead of *los*(m))

(8.46) Input: ***los*** *grandes casas*
Output: ***los*** *grandes casas* (*los*(m) instead of *las*(f)).

We can solve this problem if we consider the gender of the noun. With this solution we improved the detection rules for disagreements between determiners and adjectives: if the gender of the adjective is tagged with *C*, the grammar checker uses the gender of the noun to detect a disagreement. To avoid false positives in the detected disagreements, the grammar checker applies subsequently the following rule: if the noun does not appear in the corresponding line of the English source text, the grammar checker corrects the disagreement. This ensures the correction of the sentences 8.45 and 8.46 of the Test Suites.

We also tested the improved rules for the correction of disagreements between determiners and adjectives in in the SUMAT test set (version 2012) translated with the SUMAT system (version 2012). The grammar checker detected 5 additional disagreements (all true positives), which means that the improvement of the rules

augmented the recall. Therefore, we decided to include the improved rules in the final version of our grammar checker.

8.6. Disagreements with Verbs

8.6.1. Restricted Error Analysis

We considered including corrections of disagreements with verbs in the grammar checker. We tried to make an error analysis restricted to the following cases of disagreements with the verbs:

- disagreements of person and number, if a conjugated verb directly follows a personal pronoun in the nominative case
- disagreements of number, if a conjugated verb directly follows a noun in the nominative case

We did not include proper nous (e.g. names) in the error analysis, because Freeling tags them as proper nouns but applies no further morphologic analysis.

Freeling indicates the case only if no alternative case is possible which is never the case for nouns and only for some personal pronouns in singular. Consequently, the recall would be too low. Therefore, we did not include the condition that the pronoun and noun have to be in the nominative case in the script for the detection of disagreements with verbs. Thus the restricted error analysis considers the following disagreements:

- disagreements of person and number, if a conjugated verb directly follows a personal pronoun (based on a list with personal pronouns)
- disagreements of number, if a conjugated verb directly follows a noun

For the restricted error analysis we used the SUMAT test set (version 2012) translated by the SUMAT system. Our script found 20 disagreements of person between pronouns and verbs (see table 38). 4 of the 20 cases are true positives, 15 are false positives and for 1 case it is "impossible to decide". Most of the false positives occur because of a wrong morphologic analysis of the verb. In example 8.47 the verb *tocaba* can refer to the first or to the third person. Freeling assigns the third person but in fact the verb refers to the first person. Other false positives occur because of personal pronouns that are not the subject (they are not in the nominative case). In example 8.48 *nosotros* is not the subject but a tonic pronoun in the prepositional

phrase.

Disagreement:	Detected:	True Positives:	False Positives:	Impossible to decide:
Pronoun/verb (person):	20	4	15	1
Pronoun/verb (number):	11	2	9	0
Noun/verb (number):	74	15	49	10
Total:	105	21	73	11

Table 38.: Restricted error analysis: Disagreements between personal pronouns and verbs

An analysis of the true positives and a comparison with the English source text showed that for the correction of the disagreement, the grammar checker would have to adjust the verb form. The development of own rules for the correction of the verb is costly and time-consuming because of the complexity of the verb paradigm. Instead we could use a morphological generator for Spanish verb forms. Compared to the low number of disagreements that the grammar checker would correct with this solution, the effort for the integration (or even development) of a morphological generator in our grammar checker is too high. In this test set the grammar checker corrected only 4 disagreements and we suspect that in the translations of better systems (e.g. with system 9) the number of disagreements is probably even lower. Therefore we decided not to include the correction of disagreements of person between pronouns and verbs in the grammar checker.

(8.47) ***yo tocaba*** *cada nota*

(8.48) *¿ qué es importante* ***para nosotros*** *es encontrar clínica músicos .*

(8.49) ***iggy pop y yo éramos*** *un par de muy chicos malos .*

In total, our script detected 11 disagreements of number between pronouns and verbs (see table 38). We classified 2 cases as true positives and 9 cases as false positives. We observe that the majority (8 of 11 cases) of the detected disagreements contain the personal pronoun *yo*. The false positives containing *yo* occur because of nominal phrases containing two or more agents combined with a copula (generally *and*). In example 8.49, the agents are *Iggy Pop* (proper name) and *yo* (= personal pronoun *I*). This combination of agents requires the use of a plural verb. Our script considers only the pronoun that directly precedes the verb and therefore the script detects a disagreement. The reason for the remaining false positives are again personal pronouns that are not atonic personal pronouns in the nominative case but tonic personal pronouns of prepositional phrases.

We observe again that the number of true positives is low. Just as for the disagreements of person between pronouns and verbs we decided not to include the correction of disagreements of number between pronouns and verbs in the grammar checker.

(8.50) *el verdadero cambio en nuestras* ***vidas surgió*** *cuando nuestro primogénito nací ,*

(8.51) *hay muchas* ***historias pudiera*** *contártelo .*
(missing relative pronoun: *historias que pudiera*)

(8.52) *recordar algunos de los otros* ***temas hice*** *.*
(missing relative pronoun: *temas que hice*)

Our script detected 72 disagreements of number between nouns and the following verb. We manually classified 15 of the 72 detected disagreements as true positives, 49 as false positives and 10 as "impossible to decide". Most of the false positives occur because the noun that precedes the verb is not the subject and, therefore, this noun and the inflected verb do not have to agree. Other reasons for false positives are more complex nominal phrases containing prepositional phrases (see ex. 8.50), subordinate clauses before the verb and missing relative pronouns in the translation (see ex. 8.51 and 8.52). We did some experiments to find out how we can distinguish with the grammar checker between the true and false positives. None of our experiments was successful. Therefore, we decided not to include the correction of the disagreements of number between nouns and the following verb in the grammar checker.

Additionally, we observed that the verb form is often still wrong after the correction of the number in the detected disagreements. In example 8.54, the grammar checker would have to change the verb form not only from singular to plural but also from imperative to past tense. In example 8.55, the grammar checker would have to change the participle into a relative phrase. Although the correction of the number would improve the sentence, the effort for the translator or proofreader would remain the same because the verb has to be replaced anyway.

(8.53) *los* ***niños volverá*** *a casa de sus varias actividades* (plural form *volverán* instead of *volverá*)

(8.54) *y esto es algo que , como los* ***médicos dime*** *, (me dijeron* instead of the imperative *dime*)

(8.55) *los dos* ***hombres conocido*** *en moscú* (might probably be changed into a relative phrase)

8.7. Prepositions Demanding Infinitives

8.7.1. Restricted Error Analysis

Some Spanish prepositions require an infinitive if they are followed by a verb. Frequent prepositions which demand infinitives are *para*, *a* and *de*. In order to ensure that these prepositions are followed by an infinitive if the following word is a verb, we try to develop rules for the grammar checker.

In this section we make a restricted error analysis with the SUMAT test set (version 2012) translated by the SUMAT system (version 2012).

Disagreement:	Detected:	True Positives:	False Positives:	Impossible to decide:
peposition[41] + inf.	24	12	5	7

Table 39.: Restricted error analysis: Prepositions demanding infinitives

In total, our script detected 24 errors in which the considered prepositions (*de, a, para*) are followed by other verb forms than the infinitive (see table 39). We classified 12 of 24 cases as true positives, 5 as false positives and 7 as "impossible to decide". The reason why the script detected false positives is the wrong tagging of proper names. We can see that Freeling tagged the proper names (*Sebastian*) and *Frolovskoye* incorrectly as verbs in the examples 8.56 and 8.57.

(8.56) *de dos partitas fledermaus* ***de sebastian bach*** *, incluyendo el gran chaconne ,*

(8.57) *en abril se mudó a un nuevo " country house , en las afueras* ***de frolovskoye klin*** *,*

8.7.2. Development of the Rules

We complemented the rules used for the error analysis with additional selection rules to avoid false positives. The restricted error analysis showed that proper names with incorrect tags cause false positives. To exclude the false positives we have to distinguish between verbs which Freeling tagged correctly and verbs which Freeling tagged incorrectly and which are in fact proper names. To achieve this, we assume that the majority of the proper names are identical in English and Spanish. In addition, we assume that verbs are never identical in English and Spanish. We then compare the Spanish sentence with the corresponding line of the English source

text. If the verb that follows the preposition in the Spanish sentence also occurs in the corresponding English line, it is probably a proper name which is tagged wrongly and we do not correct anything. If the verb does not exist in the corresponding line of the English source text, we assume that the tagging is correct and we change the verb form to the infinitive.

For the correction of the detected errors the grammar checker replaces the verb form with the corresponding infinitive. We can get the infinitive from the Freeling output in which the lemma for each token is determined.

Sometimes the grammar checker must replace gerunds which are combined with a pronoun (e.g. *preguntándose*, *se* = pronoun) with the corresponding infinitive. We discovered that the grammar checker does not correct these errors. The reason is that Freeling splits the gerund with the pronoun into two tokens (e.g. *preguntándose* is split in *preguntado* and *se*). The splitting causes the loss of the accent at the gerund and the mapping between the output of Freeling and the original translation fails. Therefore we developed an additional rule which enables the mapping between verb forms which are identical except for the accent. We adjusted the rules in order to ensure that the grammar checker combines the resulting infinitive with the pronoun of the gerund.

8.7.3. Evaluation

For the evaluation we used the same method and the same test sets (SUMAT, VSI and OpenSubtitle test set) as for the other error classes (see section 8.3.3.1). We did not use Test Suites, because almost all of the possible combinations also exist in the test sets.

Disagreement:	**True Positives:**	**False Positives:**	**Impossible to decide:**	**Precision:**
SUMAT test set (SUMAT syst.):	12	0	6	100%
SUMAT test set (syst.9):	4	4	5	50%
VSI test set (syst.9):	2	2	2	50%
OpenSubtitle test set (syst.9):	0	0	0	-
In total:	18	6	13	75%

Table 40.: Evaluation: Prepositions demanding infinitives

In the SUMAT test set (version 2012) translated with the SUMAT system (version 2012) the grammar checker found 18 errors (see table 40). We classified 12 of 18

cases as true positives and 6 as "impossible to decide". We did not find any false positives; the precision is 100%.

In the translation of the SUMAT test set with system 9, the grammar checker found 12 cases in which the prepositions *para*, *a* or *de* are followed by a non-infinitive verb form (see table 40). We classified 4 of 12 cases as true positives, 4 as false positives and 5 into the category "impossible to decide". The resulting precision is 50%, which is below our recommended minimum of 75%.

The grammar checker found 6 errors in the VSI test translated with system 9 (see table 40). We classified 2 cases as true positives, 2 as false positives and 2 as "impossible to decide". The reason for the false positives are nouns which are incorrectly tagged as verbs. Also for this test set the precision is 50% which is considerably below 75%.

In the translation of the OpenSubtitle test set the grammar checker did not find a case in which prepositions *a*, *para* or *de* are followed by a non-infinitive verb form.

In total, we classified 6 of the detected errors as false positives. Three of them occur because of nouns that are erroneously tagged as verbs. To avoid the correction of these false positives we developed a rule that denies the correction of words which are tagged as verbs but which can also be tagged as nouns. We get the alternative part-of-speech tags from the Freeling output. The three remaining false positives occur because of past participles which are positioned erroneously after the preposition. To avoid the correction of these false positives, we developed a rule that denies the correction of past participles. With these additional rules we achieved a precision of 100% (see table 41). This result should be confirmed with bigger test sets because of the low number of corrections in our evaluation.

The results show again that the improvement of the translation system reduces the number of detected and corrected errors considerably (see table 41).

