UNIVERSIDADE DE LISBOA

FACULDADE DE LETRAS

QUALITY IN MACHINE TRANSLATION AND HUMAN POST-EDITING:

ERROR ANNOTATION AND SPECIFICATIONS

LUCIA COMPARIN

Trabalho de projeto orientado pela Professora Doutora Sara Mendes, especialmente elaborado para a obtenção do grau de Mestre em TRADUÇÃO

2016

ABSTRACT

Machine translation (MT) has been an important field of research in the last decades and is currently playing a key role in the translation market. The variable quality of results depending on various factors makes it necessary to combine MT with post-editing, to obtain high-quality translation. Post-editing is, nonetheless, a costly and time-consuming task. In order to improve the overall performance of a translation workflow involving MT, it is crucial to evaluate the quality of results produced to identify the main errors and outline strategies to address them. In this study, we assessed the results of MT and after the first human post-edition at Unbabel, a Portuguese startup that provides translation services combining MT with post-editing performed online by a community of editors. A *corpus* of texts translated at Unbabel from English into Italian was annotated after MT and after the first post-edition step. The data collected allowed us to identify three types of errors that are frequent and critical in terms of quality, namely “word order”, “agreement”, and “tense/mood/aspect”. Hence, correcting the errors belonging to these categories would have a major impact on the quality of translation and turn the post-editing process more accurate and efficient. The errors annotated in the *corpus* were analyzed in order to identify common patterns of errors, and possible solutions to address the issues identified were outlined. The MT system used at Unbabel and the tools available determined the choice to integrate information retrieved by error analysis in the Smartcheck, the tool used at Unbabel to automatically detect errors in the target text produced by the MT system and provide relevant messages to the editors. Therefore, our study focused on the definition and integration of rules in the Smartcheck to detect the most frequent and critical errors in the texts, in order to provide informative and accurate messages to the editor to aid him/her in the post-editing process.

Keywords: machine translation, post-editing, error annotation, automatic error detection, human editor

RESUMO

A tradução automática tem vindo a assumir uma grande importância no mercado da tradução e representa atualmente uma importante área de investigação. Durante os últimos cinquenta anos, vários sistemas de tradução automática foram desenvolvidos com base em paradigmas e abordagens diferentes. Os sistemas de tradução automática podem ser divididos entre sistemas baseados em conhecimento linguístico em forma de regras e sistemas baseados em *corpora* de textos, como os estatísticos e os baseados em exemplos. Além disso, nas últimas décadas, paradigmas diferentes foram combinados para desenvolver sistemas híbridos que utilizam *corpora* de textos, como nos sistemas estatísticos ou nos baseados em exemplos, mas integram regras e princípios linguísticos, como nos sistemas baseados em conhecimento, para resolver dificuldades gramaticais ou lexicais. Os sistemas de tradução automática são cada vez mais utilizados no processo de tradução, devido ao crescente volume de textos para traduzir e aos curtos prazos estabelecidos. Apesar de haver diferentes sistemas, os resultados são variáveis no que diz respeito à qualidade, dependendo do paradigma e do grau de especialização do sistema e dos textos a traduzir num determinado domínio. Estes factos impõem a necessidade de realizar uma edição dos textos, que pode ocorrer antes da tradução (pré-edição) ou depois (pós-edição). No primeiro caso, do texto de partida são eliminadas as estruturas ou palavras que representam dificuldades para a tradução automática realizada por um sistema em particular. No segundo caso, o texto traduzido pelo sistema é controlado e corrigido por um revisor humano. Para que este tipo de processo possa ser utilizado em grande escala no mercado da tradução, é importante reduzir os custos que lhe são inerentes e agilizá-lo. Além da pré-edição ou pós-edição, em função do paradigma considerado, integrar mais informação linguística ou atualizar os recursos lexicais utilizados permite melhorar os resultados da tradução automática.

O presente trabalho tem como objeto de estudo o controlo de qualidade na área da tradução automática, mais especificamente, na fase de pós-edição. O estudo e a análise dos resultados da tradução automática e da fase de pós-edição permitem delinear estratégias para intervir em dois sentidos: por um lado, melhorar os resultados do sistema de tradução automática graças à integração de mais informação no sistema; por outro lado, apoiar o trabalho do revisor na pós-edição, destacando erros prováveis ou assinalando pontos

críticos. A avaliação dos resultados da tradução automática inclui uma fase de análise dos erros presentes no texto de chegada e uma classificação dos mesmos, de acordo com uma tipologia de categorias de erros. No estudo da fase de pós-edição, a análise dos erros mais frequentemente corrigidos pelos revisores permite identificar que tipo de informação deve ser integrada no sistema de tradução automática e que instruções podem ser úteis aos revisores. Para a realização desta análise, adotou-se um sistema de classificação a fim de categorizar os erros e, portanto, de realizar uma avaliação quantitativa da qualidade da tradução.

O presente trabalho de projeto foi realizado em colaboração com a Unbabel, uma startup portuguesa que oferece serviços de tradução quase em tempo real, combinando tradução automática com uma comunidade de revisores. O *corpus* que é utilizado para a realização do trabalho que aqui se propõe é formado por textos em língua inglesa, traduzidos para italiano através de um sistema de tradução automática, corrigidos e editados por vários revisores humanos. São analisados os erros presentes nos textos de chegada após a tradução automática e a primeira revisão. A identificação e a análise dos erros permite chegar a generalizações sob a forma de regras a ser implementadas no processo tradução e pós-edição de textos realizado pela Unbabel. Em particular, as regras destinam-se à integração numa ferramenta que identifica automaticamente os erros no texto de chegada de algumas categorias específicas, depois da tradução automática e durante o processo de pós-edição. A ferramenta assinala o erro e, em função do tipo de problema, sugere ao revisor uma correção ou dá-lhe indicações para prestar atenção a um aspeto particular da sequência assinalada, pois é provável que contenha um erro.

O presente trabalho divide-se em oito capítulos em que são abordados os temas fundamentais envolvidos na realização do trabalho. No primeiro capítulo apresenta-se o objeto de estudo, a motivação do trabalho de projeto, a abordagem metodológica adoptada e a organização do documento. No segundo capítulo apresenta-se a fundamentação teórica em que se baseou o estudo. Aborda-se brevemente a história da tradução automática, desde as suas primeiras tentativas em meados do século XX, até aos mais recentes sistemas da primeira década do século XXI. Após a apresentação da história, são descritas algumas dificuldades linguísticas e operacionais relacionadas com a tradução automática e apresenta-se uma descrição dos diferentes sistemas de tradução automática, nomeadamente os baseados em conhecimento linguístico, os baseados em *corpora* e os híbridos. No terceiro capítulo apresenta-se o processo de tradução automática utilizado na Unbabel,

fazendo-se uma breve descrição dos passos que o compõem, o sistema de tradução automática usado para a tradução dos textos do *corpus* e as ferramentas utilizadas na fase de pós-edição para a deteção de erros e para os testes de qualidade. No quarto capítulo introduz-se a tarefa da anotação de erros descrevendo-se, em primeiro lugar, a tipologia de erros adotada na análise e a ferramenta usada para a tarefa. Seguidamente, é apresentado o *corpus* de textos considerado neste estudo e são apresentados os dados recolhidos, nomeadamente o número de erros anotados nos textos depois da tradução automática e depois da primeira fase de pós-edição. Uma análise do número de erros anotados nas várias categorias de erros segue-se a apresentação dos dados e justifica a escolha de algumas categorias de erros para as quais são propostas soluções. Nos três capítulos seguintes são analizados os erros que pertencem às três categorias escolhidas, nomeadamente “word order” (ordem de palavras), no quinto capítulo, “agreement” (concordância), no sexto capítulo, e “tense/mood/aspect” (tempo/modo/aspeto), no sétimo capítulo. Em primeiro lugar, para cada categoria de erro, são abordadas as linhas gerais que caraterizam o fenómeno linguístico em inglês e italiano, e em seguida, os erros anotados são analisados e divididos em sub-categorias. Isto permite encontrar padrões de erros frequentes e generalizá-los, de maneira a poder propor soluções gerais que dêem conta de todos os erros do mesmo tipo. No último capítulo apresentam-se as conclusões e o trabalho futuro que pode ser realizado como continuação do presente estudo e aproveitando aspetos que não foi possível explorar no âmbito do trabalho de projeto aqui apresentado.

Em suma, o presente trabalho centra-se na identificação de questões problemáticas e na proposta de soluções para a melhoria da qualidade dos resultados no processo de tradução automática, na fase de pós-edição, constituindo um importante contributo não só para a formação da mestranda no âmbito dos sistemas de tradução automática e do seu funcionamento, como também para a melhoria do desempenho do sistema de trabalho específico levado a cabo na Unbabel.

Palavras chave: tradução automática, pós-edição, anotação de erros, revisor humano, deteção automática de erros

Acknowledgments

First of all, I would like to thank my supervisor, Sara Mendes. Without her enthusiasm and patience this work wouldn’t have been possible. Not only she helped me to address the linguistic aspects of this thesis in an accurate way, but also, and more important, to develop an interest in this field and in the related areas.

This study was carried out with the collaboration of Unbabel. I would like to thank João Graça and Helena Moniz for making it possible and for their continue support. A special thanks to André Martins and André Silva, for their precious help and patience in answering all my questions. My gratitude also goes to all the colleagues and friends at Unbabel who contributed in making this such a great experience.

Last but not least, a special thanks to my family for their passionate support during these years in Lisbon, everyone in a special way and from a different place. Finally, I would like to thank my friends in Italy, Portugal, and many other countries, for sharing with me this experience and making it more enjoyable.

Contents

List of Abbreviations and Acronyms

ADJ: adjective

CBMT: *corpus*-based machine translation

EBMT: example-based machine translation

MOD: modifier

MT: machine translation

N: noun

NN: noun-noun

NP: noun phrase

POS: part-of-speech

PP: prepositional phrase

PROPN: proper noun

RBMT: rule-based machine translation

SMT: statistical machine translation

SPR: specifier

V: verb

VP: verb phrase

1 INTRODUCTION

Machine translation (henceforth MT) has been an important field of research since the second half of the 20th century. The work done in the area enabled improvements in the results, and the development of different systems that are able to perform MT. At the same time, research in MT encouraged the work in related areas, such as computational linguistics and machine learning. Thanks to research in these fields and to the improvements achieved, MT has become an important part of the translation process in the current market, as it plays a key role in handling the increasing volume of translation needed and the short time available to deliver it that has characterized the translation market over the last decades. Although the use of MT in the translation market is increasing, the quality of the results is still variable and dependent on several aspects such as the paradigm of the MT system used. Additionally, the MT systems currently available are numerous and their performance is not alike in terms of quality. These two aspects, namely the variability of the results and the number of different types of MT systems, make the evaluation of the systems a necessary step not only to define how MT systems can be improved, but also to accurately characterize the different MT systems currently available in the market and their performance, on the basis of the quality of the results.

In this thesis, we have studied quality assessment in machine translation and in post-edited texts. Quality assessment is the evaluation of the performance of a MT system in terms of the quality of the translated texts. It consists of an analysis of the results of the translation process, that is to say the output text. Quality assessment allows not only to understand whether the system produces satisfactory results, but also to identify the aspects that have to or can be improved. Assessing the quality of the results produced by a MT system also helps to define the advantages and disadvantages of the approach adopted in the translation process, and whether it is the most appropriate to translate a given type of texts. Such an analysis can be performed through the qualitative identification of errors, i.e. annotation, or through quantitative methods (BLEU, METEOR). While assessing the quality of results of a MT system, it is also useful to consider the post-editing process and the steps it consists of.

Post-editing is often combined with MT to produce high-quality results. Analyzing more thoroughly the post-editing process helps not only to assess the quality of results produced by a MT system, but also to identify the steps performed in post-edition and the improvements that can be made in order to turn the post-editing step more efficient.

The present study, which focuses on quality assessment in MT, has been carried out in collaboration with Unbabel, a Portuguese startup that offers almost real-time translation services, combining MT with crowd post-edition, done online on a platform developed by the company. In this thesis we have studied the error annotation process and the quality of the results obtained after MT and after the first human post-edition, in the language pair English-Italian. This way, we were able to identify and categorize the most common errors. The analysis of the results allowed us to find error patterns and to outline solutions to address specific issues, or to elaborate rules to automatically detect the errors and provide a warning to the post-editor. By doing so, the work presented here contributed to improving the post-editing process at Unbabel and thus to obtain higher quality results in a more cost and time efficient translation process.

1.1 MOTIVATION

As previously mentioned, research in MT is motivated by its key role in the current translation market, and by the fact that the quality of machine translated texts is not consistent, due to the fact that it depends on the paradigm of the MT system used and on whether the system is domain-specific and adapted to the texts being translated. Therefore, given the current state-of-art in the field, post-editing is necessary, even if it can be costly and time-consuming. The automatization of the post-editing process is therefore, at least in part, a necessary step. The analysis of post-editing operations allows us to work in two directions: on the one hand, it helps to improve the results of MT systems by identifying systematic errors and to take action to prevent them (possibly by integrating additional information in the translation model); on the other hand, it can aid human post-editing and make it easier, by highlighting problematic sentences and potential errors, likely to be present in the text.

As we already said, post-editing plays a key role in current MT processes. For post-edition to guarantee high-quality results, it is essential that it is done in a precise and accurate way. However, it should not be time-consuming, in order for the client to continue to benefit from the speed of MT. Automatization of the post-editing process can not only make the task more cost and time efficient, but can also contribute to preventing certain errors from going by unnoticed by the editor.

Analyzing errors in MT in the language pair English-Italian can provide useful insight not only to improve the performance of the specific MT process considered in this study, but also that of other systems, as the generalization work done for this language combination can be applied to any application involving this language pair. The choice for the language pair English-Italian is based on the fact that it is one of the most important at Unbabel in terms of volume of translation. Additionally, several linguistic phenomena observed in Italian are also characteristic of other languages, in particular Romance languages. Being so, the work done in this study can therefore be applied, with the necessary adaptations, to other language pairs.

1.2 OBJECTIVES

The general objective of this study consists in contributing to improving the quality of the results in texts translated from English into Italian by the MT system used at Unbabel. By analyzing the categories of errors that are more frequent in translated texts, and the post-editing operations performed by the editors, we aim at accomplishing the following specific goals of this study: to identify which information the system needs to integrate and what kind of guidelines can be given to editors for results to be improved, both in terms of general quality and efficiency of the translation process.

As a consequence of these general and specific goals of this study, the tangible results of the work presented here consisted in defining a set of rules to improve the performance of the Smartcheck, i.e. the tool that automatically detects errors after a text is machine translated at Unbabel. This tool provides information to the editor in order to guarantee better results after post-edition.

The main objective of this study and the methodological approach adopted will allow us to provide a thorough analysis of a number of linguistic issues in MT from English into Italian. Both the data and the analysis of linguistic issues presented in this study can be used for improving the quality of results produced by other MT systems for the same language combination.

1.3 METHODOLOGICAL APPROACH OF THIS STUDY

Literature in the field was the starting point for studying MT, the challenges it poses, and the issues it has to address. Additionally, the study of the possible paradigms adopted in MT systems helped us to identify the advantages and disadvantages of the system used at Unbabel, characterize the way it works, the results that are expected, and whether it is possible to integrate linguistic information directly in the system or if other methods should be used to overcome some of its shortcomings. In addition to more general literature on MT and MT systems, the work already published regarding the evaluation of MT systems and post-editing was considered in order to define the work that could be done in our specific case.

In order to study the errors in machine translated texts, we collected a *corpus* of texts translated at Unbabel from English into Italian and post-edited by human editors. The texts were then annotated and the errors were categorized. The data were studied in order to find repeated patterns and to identify the most common errors. When it was possible, a rule to automatically detect issues in the post-editing stage was provided. When this was not possible, given to specific limitations described for each case, an outline of possible future work was provided.

1.4 ORGANIZATION

The work presented in this document is organized as follows. In the second chapter of this study we will present a theoretical overview on the area of this study, in particular we will take into account MT, its history, the challenges it entails, and the paradigms used in MT systems. In chapter three, we will provide more details on how the translation process is done at Unbabel, on the specific MT system used by the company and the tools used to provide the service. In chapter four, we will introduce the first part of the empirical study developed, present the error annotation process and the data collected. In chapter five, we will present an analysis of word order errors annotated in the *corpus* and provide possible solutions to address the issue. The same will be done in chapter six, for agreement errors, and in chapter seven, for errors involving tense, mood, or aspect of the verb. The final chapter of this thesis will present some conclusions and future work.