Disagreement:	**True Positives:**	**False Positives:**	**Impossible to decide:**	**Precision:**
SUMAT test set (SUMAT syst.):	11	0	6	100%
SUMAT test set (syst.9):	2	0	3	100%
VSI test set (syst.9):	2	0	1	100%
OpenSubtitle test set(syst.9):	0	0	0	-
In total:	15	0	10	100%

Table 41.: Evaluation: Prepositions demanding infinitives, after the improvement

8.8. Evaluation of the Grammar Checker

In total we tried to include five different error classes in the grammar checker:

- disagreements between determiners and nouns (see section 8.3)
- disagreements between adjectives and nouns (see section 8.4)
- disagreements between determiners and adjectives (see section 8.5)
- specific cases of disagreements between pronouns/nouns and verbs (see section 8.6)
- prepositions *a*, *para* and *de* followed by a non-infinitive verb form (see section 8.7)

For each of this error classes we made a restricted error analysis, developed the rules and evaluated the results with test sets and Test Suites. We showed that the detection and correction of disagreements between pronouns/nouns and verbs did not succeeded and, therefore, we did not include this error class in the grammar checker. The evaluations of the other error classes showed a precision above 75%, so we included them in the grammar checker. We observed in all evaluations that the improvements of the translation systems reduce the number of grammatical errors which our grammar checker detects in the translations considerably.

In this section we apply the final version of the grammar checker and evaluate the corrections (see section 8.8.1). In addition, we run the grammar checker with another type of text to test if the grammar checker is text-type specific.

8.8.1. Evaluation with Subtitle Test Sets

Test set:	Corrections:	True Positives:	False Positives:	Impossible to decide:	Precision:
SUMAT (SUMAT syst.):	206	165 (80.10%)	19 (9.22%)	20 (9.71%)	89.67%
SUMAT (syst.9):	103	85 (82.52%)	7 (6.80%)	11 (10.68%)	92.39%
VSI (syst9):	116	91 (78.45%)	9(7.75%)	16 (13.79%)	91%
Opensubtitle (syst.9)):	6	4 (66.67%)	1(16.67%)	1 (16.67%)	80%
total:	431	346 (80.28%)	43 (9.98%)	48 (11.37%)	90.76%

Table 42.: Overall evaluation of the grammar checker for all considered test sets and error classes

In all test sets the final version of the grammar checker corrected in total 431 gram-

matical errors (see appendix B). We classified 344 of them (79.81%) as true positives and 35 (8.12%) as false positives. The precision is 90.76% which is considerably above 75% (see table 43. The precision is very similar (about 90%) for all test sets. For the OpenSubtile test set, in which the grammar checker corrected only few errors, the precision is even 80%. In some sentences the grammar checker corrected more than one error.

8.8.2. Application of the Script to another Type of Text

In this section we test if the grammar checker can also correct other types of text. Therefore, we use a selection of 512 parallel sentences of a DVD user manual contained in the SMULTRON[42] corpus. In total, the English version (source text) of this DVD user manual contains 10'283 tokens, in average these are 20.08 tokens per line. This is almost twice the amount of tokens per line as in the subtitle corpora and test sets.

For the experiments the sentences of the DVD user manual were not translated with the system 9 of this project, because this system is only appropriate for the translation of subtitles. Therefore, we used the Google Translator[43] for the translation. Then we analyzed the translated sentences with Freeling and applied LanguageTool. We used the translation and the output of Freeling and LanguageTool as input for our grammar checker.

Test set:	Corrections:	True Positives:	False Positives:	Impossible to decide:	Precision:
DVD user manual:	17	4	10	3	28.57%

Table 43.: Application of the grammar checker to another type of text

The grammar checker corrected 17 lines in total (see appendix C). We classified 4 of the corrected sentences as true positives (see ex. 8.58) and 10 as false positives (see ex. 8.60 and 8.61). We classified 3 corrections into the category "impossible to decide" (see ex. 8.59). The precision is 28.57% which lets us assume that our grammar checker is text-type specific and should not be run on other types of texts. This must be investigated in more detail in coming projects.

The main reasons for the false positives are wrong morphological analyses (see ex. 8.60) and nominal phrases consisting of more than one noun (see ex. 8.61), which

[42] http://mailman.uib.no/public/corpora/2009-December/009693.html
[43] http://translate.google.ch/?hl=entab=wTauto/es/

our grammar checker does not consider.

(8.58) **Input:** *este recordatorio se proporciona —- llamar la atención del instalador del sistema catv s de la sección 820-40 de nec, que - proporciona guías para el aterrizamiento adecuado y, en particular, especifica que se debe conectar el cable a tierra -* ***a la sistema*** *de tierra del edificio, tan cerca - hasta el punto de entrada del cable como sea práctico.*
Output:*este recordatorio se proporciona —- llamar la atención del instalador del sistema catv s de la sección 820-40 de nec, que - proporciona guías para el aterrizamiento adecuado y, en particular, especifica que se debe conectar el cable a tierra -* ***al sistema*** *de tierra del edificio, tan cerca - hasta el punto de entrada del cable como sea práctico.*

(8.59) **Input:** *el panel frontal y "disp"* ***botón alterna htr4*** *'s entre el tiempo / pistas / capítulos y subtítulos / ángulos / configuración de audio (sólo dvd).*
Output: *el panel frontal y "disp"* ***botón alterno htr4*** *'s entre el tiempo / pistas / capítulos y subtítulos / ángulos / configuración de audio (sólo dvd).*

(8.60) **Input:** *este enchufe encajará en* ***el tomacorriente*** *de una sola manera.* (correct determiner)
Output: *este enchufe encajará en* ***la tomacorriente*** *de una sola manera.* (incorrect determiner)
Reason: Freeling tagged *tomacorriente* erroneously as a feminine noun

(8.61) **Input:** *18 de objetos y líquidos — nunca introduzca objetos de ningún tipo en este producto a través de las aberturas, ya que podrían tocar* ***puntos de tensión peligrosos*** *o cortocircuitar piezas que - podría provocar un incendio o una descarga eléctrica.* (correct form of the adjective)
Output: *18 de objetos y líquidos — nunca introduzca objetos de ningún tipo en este producto a través de las aberturas, ya que podrían tocar* ***puntos de tensión peligrosa*** *o cortocircuitar piezas que - podría provocar un incendio o una descarga eléctrica.* (incorrect form of the adjective)
Reason: The grammar checker corrects a disagreement between the adjective and the noun*tensión*. In fact the adjective has to agree with the noun *puntos* and not with *tensión*

9. Conclusion

The goal of this project was to answer the following main research questions:

1. Does the combination of a statistical system with a rule-based grammar-checker improve the quality of the translation?
2. Is the combination of a statistical system with a rule-based grammar-checker recommendable in terms of the cost-benefit ratio?
3. How can we improve our approach in order to improve the translation quality?

The manual evaluation of the results of the grammar checker showed that the majority of the corrections made by the grammar checker improve the translated sentences. The precision of the grammar checker is about 90.76% (see section 8.8.1). We conclude that the corrections made by our grammar checker improve the quality of the translation and that our grammar checker can be applied without the risk of deteriorating the subtitle translations.

Considering the cost-benefit ratio we recommend the use of our grammar checker to correct the translations of our final system. Our suggestions for the improvement of the corpus (see section 5) and the experiments of SUMAT indicate that the SMT system can still be improved. We observed that an improvement of a SMT system results in a significant reduction of corrected grammatical errors (see section 8). Therefore, a cost-benefit ratio analysis must follow each improved system in order to estimate if further development of the grammar checker is recommendable. The cost-benefit ratio could possibly be improved by detecting and marking grammatical errors instead of additionally correcting the errors. By not correcting the errors the development effort would be reduced.

We observed significant differences in the number of corrections between the test corpora: the grammar checker corrects a considerable number of errors in the VSI test set and the SUMAT test set translations, in contrast it detects and corrects only few grammatical errors in the OpenSubtitle test set. This is probably due to the high translation quality of the OpenSubtitle test set compared to the VSI and SUMAT test sets. The application of the grammar checker on translated VSI and

SUMAT subtitles is more beneficial compared to its application on OpenSubtitles. Therefore we recommend further development of the grammar checker only if it is applied to VSI, SUMAT and similar subtitles.

In order to answer the third question, we had to consider the improvement of the SMT system as well as the improvement of the grammar checker. For the improvement of the SMT system, we tried to answer the following questions:

- What should be considered in the composition of the training, development and test corpora for the training of the SMT-system?
- How can the training data be improved?
- Is the combination of amateur and professional subtitles recommendable?

We discussed that our corpus for the training corpus consists of two heterogeneous parts: the VSI subtitles and the OpenSubtitle corpus (see section 5). A significant discrepancy between these two parts of the corpus is, for example, that one is sentence- and the other subtitle-based. Furthermore, the two parts differ in the number of tokens per line, their quality and the number of non-coincidental repetitions of complete subtitle sequences. The comparison with the results of the SUMAT project demonstrated that the inclusion of amateur subtitles and their combination with professional subtitles is recommendable, although the heterogeneity between professional and amateur subtitles is high. We did not weight the corpora differently in our project. Further projects might investigate if a weighted combination of these two corpora can improve the translation quality. Furthermore, tests might verify which is the optimal composition of the development set: either the use of amateur subtitles, which are similar to our training corpus, or the use of professional subtitles, which have a high quality.

A lot of improvement possibilities exist for the OpenSubtitle corpus: the correction of OCR errors can be improved, the extraction of repetitions (because of different film versions) might be improved and non-literal translations might be filtered out. We suggest only few improvements for the VSI corpus, such as filtering out non-literal translations, because the quality of the VSI subtitles is already high.

With reference to the improvement and extension of the grammar checker, we discussed the following research questions:

- Which kind of errors can be corrected with the grammar checker?
- or which error classes with low precision do the grammar checker rules have to be improved?

- Which possibilities do exist to extend the grammar checker?

We observed that the grammar checker works especially well for the correction of disagreements between substantives, adjectives and articles (see section 8). Further experiments should try to improve the recall and precision by improving the detection of disagreements. Our experiments showed that the detection of disagreements between verbs and nouns or pronouns is difficult. We did not achieve precision above 75% for the detection of disagreements with verbs. To improve the detection of disagreements between verbs and nouns or pronouns, we would need syntactical information, which was not available for our experiments. For the correction step we additionally would have to introduce a tool to generate the verb forms needed to correct the disagreements. The grammar checker also yields high precision for the changes of different verb forms into infinitives after selected prepositions.

Overall, further syntactical information is probably able to still improve the results of our grammar checker for all considered error classes. To obtain syntactical information, we recommend to parse or chunk the test sets. A performance test must show if the parsing of automatically translated texts is recommendable. In the present project we focused on the precision of our grammar checker. For an improvement and extension of our grammar checker, we might also take into account the recall.

Further projects might investigate if the corrections of our grammar checker are helpful and time-saving for proofreaders of automatically translated subtitles. For example, completely wrongly translated sentences containing grammatical errors must be retranslated manually anyway and we achieve no benefit with an automatic correction. Therefore, these projects must consider the context to decide if an automatic correction with the grammar checker saves time in the post-editing step and if the grammar checker should be further improved.

References

N. Bertoldi, B. Haddow, and J.-B. Fouet. Improved Minimum Error Rate Training in Moses. *Prague Bull. Math. Linguistics*, 91:7–16, 2009.

C. Callison-Burch and M. Osborne. Re-evaluating the role of BLEU in machine translation research. In *In EACL*, pages 249–256, 2006.

K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer, editors. *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Spektrum, Heidelberg, 3. edition, 2010.

T. R. Center. Freeling User Manual 3.0. may 2012.

J. Díaz Cintas. *Audiovisual translation: language transfer on screen*. Palgrave Macmillan, 2009a.

J. Díaz Cintas. Introduction - Audiovisual Translation: An Overview of its Potential. In J. Díaz Cintas, editor, *New Trends in Audiovisual Translation*, pages 1–18. Multilingual Matters, 2009b.

J. Díaz Cintas and A. Remael. *Audiovisual translation: subtitling* . St. Jerome Publishing, 2007.

B. Eckel. Evaluation Maschineller Übersetzungssysteme - Theorie und Praxis. Diplomarbeit, 1998.

M. Fishel, Y. Georgakopoulou, S. Penkale, V. Petukhova, M. Rojc, M. Volk, and A. Way. From Subtitles to Parallel Corpora. In *Proceedins of the 16th Annual Conference of the European Association for Machine Translation EAMT'2012*, pages 3–6, Trento, Italy, 2012.