2 THEORETICAL OVERVIEW

2.1 INTRODUCTION

Machine translation is a process consisting of the translation from one natural language into another performed by a computer system (Dorr et al., 1989:1). In a machine translation process the input text is inserted in the system which generates an output text corresponding to the translation. MT systems do not involve human translator's intervention. This is the main difference between MT and computer-aided translation, in which the human translator has an active role in the translation process and is assisted in the task by one or more tools, such as dictionaries, translation memories, glossaries, and terminology databanks.

The results of MT are variable and depend on factors including the kind of text translated, the purpose of the communication, the domain of the source text, the system that provided the translation, and the lexicon and syntax in the source text. The variable quality of the results is the main reason why MT is often blamed of producing poor-quality translation results. In order to obtain better results, it is necessary to either pre-edit or post-edit, respectively, the input or the output text, as we will discuss in more detail in section 2.3. Despite its shortcomings, over the last decades MT started to be used more and more, in order to deal with the increasing volume of translations needed in many fields and the short time available to deliver them. This way, MT is used to aid human translation and is integrated in the "traditional" translation process: MT systems provide the human translator with a first version of the target text that has to be edited to produce a quality translation. As a result, the translation process is accelerated and the cost reduced. However, the use of MT is not generalized among professional translators due to the fact that, when the quality of the results is poor, the translator would spend more time in correcting or re-writing a sentence than in translating it from scratch. On the contrary, MT can be used, either with or without post-editing, when the aim is not producing a high-quality translation, but rather accessing the meaning of the source text. Additionally, it is possible to adapt MT systems to the user by making them domain-specific and by integrating glossaries, and thus producing better results.

In section 2 of this chapter, we will present a brief history of MT to arrive at the present state-of-the-art. In section 3, we will consider the main challenges MT presently faces, and, in section 4, we provide a general classification of MT systems and discuss their advantages and disadvantages.

2.2 A BRIEF HISTORY OF MACHINE TRANSLATION

Even if the first idea of a mechanical translation of one text from one language into another dates back to the beginning of the 17th century, the first research and attempts to develop a working machine translation system were conducted in the second half of the 20th century, after the Second World War, when state-of-the-art technology was able to help in the task (Hutchins, 1978:119). Research focused on the creation and improvement of tools that could aid translation, such as bilingual dictionaries and terminology databases. The availability of computers encouraged research in the field and generated considerable optimism regarding the possibility of achieving a complete mechanical translation. The memorandum sent in 1949 by the American scientist and mathematician Warren Weaver to several acquaintances of his contributed to drawing the attention of researchers to MT (Hutchins, 1986: 6-7). On one side, Weaver highlighted the importance of MT as a scientific activity and research field; on the other side, he mentioned the issues it involved, including handling ambiguity and multiple meanings. In the following decades, until the mid-1960s, prototypes of MT helped in raising expectations and optimism, as researchers forecast the development of commercially available MT systems within five years (Hutchins, 1978: 119). During these years the work done in the field was extensive and the translation approach that was generally adopted was direct translation (see section 2.4 for a description of the main translation approaches followed in MT systems). At the same time, the first attempts to more accurate and elaborate approaches were made. Lexical resources were improved in those years, thanks to updated and more complete bilingual dictionaries, and new glossaries. As the research continued, the complexity of linguistic problems started to become apparent. In 1964, due to the huge investment and effort in the field, the American National Science Foundation set up a committee, the Automatic Language Processing Advisory Committee (ALPAC) to analyze the efforts done and the opportunities in MT. In 1966, the ALPAC report was issued, causing a significant reduction of funding in R&D in the MT field. The committee criticized the lack of speed and accuracy, and the high cost of MT, compared to human translation (ALPAC report, 1966: 16-20). The results of MT were considered poor in terms of quality. Instead, the ALPAC report recommended the development of machine aids for translators (ALPAC report, 1966: 32-34). An aspect that contributed to the negative opinion expressed by the

ALPAC on MT was the fact that expectations in the 1950s and 1960s were too high, considered the fact that the theoretical foundation in the field at that time was not enough to allow good results in a short period of time. Despite the report and the reduction of funding, research continued, focusing on different areas related to MT, such as computational linguistics and artificial intelligence. At the same time interactive systems were being developed, to assist translators in the translation process, but without providing a MT of the text. They included, for instance, translation memories and information about terminology.

As a consequence of the ALPAC report, in the 1960s and 1970s, research in the USA continued mainly in the translation of Russian technical and scientific texts into English, with less ambitious purposes than those of the previous decade (Hutchins, 1978: 120). However, research continued and increased in other countries as it was encouraged by specific reasons. In Canada, bilingualism boosted research in MT, due to the need of translating official documents from English into French and vice-versa. The project TAUM (Traduction Automatique de l'Université de Montréal) started in 1965 at the University of Montréal, and the TAUM-METEO MT system was presented in 1977 (Slocum, 1985: 12). In Europe, the European Communities needed to provide translations in all national languages of its member states for all the documents issued by the European Commission, and for technical and economic material. Due to the volume of translation, to the short time to deliver it, and to the limited resources, MT was considered helpful in the translation process. The EUROTRA project was launched with the goal of achieving complete and satisfactory translation in the combination of all the languages in the European Communities (EUROTRA: 2-4). At that time nine languages were included in the project. The aim of the project was to create a multilingual transfer system that integrated lexical, syntactic, and semantic information. The kind of texts to be translated was not strictly defined, but included documents issued by the European Commission and Council, and working material. In the late 1980s the project ended without achieving its primary goal, but was able to boost research in the field of MT across Europe.

Following the ALPAC report and the consequent reduction of funding in MT in the USA, the first operational MT systems appeared in the 1970s. One of them was SYSTRAN, developed by Peter Toma in California, for Russian-English translation and used by the United States Air Force and by NATO in the Apollo-Soyuz space project (Hutchins, 2001: 8). Made available in 1970, SYSTRAN was purchased by the European

Communities in 1976 for English-French. The design was later modified and more languages were added to serve the internal purposes of the European Communities. According to Hutchins (1978: 130), SYSTRAN's main competitors were LOGOS, that released an English-Vietnamese MT system in 1972, and METAL, for the language pair German-English. Additionally, several special-purpose MT systems were developed in these years. A characteristic of the systems developed in the 1970s was that their objectives were less ambitious than those set before the ALPAC report. Researchers agreed that MT systems should convey the meaning of the message, even if they did not achieve high-quality results.

In the following decade more effort was put into natural language processing and artificial intelligence, as researchers believed these areas could improve MT quality. In the end of the 1980s a team at Carnegie-Mellon University developed the KANT system (Nyberg, Mitamura, Carbonell, 1997: 1), a rule-based MT system that used lexicon, grammar and semantic resources. The system used an intermediate representation, that was called "pivot", which is characteristic of the interlingua architecture, as we will see in section 2.4.

In the 1990s, new methods dominated the MT field, while rule-based approaches started to be used less. The new systems were *corpus*-based and statistic-based. These approaches do not integrate linguistic rules manually defined by researchers, but deduct rules from a *corpus*. Analysis and generation are based on statistical methods. In *corpus*-based statistic MT systems, translation units are aligned in a *corpus* of parallel texts, and then the matching probabilities are calculated. An example of such systems is Candide, that was developed by a group at the IBM center in 1989 (Hutchins, 1984-1994: 4). In the same years the example-based approach was developed. In this approach, given a database of parallel *corpora*, the system processes, extracts, and selects equivalent phrases that are previously aligned by a statistical or rule-based method. Semantic information or statistical data on lexical occurrences are used in matching the selected phrases. One of the advantages of such systems is the accuracy of the results, as the examples are extracted from human translated texts. However, a disadvantage is that the system does not recognize the input text when it is not part of the examples in the *corpus*, and, since the system does not rely on linguistic rules to perform the translation, it is not able to translate it correctly.

Since the mid 1990s, MT was heavily influenced by the Internet (Hutchins 2010: 17). Several MT softwares specialized in the translation of web content such as web pages, emails, and chat room messages. At the same time, more MT software for personal computers were made available (Hutchins, 2005: 4) and used mainly by large corporations to translate working material and documents. According to Hutchins (2010: 17), an example of such software is SYSTRAN. Online and free MT services to be used on the Internet also appeared in these decades, such as the application Babel Fish, first created by AltaVista and then sold to Yahoo (Hutchins, 2005: 4). These systems are often used to translate articles and web content that is not necessarily syntactically correct or well-written. At the same time, several electronic dictionaries were made available on the Internet. Additionally, more Internet applications have been developed to provide MT directly on a web page or in emails (Hutchins, 2005: 4). According to Hutchins (2010:17), overall, the quality of machine translated texts on the Internet is usually poor, but often sufficient to meet the needs of the user, i.e. understanding the general meaning of a text written in another language.

With regard to research, in the first decade of the 21st century, MT research focused mainly on hybrid paradigms that combined the advantages of linguistic rules and statistical methods, in order to achieve a better fluency in target texts, as we will see in section 2.4. Research in related areas, such as natural language processing, artificial intelligence, and speech recognition was often combined with research in MT and applied to many processes, such as speech translation and multilingual summarization.

The analysis and study of the development of different MT systems during the 20th and 21st century and the emergence of different approaches to MT allowed us to understand that these two aspects are strictly related to the need to overcome several issues that arise in MT and that we will discuss in the next section.

2.3 CHALLENGES IN MT

MT is a complex field that combines both linguistic and computational aspects, requiring a knowledge of both for a system to operate (Dorr et al., 1998: 4). As a consequence, there are several difficulties in MT regarding linguistic phenomena and operational aspects. Linguistic issues are particularly due to the fact that natural languages can be ambiguous and that the specific meaning of particular words can be strictly related to the context. Operational issues, on the other side, are related to the architecture of the system itself, to the way it handles information, and to its maintenance. Solving issues in both areas contributes to achieving high-quality automatic translations.

With regard to linguistic challenges, problems can arise in MT either in source language understanding or in target language generation. A full understanding of the source language and a full ability to generate outputs in the target language are not always necessary to produce acceptable results in MT (Dorr et al., 1989: 4). Linguistic problems can be reduced by translating texts from a restricted domain, or by editing the text in order to solve problematic issues in the target text and improve the results. Editing can be performed on the input text (pre-editing) or on the target text (post-editing). The former consists of using a controlled natural language or simplifying lexicon and syntax, to make sure that the system correctly analyzes the source sentences and thus has a bigger probability of producing quality translation. Post-editing consists of correcting the errors in the target text and can be performed by a human editor or by an automatic system. In addition to the editing of the source or target text, in MT systems based on linguistic information, more rules related to a given language pair can be prepared in order to solve some issues, and the specifications can be integrated in the system to improve its performance.

The main linguistic difficulty in MT is dealing with ambiguity, i.e. the fact that a constituent can have more than one meaning or function in the sentence. Since MT systems do not integrate any context information or understanding of the world, they can only rely on the information in the sentence to solve any ambiguity. Ambiguity depends on the source and the target language. There are cases in which ambiguity does not need to be solved, because it is possible to preserve it in the target text (Dorr et al., 1998: 5). Often, however, it is necessary to disambiguate and there is not enough information available in the source text to do so. Ambiguity can arise due to specific syntactic structures or certain lexical choices. Syntactic ambiguity regards the structure of the sentence and the dependency between constituents. Ambiguity may be due, for instance, to different possible dependencies of a prepositional phrase (henceforth PP) (see example 1) or to several possibilities in terms of the coordination of clauses and verbs (see example 2)[1].

[1] Please note that, throughout this document, when the example presented is related to a translation issue, it will be provided both in English and in Italian, in this order, since this is the language combination considered in this study. The incorrect translation provided by the system will be preceded by an *, while the correct translation will be provided below. When the example illustrates a language-specific phenomenon,

1. I saw the man on the hill with the telescope.
   Ho visto l’uomo sulla collina con il cannocchiale.
2. I know you disagree with me and feel sorry.
   So che non sei d’accordo con me e mi dispiace.
   So che non sei d’accordo con me e che ti dispiace.

In 1, the PP “with the telescope” may refer either to the NP “the man on the hill”, or to the verb “saw”. The ambiguity cannot be solved without more information about the context, but, as it can be maintained in Italian, it does not pose a problem to automatic systems. In 2, however, the syntactic ambiguity lies in the coordination of the clause “feel sorry”, that can be coordinated either to the main clause “I know”, or to the subordinate clause “you disagree with me”. The translation in Italian is different in the two cases. Since it is necessary to disambiguate in the translation into Italian, and an equivalent ambiguous structure in Italian does not exist, such cases can be problematic for MT systems.

Lexical ambiguity[2] is related to the meaning or different possible meanings of a word. It can include cases of homography and polysemy, i.e. cases in which a single word form has two different meanings (homography), and cases in which a single word has two or more meanings that are related (polysemy). There are sentences in which the correct translation of a word depends on the correct identification of its part-of-speech (henceforth POS). In such cases, enough syntactic information and an accurate POS tagging can provide relevant information about the function of that word in the sentence and, therefore, allow for a correct translation (see example 3). However, there are cases in which lexical ambiguity does not depend on the POS but on the context and on the use of a specific word in a language (see example 4). In such cases, only a more fine-grained lexical resource can allow for solving the issue[3].

3. Hope_{N} - speranza
   Hope_{V} – sperare

examples will be provided only in the relevant language. With regard to Italian examples, a gloss in English will be provided in square brackets.

[2] Dorr, Jordan, and Benoit in “A Survey of Current Paradigms in Machine Translation” (Dorr et al., 1998: 5) distinguish lexical selection ambiguity and semantic ambiguity. In the context of this work, we do not make this distinction and consider a single kind of ambiguity, that we designate as “lexical ambiguity”, besides syntactic ambiguity mentioned above, naturally.

[3] A reference to the tags used in this study can be found in the list of abbreviations and acronyms at the beginning of this work.

4. Row_N (line) - fila
   Row_N (street) – via
   Row_N (argument, British English) - litigio

In addition to syntactic and lexical ambiguity, there are also cases in which the ambiguity is due to the lack of information in the source text (Dorr et al, 1998:6). It occurs when it is not possible to fully understand the meaning of a sentence simply by considering the information in it, for example because the referent of a given constituent cannot be clearly identified. In such cases contextual information is needed to solve the ambiguity.

5. I am tired.
   Sono stanco.
   Sono stanca.

As we can see in example 5, the adjective "tired" can combine with both a feminine and a masculine noun, or pronoun, as in the case in 5, and it is impossible to retrieve information regarding the gender from the pronominal subject, as gender is not morphologically expressed in it. However, the information is crucial for the translation, as the correct masculine or feminine form of the adjective must be selected for in the translation into languages such as Italian.

As previously mentioned, linguistic challenges such as ambiguity can arise not only in language understanding, but also in language generation. In this case, it is believed that complete ability to generate natural language sentences is not necessary in MT systems, since the source text provides the majority of the information needed. Yet, there are cases in which the generation requires more information than that present in the source text, for example in tense selection when the target language has a richer tense system than the source text, or in gender agreement, as we saw in 5.

Target language generation difficulties can be due to the lack of information provided by the source text analysis to select the correct constituents in the target language, or to the lack to information regarding the use of particular words in the target language. One of the difficulties in target text generation is lexical selection, when a word in the source text can have two meanings (homography or polysemy) but only one is appropriate in a given context. This difficulty is due to lexical ambiguity in the source text and can be avoided by integrating more information in the lexical resource for the target language. Tense generation, as we already mentioned, is also a problem in many cases, specially when the target language has a more complex verbal system than the source text, and therefore the information about the correct verb form in that context cannot be extracted from the source text analysis.

Depending on the source and target languages, there can also be cases in which different sentence structures are used in the source and in the target text. For example, the differences in the two languages can regard different syntactic functions for the same semantic argument, which is syntactically realized either as a subject or as an indirect object (example 6), or either as an object or as an indirect object (example 7).

6. John likes swimming.
   A John piace nuotare.
7. John told Ann that story.
   John ha raccontato a Ann quella storia.

In addition to linguistic challenges, such as those we already mentioned above, there are operational challenges regarding computational aspects. According to Dorr et al. (1998: 10), among operational challenges we can include the need to improve the system and the syntactic and lexical resources in order to handle new domains and text styles, as well as the need to include more languages and to evaluate the performance of the MT system (Dorr et al., 1998: 11-12). Including more languages in the MT system requires a significant effort, since the resources needed to perform analysis and generation must be acquired. The effort depends on the kind of approach adopted in the MT system, as the resources needed to perform translation can be either linguistic rules, when the MT system is rule-based, or *corpora* of translated texts, when the MT system is statistical or based on examples. In the first case, when linguistic rules must be acquired in order to include a new language in the MT system, the set of rules needed depends on the language combination, on the translation direction, and on the level of analysis performed by the system. For every new language added, rules must be defined to analyze the source text in the new language and to generate the target language. Additionally, a set of rules is needed to perform the transfer from the source language to the target language, depending on the language combination. In the second case, when the system is statistical or based on examples, the difficulty can be related to the fact that not enough data is available for the new language pair considered and, therefore, it is not possible to create a *corpus* to perform high-quality translations.

Among operational challenges, the maintenance of the system also requires an important effort, for instance because lexical resources have to be updated. The task is complex, since the relevant lexical information must be selected to be integrated in the lexical resource and the different uses of words must be accounted for. Additionally, it is also possible that more information has to be integrated in the tools performing the syntactic analysis, if different kinds of texts are translated with the MT system. With regard to the domain, translating texts from different domains is challenging mainly because an accurate lexical resource is needed, in order for the terminology to be translated correctly. In fact, the use of specific terms and expressions is crucial in domain-specific texts, and the MT system is not able to produce good results in translation, if it does not integrate information on the distinctive characteristics of a particular domain. Not only terminology varies from domain to domain, but also a specific syntax is used in some kinds of texts: general-purpose texts, for instance, tend to use more complex syntactic constructions, while in domain-specific documents, such as legal texts, often the syntactic constructions are more repetitive and even formulaic. On the other side, restricting the domain, apart from improving the linguistic results of MT, is also convenient from the operational point of view. Restricting the domain allows not only to reduce the size of the lexicon that needs to be acquired, but also to reduce the cost and effort of maintaining such a resource. Finally, another operational challenge cited is the evaluation of MT systems. It regards the assessment of the quality of the results of the system and whether it is properly working. Therefore, it is crucial to understand if the approach used is the most appropriate. We will discuss the evaluation of MT systems more thoroughly in chapter 4.

Linguistic and operational challenges make MT a complex process and help to determine which MT system is the most appropriate to be used. Depending on the linguistic or operational issues that may occur in the translation process, a specific approach in a MT system can be the most adequate to be used, instead of a different one. The different approaches and paradigms that can be adopted in MT systems will be discussed in the next section.

2.4 CATEGORIZATION OF MT SYSTEMS

Among MT systems, one major distinction can be made between those that are knowledge-based and those that are data driven. This distinction is crucial, as it regards the paradigm of the MT system, that is to say the kind of information that enables the translation process, and the way the translation is performed. Therefore, it is also related to the quality of the results. In rule-based (or linguistic-based) MT systems the information is expressed in the form of linguistic rules, that can account for morphological, syntactic, or semantic phenomena. In data-driven MT systems, the information used to perform a translation is extracted from *corpora* (bilingual and/or monolingual) by an automatic system. We can also mention the existence of hybrid MT systems, that combine more than one MT paradigm. In the next sections, we will generally characterize these three types of MT systems describing the way they perform translation, their advantages and disadvantages. This information is useful to evaluate the performance of a MT system, and to understand how to improve it, depending on its approach.

2.4.1 Rule-Based MT Systems

In rule-based MT (RBMT) systems, translation is based on linguistic principles that account for syntactic and semantic phenomena. They include several MT paradigms, depending on the way the principles are formulated and the way translation is performed. As general characteristics of these systems, we underline, on one side, the fact that RBMT systems produce high-quality translation results, whereas, on the other side, their coverage is reduced, because accounting for a large amount of information covering a wide range of linguistic phenomena requires a lot of effort and is very costly.

In RBMT systems we can distinguish three architectures, that is to say three ways in which the translation process can be designed. The architecture of a MT system accounts for the general design of the translation process and the way it is performed by the system. The three architectures – direct, transfer, and interlingua – can be distinguished on the basis of the analysis that is involved in the translation process. These different levels of analysis are represented in the Vauquois triangle below.

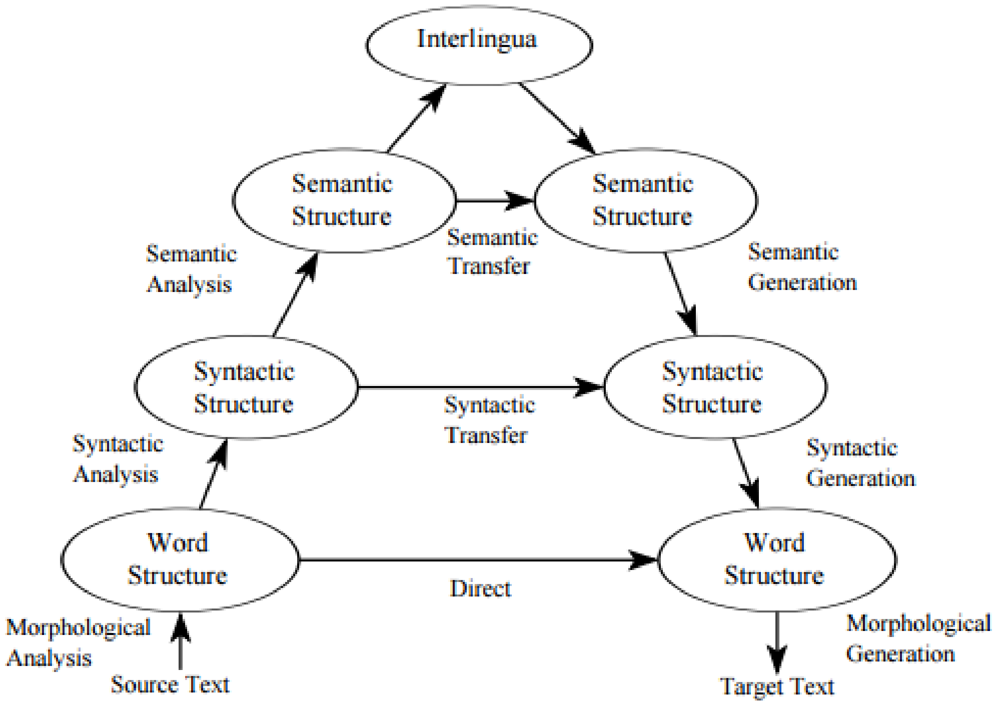


Figure 1. The Vauquois triangle

In Figure 1, we can see the levels of analysis involved in the three architectures that can be used in a MT system. Independently of the kind of architecture, the process always includes the analysis of the source language text and the generation (or synthesis) of the target language text. The analysis provided by the system is quite different in the three architectures and determines the kind of transfer that is performed in the translation process. In a direct approach, the analysis is limited to the word level, and the transfer is word-to-word. In the transfer architecture, the analysis of the source text can be syntactic and semantic, and the syntactic and lexical information retrieved by the source text analysis is transferred to syntactic and lexical information needed for the target text generation. In the interlingua architecture, the analysis of the source text is the most accurate one and provides a deep representation of the meaning of the source language text which is then used to generate the target language text. Both in the transfer and in the interlingua approach, an intermediate representation is considered necessary for the translation to take place.

The tools used in the translation task to acquire the information needed for the automatic system to work include dictionaries, that contain morphological, semantic, and syntactic information. The syntactic analysis is done by a parser, that identifies not only the syntactic structure, but also POS, phrases, and clauses (Hutchins, 1978: 122). The information provided by the parser is relevant not only in the analysis, but also in target language generation, as it helps to retrieve syntactic and lexical information that is used by the system in the translation of the constituent. The features of the constituents (e.g. “animate”, “inanimate”, “male”, “female”, etc.) included in the lexical resource are useful in semantic analysis, as well information regarding semantic relations between predicates and arguments that can be provided, for example, by a semantic parser. Semantic information is crucial in the translation of a sentence because it is often the only way to solve ambiguity cases, for example in the cases of homography and polysemy. As a consequence, semantic analysis in many cases is restricted to ambiguity resolution persisting after morphological and syntactic analysis (Hutchins, 1978: 122).

When a MT system is based on the direct architecture, the result of the translation process is a string of target-language words that maintains the same order of the source-language words. There are some systems with direct architecture that recognize some simple syntactic structures and are able to derive the order of the target language words in the translation of such structures. Only simple syntactic structures are recognized, as the analysis is not based on a complete syntactic analysis. The analysis of the source language is limited to the one that is needed to produce a target text that is acceptable and, therefore, it is determined by the target language (Hutchins, 1978: 121). According to Hutchins, for example, if a word in the source language can only be translated into one word in the target language, it is not relevant if the target language word has also other meanings. The resources used are generally limited to a bilingual dictionary, to find equivalents in the target language and, in some cases, a list of semantic and syntactic specifications, which does not allow the system to perform an accurate analysis (Hutchins, 1978:122). With regard to the quality of the results, it is often difficult or impossible for the readers to understand the target text if they do not know the source language and its characteristics. The lack of source text analysis makes it difficult to successfully handle lexical ambiguities in the majority of the cases. In the cases of lexical ambiguity in which the translation of a word depends on its POS, already mentioned in section 2.3, the system based on a direct architecture is unable to handle the ambiguity, since the POS tagging is not performed and the system cannot determine the word class and function of the word. In such cases, nevertheless, the system selects an equivalent in the target language and this could be either the word that is the most frequent, or the first meaning that is included in the lexical entry. Since the selection of the POS of the equivalent in the target language is not based on POS tagging information, the system often generates an error in the target text. Despite the lack of linguistic information to support the translation process, the results of a direct MT system can be acceptable when the text is simple, the syntax is not critical, and the domain is specific. Given all the aspects mentioned above, direct MT systems rely entirely on post-editing to produce acceptable results. The direct approach was the first adopted in the SYSTRAN MT system (Hutchins, 1978: 26-27), which was eventually combined with other approaches to improve the quality of the results achieved by the system.

As we can see from Figure 1, transfer systems range from systems using a direct architecture and those adopting the interlingua architecture. In the transfer architecture, a translation process involves three stages: the analysis of the source language, transfer, and generation of target language. The source language representations obtained through the analysis of the source text are transformed, or transferred, into the corresponding target language representations. This can be done both at the syntactic and semantic level. Three sets of rules are needed in the process: source-language analysis rules, transfer rules, and target language generation rules (Hutchins 1978: 130). These depend both on the source and on the target language. Source language analysis rules and target language generation rules are able to map the surface text and source and target representations, respectively, while transfer rules map these two types of representations with each other. Semantic analysis and rules can be added to the system in order to improve the results by having a deeper source text analysis. When MT systems incorporate this kind of information, the relevant semantic and syntactic information is combined in the representation of the source text before the transfer. The quality of the results can also be improved by a deeper analysis and more complete rules, and a rich bilingual lexical resource. With regard to ambiguity, as we saw in section 2.3, in those cases in which lexical ambiguity can be solved with POS information, transfer systems are usually able to disambiguate, because the source language analysis provides information about the POS. In those cases in which the ambiguity does not need to be solved in the target text, in some transfer systems, a set of rules is integrated to recognize the cases in which the equivalent in the target language also admits the same ambiguity. However, there are structures that are more problematic and that transfer systems are not able to solve. For instance, the system is not always able to solve ambiguity in long and complex sentences, when the syntax structure is ambiguous and difficult to identify. The disadvantage of transfer systems is that, as stated above, a large set of rules is needed. Additionally, the rules are specific to each language pair considered and direction of translation, and transfer rules have to be prepared and added when the source or the target language changes. In this case, the recognition or production part of the set of rules must be adapted to the new language pair considered (Hutchins, 1978: 130). The MT system developed under the scope of the EUROTRA project which, as previously mentioned aimed at providing multilingual translation for the languages of the European Communities member states, is an example of MT system based on the transfer architecture (Slocum, 1985: 34).

While in MT systems based on the transfer approach a translation is performed between the source language representation and the target language representation, in MT systems that adopt the interlingua architecture the analysis of the source language text results in a representation that is not language-specific. The target text is generated using this "universal" representation. In this architecture, the transfer between the source and the target text is almost absent. According to Hutchins (1978: 131), in this architecture, the translation only consists of two phases: the generation of the interlingua representation from the source text and the generation of the target text from the interlingua representation. Syntactic and semantic information is included in the interlingua representation. This architecture is based on the idea that a single concept can be derived from the meaning of a sentence, and the representation of such a concept is the same in the source and target languages, or in any language. In an interlingua MT system, in order to perform the translation, the sets of rules needed are those linking the surface text and the interlingua representation. The advantage of such an architecture is that the analysis done for one source language can be used for several language pairs and the linking rules used to generate the target language can be used in the translation from any source language. Additionally, indirect approaches (transfer and interlingua) are useful in multilingual MT systems, because only one program has to be written for source analysis and target synthesis for every language. According to Slocum (1985: 11), an example of a MT system based on the interlingua architecture is the one developed under the scope of the CETA project (Centre d'Études pour la Traduction Automatique) at Grenoble University, France, between 1961 and 1971.

Apart from the architecture of the system, RBMT systems can also be distinguished on the basis of the type of linguistic rules they incorporate. For example, we can distinguish MT systems based on rules regarding syntax or lexicon. However, and independently of all these subtypes of RBMT system, overall, the quality of the results obtained with RBMT systems is satisfactory. They are able to handle several cases of ambiguity, for example head-switching ambiguity presented in 6 and 7. The main disadvantage is that creating a system for a new language pair requires a major effort and that grammatical and lexical resources currently available are not sufficient to handle some cases of ambiguity. Additionally, the grammatical and lexical coverage of RBMT systems depends solely on the information integrated in the system and, therefore, the results are poor when the source text contains words or syntactic structures that are not covered by the system and thus which are not recognized by it.

2.4.2 *Corpus*-Based MT Systems

As we already mentioned, RBMT dominated until the 1980s, while in the end of the decade, *corpus*-based MT (CBMT) research started to be predominant, thanks to the availability of large text *corpora* and of computational systems that were able to deal with such data. In CBMT systems, the information needed in the translation process is provided by *corpus*-data. Among CBMT systems we can distinguish statistical MT (SMT) systems and example-based MT (EBMT) systems. The first results in CBMT research regarding SMT systems were published in 1988 by the IBM center, while those regarding EBMT systems were first published in 1984 by the Japanese computer scientist Nagao (Hutchins, 2015: 14). SMT research focused on the use of probability models in translation, while EBMT research preferred the use of *corpora* to train the system to obtain translation models. The use of explicit linguistic information in both paradigms is reduced. The number of rules aiding the translation process is limited, while data and translation and language models are extracted from the texts included in the *corpus*.

As mentioned, the SMT approach was first developed in 1988 at the IBM center, in the Candide project (Hutchins, 1994: 4). The research in the field was based on speech processing techniques and on a mathematical theory of probability distribution and estimation (Dorr et al. 1998: 30). A bilingual parallel *corpus* was used to acquire a translation model, while a monolingual *corpus* was used to learn a language model. This approach relies on the use of machine learning methods applied to translation. According to Hutchins (2015: 13), in the development of the MT system, the texts of the bilingual *corpus* are aligned, in order to acquire a translation model, and the monolingual *corpus* is analyzed to acquire a language model. The frequency of words in the *corpus* that is calculated in the development process, as well as the probability of their specific combination and equivalence between language pairs, are used, in the translation process, to identify the most probable equivalent of a given input word or phrase. Once the most probable translation of the words or phrases occurring in a sentence is extracted, the output is reordered based on the language model that determines the most common sequences of words in a given language, i.e. the probability of a given word being followed by another. The language model is developed using a monolingual *corpus* that contains information about word frequencies in the target language. The kind of *corpus* that is used determines the accuracy of the translation results, as these depend on the quantity, quality, domain of the data, and their closeness to the type of text being translated. The only information used in the translation process is statistically-retrieved information, extracted from the bilingual *corpus*, without using any lexicon or grammar, i.e. any external language resources. Therefore, the quality of the translation can be increased by improving the accuracy of the probabilistic models used in the translation process. In comparison to RBMT systems, SMT systems handle cases in which the input is not understood by the system, either because it is the first time it is translated, or because it is not grammatical. Thanks to the probabilistic method, the system is able to provide a translation even in such cases, while RBMT systems are unable to provide any results in such cases since no linguistic rule related to the problem is integrated in the system.