G. C. F. Fong. Let the Words Do the Talking: The Nature and Art of Subtitling. In G. C. F. Fong and K. K. L. Au, editors, *Dubbing and subtitling in a world context*, pages 91–106. Chinese University Press, 2009.

A. Goldstein. *Foreign language movies - dubbing vs. subtitling* . Kovač, 2009.

C. Hardmeier. Using Linguistic Annotations in Statistical Machine Translation of Film Subtitles . Master's thesis, Universität Basel, 2008.

R. Hillman. Spoken Word to Written Text. Subtitling. In K. Malmkjær and K. Windle, editors, *The Oxford handbook of translation studies* , pages 379–424. Oxford University Press, 2011.

P. Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

P. Koehn. MOSES. Statistical Machine Translation System. User Manual and Code Guide. Apr. 2013.

A. Lavie and A. Agarwal. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

S. Nagel, S. Hezel, K. Hinderer, and K. Pieper. *Audiovisuelle Übersetzung. Filmuntertitelung in Deutschland, Portugal und Tschechien.* Leipziger Studien zur angewandten Linguistik und Translatologie. Lang, 2009.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318. ACL, 2002.

V. Petukhova, R. Agerri, M. Fishel, S. Penkale, A. del Pozo, M. S. Maucec, A. Way, P. Georgakopoulou, and M. Volk. SUMAT: Data Collection and Parallel Corpus Compilation for Machine Translation of Subtitles. In N. C. C. Chair, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

RAE. *Diccionario de la lengua española - Vigésima segunda edición.* La Real Academia Española y las Academias de la Lengua Española, 22 edition, 2011.

R. Raya. *Gramática básica del estudiante de español.* Klett, 2008.

L. Romero. When Orality Is Less Pre-fabricated: An Analytical Model for the Study of Colloquial Conversation in Audiovisual Translation. In L. McLoughlin, M. Biscio, and M. Mhainnín, editors, *Audiovisual Translation Subtitles and Subtitling. Theory and Practice.*, pages 19–54. Peter Lang, 2011.

R. Sennrich. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In W. Daelemans, M. Lapata, and L. Màrquez, editors, *EACL*, pages 539–549. The Association for Computer Linguistics, 2012.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

SUMAT. SUMAT Annual Public Report. 2012.

J. Tiedemann. Improved sentence alignment for movie subtitles. In *In Proceedings of RANLP, Borovets*, 2007.

J. Tiedemann. News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, volume V, pages 237–248. John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, 2009.

J. Tiedemann and L. Nygaard. The OPUS Corpus - Parallel and Free: http://logos.uio.no/opus. In *LREC*. European Language Resources Association, 2004.

M. Volk, R. Sennrich, C. Hardmeier, and F. Tidström. Machine Translation of TV Subtitles for Large Scale Production. In *Proceedings of the Second Joint EM+/CNGL Workshop "Bringing MT to the User: Research on Integrating MT in the Translation Industry"*, pages 54–62. Association for Machine Translation in the Americas, 2010.

A. The Application of the Grammar Checker on the Test Suites

We applied Test Suites to check the performance of the following error classes:

- Disagreements between determiners and nouns
- Disagreements between adjectives and nouns
- Disagreements between determiners and adjectives

Each test set consists of two parts:

- Incorrect sentences which should be corrected by the grammar checker
- Correct sentences which should not be changed by the grammar checker

A.1. Disagreements between Determiners and Nouns

A.1.1. Incorrect Sentences of the Test Suites

Before the correction	**After the correction**	**Improvement**	**Comment**
*Me gusta **aquella** libro.*	*Me gusta **aquel** libro.*	yes	
*Me gusta **aquellos** libro.*	*Me gusta **aquel** libro.*	yes	
*Me gusta **aquellas** libro.*	*Me gusta **aquel** libro.*	yes	
*Me gusta **aquel** camisa.*	*Me gusta **aquella** camisa.*	yes	
*Me gusta **aquellas** camisa.*	*Me gusta **aquella** camisa.*	yes	
*Me gusta **aquellos** camisa.*	*Me gusta **aquella** camisa.*	yes	
*Me gustan **aquel** libros.*	*Me gustan **aquellos** libros.*	yes	
*Me gustan **aquella** libros.*	*Me gustan **aquellos** libros.*	yes	
*Me gustan **aquellas** libros.*	*Me gustan **aquellos** libros.*	yes	
*Me gustan **aquel** camisas.*	*Me gustan **aquellas** camisas.*	yes	
*Me gustan **aquella** camisas.*	*Me gustan **aquellas** camisas.*	yes	
*Me gustan **aquellos** camisas.*	*Me gustan **aquellas** camisas.*	yes	
*Me gusta **esta** libro.*	*Me gusta **este** libro.*	yes	
*Me gusta **estos** libro.*	*Me gusta **este** libro.*	yes	
*Me gusta **estas** libro.*	*Me gusta **este** libro.*	yes	
*Me gusta **este** camisa.*	*Me gusta **esta** camisa.*	yes	
*Me gusta **estas** camisa.*	*Me gusta **esta** camisa.*	yes	
*Me gusta **estos** camisa.*	*Me gusta **esta** camisa.*	yes	

*Me gustan **este** libros.*	*Me gustan **estos** libros.*	yes	
*Me gustan **esta** libros.*	*Me gustan **estos** libros.*	yes	
*Me gustan **estas** libros.*	*Me gustan **estos** libros.*	yes	
*Me gustan **este** camisas.*	*Me gustan **estas** camisas.*	yes	
*Me gustan **esta** camisas.*	*Me gustan **estas** camisas.*	yes	
*Me gustan **estos** camisas.*	*Me gustan **estas** camisas.*	yes	
*Nunca he visto **tales** cosa.*	*Nunca he visto **tal** cosa.*	yes	
*Nunca he visto **tal** cosas.*	*Nunca he visto **tales** cosas.*	yes	
*Te presento a **mi** hermanos.*	*Te presento a **mis** hermanos.*	yes	
*Te presento a **mis** hermano.*	*Te presento a **mi** hermano.*	yes	
*He visto a **tu** hermanas.*	*He visto a **tus** hermanas.*	yes	
*He visto a **sus** hermana.*	*He visto a **su** hermana.*	yes	
*He visto a **nuestros** hermano.*	*He visto a **nuestro** hermano.*	yes	
*He visto a **nuestro** hermanos.*	*He visto a **nuestros** hermanos.*	yes	
*He visto a **vuestras** hermana.*	*He visto a **vuestra** hermana.*	yes	
*He visto a **vuestra** hermanas.*	*He visto a **vuestras** hermanas.*	yes	
*¿**Cuántos** dinero tienes?*	*¿**Cuántos** dinero tienes?*	no	tagging error
*¿**Cuánta** hermanas tienes?*	*¿**Cuántas** hermanas tienes?*	yes	
***Los** hombre es amable.*	***Los** hombre es amable.*	no	specialized rule is not adjusted to uppercase letters
***los** hombre es amable.*	***el** hombre es amable.*	yes	
***El** hombres son amables.*	***El** hombres son amables.*	no	specialized rule is not adjusted to uppercase letters
***el** hombres son amables.*	***los** hombres son amables.*	yes	
***las** mujer es amable.*	***la** mujer es amable.*	yes	
*Tengo que hacer **varia** cosas.*	*Tengo que hacer **varias** cosas.*	yes	
*Tengo que escribir **vario** textos.*	*Tengo que escribir **vario** textos.*	no	tagging error
*Tengo **unos** libro.*	*Tengo **un** libro.*	yes	
*Tengo **unas** hermana.*	*Tengo **una** hermana.*	yes	
*Tengo que hacer **tanta** cosas.*	*Tengo que hacer **tantas** cosas.*	yes	
*Tengo que escribir **tanto** textos.*	*Tengo que escribir **tanto** textos.*	no	tagging error
*Eso es **otras** cosa.* *No he recibido **ninguna** correo electrónico.*	*Eso es **otra** cosa.* *No he recibido **ningún** correo electrónico.*	yes yes	
*No he recibido **ningunos** correo electrónico.*	*No he recibido **ningún** correo electrónico.*	yes	
*No he recibido **ningunas** correo electrónico.*	*No he recibido **ningún** correo electrónico.*	yes	
*Tiene **una** alma buena.*	*Tiene **una** alma buena.*	no	because the feminine substantive beginns with *-a*
*Tiene **una** amigo bueno.*	*Tiene **un** amigo bueno.*	yes	
*Tiene **unos** amigo bueno.*	*Tiene **un** amigo bueno.*	yes	
*Lo he dicho a **todo mi** hermanos.*	*Lo he dicho a **todo mis** hermanos.*	(yes)	*todo* might be corrected also

A.1.2. Correct Sentences of the Test Suites

Before the correction	**After the correction**	**Changes**	**Comment**
*Me gusta **aquel** libro.*	*Me gusta **aquel** libro.*	no	
*Me gusta **aquella** camisa.*	*Me gusta **aquella** camisa.*	no	
*Me gustan **aquellos** libros.*	*Me gustan **aquellos** libros.*	no	
*Me gustan **aquellas** camisas.*	*Me gustan **aquellas** camisas.*	no	
*Me gusta **este** libro.*	*Me gusta **este** libro.*	no	

*Me gusta **esta** camisa.*	*Me gusta **esta** camisa.*	no	
*Me gustan **estos** libros.*	*Me gustan **estos** libros.*	no	
*Me gustan **estas** camisas.*	*Me gustan **estas** camisas.*	no	
*Nunca he visto **tal** cosa.*	*Nunca he visto **tal** cosa.*	no	
*Nunca he visto **tales** cosas.*	*Nunca he visto **tales** cosas.*	no	
*Te presento a **mis** hermanos.*	*Te presento a **mis** hermanos.*	no	
*Te presento a **mi** hermano.*	*Te presento a **mi** hermano.*	no	
*He visto a **tus** hermanas.*	*He visto a **tus** hermanas.*	no	
*He visto a **nuestros** hermanos.*	*He visto a **nuestros** hermanos.*	no	
*He visto a **nuestro** hermano.*	*He visto a **nuestro** hermano.*	no	
*He visto a **su** hermana.*	*He visto a **su** hermana.*	no	
*He visto a **vuestra** hermana.*	*He visto a **vuestra** hermana.*	no	
*He visto a **vuestras** hermanas.*	*He visto a **vuestras** hermanas.*	no	
*¿**Cuánto** dinero tienes?*	*¿**Cuánto** dinero tienes?*	no	
*¿**Cuántas** hermanas tienes?*	*¿**Cuántas** hermanas tienes?*	no	
***El** hombre es amable.*	***El** hombre es amable.*	no	
***el** hombre es amable.*	***el** hombre es amable.*	no	
***Los** hombres son amables.*	***Los** hombres son amables.*	no	
***los** hombres son amables.*	***los** hombres son amables.*	no	
***la** mujer es amable.*	***la** mujer es amable.*	no	
*Tengo que hacer **varias** cosas.*	*Tengo que hacer **varias** cosas.*	no	
*Tengo que escribir **varios** textos.*	*Tengo que escribir **varios** textos.*	no	
*Tengo **un** libro.*	*Tengo **un** libro.*	no	
*Tengo **una** hermana.*	*Tengo **una** hermana.*	no	
*Tengo que hacer **tantas** cosas.*	*Tengo que hacer **tantas** cosas.*	no	
*Tengo que escribir **tantos** textos.*	*Tengo que escribir **tantos** textos.*	no	
*Eso es **otra** cosa.*	*Eso es **otra** cosa.*	no	
*No he recibido **ningún** correo electrónico.*	*No he recibido **ningún** correo electrónico.*	no	
*Tiene **un** alma buena.*	*Tiene **un** alma buena.*	no	
*Tiene **un** amigo bueno.*	*Tiene **un** amigo bueno.*	no	
*Lo he dicho a **todos mis** hermanos.*	*Lo he dicho a **todos mis** hermanos.*	no	