The other paradigm of MT systems that is based on a *corpus*, in addition to the paradigm adopted in SMT systems, is the one used in example-based MT systems. The first EBMT system, as we already mentioned, was developed in 1984 in Japan. According to Somers (2003: 6), the system presented by Nagao was based on the idea that translation always involves a process of finding examples that are similar to the text to be translated and that were already translated. Therefore, the fundamental steps of MT performed through an EBMT system in the system developed by Nagao were “matching fragments against a database of real examples, identifying the corresponding translation fragments, and recombining these to give the target text” (Somers, 2003: 7). In the system proposed, a bilingual *corpus* was extracted from dictionaries and pairs that set lexical equivalences. The matching between the source and the target text was done with a semantic method, through a semantic network and domain terms (Hutchins, 2015: 14). In current EBMT systems, the bilingual *corpus* consists of parallel translations. The source language sentence, or the sentence with the highest correspondence to it in the *corpus*, when the exact same sentence does not occur in it, are searched in the *corpus* and this process substitutes the source text analysis that some MT systems such as RBMT systems perform.

The matching does not rely solely on statistical probabilities like in SMT, but can adopt linguistic approaches for identifying words. In EBMT systems, after the matching is completed, the phrases in the source and target texts from the parallel *corpus* are aligned, and then modified and combined in order to obtain a translation of the sentence. In the alignment step, the system recognizes what part of the translated fragment corresponds to the matched part of the source text, and, then, combines it with other parts of different sentences extracted in the same way. As in SMT systems, accuracy and quality of the results depend on the quantity, quality, and domain of the *corpus* of parallel texts that is used to develop the system (Hutchins, 2015:14).

In the last years there has been an attempt to combine the methods used in SMT and EBMT. Therefore, SMT started to use more linguistic data and phrase-based alignment methods, while EBMT systems use more statistical techniques in the analysis of *corpora*.

2.4.3 Hybrid MT Systems

In addition to RBMT systems and CBMT systems, we can also have hybrid MT systems, that combine linguistic and non-linguistic paradigms. Linguistic information from the source text is obtained through parsing, whereas the system relies on statistical methods and example-based techniques to handle dependency issues and phrasal translation. The first hybrid MT systems were developed simply by adding a number of language tools to *corpus*-based MT systems, such as morphological analyzers, POS taggers, or syntactic analyzers. The main idea was to combine the advantages of the existing paradigms, such as the fluency and lexical selection in SMT systems, and rules able to handle syntax and long-distance dependency in RBMT systems. Hybrid MT systems often use less informative resources that are less expensive and easier to acquire, such as reduced parallel *corpora*. The hybrid MT systems developed in the last decades are based on two methods: either they combine different MT systems and try to select the best output among the results, or they select and combine fragments of the results of different MT systems, in order to produce a new target text. An example of a hybrid MT system is Pangloss-Lite (Frederking, Brown, 1996: 268), that combines an EBMT paradigm with the transfer approach and statistical language models.

In recent years, the research on combining paradigms has also focused on statistical post-editing, in order to turn the process more time and cost efficient, by automatizing it. In statistical post-editing, the output of a RBMT system is analyzed based on a statistical language model in order to detect fragments that are not fluent in the target language. This way, it is possible to actually improve the fluency of the target text.

Outlining the characteristics of different MT systems helps us to identify the differences and strengths of the several approaches and paradigms, and, therefore, the way they can be improved. We will now concentrate on the SMT system used at Unbabel in its translation process, which we will present in chapter 3, together with the way translation is performed at the company.

3 TRANSLATION AT UNBABEL

3.1 INTRODUCTION

As we already mentioned in chapter 1, this work was carried out with the collaboration of the Unbabel team. The way translation is done at Unbabel, the tools that are used, and the objectives the startup set for itself in order to improve the quality of the results were crucial aspects to take into account in defining the objective of this work. Therefore, in this chapter we will present the Unbabel workflow, the MT system used at the time the texts included in the *corpus* were translated, and the tools available at Unbabel to aid the translation process and which played a lead role in the definition of the solutions proposed in this work.

3.2 THE UNBABEL WORKFLOW

Unbabel is a Portuguese startup headquartered in the USA that provides translation services combining MT and crowdsourced translation. It offers translation services involving several language pairs and relies on a community of editors that work online on the company platform.

As all startups, Unbabel is a recently-created business providing a service or a product that was still not offered in the market, or that was offered in a different and inferior way. These characteristics lead a startup to being usually fast-growing and constantly developing new products. Being so, technical improvements are continuous and new tools are rapidly developed. This is why in this study we often mention tools that are currently not available or used at Unbabel, but that will be soon implemented. As a consequence, on one side, we were not able to test our results; but on the other side, this means that the study presented here can have a major impact on the translation process at Unbabel, and thus on the results achieved, as these are constantly being improved.

Unbabel adopts a crowd translation model, that involves multiple translators for a single translation project. The main idea behind the process used at Unbabel consists of dividing texts into small chunks, or segments, and distributing them between a number of translators. Additionally, the translation services provided by Unbabel do not rely solely on professional translators, like it is done traditionally, but combine their work with the effort of proficient speakers of a given language pair, which significantly increases the size of the community of translators available to work on a given translation project. A large community of translators as the one available at Unbabel allows the company to deliver a translation in a short time and to reduce the costs of producing it. The first advantage, i.e. increased speed, is due to the fact that crowd translation allows multiple translators to work on a text at the same time. Consequently, together, they are able to complete the task more quickly. The second advantage, i.e. reduced costs, is due to the fact that not only professional translators are involved in the process, but also bilingual individuals. This is a common criticism to crowd translation, along with the fact that it is more difficult to achieve consistency in the text when it is divided into small chunks. As we will see below, the Unbabel workflow takes the advantages of crowd translation, combining them with MT, while putting in place a combination of strategies to overcome the quality issue by doing quality checks and enabling multiple editors to review the same segment.

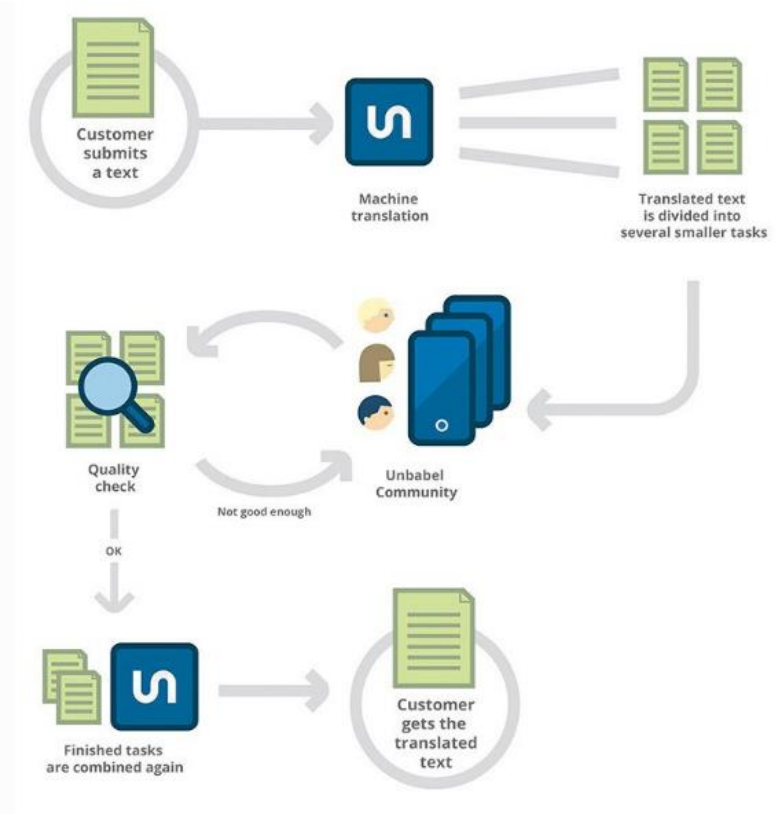

Customer submits a text
Machine translation
Translated text is divided into several smaller tasks
Quality check
Not good enough
Unbabel Community
OK
Finished tasks are combined again
Customer gets the translated text

Figure 1. Unbabel workflow

Figure 1 represents the Unbabel workflow. Before analyzing each step of the workflow independently, we have to mention that, before MT is done, the source text is pre-processed, that is to say the topic, genre, and difficulty of the text are defined, and client-specific information is added. This can include terminology, instructions on the register and style, translation memories, and the identification of proper nouns (named entities) that should not be translated. We will now analyze more thoroughly each step of the process presented in Figure 1.

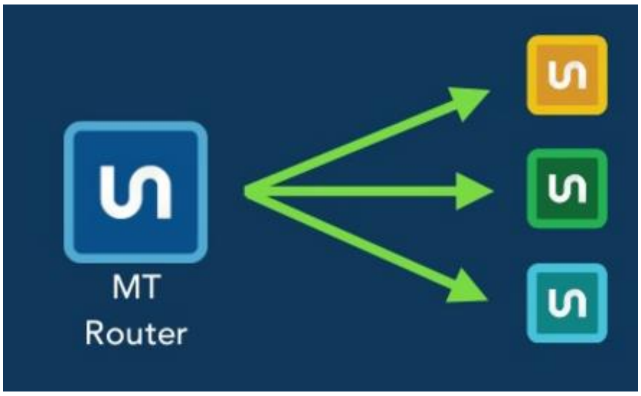


Figure 2. MT and automatic tools: step 1 of the Unbabel workflow

In step 1 (Figure 2) the text is translated by the MT system. At the time this work was done, Unbabel was using the Google SMT system. The machine translated text was then post-processed by a number of tools that detect the most common errors in the target text and correct them, or highlight them for the human editor to correct them. This step includes a spellcheck and the Smartcheck, which we will discuss in more detail in section 3.4.

Figure 3. Segmentation and post-edition: step 2 of the Unbabel workflow

Subsequently, in step 2 (Figure 3), the source and target texts are divided into small segments (paragraphs or sentences) and sent to the members of the community to be edited. The action of dividing the text into smaller chunks is called “segmentation” and, as we will see in the following chapters, it sometimes causes linguistic ambiguity, for example when, due to segmentation, information needed for the accurate translation of a constituent is lacking in the segment. After segmentation, the chunks of the translated text are made available on the platform for the editors of the target language community. On the platform, editors can access the task page, where the source and target segments are shown, together with client instructions, glossary terms, and quality warnings or suggestions made by the Smartcheck. Editors can therefore decide whether to correct the task or to skip it and receive another one. By being provided with both the source and the target text, editors are able not only to check if the target text is well-written, but also if it conveys the meaning of the original. They are also provided with tools to help them in the post-editing process, such as client instructions and glossaries. When the segment is edited, quality is automatically checked, to determine whether the segment needs to be edited once more or if it is ready to be delivered to the client. In the first case, the edited segment is made available again on the platform, for another editor to improve its quality.

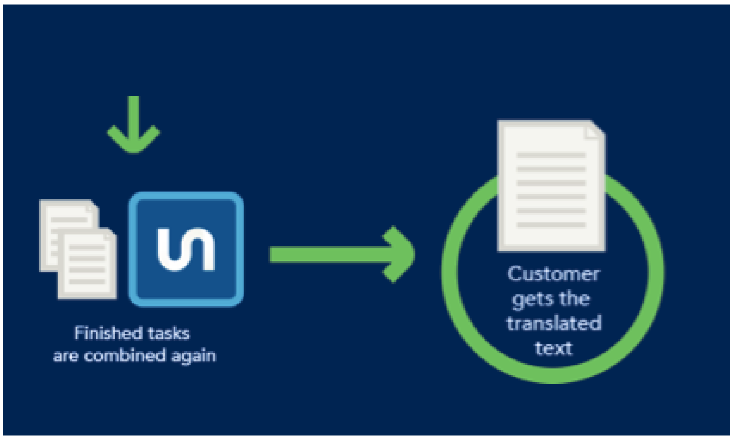


Figure 4. Delivering to the client: step 3 of the Unbabel workflow

In step 3 (Figure 4), after the text is edited, the different segments are combined again and the text is submitted to the customer. For some types of content, the complete text, i.e. the sum of segments, is sent to a senior editor before being submitted to the client. This way the entire text is reviewed once again in order to avoid inconsistencies and improve fluency. Senior editors are selected within the community based on their translation skills.

This translation process is cost and time efficient. Relying on a large community of editors that can work online enables Unbabel to deliver the translation in a short time. The fact that the post-editing is done by human editors guarantees the quality of the results.

3.3 MACHINE TRANSLATION AT UNBABEL

As we already said in the previous section, at the time the texts in the *corpus* were translated, the MT system used at Unbabel was Google Translator. Google Translator is a free SMT system available online that is currently able to translate from and into more than 70 languages. The approach is data-driven and based on web content.

The first system was launched by Google in 2001. It translated texts in five different languages into English using a third-party RBMT system. In the following two years, more languages were added and Google started to develop its own MT approach, based on the use of data. In 2005, during the machine translation evaluation organized by NIST (National Institute for Standards and Technology), Google outperformed the other systems evaluated, becoming the market leader. However, during the evaluation test, the system took 40 hours to translate 1000 sentences, which led Google to focus on speed in further improvements of the system. In 2007, the technology needed for the data-based MT system was completed, and the Google SMT system replaced entirely the third-party RBMT system used before.

The main advantage of Google Translator is that it is able to use the content available on the web. This huge amount of data allows the system to cover different domains and multiple languages. Of course, there are languages for which not enough data is available in order to guarantee high-quality results.

In September 2016, Unbabel trained a MT system for the English-Italian language pair using Moses. Moses is an open-source SMT system that enables the user to train translation models for any language pair. It is composed of a training pipeline and a decoder. The former includes the stages involved in the translation process, such as tokenization of the text (that is dividing it into smaller pieces, called tokens), alignment, acquiring a language model, and automatically selecting the best possible translations among the results of different statistical models. The decoder finds the sentence with the highest score in the target language, based on the translation model. The advantage of using Moses is that it can be customized to the needs of the user, that the text does not go through an external server, and that many additional tools can be integrated, for example to improve the analysis of the source text with POS tagging information.

The texts in the *corpus* considered in this study were translated using the Google SMT, while Unbabel is currently using the MT system trained with Moses to translate from English in to Italian.

3.4 TOOLS USED AT UNBABEL

Apart from the MT system, Unbabel uses some additional tools to analyze the data and improve the quality of the translation in different ways, from which we underline the

Smartcheck and the dependency and syntactic parser, to which we dedicate this final section of this chapter.

3.4.1 Smartcheck

The Smartcheck is a tool developed by Unbabel that checks format, grammar and style in the texts translated on the company's platform. It also includes a spellchecker. The Smartcheck analyzes the translated segments and underlines the word or the sequence of words where an error is identified, providing a suggestion to address it to the editor. The Smartcheck includes different tests in order to identify and tag different issues that may occur in the text. However, not all the checks are available for all the languages. Those available for Italian are:

- Client guidelines: checks if glossary terms in the source text are correctly and consistently translated, if there are forbidden target language words, and if the client format is respected. The corresponding error categories marked in the correction suggestions are: "client_vocabulary", and "client_format".
- Contractions: checks if there is a sequence of words that should be contracted. Error category is: "preposition_conjunction".
- Repetitions: checks if a word is repeated. Error category: "addition".
- Spellcheck: checks if there are misspelled words and if the numbers in the source text were maintained in the target text. Error category: "spelling".
- Typographical balance: checks if there are unbalanced quotes and parenthesis. Error category: "punctuation".
- Whitespace: checks if there are two or more adjacent spaces, if there is a space at the beginning of the sentence, and if there is a whitespace before punctuation. Error category: "typographical".

The Smartcheck tags the comments as warnings or errors. In the first case, the word or expression is underlined in green and the editor can submit the text on the platform without introducing any changes. When the Smartcheck detects an error, the word or expression is underlined in red and the editor has to read the message introduced by the Smartcheck and decide whether to address it or to explicitly ignore it before submitting the translation.

As we can see from the error categories listed, the Smartcheck only takes into account the target language for the checks it performs, with the exception of "Client guidelines" and “Spellcheck”. In the former, the Smartcheck verifies if a glossary term occurs in the source text, and, if so, checks if it is correctly translated in the target text. With regard to the spellcheck, the Smartcheck takes into account the source text when there is a number in the target text, to verify if it is the same that occurs in the source text.

The Smartcheck does not automatically edit the text, but only provides warnings or suggestions to the editor. The efficiency of the tool is quite important for the process at Unbabel, as well as its precision. It is important that only relevant suggestions and warnings are provided to the Editor, because it takes time for the editor to go through all the messages (specially the error messages, that require an action to be taken from the editors), and this can result in too much time being spent on the review of the task, as well as in a less accurate post-edition, if many false-positives occur. Accuracy and precision are therefore crucial for the post-editing process to be cost and time efficient. In chapter 4, after presenting the data considered in this study, we will analyze the performance of the Smartcheck in our *corpus*.

An external tool is also used at Unbabel in helping the Smartcheck to detect the errors: the Language Tool. It is an open-source program that provides a proof-reading service and detects both spelling and grammar errors. The information is provided in the form of regular expressions and is available for many languages. The Language Tool is integrated in the Smartcheck process and is used as one of the tests. However, the Smartcheck includes more information and rules than this tool, and is adapted to the company’s clients and texts.

3.4.2 Dependency and Syntactic Parser

A dependency and syntactic parser is a syntactic analyzer that provides information regarding the structure of a sentence. It identifies and classifies phrases in a sentence, such as NPs, VPs, and PPs, and identifies the syntactic relations holding between them. It also distinguishes between main and dependent clauses. A parser is therefore an important tool in the process of automatically establishing the correct syntactic dependency between constituents occurring in a sentence. It is useful in understanding which constituents modify other constituents and in solving syntactic ambiguity. The morphological information provided by the parser can also be useful in addressing lexical ambiguity, when the meaning of a constituent varies depending on the POS.

The parser used at Unbabel was developed by Martins, Almeida, and Smith in 2013 (Martins, A., Almeida, M., Smith, N.: 2013). However, the parser analysis was not yet fed to the MT system at the time the texts in the *corpus* were translated. Currently, the parser is used to analyze data in order to integrate more information in the Smartcheck. In the future, it will be used in the quality checking process and will be integrated in the Smartcheck.

When a sentence is analyzed by the parser, it provides information on the base form of the word, the POS, the value for the specific features of the word (for example, number, gender, person, mood, tense, verb form), and a dependency tree representing the syntactic structure of the sentence. More details on the parser and its analysis will be given in the analysis of the errors presented in chapters 5,6, and 7.

Presenting a theoretical overview, as well as the Unbabel workflow and the tools used at the company allows us to evaluate more accurately the translation process, through the annotation of the errors in the target text after MT and the first human post-edition. In the next chapter we will discuss error annotation and present the data collected.

4 ERROR ANNOTATION

4.1 INTRODUCTION

Error annotation is the identification, categorization and analysis of errors in a text. It can be used for different purposes, such as second language learning, quality assessment, or evaluation of a company's translation service. Depending on the project, on the source and target languages and on the translator (human or machine), a metric can be established. The annotation can be automatic (based on metrics such as BLEU or METEOR) or human. Human annotation, on one side, is expensive and time consuming. It is also more difficult to achieve consistency and objectiveness when the annotator is human. On the other side, it is more accurate and can provide a more thorough analysis of the errors. Annotation can be performed by one annotator or by multiple annotators and, in this case, the agreement among the human annotators can be calculated in order to provide additional information on the reliability of the results. Automatic annotation is preferred when a big volume of texts must be annotated and when the system's performance has to be tested regularly. The metric used in automatic annotation measures the translation closeness to a reference human translation, or a group of human translations. In this study, we performed human annotation as it is done at Unbabel and, due to the number and category of errors present in the text as well as to the goals of our study, annotation by multiple annotators was not considered necessary. However, extending the annotation task developed to multiple annotators can be part of a future work that aims, for example, to improve the taxonomy and the assessment of the severity of the errors.

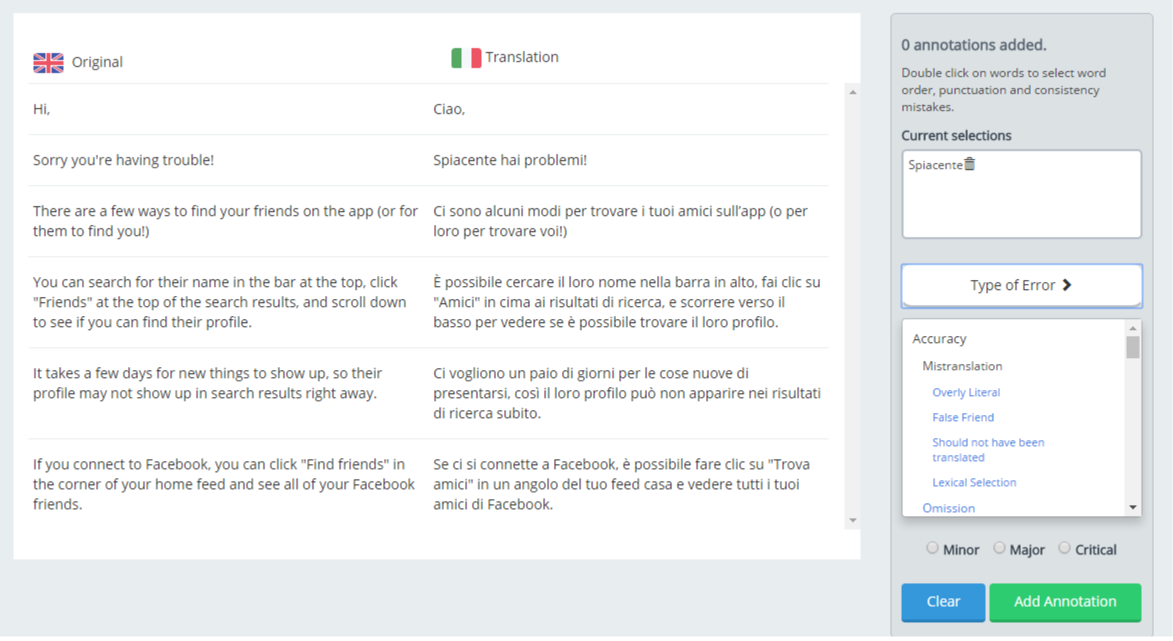
In the section 2 of this chapter we will introduce the error typology used at Unbabel and the *corpus* annotated, in section 3 we will analyze more thoroughly the error types, the penalty system and the guidelines used to annotate. In section 4 we will present the annotation data and conclude this chapter.

4.2 DEVELOPING AN ERROR TYPOLOGY AT UNBABEL

In order to develop an error typology to annotate texts at Unbabel, the documents and guidelines used as a basis were the MQM framework (Lommel, 2015) and TAUS documents (www.taus.net). The former is a model developed in the Quality Translation 21 project, as we will explain in the next paragraph, while the latter is a resource center that provides support to translation service providers through different tools, such as software, metrics, and knowledge.

Quality Translation 21 (QT21) is a machine translation project funded by the European Union’s Horizon 2020 research and innovation program. It is included in the European Single Digital Market, that was stated as a EU goal 2020, and aims to overcome barriers, in particular language barriers, in order to encourage the flow of ideas, commerce and people within the EU. The project’s goal is to improve statistical and machine-learning based translation models, enhance evaluation and learning from mistakes, by systematically analyzing quality barriers informed by human translators. The QT21 project developed a framework, the Multidimensional Quality Metrics (MQM), to define a metric and a quality score that can be used to assess MT quality. Different aspects of the quality of a translation are assessed and categorized in this framework. “The MQM framework does not provide a translation quality metric, but rather a framework for defining task-specific translation metrics” (Lommel: 2015). QT21 also cooperates with TAUS in developing annotation tools and metrics that translation providers can use to assess their position in the industry and the quality of their service.

The analysis of MQM framework and current state-of-art in annotation helped to understand what the right annotation tool and *corpus* for achieving the goals of this work could be. The *corpus* was annotated using a tool created by Unbabel (see Figure 1 and 2), and used to assess the quality of the texts delivered to clients in different language pairs. This annotation tool shows the source text, the target text, the annotations of the Smartcheck (see section 3.4.1), and the glossary terms. Errors can be annotated in the target text and then classified according to the taxonomy that is shown in the tool once a word or sequence of words containing the error is selected in the target text. Additionally, there is a bar to assess the fluency of the text, using a scale of 0 to 5. The minimum number of words of each text was set to 100, in order to ensure that the texts are long enough for the context to be understood, and the maximum was set to 700, so that the human annotation does not require too much time and effort. Please note that the annotation tool described was developed to annotate texts at Unbabel with a different purpose from that of this study, therefore there will be some specific aspects related to this study that we will present in section 4.4.



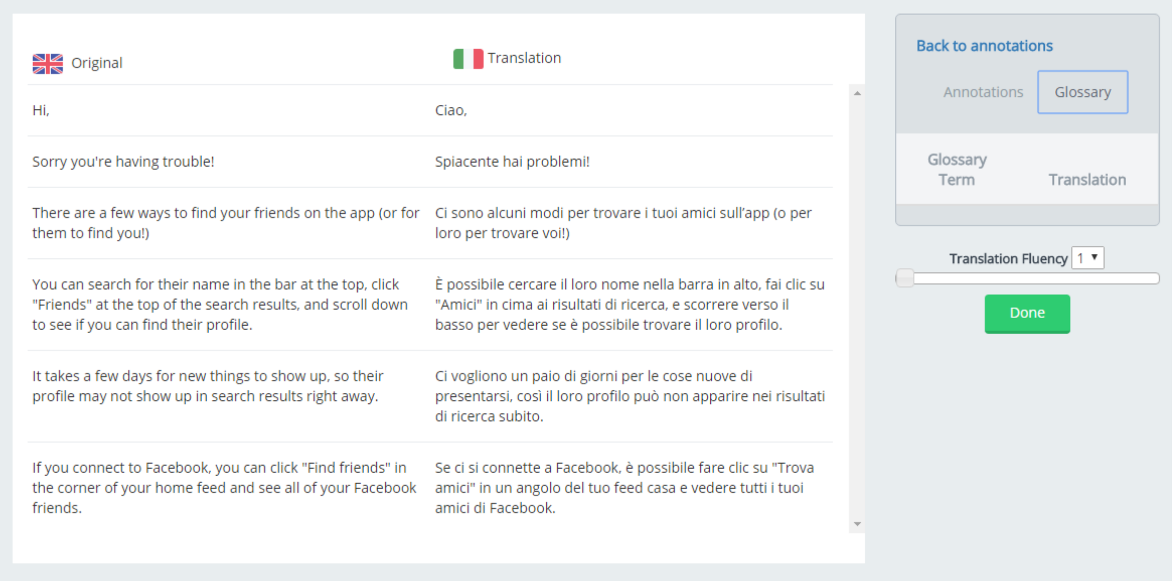
Original	Translation
Hi,	Ciao,
Sorry you're having trouble!	Spiacente hai problemi!
There are a few ways to find your friends on the app (or for them to find you!)	Ci sono alcuni modi per trovare i tuoi amici sull'app (o per loro per trovare voi!)
You can search for their name in the bar at the top, click "Friends" at the top of the search results, and scroll down to see if you can find their profile.	È possibile cercare il loro nome nella barra in alto, fai clic su "Amici" in cima ai risultati di ricerca, e scorrere verso il basso per vedere se è possibile trovare il loro profilo.
It takes a few days for new things to show up, so their profile may not show up in search results right away.	Ci vogliono un paio di giorni per le cose nuove di presentarsi, così il loro profilo può non apparire nei risultati di ricerca subito.
If you connect to Facebook, you can click "Find friends" in the corner of your home feed and see all of your Facebook friends.	Se ci si connette a Facebook, è possibile fare clic su "Trova amici" in un angolo del tuo feed casa e vedere tutti i tuoi amici di Facebook.

0 annotations added.

Double click on words to select word order, punctuation and consistency mistakes.

Current selections: Spiacente

Figure 1. Unbabel annotation tool: types of error and severity



Original	Translation
Hi,	Ciao,
Sorry you're having trouble!	Spiacente hai problemi!
There are a few ways to find your friends on the app (or for them to find you!)	Ci sono alcuni modi per trovare i tuoi amici sull'app (o per loro per trovare voi!)
You can search for their name in the bar at the top, click "Friends" at the top of the search results, and scroll down to see if you can find their profile.	È possibile cercare il loro nome nella barra in alto, fai clic su "Amici" in cima ai risultati di ricerca, e scorrere verso il basso per vedere se è possibile trovare il loro profilo.
It takes a few days for new things to show up, so their profile may not show up in search results right away.	Ci vogliono un paio di giorni per le cose nuove di presentarsi, così il loro profilo può non apparire nei risultati di ricerca subito.
If you connect to Facebook, you can click "Find friends" in the corner of your home feed and see all of your Facebook friends.	Se ci si connette a Facebook, è possibile fare clic su "Trova amici" in un angolo del tuo feed casa e vedere tutti i tuoi amici di Facebook.

Translation Fluency: 1

Figure 2. Unbabel annotation tool: glossary and translation fluency bar

While designing an error taxonomy at Unbabel, some prerequisites were taken into account. First, it should address all the issues relevant in a MT task, but it should contain a limited number of categories, in order to avoid "noise" in data annotation and to make the annotation process affordable both in terms of time dedicated to the task and in terms of its cost. Secondly, the complexity of the categories should be limited, to ensure that all annotators understand them in the same way and are able to clearly distinguish between them. Finally, the standards and the work already done in the annotation field were taken into account to define the useful categories at Unbabel and the severity system used. The amount of data already available at Unbabel and the previous error typology used in annotation within the company were also considered in the creation of a new taxonomy. Not only languages, more precisely some specific phenomena observed in particular languages, were relevant for defining the categories to include, but also the kind of texts that are usually translated at Unbabel. With regard to the languages, some categories were added specifically to account for errors in particular languages, e.g. language variety for Spanish, Portuguese and English. More categories were added in those cases in which in-house annotators had already experienced difficulty in classifying errors with the previous error typology. For example, subcategories were added in the category "spelling" and the category "awkward syntax or style" became just "awkward style", while syntax errors were included in other categories such as "coherence", "POS" and "tense/mood/aspect". In the following section we present the error taxonomy used in this study.

4.3 ERROR TAXONOMY

The error types included in the new error taxonomy defined at Unbabel and used in this work are divided into the following categories: accuracy, fluency, style, terminology, language variety, named entities, formatting and encoding. Please note that only the leaf entries of the taxonomy tree can be selected and marked as errors in the annotation tool.[4]

ACCURACY: errors in this category concern the relationship between the source text and the target text and the extent to which the latter maintains the meaning and the information of the former;

Mistranslation: wrong translation in the target language:

- **Overly literal**: direct translation of idioms, sentences and structures;

- **False friend**: mistranslation of a false friend (word or expression that has a similar form in the source and target languages, but a different meaning);

- **Should not have been translated**: content is translated instead of being left in the source language;

[4] Please note that, in the presentation of the typology, we underlined the error categories that can be selected by the annotator, as only leaf categories can be selected.

- **Lexical selection**: wrong word was used in the target language;

Omission: content is omitted in the target text;

Untranslated: content is not translated into the target language;

Addition: content is added in the target text.

FLUENCY: errors in this category regard the quality of a text, assessing whether it is well-written and easy to read, and if it accomplishes its communication purpose in the target language;

Inconsistency: a different translation of the same content is provided within the same text:

- **Word selection**: two or more different translations are provided within the text for the same lexical expression in the source text;

- **Tense selection**: the same verb tense in the source language is translated by two different tenses in the target language;

Coherence: the text is not semantically clear, logic and consistent, and, therefore, it cannot be understood by the reader;

Duplication: content is duplicated in the target text;

Spelling: misspelled word:

- **Orthography**: wrong orthography;

- **Capitalization**: wrong capitalization;

- **Diacritics**: wrong use or lack of diacritics;

Typography: wrong presentation and appearance of the target text:

- **Punctuation**: wrong use or lack of punctuation;

- **Unpaired quote marks and brackets:** the quote marks or brackets are opened but not closed, or vice-versa;

- **Whitespace**: addition or lack of a whitespace. The category was especially added for Chinese due to the high number of whitespaces that are caused by the segmentation of the sentences in MT and that should not be present in the target text, since Chinese does not have whitespaces.