A.2. Disagreements between Adjectives and Nouns

A.2.1. Incorrect Sentences of the Test Suites

Before the correction	**After the correction**	**Improvement**	**Comment**
*el hombre **amables***	*el hombre **amable***	yes	
*los hombres **amable***	*los hombres **amables***	yes	
*la casa **antiguas***	*la casa **antigua***	yes	
*la casa **antiguo***	*la casa antigua*	yes	
*las casas **antiguos***	*la casa **antigua***	yes	
*las casas **antigua***	*las casas **antiguas***	yes	
*las casas **antiguo***	*las casas **antiguas***	yes	
*las casas **antiguos***	*las casas **antiguas***	yes	
*una **viejo** mujer*	*una **vieja** mujer*	yes	
*una **viejos** mujer*	*una **vieja** mujer*	yes	
*una **viejas** mujer*	*una **vieja** mujer*	yes	

el ***grande*** *hombre*	*el* ***grande*** *hombre*	no	no disagreement error, but wrong form
el ***grandes*** *hombre*	*los* ***grandes*** *hombres*	yes	
los ***gran*** *hombres*	*los* ***grandes*** *hombres*	yes	
los ***grande*** *hombres*	*los* ***grandes*** *hombres*	yes	
el ***primera*** *día*	*el* ***primer*** *día*	yes	
el ***primero*** *día*	*el* ***primero*** *día*	no	no disagreement error, but wrong form
los ***primer*** *días*	*los* ***primeros*** *días*	yes	
los ***primera*** *días*	*los* ***primeros*** *días*	yes	
los ***primeras*** *días*	*los* ***primeros*** *días*	yes	
una cosa ***importantes***	*una cosa* ***importante***	yes	
las personas ***débil***	*las personas* ***débiles***	yes/no	correct use of the number/wrong accent
las persona ***amable*** *y* ***fuerte***	*la persona* ***amables*** *y* ***fuerte***	-yes/no	the first adjective is corrected/the second adjective is not corrected

A.2.2. Correct Sentences of the Test Suites

Before the correction	**After the correction**	**Changes**	**Comment**
el hombre ***amable***	*el hombre* ***amable***	no	
los hombres ***amables***	*los hombres* ***amables***	no	
la casa ***antigua***	*la casa* ***antigua***	no	
las casas ***antiguas***	*las casas* ***antiguas***	no	
una ***vieja*** *mujer*	*una* ***vieja*** *mujer*	no	
el ***gran*** *hombre*	*el* ***gran*** *hombre*	no	
los ***grandes*** *hombres*	*los* ***grandes*** *hombres*	no	
el ***primer*** *día*	*el* ***primer*** *día*	no	
los ***primeros*** *días*	*los* ***primeros*** *días*	no	
una cosa ***importante***	*una cosa* ***importante***	no	
las personas ***debiles***	*las personas* ***debiles***	no	
las personas ***amables*** *y* ***fuertes***	*las personas* ***amables*** *y* ***fuertes***	no	

A.3. Disagreements between Determiners and Adjectives

A.3.1. Incorrect Sentences of the Test Suites

Before the correction	**After the correction**	**Improvement**	**Comment**
unas *vieja dama*	***una*** *vieja dama*	yes	
un *vieja dama*	***una*** *vieja dama*	yes	
unos *vieja dama*	***una*** *vieja dama*	yes	
una *viejo hombre*	***un*** *viejo hombre*	yes	
unos *viejo hombre*	***un*** *viejo hombre*	yes	
unas *viejo hombre*	***un*** *viejo hombre*	yes	
la *viejas damas*	***las*** *viejas damas*	yes	
los *viejas damas*	***las*** *viejas damas*	yes	
el *viejas damas*	***las*** *viejas damas*	yes	

el *viejos hombres*	***los*** *viejos hombres*	yes	
las *viejos hombres*	***los*** *viejos hombres*	yes	
la *viejos hombres*	***los*** *viejos hombres*	yes	
la *grandes casas*	***las*** *grandes casas*	yes	
el *grandes casas*	***las*** *grandes casas*	yes	
los *grandes casas*	***las*** *grandes casas*	yes	

A.3.2. Correct Sentences of the Test Suites

Before the correction	**After the correction**	**Changes**	**Comment**
una *vieja dama*	***una*** *vieja dama*	no	
un *viejo hombre*	*un* viejo hombre	no	
las *viejas damas*	***las*** *viejas damas*	no	
los *viejos hombres*	***los*** *viejos hombres*	no	
las *grandes casas*	***las*** *grandes casas*	no	

B. Corrections Made by our Grammar Checker in the Test Sets

B.1. SUMAT Test Set (2012) Translated with the SUMAT System

B.1.1. True Positives

True positives mean: The application of our grammar checker yields an improvement of the sentence

Before the correction	**After the correction**
fue un gran empresa .	*fue una gran empresa .*
así se agotaron las película al mismo tiempo .	*así se agotaron la película al mismo tiempo .*
pero el salvadora hacía los dos espectáculos ,	*pero la salvadora hacía los dos espectáculos ,*
había dos baterías y cuántos coristas , ? una docena de ?	*había dos baterías y cuántas coristas , ? una docena de ?*
así que vino con el sello de su intención básico .	*así que vino con el sello de su intención básica .*
barry y diseñé la instrucciones .	*barry y diseñé las instrucciones .*
teníamos la guitarra con los medicamentos y el food-stuffs dentro .	*teníamos la guitarra con los medicamentos y los food-stuffs dentro .*
era un ambiente eléctrica .	*era un ambiente eléctrico .*
era los psicodélicos veces y empecé a pintar el bajo .	*era las psicodélicas veces y empecé a pintar el bajo .*
era cuestión de llamó a los amigos que sabía	*era cuestión de llamar a los amigos que sabía*
me levanté y era muy relajante porque había muchos grandes estrellas .	*me levanté y era muy relajante porque había muchas grandes estrellas .*
lo más evidente desaparecidos era eric clapton .	*lo más evidentes desaparecidos era eric clapton .*
era muy difícil para mí , sobre todo con mi circunstancias .	*era muy difícil para mí , sobre todo con mis circunstancias .*
apple marina a mi mente , necesitó mucho humildad para george en esa etapa de su vida ,	*apple marina a mi mente , necesitó mucha humildad para george en esa etapa de su vida ,*
así que necesitó mucho humildad de su parte poder llamar a la vuelta	*así que necesitó mucha humildad de su parte poder llamar a la vuelta*
así que era cuestión de trabajando en el último minuto .	*así que era cuestión de trabajar en el último minuto .*
así que , una especie de los mejores voces del rock , ¿ o sigues o callado .	*así que , una especie de las mejores voces del rock , ¿ o sigues o callado .*
estamos intentando preparar la música a este evento especial esta histórica programa ,	*estamos intentando preparar la música a este evento especial este histórico programa ,*
instrumentos india son muy sensibles con el fuerte luces ,	*instrumentos india son muy sensibles con las fuertes luces ,*
me impresionó cómo muy cálida personas .	*me impresionó cómo muy cálidas personas .*

es uno de los pocos sitios intacta exótico quedan en el mundo .	*es uno de los pocos sitios intactos exótico quedan en el mundo .*
sentado en la silla pegadizo que tener una .	*sentado en la silla pegadiza que tener una .*
con este disco hemos cubierto tantas base .	*con este disco hemos cubierto tanta base .*
y el instrumentations son tan expuesto porque son tan simple .	*y los instrumentations son tan expuesto porque son tan simple .*
no puedes esconderte detrás de percusión y estridente teclados .	*no puedes esconderte detrás de percusión y estridentes teclados .*
y vamos a ser el nuevo disco aquí por mi amigos cubana	*y vamos a ser el nuevo disco aquí por mis amigos cubana*
sí ... creo que la preparación para fue apurando la cantante gira ,	*sí ... creo que la preparación para ser apurando la cantante gira ,*
es lo que es , es un caso de pasando y repetir las piezas .	*es lo que es , es un caso de pasar y repetir las piezas .*
no puede hacer el error de preguntándose cómo el tema es	*no puede hacer el error de preguntarse cómo el tema es*
estos últimos poco jaunts en la noche han sido bueno para mí . en mis brazos	*estos últimos pocos jaunts en la noche han sido bueno para mí . en mis brazos*
un clásico frase : " no dejes que el momento escaparme " .	*una clásica frase : " no dejes que el momento escaparme " .*
veremos cómo voy cuando se trata de haciéndolo en directo	*veremos cómo voy cuando se trata de hacerlo en directo*
porque tuvimos anuncio libs y esos bv piezas que fui allí ,	*porque tuvimos anuncio libs y esas bv piezas que fui allí ,*
en mi primer o segundo sesión con biff en brighton .	*en mi primer o segunda sesión con biff en brighton .*
y tendría que adoptar se llamaba " luchador de sumo postura "	*y tendría que adoptar se llamaba " luchador de suma postura "*
y repito , no he presionado era bastante dulce comienzos	*y repito , no he presionado era bastante dulces comienzos*
a los espectador reflexionar sobre algo	*al espectador reflexionar sobre algo*
nos tenemos nuestra catástrofes	*nos tenemos nuestras catástrofes*
no sólo por su supervivencia contra aterradora probabilidades ,	*no sólo por su supervivencia contra aterradoras probabilidades ,*
y producir gloriosa resultados ,	*y producir gloriosos resultados ,*
y me dicen : " tienes estupendo manos ,	*y me dicen : " tienes estupendas manos ,*
de enviar a uno de esos instituciones , a dormir ,	*de enviar a uno de esas instituciones , a dormir ,*
puedes hacer un infrecuente situación normal de una situación muy fácilmente .	*puedes hacer una infrecuente situación normal de una situación muy fácilmente .*
ahora , yo no entraría nunca por aquí porque había demasiadas pasos , naturalmente .	*ahora , yo no entraría nunca por aquí porque había demasiados pasos , naturalmente .*
porque normalmente todos esos increíble pelotas	*porque normalmente todos esas increíbles pelotas*
escuchar las tenedorcitos siendo recogen en los camareros ,	*escuchar los tenedorcitos siendo recogen en los camareros ,*
y para convertirlo en un entorno pacíficos , cómodo sano ,	*y para convertirlo en un entorno pacífico , cómodo sano ,*
tiene que haber algún frustración	*tiene que haber alguna frustración*
el problema son son muy repugnante dietas	*el problema son son muy repugnantes dietas*
¿ quién va a tocar el partita en mi mayor de bach .	*¿ quién va a tocar la partita en mi mayor de bach .*
junto a saber de qué hablas , en términos técnico ,	*junto a saber de qué hablas , en términos técnicos ,*
creo que la trout es un gran parte de la historia .	*creo que la trout es una gran parte de la historia .*
vibraciones vibraciones musical que se iba a ...	*vibraciones vibraciones musicales que se iba a ...*
porque tocaba con una gran temperamento y dije : " esto es jackie " ,	*porque tocaba con un gran temperamento y dije : " esto es jackie " ,*
que mover el peor manera por su música .	*que mover la peor manera por su música .*
" tu zapatos son mierda " .	*" tus zapatos son mierda " .*
" ¿ estas gran amplificadores y esta gran batería .	*" ¿ estos grandes amplificadores y esta gran batería .*