- Inconsistency in character use: specially added for Chinese, to mark the inconsistency in the use of traditional and simplified characters;

Grammar: issues concerning grammar, syntax and morphology:

Function words: mistranslation, addition or omission of words that basically have a syntactic function (prepositions, conjunctions, determiners, etc.);

- **Prepositions:** mistranslation, addition or omission of a preposition;

- Conjunctions: mistranslation, addition or omission of a conjunction;

- Determiners: mistranslation, addition or omission of a determiner;

Word form: word used in a wrong form:

- Part-of-speech: wrong category (POS) of the word used (noun, verb, adjective, pronoun, conjunction, preposition, determiner);

- Agreement: lack of consistency in the number, gender, case and/or person of two or more syntactically related words;

- Tense/mood/aspect: wrong selection of tense, mood or aspect of a verb form;

Word order: wrong word order in the target language;

Sentence structure: wrong sentence structure in the target language.

STYLE: issues concerning register and fluency;

Register: an informal register was used instead of a formal one or vice-versa;

Inconsistent register: both formal and informal registers were used in the same text;

Repetitive style: presence of repetitions;

Awkward style: un-naturalness of a sentence in the target language. However, the issue is not related to literality or coherence.

TERMINOLOGY: mistranslation of terminology;

Noncompliance with client or company style guide: the style guide is not respected;

Noncompliance with the glossary and vocabulary: the information encoded in the glossary is not respected.

WRONG LANGUAGE VARIETY: use of a word or expression from a different language variety. This category was specially added for Portuguese, to distinguish European Portuguese from Brazilian Portuguese, Spanish, to distinguish European Spanish from Latin America Spanish, and English, to distinguish British English from American English.

NAMED ENTITIES: wrong translation of proper nouns;

Person: mistranslation of a person’s name;

Organization: mistranslation of an organization’s name;

Location: mistranslation of a geographical name;

Function: mistranslation of a person’s position or job;

Product: mistranslation of the name of a product;

Amount: wrong unit of measure used in the target language;

Time: wrong time format in the target language.

FORMATTING AND ENCODING: issues concerning the segmentation of sentences and paragraphs.

As we can see above, this error typology consists of 41 error categories that are included in 7 major categories (“accuracy”, “fluency”, “style”, “terminology”, “wrong language variety”, “named entities”, and “formatting and encoding”.)

4.1.1 Penalty System

Attributing a penalty to each error enables the annotation tool to calculate a numerical quality score for each translation that can be used as an indicator of its quality and of the improvements still to be made. Additionally, it is used in the industry to position the company in the market. At Unbabel, a penalty system was set up based on the system used at Google LQE (Localization Quality Evaluation) and in the MQM. Therefore, the errors annotated are divided according to their severity into minor, major and critical errors.

- Minor error: error that does not compromise the meaning of the source text, does not generate confusion nor misunderstanding. This type of error does not prevent the reader from understanding the target text but it can affect fluency or clarity. These can be punctuation mistakes, when they do not change the meaning of the sentence, spelling mistakes, repetitions, or missing hyphens. The penalty associated to this type of error is 0.5 points.
- Major error: error that generates confusion or makes the comprehension of the text more difficult. This type of error can slightly change the meaning of the target text with regard to the original. These can be errors involving lexical selection, agreement, sentence structure, tense selection, preposition selection, noncompliance with glossary, etc. The penalty associated to this type of error is 1 point.
- Critical error: prevents the reader from understanding the target text or changes the meaning of the source text. This type of error may damage a company’s reputation, and may carry health, safety or legal implications. It negatively impacts on the functionality of a product or service. The penalty associated to this type of error is 3 points.

4.3.1 Special Remarks on the Annotation Performed at Unbabel

There are a few aspects that make any annotation process considerably challenging. In the specific case of the annotation performed at Unbabel, first of all, there are cases in which an error can be categorized in different ways and the selection of the error type depends on the annotator's decision. For example, when a preposition is omitted, the error can be either marked as "omission" or as "preposition". Another difficulty consists in deciding which error should be marked when more than one occur in one word. Due to definitions implemented in the annotation tool used at Unbabel, only one category can be selected, therefore the most critical or relevant error should be marked. Naturally, within the scope of any study such as the one presented here, the purpose of the empirical study plays an important role in determining which are the most important categories. This way, for instance, when the spellcheck is being studied, every time a spelling error occurs together with another, priority is given to the category "orthography". In this particular study, given the challenges of MT described in chapter 2, it was very obvious that we would be able to address some errors but not others, such as all the phenomena related to the creative use of language, depending on their nature and specific properties. Therefore, while annotating, we gave priority to the categories "agreement", tense/mood/aspect", "word order", "sentence structure", "prepositions", "conjunctions", and "determiners". A third difficulty is related to the fact that some texts already included a sentence in Italian at the beginning of the text. This is due to the fact that Help Centers often use greeting sentences in the target language that are introduced automatically when answering a client from a particular country. These sentences were not taken into account in the annotation. A fourth difficulty is related to the fact that the taxonomy was created to annotate final translations, ready to be delivered to the client, and, therefore, did not completely suit the annotation of machine translated texts. The types of errors present in a machine translated text are different than those in a post-edited text, and some categories in the error typology were almost never used, while some others could have been useful in the annotation. For example, all the categories regarding register and style were almost never selected, while a category "wrong meaning" would have been useful to mark the cases in which one meaning of a word was considered, while the correct was another one. For example, the word "bill" was translated into "disegno di legge" (proposed legislation) instead of "conto" (amount owed). Marking the error as simply "lexical selection" did not help us distinguish such cases from those in which the correct meaning was selected, but a more appropriate word in the target text should have been used.

The annotation process in this study followed the MQM guidelines and specific guidelines related to the study that were created in order to allow a clear understanding of the categorization. The guidelines provided to the annotator in the MQM framework were helpful in defining the annotation task, but is was not always possible to follow them, because of the specifications of annotation at Unbabel and of the tool used. Among the instructions provided to the annotator, in the MQM framework, the following were considered relevant and were followed in this study:

- "Prefer more specific types to general ones. However, if a specific type does not apply, choose a general type."
- "If one word contains two errors, enter both errors separately and mark the respective word in both cases."
- "Only tag the relevant text."
- "If correcting one error would take care of the others, tag only that error."

(Burchardt, Lommel: 2014)

Even if all the five instructions listed were considered appropriate and useful for this study, only the third and fourth were actually followed, since, as previously mentioned, the annotation tool only allowed to select the leaf entries of the typology and to mark one error per word, making it impossible to follow the first two instructions. Even so, the guidelines listed above can help the annotator in the identification and categorization of errors. In particular, what is clear from the instructions given is that the annotator should be as specific as possible in the categorization, in order to facilitate the following analysis. Additionally, the annotation should be a process as clear and efficient as possible, therefore only the relevant text should be marked, in order to avoid wasting time in trying to identify the error again in the analysis stage.

4.4 ANNOTATION DATA

The *corpus* annotated in this study consists of 50 texts translated from English into Italian with the SMT system, reviewed by the editors of the community and annotated both after MT and after the first post-edition done by a speaker of the target language. The editor is not necessarily native, however in the *corpus* considered, all the tasks were done by native speakers of Italian. The reason for analyzing the machine translated text was to study the errors present, categorize them, and try to solve the most critical and most recurrent ones. The annotation of the first human post-edition is useful to provide us with information about the errors the editors correct and those that easily persist along the different stages of the translation process at Unbabel and the changes the editors introduce. The final step of post-edition, the senior editor review of the text, involves issues such as lexical selection, literal translation or collocations. Since these issues are difficult to generalize, if not impossible, through rules that can be used to automatize the post-editing process, as such errors are strictly related to the creative use of language, to idioms and to cultural references, and since that is not the goal of our project, the senior editor stage was not taken into account in this study. Since the company aims to optimize post-edition in order to minimize the cost and time spent in each translation and reduce the post-editing process to a single post-editing step, all the errors were addressed in the levels analyzed, considering that the first human edition should produce a correct and fluent text, ready to be delivered to the client. Additionally, the two levels were analyzed to compare the errors made by the machine translation system and those made by human translators and, in particular, to see which errors are corrected by human editors and which are not. In this particular study, the annotation of errors and their analysis allow us to understand the impact of distinct error types on translation quality and how many of them can be avoided by improving either the automatic tools operating on the text after MT or the MT system itself.

The fluency assessment was not considered relevant in this study because the high number of errors, and particularly of critical errors, in the translated texts has a great impact on the fluency of the target text in general and, therefore, did not allow us to make relevant distinctions regarding this aspect at this stage.

After discussing the error typology and annotation performed at Unbabel, we will now present the error annotation data of this study, i.e. the number of errors annotated in machine translated texts and in texts after the first post-edition.[5]

[5] Please note that each major category was represented in a different table, in a dark gray row. We marked in light gray the error categories that do not correspond to leaf categories in the typology and, therefore, could not be selected as error.

Accuracy error types	MT	First edition
Mistranslation		
Overly literal	9	4
False friend	0	0
Should not have been translated	18	3
Lexical selection	165	37
Omission	6	0
Untranslated	27	9
Addition	11	2
Total	236	55

Table 1.

Fluency error types	MT	First edition
Inconsistency		
Word selection	1	1
Tense selection	0	0
Coherence	2	1
Duplication	0	0
Spelling		
Orthography	1	1
Capitalization	52	19
Diacrits	0	0
Typography		
Punctuation	9	4
Unpaired quote marks and brackets	1	
Whitespace	17	5
Inconsistency in character use	0	0
Grammar		
Function words		
Prepositions	70	10
Conjunctions	12	1
Determiners	237	19
Word form		
Part-of-speech	30	1
Agreement	159	13
Tense/mood/aspect	101	3
Word order	106	4
Sentence structure	50	1
Total	848	83

Table 2.

Style error types	MT	First edition
Register		
Inconsistent register	0	1
Repetitive style	1	2
Awkward style	0	0
Total	1	3

Table 3.

Terminology error types	MT	First edition
Noncompliance with client or company style guide	0	0
Noncompliance with the glossary and vocabulary	0	14
Total	0	14

Table 4.

Wrong language variety error types	MT	First edition
Wrong language variety	0	0

Table 5.

Named entities error type	MT	First edition
Person	1	2
Organization	2	0
Location	12	9
Function	0	0
Product	2	2
Amount	0	0
Time	2	2
Total	19	15

Table 6.

Formatting and encoding error type	MT	First edition
Formatting and encoding	0	0

Table 7.

Total number of errors	MT	First edition
Accuracy errors	236	55
Fluency errors	848	83
Style errors	1	3
Terminology errors	0	14
Wrong language variety errors	0	0
Named entities errors	19	15
Formatting and encoding errors	0	0
Total	1.104	170

Table 8.

As we can see from the tables above, the number of errors annotated is high and is not evenly distributed among the different categories. This is certainly not independent of the fact that only the most relevant error was marked when there was more than one error in a word or phrase. The category with the highest number of errors annotated, in machine translated texts, is "determiners", followed by "lexical selection", "agreement", "tense/mood/aspect", and "word order". These categories include errors that can prevent the reader from understanding the text clearly, having a major or critical impact on the quality of the translation. Two categories that have a lower number of errors but are still crucial for the quality of translation results are "sentence structure" and "prepositions". In the former, the errors annotated have a huge impact on the translation because they often result in a sentence that is impossible to understand. Additionally, when post-editing such cases, the editor has to intervene on the structure of the sentence, which takes significantly more time than just changing a morpheme or a word. The editor actually has to rewrite the sentence when confronted with sentence structure errors, i.e. the editor has to translate the sentence he/she is working on. In the "prepositions" category, like in "sentence structure", the time spent in post-editing is longer, because, often, the meaning of the text cannot be understood just by reading the text produced by the MT system. This means that the editor has to go back to the source text to identify the correct translation.

With regard to the number of errors in the two stages, MT and post-editing, we can see from the table above that it decreased critically, going down from 1.104 to 170, which roughly corresponds to a 85% error reduction. The number of errors decreased in the majority of the categories, it remained the same in some of them and increased in one ("noncompliance with client's glossary and vocabulary"). As only one error could be marked when two errors occurred in the same word, some errors were not considered essential in the annotation of texts produced by the MT system. For instance, at the MT level, an error occurred in the translation of some glossary terms, which moreover did not occur in the correct word order, and the "word order" error was marked, because it was considered more relevant as a translation error. In the edited texts, the errors that become the most relevant are those related to the creative use of language and style, for example the use of correct determiners and an adequate lexical selection.

As previously mentioned, the annotation tool shows the errors and warnings detected by the Smartcheck and the message provided to the editor. Analyzing the errors detected by the tool and the message provided, we were able to calculate not only the number of errors detected, but also the number of times the words selected were actually an error, and those which corresponded to a false positive. The number of false positives was limited to a couple of agreement errors, while, as we will see below, there were several cases in which the suggestion provided was wrong, due to incorrect or incomplete instructions. This occurred mainly in style and tense agreement suggestions. The labels and messages for the annotated *corpus* were the following[6]:

- Misspelled word.
- Preposition_conjunction: l'uso della d eufonica dovrebbe essere limitato ai casi di incontro della stessa vocale (the use of the euphonic "–d" should be limited to the cases in which the same vowel occurs.)
- Agreement: controllare il tempo dei verbi utilizzati nella frase (check the verb tenses in the sentence.)
- Spelling: forse volevi dire […] (maybe you meant […].)

[6] Please note that some messages given by the Smartcheck were in English, while some others in Italian. An English translation of those in Italian is provided in brackets.

- Style: si consiglia di non iniziare una frase con una congiunzione (it is recommended not to start a sentence with a conjunction.)
- Typographical: extra whitespace next to.
- Agreement: l’articolo non concorda, usare: […] (the article does not agree, use: […].)
- Preposition_conjunction: use ‘col’ instead of ‘con il’.
- Agreement: l’aggettivo ‘chiunque’ richiede il congiuntivo (the adjective “chiunque” requires the “congiuntivo”.)
- Spelling: si consiglia di sostituire la preposizione ‘di’ con la forma apostrofata ‘d’’ (it is recommended to substitute the preposition “di” with the form “d’” followed by an apostrophe.)

The errors detected by the Smartcheck in the annotated *corpus* can be divided into the following three categories:

- Useful instructions, like “spelling” or “typographical”. It is likely that the error will be corrected.
- Useless instructions, like in “agreement” and “preposition_conjunction.” The highlighting provided by the Smartcheck helps the editor to see the error, the general category is usually correct but the suggestion provided by the tool is not useful to correct the error.
- Wrong instructions that provide misleading information, such as those regarding style.

We used the categories listed above to analyze the Smartcheck suggestions in the *corpus* annotated

Classification of Smartcheck instructions	
Useful instructions	95
Useless instructions	66
Wrong instructions	79
Total	240

Table 9.

In Table 9 we can see that 240 errors were detected by the Smartcheck, but only in 95 cases the error was actually an error and the instruction provided to the editor was useful.

In 66 cases the message did not provide useful information to the editor, but the constituents selected by the Smartcheck correspond to an error. And, finally, in 79 cases the instruction provided to the editor was simply wrong. For example, when in the instruction it was stated that the words “qualcuno” and “chiunque” require the “congiuntivo” mood, the instruction was considered wrong because the two words can be followed either by the “congiuntivo” mood or by the “indicativo” mood. This means that the rule integrated in the Smartcheck is not correct, because it does not admit the cases in which “qualcuno” and “chiunque” are followed by the “indicativo” mood. Other cases of wrong instructions were the false positives in the detection of agreement errors.

Useful instructions per error category generated by the Smartcheck	
Typographical	18
Gender agreement	50
Spelling	14
Tense agreement	11
Prepositions	2
Total	95

Table 10.