que ni siquiera mirar las máquina letras .	*que ni siquiera mirar la máquina letras .*
para escribir su propia sencillos , así que le dieron a mí .	*para escribir sus propios sencillos , así que le dieron a mí .*
eran singularly las peores días de mi vida .	*eran singularly los peores días de mi vida .*
vamos a jugar algunos mayores canciones , un hombre llamado sol ser uno de ellos .	*vamos a jugar algunas mayores canciones , un hombre llamado sol ser uno de ellos .*
estás en una mentalidad creativo	*estás en una mentalidad creativa*
y no es un verdadero representación de lo que el grupo parece .	*y no es una verdadera representación de lo que el grupo parece .*
pero cuando estás en el escenario a todo volumen con tu amplificadores	*pero cuando estás en el escenario a todo volumen con tus amplificadores*
vi la piedra rosas tocar la emperatriz ballroom .	*vi la piedra rosa tocar la emperatriz ballroom .*
y hay muchos virtudes de dvd ,	*y hay muchas virtudes de dvd ,*
al espectador ha instante acceso a todos los grandes secuencias en el disco .	*al espectador ha instante acceso a todos las grandes secuencias en el disco .*
y junto con los avances técnico ,	*y junto con los avances técnicos ,*
es uno de los más duradero cosas en la sociedad .	*es uno de los más duraderas cosas en la sociedad .*
que es una película sobre cuánto música puede a la gente	*que es una película sobre cuánta música puede a la gente*
incluso en los direst de circunstancias ,	*incluso en el direst de circunstancias ,*
primero , el retrato la zumos vital es ruso .	*primero , el retrato los zumos vitales es ruso .*
vladimir ashkenazy : la zumos vital es ruso .	*vladimir ashkenazy : los zumos vitales es ruso .*
donde él estudió al famoso música diez años .	*donde él estudió a la famosa música diez años .*
y con poca descanso hasta junio de 1968 ,	*y con poco descanso hasta junio de 1968 ,*
era muy romántico circunstancias , muy romántico , debo decir .	*era muy románticas circunstancias , muy romántico , debo decir .*
y creíamos que podíamos organizábamos nuestras muebles y y cosas	*y creíamos que podíamos organizábamos nuestros muebles y y cosas*
pero nada ha llegado es vacía habitaciones .	*pero nada ha llegado es vacías habitaciones .*
vida familiar tiene algún extraño connotaciones que se sienta delante de un fuego	*vida familiar tiene algunas extrañas connotaciones que se sienta delante de un fuego*
no puedes oír mi notas altas porque eres muy alto .	*no puedes oír mis notas altas porque eres muy alto .*
el equilibrio varía porque en estos grandes espesa acordes , ya sabes , es ...	*el equilibrio varía porque en estos grandes espesos acordes , ya sabes , es ...*
y alguien vendrá : " ¿ tenemos cinco minutos de dices algo ? "	*y alguien vendrá : " ¿ tenemos cinco minutos de decir algo ? "*
" no tengo la sensación de satisfacción duradero . "	*" no tengo la sensación de satisfacción duradera . "*
no lo entienden . no tienen ningún tradición .	*no lo entienden . no tienen ninguna tradición .*
es mecánico y ejercicios matemático .	*es mecánico y ejercicios matemáticos .*
y emotivos y el mejor música en el mundo ,	*y emotivos y la mejor música en el mundo ,*
no , los cuernos . tu cuernos éramos muy buenos ,	*no , los cuernos . tus cuernos éramos muy buenos ,*
busca , irónicamente , que los rusos están vendiendo su grabaciones en el oeste .	*busca , irónicamente , que los rusos están vendiendo sus grabaciones en el oeste .*
la música es el mayor expresión de la mente humana .	*la música es la mayor expresión de la mente humana .*
¿ cuánto importancia puedes conectar a las críticas	*¿ cuánta importancia puedes conectar a las críticas*
son sus zumos ruso	*son sus zumos rusos*
especialmente porque cierto leyenda , es :	*especialmente porque cierta leyenda , es :*
e incluso el único lírica pasaje , que es ...	*e incluso el único lírico pasaje , que es ...*
y se refleja en su increíble armonías .	*y se refleja en sus increíbles armonías .*
era un gran alumna en armonía . tenía 5 + siempre .	*era una gran alumna en armonía . tenía 5 + siempre .*
intentó algunos feliz música , redactar un feliz música ,	*intentó alguna feliz música , redactar una feliz música ,*
bailes sinfónica el mismo .	*bailes sinfónicos el mismo .*

no sé si es una observación superficiales . para mí no .	*no sé si es una observación superficial . para mí no .*
su gran puntos son generosidad de expresión ,	*sus grandes puntos son generosidad de expresión ,*
era de ninguna manera el prototipo de un grupo de variaciones clásica .	*era de ninguna manera el prototipo de un grupo de variaciones clásicas .*
una metamorfosis que hacía justicia a los potencial de esas canciones	*una metamorfosis que hacía justicia al potencial de esas canciones*
hay muchos rápido variaciones . por supuesto , hay despacio .	*hay muchas rápidas variaciones . por supuesto , hay despacio .*
pero el lento variantes son muy importantes	*pero las lentas variantes son muy importantes*
todo el tiempo entre este drama y conducir del rápido variantes ,	*todo el tiempo entre este drama y conducir de las rápidas variantes ,*
es la única tono mayor episodio en las variaciones .	*es el único tono mayor episodio en las variaciones .*
si no hay un rayo de esperanza , no creo que es un esperanzas episodio ,	*si no hay un rayo de esperanza , no creo que es unas esperanzas episodio ,*
el muy reveladora sobre su más íntimas sentimientos .	*el muy reveladora sobre su más íntimos sentimientos .*
se sintió luchando para asumir el gran tradiciones del oeste ,	*se sintió luchando para asumir las grandes tradiciones del oeste ,*
a ser un reconocido director de la mayor rango ,	*a ser un reconocido director del mayor rango ,*
así que cometí un montaje de secuencias de algunos de esos compositor películas ,	*así que cometí un montaje de secuencias de algunos de ese compositor películas ,*
un poco más de los primeros violines , menos del segundos .	*un poco más de los primeros violines , menos de los segundos .*
pero se puede oír una la bemol menor en el silencio . vale .	*pero se puede oír una el bemol menor en el silencio . vale .*
es un incomprensible profundidad en la música ,	*es una incomprensible profundidad en la música ,*
aunque parece un contradicciones .	*aunque parece unas contradicciones .*
y apresurado ideas suelen producir interesante resultados	*y apresurado ideas suelen producir interesantes resultados*
estoy descubrir durante los últimos giras ,	*estoy descubrir durante las últimas giras ,*
" la mayoría solo preocúpate de golpear la nota correcta ... "	*" la mayoría sola preocúpate de golpear la nota correcta ... "*
hubo un charlas sobre : " ¿ continuamos ? "	*hubo unas charlas sobre : " ¿ continuamos ? "*
ese concierto en atenas está muy bien . tiene un gran área detrás del escenario ,	*ese concierto en atenas está muy bien . tiene una gran área detrás del escenario ,*
y llevar a era muy importante .	*y llevar a ser muy importante .*
no hay micro está excepto el overheads y el bass-drum micros .	*no hay micro está excepto los overheads y el bass-drum micros .*
para interesante temas que te gustaría hacer .	*para interesantes temas que te gustaría hacer .*
y luego cambiar o de evolucionan , cosas así	*y luego cambiar o de evolucionar , cosas así*
así que tengo mi cascos y estoy aprendiendo en el avión .	*así que tengo mis cascos y estoy aprendiendo en el avión .*
y nos quedamos en buen hoteles y todo .	*y nos quedamos en buenos hoteles y todo .*
y de su intento de suicidio , eso no habría curado cicatrices izquierdo ,	*y de su intento de suicidio , eso no habría curado cicatrices izquierdas ,*
el sueco orquesta sinfónica de radio	*la sueca orquesta sinfónica de radio*
y era encontrar consuelo en su extraño relación con esta mujer inusual .	*y era encontrar consuelo en su extraña relación con esta mujer inusual .*
la relación normalmente tomé una mayor significado	*la relación normalmente tomé un mayor significado*
empieza a ceda ante una preocupación introspectivo	*empieza a ceder ante una preocupación introspectiva*
un torrente de los más íntimamente revelar declaraciones de sus intenciones artística .	*un torrente de los más íntimamente revelar declaraciones de sus intenciones artísticas .*
en una más intenso correspondencia ,	*en una más intensa correspondencia ,*
" quizá no soy una persona íntimo contigo ,	*" quizá no soy una persona íntima contigo ,*
" ¿ sabes cuántas feliz momentos que pagar ,	*" ¿ sabes cuántas felices momentos que pagar ,*

la relación tenía rápido elegido un extraño intensidad ,	*la relación tenía rápido elegido una extraña intensidad ,*
” algunos creciente , radiante imagen humana se apresura a destella incitante lejos ” .	*” algún creciente , radiante imagen humana se apresura a destellar incitante lejos ” .*
” ¿ distante ahora suena el primer tema de la allegro obsesivo .	*” ¿ distante ahora suena el primer tema del allegro obsesivo .*
” y toda la vida es una realidad duro alternation ininterrumpido	*” y toda la vida es una realidad dura alternation ininterrumpido*
y aunque es un verdadero pieza biografía emocional ,	*y aunque es una verdadera pieza biografía emocional ,*
” entre esos pocos excepciones .	*” entre esos pocas excepciones .*
y balakirev en un espíritu de nuevo fervor religiosa ,	*y balakirev en un espíritu de nuevo fervor religioso ,*
” y eso hará que quizá sea el mejor de mi sinfónico composiciones ” .	*” y eso hará que quizá sea el mejor de mis sinfónicas composiciones ” .*
y algunas masculinos y la música	*y algunos masculinos y la música*
” y está destrozado por funesta preguntas de existencia .	*” y está destrozado por funestas preguntas de existencia .*
las compañeros constante desde su infancia ,	*los compañeros constantes desde su infancia ,*
” la obra es muy difícil y ondean un gran orquesta .	*” la obra es muy difícil y ondean una gran orquesta .*
y con esos temeroso palabras ,	*y con esas temerosas palabras ,*
estaba en el comienzo de éxito pública	*estaba en el comienzo de éxito público*
” para que el ruso es un vago persona por excelencia .	*” para que el ruso es una vaga persona por excelencia .*
si me pregunta , ¿ debo entender el omnipotencia divina	*si me pregunta , ¿ debo entender la omnipotencia divina*
” no peor que su predecessors .	*” no peor que sus predecessors .*
” tras dos representaciones de mi nuevo sinfonía en san petersburgo	*” tras dos representaciones de mi nueva sinfonía en san petersburgo*
” hay algo repulsivo , un insinceridad y justamente gracias a lo artificial .	*” hay algo repulsivo , una insinceridad y justamente gracias a lo artificial .*
” y era evidente que el ovations que recibí	*” y era evidente que los ovations que recibí*
” nuestros symphony !	*” nuestro symphony !*
con esos sentimental palabras ,	*con esas sentimental palabras ,*
con los años de sus mejores desarrollo .	*con los años de su mejor desarrollo .*
y que a pesar de sus ofertas para comenzar la subsidio halagos ,	*y que a pesar de sus ofertas para comenzar el subsidio halagos ,*
era una apasionada mujer y no eres ajena al culpa .	*era una apasionada mujer y no eres ajena a la culpa .*
tchaikovski no podía es expresar su propia herido	*tchaikovski no podía es expresar su propio herido*
y comprobado ambos su edad y un melancolía bordering en amargura .	*y comprobado ambos su edad y una melancolía bordering en amargura .*
en 1892 se puso a trabajar en un nuevo symphony , pero con poco condena .	*en 1892 se puso a trabajar en un nuevo symphony , pero con poca condena .*
” estoy encantado de haber venido a este irrevocable decisión ” .	*” estoy encantado de haber venido a esta irrevocable decisión ” .*
” habrá muchos formal innovaciones ,	*” habrá muchas formal innovaciones ,*
” el vaquero esencia de la sinfónica el plan es vida .	*” la vaquera esencia de la sinfónica el plan es vida .*
” primer movimiento , todos impulsivo pasión , confianza ,	*” primer movimiento , toda impulsiva pasión , confianza ,*

B.1.2. False Positives

False positives mean: The application of our grammar checker causes a worsening of the sentence

Before the correction	**After the correction**
nadie lo había hecho un gran música beneficio concierto ,	*nadie lo había hecho una gran música beneficio concierto ,*
los cámaras empezó todo al mismo tiempo ,	*las cámaras empezó todo al mismo tiempo ,*
que la grabación exclusivos agreements con sellos diferentes en el mismo disco .	*que la grabación exclusiva agreements con sellos diferentes en el mismo disco .*
desde la matiné o de la tarde .	*desde el matiné o de la tarde .*
la crisis deepened cuando inundaciones enorme éxito la región .	*la crisis deepened cuando inundaciones enorm éxito la región .*
es solo cuestión de intentando coger cosas que el mundo lo sabía .	*es sola cuestión de intentar coger cosas que el mundo lo sabía .*
tocamos unos jazz , tocamos algunos reggae , tocamos un alma .	*tocamos un jazz , tocamos algunos reggae , tocamos un alma .*
cómo nos diseñarla , si es una lenta construir y un gran final	*cómo nos diseñarla , si es una lenta construir y una gran final*
como una droga fue un proceso muy sencillo porque la demo terminó .	*como una droga fue un proceso muy sencillo porque el demo terminó .*
y oí la demo ,	*y oí el demo ,*
número dos , es muy agradable ayudar a la gente ayuda jóvenes ,	*número dos , es muy agradable ayudar a la gente ayuda jóvene ,*
por suerte , teníamos legítimo cámaras y director	*por suerte , teníamos legítimas cámaras y director*
que hizo su propia inglés letra	*que hizo su propio inglés letra*
toca gran acordes , que no estaba antes .	*toca grande acordes , que no estaba antes .*
sergei rachmaninov variaciones sobre un tema de corelli	*sergei rachmaninoves variaciones sobre un tema de corelli*
pero dado de rachmaninov atractivo para muchos scherzando cosas ,	*pero dado de rachmaninov atractiva para muchos scherzando cosas ,*
la mayoría solo preocúpate de golpear la nota correcta !	*la mayoría sola preocúpate de golpear la nota correcta !*
simplemente peters bastardo geriátricos en una especie de	*simplemente peters bastardo geriátrico en una especie de*
nadezhda von meck la viuda de una promotora exprés	*nadezhda von meck la viuda de un promotoro exprés*

B.1.3. Impossible to Decide

Impossible to decide means: The application of our grammar checker causes neither an improvement nor a worsening of the sentence x

Before the correction	**After the correction**
el george harrison bob dylan todas hierba en el concierto	*el george harrison bob dylan toda hierba en el concierto*
de dieron .	*de dar .*
y muchas triste que sucedía en bangladesh	*y mucha triste que sucedía en bangladesh*
¿ qué es importante para nosotros es encontrar clínica músicos .	*¿ qué es importante para nosotros es encontrar clínicos músicos .*
y si mi whoos solo ...	*y si mis whoos solo ...*
porque es otro de esos capa sobre capa tipo canciones .	*porque es otro de esa capa sobre capa tipo canciones .*
es un gran euros . me transportó a , y espero que no ofenderías ,	*es unos grandes euros . me transportó a , y espero que no ofenderías ,*
y hacer unas rápido pasa de bv .	*y hacer un rápido pasa de bv .*
por eso estoy contento de que no era una niña prodigio	*por eso estoy contento de que no era un niño prodigio*

y debo decir , no tantas graffiti en la pared .	*y debo decir , no tanto graffiti en la pared .*
y puedo tocar menos en el f un sharps .	*y puedo tocar menos en la f unos sharps .*
lo más importante de tocando conciertos	*lo más importante de tocar conciertos*
y debía ser mucha staccato allí ,	*y debía ser mucho staccato allí ,*
vamos al negocio de ofreciendo una fiesta .	*vamos al negocio de ofrecer una fiesta .*
tardé cinco días para zumbando después londres concierto .	*tardé cinco días para zumbar después londres concierto .*
sigue sale proclama generoso , dando a todo lo que tiene ,	*sigue sale proclama generosa , dando a todo lo que tiene ,*
y dave solo para los llama , lib .	*y dave solo para la llama , lib .*
ya sabes , ese tipo de va con nosotros .	*ya sabes , ese tipo de ir con nosotros .*
que refleja su verdadero tener un estrecha con una mujer	*que refleja su verdadero tener una estrecha con una mujer*
su identificación con el destino de su joven heroínas	*su identificación con el destino de sus jovenes heroínas*
" he tenido que expresar con palabras y frases pensamientos y musical imágenes musical .	*" he tenido que expresar con palabras y frases y pensamientos y musicales imágenes musicales .*

B.2. SUMAT Test Set (2012) Translated with our Final System

B.2.1. True Positives

True positives mean: The application of our grammar checker yields an improvement of the sentence

Before the correction	**After the correction**
entonces supongo que nadie llegué a abordando este asunto	*entonces supongo que nadie llegué a abordar este asunto*
y entonces vino con el sello de su intención básico .	*y entonces vino con el sello de su intención básica .*
era la psicodélica veces y empecé a pintar mi bajo .	*era las psicodélicas veces y empecé a pintar mi bajo .*
entré y fue muy relajante , porque había todos estos grandes estrellas .	*entré y fue muy relajante , porque había todos estas grandes estrellas .*
un montón de último momento logística ...	*un montón de último momento logístico ...*
eso duro drogas muy involucrado en la industria musical y el mundo del espectáculo .	*eso duras drogas muy involucrado en la industria musical y el mundo del espectáculo .*
también ayudó con los trabajo pionero con cólera ,	*también ayudó con el trabajo pionero con cólera ,*
usaría este gran habilidad tenía otra razón .	*usaría esta gran habilidad tenía otra razón .*
probablemente llegó aquí con un típico vista de lo que sería como cuba .	*probablemente llegó aquí con una típica vista de lo que sería como cuba .*
acabamos de disparar una ronda de los músicos práctica áreas .	*acabamos de disparar una ronda de los músicos prácticos áreas .*
y el instrumentations son tan expuesto porque son tan simples .	*y los instrumentations son tan expuesto porque son tan simples .*
y vamos a realizar el nuevo álbum aquí por mi cubanos amigos	*y vamos a realizar el nuevo álbum aquí por mis cubanos amigos*
es que sigues viendo edificios en vez de comercial carteles .	*es que sigues viendo edificios en vez de comerciales carteles .*
es lo que es , entonces es sólo un caso de pasando y repetir las partes .	*es lo que es , entonces es sólo un caso de pasar y repetir las partes .*
lo hice una vez , entonces volví a hacer algunas cosas diferente .	*lo hice una vez , entonces volví a hacer algunas cosas diferentes .*

porque estaba cansado de trabajar en sórdido estudios en londres	*porque estaba cansado de trabajar en sórdidos estudios en londres*
es muy divertido prenderse al golpes impares .	*es muy divertido prenderse a los golpes impares .*
y como digo , no estaba presionado , fue muy gentil comienzos	*y como digo , no estaba presionado , fue muy gentiles comienzos*
hay un chino máxima ,	*hay un chino máximo ,*
no sólo por su supervivencia contra aterrador probabilidades ,	*no sólo por su supervivencia contra aterradores probabilidades ,*
y producir resultados glorioso en el camino ,	*y producir resultados gloriosos en el camino ,*
y todos me dicen , " tienes estupendo manos ,	*y todos me dicen , " tienes estupendas manos ,*
porque normalmente había todas esas bolas increíble	*porque normalmente había todas esas bolas increíbles*
no se han quejado de los rollos duro o el pollo relleno ,	*no se han quejado de los rollos duros o el pollo relleno ,*
así que decidimos hacer el sonatas de beethoven ,	*así que decidimos hacer las sonatas de beethoven ,*
y luego eso la hizo un increíble personalidad ,	*y luego eso la hizo una increíble personalidad ,*
con una especie de imitación de muddy waters grupo llamado los chicos masculina .	*con una especie de imitación de muddy waters grupo llamado los chicos masculinos .*
dos de los peores líneas que ha escrito .	*dos de las peores líneas que ha escrito .*
recuerdo una mañana , después un particularmente dañinos noche	*recuerdo una mañana , después un particularmente dañina noche*
para que escriban su propia solteros , así que fue devuelto a mí .	*para que escriban sus propios solteros , así que fue devuelto a mí .*
desafortunadamente , nos separamos cuando estábamos en el pico con urbana himnos	*desafortunadamente , nos separamos cuando estábamos en el pico con urbanos himnos*
será puro celebración para nosotros y los fans . no puedo esperar .	*será pura celebración para nosotros y los fans . no puedo esperar .*
dos de los grandes virtudes de dvd ,	*dos de las grandes virtudes de dvd ,*
y , por los menús y los chaptering ,	*y , por los menús y el chaptering ,*
el espectador tiene acceso instantáneo a cada gran secuencia de la disco .	*el espectador tiene acceso instantáneo a cada gran secuencia del disco .*
que es una película sobre cuánto música puede decir a la gente	*que es una película sobre cuánta música puede decir a la gente*
incluso en el terrible circunstancias -	*incluso en las terribles circunstancias -*
mi familia aún vive en rusia , y tengo una especie de práctica corbatas , sí .	*mi familia aún vive en rusia , y tengo una especie de prácticas corbatas , sí .*
fue muy romántico circunstancias , muy romántico , debo decir .	*fue muy románticas circunstancias , muy romántico , debo decir .*
la vida familiar tiene connotaciones gracioso , que te sientas en frente de un fuego	*la vida familiar tiene connotaciones graciosas , que te sientas en frente de un fuego*
no puedes oír mi notas altas porque eres tan fuerte .	*no puedes oír mis notas altas porque eres tan fuerte .*
está jugando gran acordes allí que no estaba antes .	*está jugando grandes acordes allí que no estaba antes .*
la actitud a ese gran música es bastante condescendiente en rusia , creo .	*la actitud a esa gran música es bastante condescendiente en rusia , creo .*
todo es mecánico , matemática ejercicios .	*todo es mecánico , matemáticos ejercicios .*
dicho de bach es fugues que sólo estaban matemática ejercicios	*dicho de bach es fugues que sólo estaban matemáticos ejercicios*
mi más valiosa crítico también .	*mi más valioso crítico también .*
sergei rachmaninov compuso su variaciones sobre un tema de corelli .	*sergei rachmaninov compuso sus variaciones sobre un tema de corelli .*
y se arrojó de un roca y se suicidó .	*y se arrojó de una roca y se suicidó .*
y refleja incluso en su increíble armonías .	*y refleja incluso en sus increíbles armonías .*
y esto es de la tercer concierto para piano ,	*y esto es del tercer concierto para piano ,*
este es uno de los central melodías en el segundo movimiento de la tercera sinfonía	*este es uno de las central melodías en el segundo movimiento de la tercera sinfonía*