In Table 10 we analyzed more thoroughly the 95 cases in which the Smartcheck correctly identifies an error and provides a useful instruction. We can notice that the tool correctly detects typographical errors and gender agreement errors, in particular when the error occurs in agreement between the article and the noun. Among the 14 spelling errors detected, 7 regarded the use of the letter “d” in a word that ends with a vowel when the next word starts with the same vowel, while the others were orthography errors. In 11 cases the Smartcheck detected an error in the tense of the verb and provided a useful instruction and, finally, in only two cases the tool detected an error in the choice of the preposition.

As we can see comparing the number of errors annotated and those detected by the Smartcheck (tables 1, 2, and 10), there are some categories in which the errors detected by the Smartcheck are more than those annotated, for example in orthography only one error was annotated, while 14 were detected by the Smartcheck. This is due to the fact that spelling errors occurred in a word together with another error and it was the other error that was marked in the annotation.

Because of the often useless information that the Smartcheck provides and because of the quality of the MT, the editor, in the majority of the cases, actually has to re-do the translation, instead of just reviewing it. This results in a costly and time-consuming process. If the editor receives a MT translation with a high number of errors, but with useful instructions on how to correct them, the risk of not correcting an error or not noticing it is substantially reduced.

In order to improve MT by integrating information in the system, we can either work on source language understanding or on target language generation. Since the Smartcheck works mainly on the target language, the first issues addressed will be those regarding the target text. In the following chapters, we will focus on three categories of errors, namely "word order", "agreement", and "tense/mood/aspect". The choice is due to the fact that these are frequent errors (respectively 9%, 14%, and 9% of the total number of errors annotated) which have a significant impact on the translation quality. Addressing these issues automatically would help to reduce significantly the post-editing effort, accelerating the process and minimizing its cost. Additionally, Unbabel does not have the tools to solve all the problems identified yet. This way, we selected the errors with the higher probability of being addressed in the current architecture of the system used at Unbabel. Even so, and despite the fact that solutions will be presented for addressing the three categories mentioned above in the following chapters, before we will briefly analyze two categories that will not be studied in detail, as we know that the solution cannot be implemented at Unbabel yet. The reason for presenting this preliminary analysis is that the errors in these two categories were frequent and it was possible to generalize them and to find a pattern in the annotated errors. The analysis will be useful for future work, when more tools to address the issues are available.

4.4.1 Determiners

This category included errors that occurred in the translation or use of determiners. We included in this category the following constituents: articles, personal pronouns, possessives, and demonstratives. Therefore, the category annotated two main types of errors presented below.

4.4.1.1 Incorrect or Missing Article or Personal Pronoun

These errors are particularly common in the language pair English-Italian because of the contrast in the amount and use of articles, personal pronouns, and possessives in English and Italian, which amounts to the fact that often there is not a one to one mapping in these constituents for this language pair. Whereas in English the subject of the verb must always be overtly expressed, in Italian it can be omitted if it can be assessed through the verb form. Additionally, in Italian articles contain more morphological information than in English, as they specify the gender and the number of the noun they precede and their form depends on the first letter or morpheme of the noun. The articles we can find in Italian are:

Articles in Italian		
Definite articles		
	Masculine	Feminine
Singular	Lo, il, l'	La, l'
Plural	I, gli	Le
Indefinite articles		
	Masculine	Feminine
Singular	Uno, un	Una, un'
Plural	Dei, degli	Delle

Table 11.

This means that from a single form in English, the MT system has to select one of the several possibilities in Italian, which almost always constitutes a problem for the MT system. We also included in this category the errors concerning other determiners, such as personal pronouns, possessives and demonstratives. While in English the subject must be expressed, in Italian the personal pronoun can be omitted. The MT system automatically translates the pronoun in Italian, but cannot distinguish the cases in which the pronoun is needed from those in which it is not. Additionally, in English it is common to use demonstratives and possessives in sentences such as "thank you for your answer" or "thank you for your patience". In Italian, the possessive is omitted: "grazie della risposta", "grazie della pazienza". This kind of error is strictly related to language use and therefore was not considered relevant in this study, since a human editor detects and corrects it easily, while it is significantly complex to define the restrictions involved without over generating so that the machine is able to determine the cases in which it has to be omitted and those in which it should not.

4.4.1.2 The Translation of the Pronoun “you”

The pronoun “you” in English can be translated into Italian as “tu” for the second person of the singular, or “voi” for the second person of the plural. When the number is not specified, both forms are acceptable. Additionally, in formal contexts, it can be translated into “Lei”. The texts annotated were either Help Center emails (and the singular is commonly used in Italian in this context), or texts in which the client explicitly asked to use the singular form to keep an informal register. However, since the use of the informal pronoun was not included in the client’s instructions of the tickets, the use of the pronoun “Lei” was not marked as an agreement error when it was consistent throughout the text.

Additionally, the system usually translates “you” by “si”, using therefore an impersonal structure. The particle “si” has different functions in Italian and can be used in impersonal structures, when the subject of the action expressed by the verb is generic and indefinite. The use of the impersonal structure with the function word ‘si’ is admitted in Italian and, sometimes, corresponds to an English sentence containing the pronoun “you”. However, the number of cases in which this is possible is reduced, and the systematic translation of the pronoun “you” by “si”, performed by the MT system at Unbabel, can generate many errors.

While the first type of issues related to the use of determiners is usually corrected by the editor, the pronoun is not always changed in the post-editing stage. It was not possible to solve this error in this study, because the choice of the pronoun depends on the client and on the text. Having a rule establishing that the pronoun has to be always “tu” would not result in a correct translation in the plural, nor allow for a formal register, that could be appropriate in specific contexts.

4.4.2 Sentence Structure

This category included errors in the translation of English structures into Italian. The errors occurred, in the majority of the cases, in the selection of the right tense or mood of the verb and in POS selection, but were annotated as sentence structure errors because we noticed that the error occurred when a particular sentence structure was used in the source text. Several different structures were recurrently not translated correctly, the most common being:

1. Relative clause without the relative pronoun:
   Steps you take before running into problems:
   *Passaggi da seguire prima di incappare nel problema:
2. Relative clause without the relative pronoun and with the preposition at the end:
   What about the features I paid for?
   *Che dire delle caratteristiche che ho pagato per?
3. Completive clause without the conjunction “that”:
   Help us to ensure our products are the best.
   *Aiutarci a garantire i nostri prodotti sono i migliori.
4. Nonfinite completive clause:
   It may cause our site not to function properly.
   *Può causare il nostro sito per non funzionare correttamente.
5. Translation of clauses introduced by “sorry”:
   Sorry that you are having troubles.
   *Spiacente hai problemi.
6. Translation of the infinitive clause introduced by “for”:
   They will arrange for someone to help you.
   *Si provvederà per qualcuno di aiutarvi.
7. Translation of the clause introduced by “after”:
   You will remember your trip long after you get home.
   *Vi ricordate il vostro viaggio molto tempo dopo si arriva a casa.
8. Translation of the gerundive clause introduced by “from”:
   It should not prevent your phone from working.
   *Non dovrebbe evitare che il telefono di lavoro.

In this study, it was not possible to solve the errors in these categories, namely “determiners” and “sentence structure”, because it is currently not possible to integrate the linguistic rules needed to detect the difficult structure in English and provide their correct translation in Italian in the MT system used. However, the analysis provided and the generalization done will be helpful in future work regarding the automatic detection and solution of errors pertaining to these error categories. Such work would have a major impact on the quality of the results since, as previously mentioned, determiner errors are frequent and represent 21% of the total number of errors annotated in this study, while sentence structure errors make it impossible to understand the target text.

In this chapter we presented the annotation data and their classification, and introduced the error categories we will concentrate on, i.e. “word order”, “agreement”, and “tense/mood/aspect”. In chapter 5 we will take into account word order errors, analyzing them and providing possible solutions.

5 WORD ORDER

5.1 INTRODUCTION

Word order is a crucial aspect of language, often playing a decisive role in the grammar of specific languages, and takes into account the position of syntactic constituents in a sentence. We can say that languages have a flexible (or free) word order when the constituents can occupy different relative positions in the sentence, without changing its meaning, as it happens in many languages with a case-system. Strict word order languages are those in which the syntactic constituents follow a specific order. We can talk about fixed or free word order both in the sequence of phrases in a sentence and within the constituents of a phrase itself. Both in English and Italian, for example, the NP that is the subject of the verb precedes the VP[7]. Additionally, in an NP, determiners, adjectives and nouns should follow a specific order for the phrase to be acceptable. For example, in both English and Italian, the sentence is not acceptable if the determiner follows the noun. The same happens in a PP, for determiners, prepositions, adjectives, and nouns.

English is a strict word order language, as the constituents have to follow a predetermined order in a sentence for a given meaning to be obtained: the SVO order establishes the position of the subject, the verb, and the object, while the other complements or arguments can follow the object, or precede it if they are time and place complements. Additionally, subjects must be overtly produced. Different word orders (within limitations) are accepted in poetry, in literature, or in topicalization, when a constituent is emphasized by changing the position in the sentence.

Italian, on the contrary, has a more flexible word order when it comes to the position of phrases in the sentence. However, the order is considerably strict within phrase boundaries: for instance, the determiner always precedes the noun, the adjective precedes or follows the noun according to predetermined rules, the preposition always precedes the noun under its scope, and so on. Different word orders are also accepted in topicalization, poetry, or literature.

[7] In Italian, sentences in which the subject follows the verb are admitted in literary texts and in topicalized sentences in which the focus is on the VP. In this case, the subject is often preceded by a comma.

a. Si alzava il sole.
[Was rising the sun]

b. Piangeva, Maria.
[Was crying, Maria]

In this study, we decided to focus on word order errors in noun modification structures, due to the fact that these were the most common in our *corpus*, as we will see below. We did not take into account determiners and quantifiers. The reason for that is, firstly, that these constituents have a different syntactic function from that of modifiers, and, in the second place, that they do not pose difficulties in the translation from English into Italian, i.e. they are generally well dealt with by MT systems. Their position is the same in the two languages and the MT system systematically produces the correct order. With regard to the syntactic function, determiners and quantifiers are considered specifiers of the noun, they occur before it, and together they form an NP. On the contrary, modifiers can occur before and after the noun and limit and either introduce or underline restrictions to its denotation. Although in English grammar genitives are considered specifiers, such constructions were taken into account in this study, since a significant number of errors occurred in phrases involving this structure. This is the single exception to the delimitation of our object mentioned above.

In the second section of this chapter we will analyze noun modification involving different constituents both in English and Italian, while in the third we will consider the specific errors annotated and present possible solutions to address them.

5.2 NOUN MODIFICATION

Noun modification is an important aspect of natural languages. It allows the speaker to convey additional information about the subject or the complement. There are different ways of modifying a noun in languages like English:

- With an adjective: red table, old dog, big house;
- With a past participle: published article, broken chair;
- With a prepositional phrase: basket of oranges, man with the hat;
- With another noun: history book, leather suitcase;
- With a verb in the –ing form: accompanying adult, singing bird;
- With a relative clause: the girl who was painting.

A noun can have more than one modifier, for example an adjective and a noun, a past participle and a prepositional phrase, or one combination of the modifiers mentioned above: the old leather sofa, my blue scarf with dots, the dish left on the table.

Please note that, for the sake of the work presented here, we do not discuss the contrast between complements and adjuncts within the NP and will generally designate all these constituents as modifiers. The motivation for this option is the fact that this distinction has no impact on the translation results for the language pair we studied.

5.2.1 Adjective Noun Modification in English and Italian

When the modifier is an adjective, it usually precedes the noun in English. If there is more than one adjective, they all precede the noun and their relative order is determined by their semantic content: opinion, size, physical quality, shape, age, color, origin, material, type, purpose.

1. Beautiful young white pony

On the contrary, in Italian, the adjective usually follows the noun, as the information that is added usually is added on the right. However, the position depends on the type of adjective and on the relation it establishes with the noun. The adjective necessarily follows the noun when it is a relational adjective (i.e. a non-descriptive adjective, that is morphologically derived from a noun, maintains a semantic relation with it, and cannot be used in predicative position) (see example 2), when it is in the comparative or superlative form (example 3), when it denotes an objective quality of the noun (example 4), and when it has a complement (example 5). Please note that prenominal position is accepted in literary texts in some cases: it is admitted in the example 4, while it is not accepted in the other examples below, even when they occur in a literary context.

2. Situazione economica
   [Situation economic]
3. Una casa più grande
   [A house bigger]
4. Capelli neri
   Neri capelli (literary)
   [Hair black]
5. Un titolo facile da ricordare
   [A title easy-to-remember]

The adjective precedes the noun when it expresses a subjective quality of the noun (see example 6) or when the speaker wants to emphasize it (example 7).

6. Strane cose
[Strange stuff]
7. La dorata luce del tramonto
[The golden light of the dusk]

There are also some adjectives, such as “vecchio”, “buono”, “alto”, and “grosso”, that change their meaning depending on the position they appear in. In the postnominal position, they denote a physical characteristic of the referent, while, when they occur in the prenominal position, they refer to a non-physical aspect.

8. Due vecchi amici – two people that have been friends for a long time
Due amici vecchi – two friends that are old
9. Un alto funzionario – an officer that has an important function
Un funzionario alto – an officer that is tall
10. Un grande dipinto – a painting that has great artistic value
Un dipinto grande – a big painting

When there is more than one adjective, the relative order of the adjectives depends on the type of each one of them. If one is relational and one is a quality adjective, the former should be the one immediately after the head noun, and the latter should follow it (example 11). It is also possible to have the quality adjective preceding the noun (example 11.c). If the two of them express a quality, the conjunction “e” or the comma should be used (example 12). It is also possible to have one of them in the prenominal position, if it is emphasized (example 13).

11. a. Un documento politico recente
[A document political recent]
b. *Un documento recente politico
[A document recent political]
c. Un recente documento politico
[A recent document political]
12. a. Un quaderno nero e grande
[A notebook black and big]
b. Un quaderno nero, grande
[A notebook black, big]

13. a. Un ragazzo giovane biondo
[A boy young blond]
b. Un giovane ragazzo biondo
[A young boy blond]

When there are more than two modifying adjectives, the order in Italian is completely different than the one in English, as we can see in example 14.

14. A beautiful old blue watch
Un bel vecchio orologio blu

The more fluent and natural order in Italian is the one presented in 14, where some of the adjectives are emphasized, but the structure “un orologio blu vecchio e bello” is also acceptable. Different orders are possible if the speaker wants to highlight a particular adjective. In a translation without any adjective emphasized, like “un orologio blu vecchio e bello”, we can see that the adjectives follow the noun in the opposite order of the English NP.

5.2.2 Noun-Noun Modification in English and Italian

Noun-Noun (henceforth NN) modification is a common strategy in English, while it is rare in Romance languages like Italian. In NN modification, in English, the nominal modifier precedes the modified noun and some cases in which it can be used are the following:

- To indicate the material: a gold frame, a leather bag;
- To indicate a part-whole relation: village market, the table leg;
- When measures, age or values are used as modifiers: a twenty-kilogram suitcase, a fifty-year-old lady, a three-euro coffee.

If the modifiers are an adjective and a noun, both modifying the head noun, then the modifying noun is necessarily adjacent to the modified noun.

15. A big metal box

When there is a sequence of modifiers that includes nouns and adjectives, the order suggested by the Cambridge Dictionary is the same as the one for adjectives: opinion, size, physical quality, shape, age, color, origin, material, type, purpose.

16. A precious big old wedding ring

In Italian, a noun modifies another noun only in a couple of very rare and crystallized cases, such as “indirizzo email”. There are some words that etymologically come from a case of NN modification, such as “ferrovia” (railroad) or “pescespada” (swordfish). However, they are not considered as examples of NN modification, because they are now a single character sequence and, therefore, a single word.

5.2.3 Other Types of Noun Modification in English and Italian

A noun can also be modified in English by a past participle (example 17) or a verb in the –ing form (example 18). In both cases the modifier precedes the noun if they have a non-predicative value.

17. The broken window
18. The swimming competition

In Italian present and past participles can be modifiers (see examples 19 and 20), while the gerund cannot.

19. Un uccellino cantante
    [A bird singing]
20. Una finestra rotta
    [A window broken]

An English verb in the –ing form can be translated into Italian either by a present participle or a gerund. However, when the verb in the –ing form modifies a noun, it has to be translated into an adjective, if there is an equivalent one (example 21), into a PP (example 22) or into a relative clause (example 23). A gerund cannot be used in the translation, since the modification of a noun with a gerund is not admitted in Italian, while the present participle can be used only in some cases, such as 23.