no sé si es un observaciones superficiales . no es para mí .	*no sé si es unas observaciones superficiales . no es para mí .*
si estos no son elementos esenciales para nuestra vidas , yo no sé qué es ,	*si estos no son elementos esenciales para nuestras vidas , yo no sé qué es ,*
y en que creó a su estilo idiomática .	*y en que creó a su estilo idiomático .*
hay muchas variaciones rápido . por supuesto que son lentas .	*hay muchas variaciones rápidas . por supuesto que son lentas .*
pero el lento variaciones son muy importantes	*pero las lentas variaciones son muy importantes*
y luego , con un maravilloso inventiva ,	*y luego , con una maravillosa inventiva ,*
si no hay un rayo de esperanza , porque no creo que es una buena episodio ,	*si no hay un rayo de esperanza , porque no creo que es un buen episodio ,*
nos lleva de regreso al hielo congelado del original melodía	*nos lleva de regreso al hielo congelado de la original melodía*
creo que es sólo la cálida , conmovedora episodio adjunto de alguna manera ,	*creo que es sólo la cálida , conmovedoro episodio adjunto de alguna manera ,*
entonces , muy aproximadamente , hay un conjunto de variaciones rápido	*entonces , muy aproximadamente , hay un conjunto de variaciones rápidas*
y hicimos un montaje de secuencias de algunos de esos compositor de películas ,	*y hicimos un montaje de secuencias de algunos de ese compositor de películas ,*
pero puedes oír un la bemol acorde menor en el silencio , ¿ sabes ? bien .	*pero puedes oír un el bemol acorde menor en el silencio , ¿ sabes ? bien .*
pero poco después de eso , ¿ qué es para mí es una completa tipo de alcantarilla .	*pero poco después de eso , ¿ qué es para mí es un completo tipo de alcantarilla .*
ese concierto en atenas está muy bien . tiene un gran camerinos ,	*ese concierto en atenas está muy bien . tiene unos grandes camerinos ,*
el griego gente apasionado , para poder oír el canto .	*el griega gente apasionado , para poder oír el canto .*
lo que significaba que teníamos que cancelar dos conciertos al final de la pierna europeo .	*lo que significaba que teníamos que cancelar dos conciertos al final de la pierna europea .*
cada banda diferentes . algunas bandas , cada uno ...	*cada banda diferente . algunas bandas , cada uno ...*
tiene mucho más física aspecto a su actuación	*tiene mucho más físico aspecto a su actuación*
la derecha es el principal patada .	*la derecha es la principal patada .*
por interesante pistas que te gustaría hacer .	*por interesantes pistas que te gustaría hacer .*
no creo que haya una noche donde no disfruto jugar la clásica canciones .	*no creo que haya una noche donde no disfruto jugar las clásicas canciones .*
su primer contacto con el compositor llegó en forma de un generoso comisión	*su primer contacto con el compositor llegó en forma de una generosa comisión*
en tchaikovsky es creativa desvelos .	*en tchaikovsky es creativos desvelos .*
la relación había llevadas en un peculiar intensidad ,	*la relación había llevadas en una peculiar intensidad ,*
byron sugirió es manfred como tema para otro poema sinfónica .	*byron sugirió es manfred como tema para otro poema sinfónico .*
y que tal vez sea la mejor de mis composiciones sinfónica " . .	*y que tal vez sea la mejor de mis composiciones sinfónicas "*
y el más duro y más masculino música	*y el más duro y más masculina música*
atormentado por inútil anhelos y recuerdos de su pasado culpable .	*atormentado por inútiles anhelos y recuerdos de su pasado culpable .*
la pieza es muy difícil y exige un gran orquesta .	*la pieza es muy difícil y exige una gran orquesta .*
ahora , sin embargo , hace implica cierta confianza en máximo clemencia .	*ahora , sin embargo , hace implica cierta confianza en máxima clemencia .*
y desde su artístico misión significaba tanto para él ,	*y desde su artística misión significaba tanto para él ,*
habrá muchas innovaciones formal ,	*habrá muchas innovaciones formales ,*
" primer movimiento , todo impulsiva pasión , confianza ,	*" primer movimiento , toda impulsiva pasión , confianza ,*
él se dirigió a la central problema emocional de su vida completa square ,	*él se dirigió al central problema emocional de su vida completa square ,*

B.2.2. False Positives

False positives mean: The application of our grammar checker causes a worsening of the sentence

Before the correction	**After the correction**
y los dos bateristas son sólo truenos , ringo y keltner son solo truenos .	*y los dos bateristas son sólo truenos , ringo y keltner son solos truenos .*
¿ cómo podemos diseñarla , o no es un lento y construir un gran final	*¿ cómo podemos diseñarla , o no es un lento y construir una gran final*
porque es otro de estos capa sobre capa de canciones .	*porque es otro de esta capa sobre capa de canciones .*
y debo decir , no tantos graffiti en la pared .	*y debo decir , no tanto graffiti en la pared .*
afortunadamente , teníamos legítimo cámaras y director	*afortunadamente , teníamos legítimas cámaras y director*
vi los stone roses jugar la emperatriz de baile .	*vi el stone roses jugar la emperatriz de baile .*
y creo que ahora me las arreglo para expresar esa comprensión	*y creo que ahora me el arreglo para expresar esa comprensión*

B.2.3. Impossible to Decide

Impossible to decide means: The application of our grammar checker causes neither an improvement nor a worsening of the sentence

Before the correction	**After the correction**
y para cuando la segunda vino ,	*y para cuando el segundo vino ,*
pero definitivamente muchas operístico tipo , sabes ,	*pero definitivamente mucho operístico tipo , sabes ,*
muy rápido y una vez oí una grabación de eso en muchas eco de hall ,	*muy rápido y una vez oí una grabación de eso en mucho eco de hall ,*
y también puedo jugar menos en el f sharps .	*y también puedo jugar menos en la f sharps .*
lo más importante de dando conciertos	*lo más importante de dar conciertos*
50 salvaje , punks demente saltó en esta pared	*50 salvaje , punks dementes saltó en esta pared*
y alguien vendrá , " ¿ podríamos tener cinco minutos de quiere decir algo ? "	*y alguien vendrá , " ¿ podríamos tener cinco minutos de querer decir algo ? "*
sabes , ese tipo de sigue con nosotros también .	*sabes , ese tipo de seguir con nosotros también .*
su identificación con el destino de su joven heroínas	*su identificación con el destino de sus jovenes heroínas*
" he tenido que poner en palabras y frases musical musical pensamientos e imágenes .	*" he tenido que poner en palabras y frases musicales musicales pensamientos e imágenes .*
" puedo pagarle la olvido que busca en vano .	*" puedo pagarle el olvido que busca en vano .*

B.3. VSI Test Set Translated with our Final System

B.3.1. True Positives

True positives mean: The application of our grammar checker yields an improvement of the sentence

Before the correction	**After the correction**
tener un juegos olímpicos .	*tener unos juegos olímpicos .*
no puedo evitar pensar en cómo mi olímpico sueños empezó	*no puedo evitar pensar en cómo mis olímpicos sueños empezó*
dawson derribando enorme genial trucos !	*dawson derribando enorme geniales trucos !*
monjes mantener el país es rico historia y cultura vivo	*monjes mantener el país es rica historia y cultura viva*
en parte realizando tradicional ceremonias del té	*en parte realizando tradicionales ceremonias del té*

me llenó de un raro y genuino felicidad .	*me llenó de un raro y genuina felicidad .*
porque mi papá , tiene como una gran patillas	*porque mi papá , tiene como unas grandes patillas*
más ligero distritos de compras ,	*más ligeros distritos de compras ,*
y mudarnos de una surf paraíso a otro ,	*y mudarnos de un surf paraíso a otro ,*
vivo de esto , te tengo una gran patrocinadores que me apoyan ,	*vivo de esto , te tengo unos grandes patrocinadores que me apoyan ,*
josh puede retirado de su activo ola en carrera	*josh puede retirado de su activa ola en carrera*
y consiguió su primer contratos con sponsors .	*y consiguió sus primeros contratos con sponsors .*
básicamente , era , supongo , la humillación de diciendo ,	*básicamente , era , supongo , la humillación de decir ,*
ali baba , es el más mágico ola ,	*ali baba , es el más mágica ola ,*
es el jugo que mantiene nuestro baterías corriendo .	*es el jugo que mantiene nuestros baterías corriendo .*
josh y kauli tener legendario batallas	*josh y kauli tener legendarias batallas*
a lo más increíble condiciones	*a lo más increíbles condiciones*
¿ dónde vas a ser empujado para tu absoluta límites .	*¿ dónde vas a ser empujado para tus absolutos límites .*
creo que eso es lo que está empezando a hacer el cabo de un difícil carrera .	*creo que eso es lo que está empezando a hacer el cabo de una difícil carrera .*
y tiene la oportunidad de conseguir el mejor olas .	*y tiene la oportunidad de conseguir las mejores olas .*
y puedo empujar la ascensos .	*y puedo empujar el ascensos .*
anfitriones el prólogo del 2011 absa capa épico .	*anfitriones el prólogo del 2011 absa capa épica .*
listos para la fase 1 del 2011 absa capa épico .	*listos para la fase 1 del 2011 absa capa épica .*
del 2011 absa capa épico .	*del 2011 absa capa épica .*
dominando la absa capa épico de 2011	*dominando la absa capa épica de 2011*
es uno de los más etapas en la historia de absa capa épico .	*es uno de los más etapas en la historia de absa capa épica .*
conservar energía para las largas trepa	*conservar energía para la larga trepa*
y en la última extenuante escalar el exquisito groenlandberg ,	*y en el último extenuante escalar el exquisito groenlandberg ,*
los mejores equipos internacional está liderando el campo .	*los mejores equipos internacionales está liderando el campo .*
con sólo una etapa más que ir , carrera tácticas son importantes ,	*con sólo una etapa más que ir , carrera táctica son importantes ,*
del 2011 absa capa épico .	*del 2011 absa capa épica .*
están desesperados por alcanzar la meta primero	*están desesperados por alcanzar la meta primera*
en todos los africanos categoría ,	*en todos el africana categoría ,*
para ganar el prestigioso absa capa épico .	*para ganar el prestigioso absa capa épica .*
así que es eso de la capa épico para este año .	*así que es eso de la capa épica para este año .*
es actualmente 6 minutos detrás del general líderes	*es actualmente 6 minutos detrás de los generales líderes*
y hay algo de agua fría lugares	*y hay algo de agua frío lugares*
he tenido mucha diversión olas libre sólo surf ,	*he tenido mucha diversión olas libres sólo surf ,*
william , sin embargo , destrozó las olas con sus poderosas estilo	*william , sin embargo , destrozó las olas con su poderoso estilo*
la última calor de los cuartos de final ve miguel pupo	*la último calor de los cuartos de final ve miguel pupo*
sus dos olas tenía todo barriles , poderoso se vuelve ,	*sus dos olas tenía todos barriles , poderoso se vuelve ,*
alimentamos mantarrayas aquí salvaje mantarrayas .	*alimentamos mantarrayas aquí salvajes mantarrayas .*
uno de los muchos templos de oro magnífica	*uno de los muchos templos de oro magnífico*
y yo sólo tenía una increíble año y estoy tan feliz .	*y yo sólo tenía un increíble año y estoy tan feliz .*
incluso a muchos de los mejores competidores masculino .	*incluso a muchos de los mejores competidores masculinos .*
todos hacen su trucos muy impulsado ahora	*todos hacen sus trucos muy impulsado ahora*
pero todavía tiene que lidiar con los difíciles condiciones de viento .	*pero todavía tiene que lidiar con las difíciles condiciones de viento .*
porque creo algo de bromeando ,	*porque creo algo de bromear ,*

y a intentar terminar el dobles hoy .	*y a intentar terminar los dobles hoy .*
y estamos teniendo una bonita y estable condiciones ahora .	*y estamos teniendo una bonita y estables condiciones ahora .*
muy remoto gargantas , o jugar olas .	*muy remotas gargantas , o jugar olas .*
y estaban usando estas muy vieja manualidades	*y estaban usando estas muy viejas manualidades*
y pusieron un gran placa ahí arriba	*y pusieron una gran placa ahí arriba*
había tantos abrumadora impresión .	*había tanto abrumadora impresión .*
¿ dónde impresionante témpanos flotando pintoresco glacial lagunas ,	*¿ dónde impresionantes témpanos flotando pintoresco glacial lagunas ,*
tenemos estos grandes cataratas ese flujo directamente	*tenemos estas grandes cataratas ese flujo directamente*
y luego otro catarata y luego otra buena recuperación , piscina	*y luego otra catarata y luego otra buena recuperación , piscina*
último par de días , ha sido increíble olas ,	*último par de días , ha sido increíbles olas ,*
ahí parado , poniendo perfecta barriles ,	*ahí parado , poniendo perfectos barriles ,*
te caíste , caída seco !	*te caíste , caída seca !*
para algunos de los muchachos brasileño .	*para algunos de los muchachos brasileños .*
fue un 1-way calle en la número uno	*fue un 1-way calle en el número uno*
y no había ningún barriles antes o después de eso .	*y no había ningunos barriles antes o después de eso .*
en el prestigioso destinos alrededor del mundo .	*en los prestigiosos destinos alrededor del mundo .*
su reinado sobre el steeps no puede ceder ,	*su reinado sobre los steeps no puede ceder ,*
estoy deseando otro concursos , presione más .	*estoy deseando otros concursos , presione más .*
con ducroz se oro en esquí con rápido y técnico línea .	*con ducroz se oro en esquí con rápido y técnica línea .*
también el hogar de corvatsc pico donde el famoso cara norte .	*también el hogar de corvatsc pico donde la famosa cara norte .*
mantiene la velocidad y fluidez intacto .	*mantiene la velocidad y fluidez intacta .*
se dirige a otra doble precipicio .	*se dirige a otro doble precipicio .*
y tiene un disparo en su cuarta título si puedo sacarlo de nuevo hoy .	*y tiene un disparo en su cuarto título si puedo sacarlo de nuevo hoy .*
comienza con un gran salto y una impecable aterrizaje .	*comienza con un gran salto y un impecable aterrizaje .*
nadie hace piruetas mejor que él en un freeride cara .	*nadie hace piruetas mejores que él en un freeride cara .*
ducroz se cuelga en el tercer lugar , a pesar de sus underperformance .	*ducroz se cuelga en el tercer lugar , a pesar de su underperformance .*
el mundo gira novato es el mayor juventud snowboard serie	*el mundo gira novato es la mayor juventud snowboard serie*
con el primer centro comercial deporte sudamérica novato festival :	*con el primer centro comercial deporte sudamérico novato festival :*
el gran rampas están en perfectas condiciones .	*el grandes rampas están en perfectas condiciones .*
para pagar ecológico créditos .	*para pagar ecológicos créditos .*
el gran argentinos de fútbol .	*el grandes argentinos de fútbol .*
y africano héroes liderada por el rey ,	*y africanos héroes liderada por el rey ,*
maradona es el siguiente misión era la copa mundial de 1990 .	*maradona es la siguiente misión era la copa mundial de 1990 .*
volverse famoso por su stepovers sublime y habilidad .	*volverse famoso por sus stepovers sublime y habilidad .*
alta calidad está en constante evolución , sólo un ciclistas pueden hacerlo .	*alta calidad está en constante evolución , sólo unos ciclistas pueden hacerlo .*
él demuestra ser uno de los mejores esquiadores perfecta .	*él demuestra ser uno de los mejores esquiadores perfectos .*
otra parada , el ruso aventura .	*otra parada , la rusa aventura .*
golpea una doble precipicio , máxima velocidad , se siente como el campeón responde .	*golpea un doble precipicio , máxima velocidad , se siente como el campeón responde .*
bajando este cara de silverado .	*bajando esta cara de silverado .*

después de un sólido primera mitad , choca con su gran salto . *hacia los grandes precipicios .*	*después de un sólido primero mitad , choca con su gran salto . hacia las grandes precipicios .*
muchos sluff bajando la cara ahora porque es extremadamente caro sección .	*muchos sluff bajando la cara ahora porque es extremadamente cara sección .*
es hora de presentar el podios de verbier xtreme ,	*es hora de presentar los podios de verbier xtreme ,*
de le rue restos intocable y gana su tercer título del campeonato mundial , directo	*de le rue restos intocables y gana su tercer título del campeonato mundial , directo*

B.3.2. False Positives

False positives mean: The application of our grammar checker causes a worsening of the sentence

Before the correction	**After the correction**
a 600 km de la costa oeste africana .	*a 600 km de la costa oeste africano .*
nueva zelanda compañero , marc jacobs , él pasó por su calor .	*nueva zelando compañero , marc jacobs , él pasó por su calor .*
y después de uno y medio días podríamos conducir .	*y después de uno y medios días podríamos conducir .*
entonces tienes dificultad técnica justo después de eso .	*entonces tienes dificultad técnica justa después de eso .*
al final , fue riou es astuto y verticales revés ataque	*al final , fue riou es astuto y vertical revés ataque*
es un típico montaña grande esquiador , técnicamente muy sólido .	*es una típica montaña grande esquiador , técnicamente muy sólido .*
en val senales y kronplatz , 2 de los mejores italiano nieve parks .	*en val senales y kronplatz , 2 del mejor italiano nieve parks .*
fue un humillante final para su séptimo hechizo en nápoles .	*fue una humillante final para su séptimo hechizo en nápoles .*
es conocido por es abundante nieve , ideal terreno alta calidad .	*es conocido por es abundante nieve , ideal terreno alto calidad .*

B.3.3. Impossible to decide

Impossible to decide means: The application of our grammar checker causes neither an improvement nor a worsening of the sentence

Before the correction	**After the correction**
los 36 años viejo hawaiano ya está arreglado en cabo verde	*los 36 años viejos hawaiano ya está arreglado en cabo verde*
la mágica y salvaje africano montaña carrera	*la mágica y salvaje africana montaña carrera*
para el jersey amarillo con un hueco 3 minutos a pié .	*para el jersey amarillo con un hueco 3 minutos a piar .*
equipo 36one-songo-specialized otra vez cruza la línea primero	*equipo 36one-songo-specialized otra vez cruza la línea primera*
la mágica y salvaje africano montaña carrera	*la mágica y salvaje africana montaña carrera*
medallista olímpico y múltiples campeón mundial ,	*medallista olímpico y múltipl campeón mundial ,*
por ser la peor serie de eventos la wqs .	*por ser la peor serie de eventos las wqs .*
de las diferentes canoa o estos waka alrededor ,	*de la diferente canoa o este waka alrededor ,*
hice unos pkra no tantos .	*hice un pkra no tantos .*
por ser la peor serie de eventos la wqs .	*por ser la peor serie de eventos las wqs .*
uno de la parte más difícil partes de la cara .	*uno de la parte más difíciles partes de la cara .*
creando una popular gol celebración desde entonces .	*creando un popular gol celebración desde entonces .*
aparentemente imposible pases impresionantes carreras ,	*aparentemente imposibles pases impresionantes carreras ,*

candide thovex , una leyenda viviente y múltiples ganador de los x games , quot; los finalistas son clavos excelente corre	*candide thovex , una leyenda viviente y múltipl ganador de la x games , quot; los finalistas son clavos excelentes corre*

B.4. OpenSubtitle Test Set Translated with our Final System

B.4.1. True Positives

True positives mean: The application of our grammar checker yields an improvement of the sentence

Before the correction	**After the correction**
todavía hay nuestro famoso pregunta , querida .	*todavía hay nuestra famosa pregunta , querida .*
¿ quién sería esa persona selecto ?	*¿ quién sería esa persona selecta ?*
su enorme talentos se desperdician aquí , ¿ verdad ?	*su enormes talentos se desperdician aquí , ¿ verdad ?*
¡ viejo trotaconventos .	*¡ vieja trotaconventos .*

B.4.2. False Positives

False positives mean: The application of our grammar checker causes a worsening of the sentence

Before the correction	**After the correction**
¿ quién quiere elegir una persona primero ?	*¿ quién quiere elegir una persona primera ?*

B.4.3. Impossible to Decide

Impossible to decide means: The application of our grammar checker causes neither an improvement nor a worsening of the sentence

Before the correction	**After the correction**
en la primera andrógino siendo ...	*en el primero andrógino siendo ...*

C. Corrections Made by our Grammar Checker in the DVD User Manual

C.1. True Positives

True positives mean: The application of our grammar checker yields an improvement of the sentence

Before the correction **Improvement**	**After the correction** **Comment**
este recordatorio se proporciona —- llamar la atención del instalador del sistema catv s de la sección 820-40 de nec, que - proporciona guías para el aterrizamiento adecuado y, en particular, especifica que se debe conectar el cable a tierra - ***a la sistema*** *de tierra del edificio, tan cerca - hasta el punto de entrada del cable como sea práctico.*	*este recordatorio se proporciona —- llamar la atención del instalador del sistema catv s de la sección 820-40 de nec, que - proporciona guías para el aterrizamiento adecuado y, en particular, especifica que se debe conectar el cable a tierra -* ***al sistema*** *de tierra del edificio, tan cerca - hasta el punto de entrada del cable como sea práctico.*
si no aparece el menú de configuración, - comprobar su "monitor out" ***aplicable conexiones*** *- o salidas "s-video" "composite" o.*	*si no aparece el menú de configuración, - comprobar su "monitor out"* ***aplicables conexiones*** *- o salidas "s-video" "composite" o.*
los ajustes de bajos / agudos afectan sólo a ***los canales izquierdo*** */ derecho, no el subwoofer.*	*los ajustes de bajos / agudos afectan sólo a* ***los canales izquierdos*** */ derecho, no el subwoofer.*
la entrada de audio asignada siempre será recordado - siempre que se seleccione esa entrada - a través del panel frontal input select o ***la entrada-botones de selección*** *de la htr4 remotas 's -.*	*la entrada de audio asignada siempre será recordado - siempre que se seleccione esa entrada - a través del panel frontal input select o* ***los entrada-botones de selección*** *de la htr4 remotas 's -.*

C.2. False Positives

False positives mean: The application of our grammar checker causes a worsening of the sentence

Before the correction	**After the correction**
este enchufe encajará en ***el tomacorriente*** *de una sola manera.* *si el enchufe no - para encajar, - llame a un electricista - para reemplazar* ***el tomacorriente*** *obsoleto.*	*este enchufe encajará en* ***la tomacorriente*** *de una sola manera.* *si el enchufe no - para encajar, - llame a un electricista - para reemplazar* ***la tomacorriente*** *obsoleta.*
17 sobrecarga — no sobrecargue ***los tomacorrientes*** *de pared, cables de extensión o receptáculos integrales, ya que esto puede resultar en un riesgo de incendio o descarga eléctrica.*	*17 sobrecarga — no sobrecargue* ***las tomacorrientes*** *de pared, cables de extensión o receptáculos integrales, ya que esto puede resultar en un riesgo de incendio o descarga eléctrica.*

18 de objetos y líquidos — nunca introduzca objetos de ningún tipo en este producto a través de las aberturas, ya que podrían tocar ***puntos de tensión peligrosos*** *o cortocircuitar piezas que - podría provocar un incendio o una descarga eléctrica.*	*18 de objetos y líquidos — nunca introduzca objetos de ningún tipo en este producto a través de las aberturas, ya que podrían tocar* ***puntos de tensión peligrosa*** *o cortocircuitar piezas que - podría provocar un incendio o una descarga eléctrica.*
19 daños que requieren — desenchufe este producto ***del tomacorriente*** *y acuda al personal de servicio calificado bajo las siguientes condiciones:*	*19 daños que requieren — desenchufe este producto* ***de la tomacorriente*** *y acuda al personal de servicio calificado bajo las siguientes condiciones:*
- conectar el equipo a ***un tomacorriente*** *en un circuito diferente de aquel al que está conectado el receptor —-*	*- conectar el equipo a* ***una tomacorriente*** *en un circuito diferente de aquel al que está conectado el receptor —-*
18 teléfonos: - acepta ***auriculares estéreo*** *con una clavija estéreo 1/4-inch estándar (- utilizar un adaptador adecuado para los auriculares equipados - con un enchufe más pequeño).*	*18 teléfonos: - acepta* ***auricular estéreo*** *con una clavija estéreo 1/4-inch estándar (- utilizar un adaptador adecuado para los auriculares equipados - con un enchufe más pequeño).*
por favor - compruebe que los altavoces están clasificados —- sea ***8 ohmios mínimo*** *por altavoz.*	*por favor - compruebe que los altavoces están clasificados —- sea* ***8 ohmios mínimos*** *por altavoz.*
- abra la palanca ***del terminal pinza****;*	*- abra la palanca* ***de la terminal pinza****;*
hay ***dos tamaños de disco diferentes****.*	*hay* ***dos tamaños de disco diferente****.*

C.3. Impossible to Decide

Impossible to decide means: The application of our grammar checker causes neither an improvement nor a worsening of the sentence

Before the correction	**After the correction**
el botón de volumen también se utiliza —- para aumentar / disminuir graves / ***agudos nivel****, balance izquierda / derecha, ajustar el nivel de la rdc y activar / desactivar el srs.*	*el botón de volumen también se utiliza —- para aumentar / disminuir graves /* ***agudo nivel****, balance izquierda / derecha, ajustar el nivel de la rdc y activar / desactivar el srs.*
- seleccione la entrada de video 4 - con el ***htr4 remoto*** *oa través del panel frontal "seleccionar entrada" botón.*	*- seleccione la entrada de video 4 - con el* ***htr4 remota*** *oa través del panel frontal "seleccionar entrada" botón.*
el panel frontal y "disp" ***botón alterna htr4*** *'s entre el tiempo / pistas / capítulos y subtítulos / ángulos / configuración de audio (sólo dvd).*	*el panel frontal y "disp"* ***botón alterno htr4*** *'s entre el tiempo / pistas / capítulos y subtítulos / ángulos / configuración de audio (sólo dvd).*