21. An accompanying adult
    *Un adulto accompagnando
    *Un adulto accompagnante
    Un adulto accompagnatore

22. The swimming competition

*La gara nuotando

*La gara nuotante

La gara di nuoto

23. The flying insect

*L’insetto volando

L’insetto volante

L’insetto che vola

With regard to the other modifying structures mentioned above, a PP can be a modifier both in English and Italian and it follows the modified noun in both languages (example 24). The same happens with relative clauses (example 25).

24. A group of people

Un gruppo di persone

25. The person who gave me this present was very kind.

La persona che mi ha fatto questo regalo è stata molto gentile.

5.3 WORD ORDER ERRORS IN THE *CORPUS*

In our *corpus*, we find many errors annotated as belonging to the word order category. The issue is crucial because, even if there are errors of this type in which the editor easily and quickly understands the correct word order, some of the translated structures are ambiguous, leading the editor to spend a considerable amount of time to produce the correct translation structure in the target language. To address the problems in this category, we divided them in different subcategories, which are presented in the table below. Please note that the categories are listed independently of the noun occurring within an NP or within an NP nested inside a PP.

Word order errors	
Number of unique errors	68
Number of errors in noun modification	65
Number of errors involving other structures	3
Total	106

Table 1.

As we can see from Table 1, among the 106 errors annotated in translation results at the MT level, we could distinguish 68 unique errors, while the remaining 38 errors were repeated. We divided the 68 unique errors into those involving a noun modification structure, and those involving other structures. The three errors of this subcategory are the following:

- Wrong position of the adverb that modifies a verb;

26. We <u>highly</u> suggest
    *<u>Altamente</u> suggeriamo
    Suggeriamo <u>calorosamente</u>

- Wrong position of the negation;

27. Or might <u>not</u> have been credited
    *O <u>non</u> potrebbe essere stato accreditato
    O potrebbe <u>non</u> essere stato accreditato

- Wrong relative position of the verb form and two adjectives in predicative position.

28. I do understand how <u>annoying and frustrating</u> it can be.
    *Io capisco come <u>fastidioso e</u> può essere <u>frustrante.</u>
    Capisco che possa essere <u>fastidioso e frustrante.</u>

In this study, we decided not to address problems belonging to this last subcategory, because of their low occurrence and dispersion. On the contrary, we decided to focus on errors involving noun modification structures.

5.3.1 Noun Modification Errors in the *Corpus*

Word order errors in noun modification	
Errors in noun-noun modification	29
Errors in adjective-noun modification	4
Errors in noun modification with both noun(s) and adjective(s)	32
Total	65

Table 2.

The 65 unique errors involving a noun modification structure were divided into the subcategories considered in Table 2, based on the analysis presented in sections 5.2.1, 5.2.2, and 5.2.3. In addition to the categorization presented in the mentioned sections, we can notice that, among the 65 errors occurring in noun modification structures, in 27 cases the head noun of the NP is a named entity, i.e. a proper noun: 17 cases occurring in NN modification structures, one in an adjective-noun modification structure, and 9 in modification structures with both adjectives and nouns. The high occurrence of errors annotated involving a named entity highlights the importance of a solution regarding these constituents that are frequent and problematic. Named entities often represent a challenge in MT because their use in language is idiosyncratic. Additionally, they are often not included in lexical resources due to the low occurrence of them, and are often not present in the *corpora* used to train MT systems for the same reason.

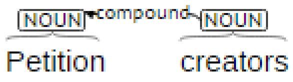
5.3.1.1 Noun-Noun Modification

Errors in the noun-noun modification subcategory involved wrong word order in the translation into Italian of a noun modified in English by another (or several other) noun(s) (see examples 29 and 30). We included in this subcategory the incorrect translation of the possessive case (example 31).

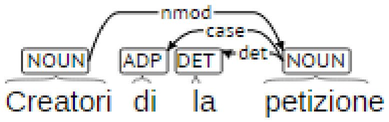
29. Petition creators
*Petizione creatori
Creatori della petizione
30. Summer residence
*Estate residenza
Residenza estiva
31. Petition’s goal
*Petizione’s obiettivo
Obiettivo della petizione

The syntactic trees corresponding to examples such as those presented in the examples 29 and 30 are the following, for English and Italian[8]:

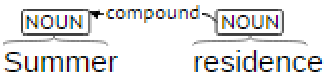
English

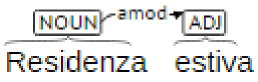


Italian



[8] The syntactic trees presented in this chapter were produced by the parser developed by Martins, Almeida, and Smith in 2013, that is currently used at Unbabel (see section 3.4.2). The tags used in the parser analysis are those of the Stanford dependencies representation. For a reference, see Marneffe and Manning (2008).

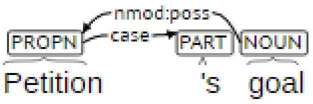




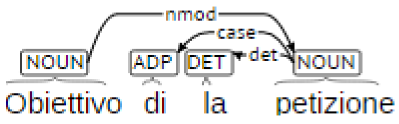
As we can see, depending on the specific example, two different structures are used in Italian to translate a NN modification structure in English. If the modifying noun in English is translated as a PP in Italian, the head noun of the NP nested in the PP has to be the translation of the modifying noun in English (see example 29, petition→petizione). If an adjective is used in Italian to translate the modifying noun in English, then this adjective must be semantically related to the modifying noun in English (see example 30, summer$_{N}$→estiva$_{ADJ}$).

With regard to the possessive case, as we already said, it is considered to be a specifier by the English grammar, thus corresponding to a different syntactic structure, as illustrated in the syntactic trees below, that represent example the 31:

English



Italian



In Italian, the structure used to translate examples such as 31, is a PP, headed by the preposition “di”, the nested noun being the one marked with the genitive case in English (petition→petizione).

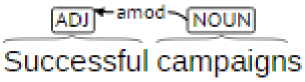
5.3.1.2 Adjective Noun Modification

Errors in this subcategory correspond to the generation of the wrong word order in the translation of nouns modified by one or more adjectives (example 32). We included in this subcategory the incorrect translation of nouns modified by past participles or gerunds used as adjectives (example 33). In our *corpus*, there are only 4 occurrences of adjective-noun modification errors. The low occurrence of this kind of error shows that, in general, this type of modification structures is usually correctly dealt with by the system.

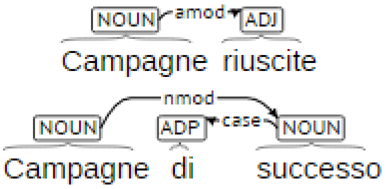
32. Successful campaigns
    *Di successo campagne
    Campagne di successo
33. Newly-created petition
    *Appena creato petizione
    Petizione appena creata

The syntactic trees representing the simplest case, i.e. an adjective noun modification structure involving a single adjective, in English and Italian are represented below. Please note that, in the translation into Italian, both the adjective and the PP are admitted and the use of the former or the latter depends on whether an equivalent adjective exists in Italian or not.

English

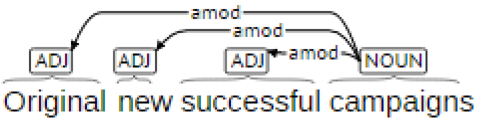


Italian

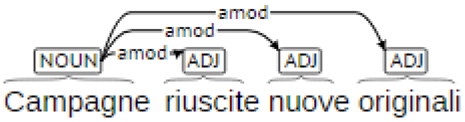


A previously mentioned, it is possible to have more than one adjective modifying a head noun. In such cases, e.g. in the NP "original new successful campaigns" translated into the Italian NP "campagne riuscite nuove originali", the syntactic trees are the following[9]:

English

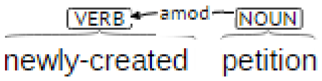


Italian

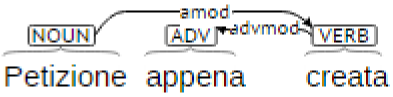


With regard to example 33, the head noun is modified by a past participle that is preceded by an adverb. The syntactic trees representing the English structure and its translation into Italian are the following:

English



Italian



As we can see from the syntactic trees, the verb form (past participle) establishes an adjective noun modification with the head noun, while the adverb modifies the verb form.

5.3.1.3 Noun Modification with both Noun(s) and Adjective(s)

Errors in this subcategory involved wrong word order in the translation of nouns simultaneously modified by one or more nouns and one or more adjectives. This subcategory includes both the cases in which all the constituents modify a single head noun (see example 34), and the cases of inlaid modification, in which one or more modifying constituents modify a modifier of the head noun (examples 35, 36, and 37).

[9] Note that, however, in our *corpus* no errors were annotated in adjective noun modification structures with more than one adjective modifying the head noun.

34. 90-minute introductory tour

*90 minuti tour introduttivo

Tour introduttivo di 90 minuti

35. Medici family political court

*Famiglia tribunale politico dei Medici

Tribunale politico della famiglia Medici

36. Glass covered pyramidal tower

*Coperta torre piramidale in vetro

Torre piramidale coperta di vetro

37. Limited editions souvenirs

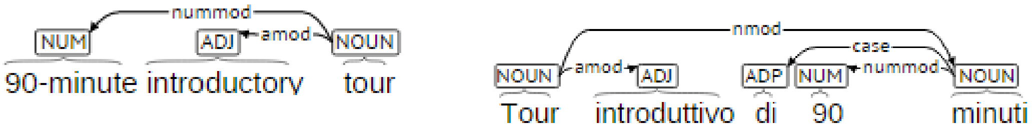
*Edizioni limitate souvenir

Souvenir a edizione limitata

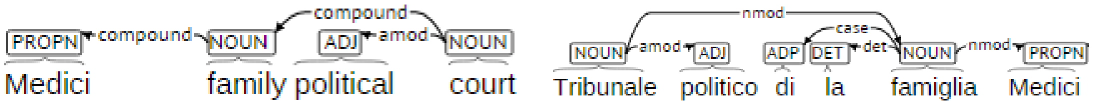
In the four examples above we have different sequences of modifiers and different syntactic trees corresponding to the structures of the examples:

English Italian

(34)

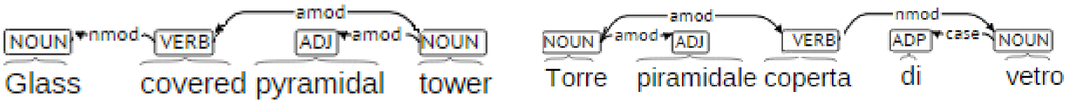


(35)

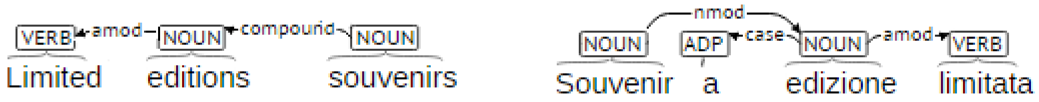


As we can see in 34 and 35, in terms of syntactic tree, there is no distinction in the structure of the two examples in Italian. However, we can notice that, in 34, the PP “di 90 minuti” modifies the noun “tour”, while, in 35, the PP “della famiglia Medici” modifies “tribunale politico”. The distinction is not apparent in the tree, because the parser does not mark whether the dependency is on a group of constituents or simply on one constituent.

(36)


NOUN
nmod
VERB
amod
ADJ
amod
NOUN
Glass covered pyramidal tower
NOUN
amod
ADJ
amod
VERB
nmod
ADP
case
NOUN
Torre piramidale coperta di vetro

(37)



Many different sequences of modifiers are possible and the syntactic tree is almost invariably significantly different in English and in Italian. Not only the order of the constituents varies, but also the category they correspond to, since a modifying noun is often translated as a PP.

Among the 32 errors belonging to this subcategory, the errors occurred in the sequences listed below:

- N1 + ADJ + N2, where the first noun (N1) modifies the adjective that modifies the head noun (N2);

38. Plague-ravaged houses
*Case peste devastata
Case devastate dalla peste

- N1 + ADJ + N2, where the fist noun (N1) and the adjective both modify the head noun (N2);

39. Milano top attractions
*Milano top attrazioni
Migliori attrazioni di Milano

- ADJ + N1 + N2, where the adjective is related to the modifying noun (N1);

40. Limited editions souvenirs
*Edizioni limitate souvenir
Souvenir a edizione limitata

- ADJ + N1 + N2, where the adjective is related to the head noun (N2);

41. Original wartime bunker
*Tempo di guerra bunker originale
Originale bunker di guerra

- ADJ1 + ADJ2 + N1 + N2, where the two adjectives (ADJ1 and ADJ2) and the modifying noun (N1) modify the head noun (N2);

42. Hungarian unique folklore traditions
*Ungheresi tradizioni folcloristiche uniche
Tradizioni folcloriche ungheresi uniche

– N1 + N2 + ADJ + N3, where the first noun (N1) is related to the noun (N2) modifying the head noun (N3) and the adjective modifies the head noun;

43. The Medici family political court
*La famiglia tribunale politico dei Medici
Il tribunale politico della famiglia Medici

– N1 + ADJ1 + ADJ2 + N2, where the first noun (N1) modifies the first adjective (ADJ1) that, together with the second adjective (ADJ2), modifies the head noun (N2);

44. Glass covered pyramidal tower
*Coperta torre piramidale in vetro
Torre piramidale coperta di vetro

– ADJ1 + ADJ2 + N1 + ADJ3 + N2, where all the modifiers modify the head noun (N2);

45. The most important Catalan art-nouveau, or modernista buildings
*La più importante catalana in stile liberty, o di edifici modernisti nella vita del paese
I più importanti edifici catalani in stile liberty o modernisti

– N1 + N2 + ADJ + N3 + N4, where the first three modifiers (N1, N2, and ADJ) modify the third noun (N3) that modifies the head noun (N4);

46. Dance, house and commercial music sessions
*Danza, casa e sessioni di musica commerciale
Sessioni di musica dance, house e commerciale

In this study, we focus on the subcategories that include frequent errors to identify generalizations and put forth possible solutions to address the shortcomings of the MT system. A way to have the editors check every case, would be having the Smartcheck highlight all the sequences of nouns and adjectives. However, the strategy would be counterproductive, due to the high number of occurrences of such sequences in a text – it would basically mean that we would be highlighting every other NP – and to the fact that editors almost always identify and correct word order errors. Therefore, having too many alarms would reduce the precision and effectiveness of the Smartcheck. In the next section we will discuss and present possible solutions to the errors generalized, and provide rules that can be applied to the Smartcheck and to MT.

5.3.2 Possible Solutions

In trying to solve the problems related to word order within an NP, there are several difficulties. First of all, understanding whether in the two languages there is a rule that can be applied to each case observed to solve the issue, without introducing problems in other examples, i.e. without over generating. A general rule in the modification with quality adjectives could be that for Italian the order is the opposite: the English pair ADJ + N should be N + ADJ in the target language, and, when there is a sequence of adjectives, the relative order of the constituents in Italian is the opposite of the English. However, deriving the correct order is more complicated when there is more than one modifier because, as we saw, these can modify different constituents, namely another modifier, instead of modifying the head noun.

Second, the parser used at Unbabel, described in chapter 3 (section 3.4.2) does not distinguish between different types of adjectives. It only recognizes degree (comparative or superlative). This can be a difficulty in Italian, because an accurate translation should take into account the type of adjective in order to establish the correct sequence of modifiers.

Third, as we are intervening on the target text, we cannot use the syntactic parser in English to understand the correct dependency, we are only working with the target text, from an operational point of view. The parser in Italian, obviously, does not always recognize the correct dependency by analyzing a translation with an incorrect word order. In fact, when an incorrect translation produced by the MT system is analyzed by the parser, not only does the parser often fail to identify correctly the dependency between the constituents, but also it sometimes fails to accurately determine their POS, as it classifies adjectives as verbs and pronouns, for instance.

Another difficulty is that, when the Italian translation of a modifying adjective or noun is a PP, the correct preposition has to be selected. The choice is difficult because of the language-specific use of prepositions and the wide number of exceptions. In Italian, for example, the preposition “di” is used in examples in 47 and 49, but the preposition “da” is used in 48. Additionally, a determiner is in some cases needed before the noun (example 47).

47. The apartment keys
Le chiavi dell'appartamento

48. Baseball field
Campo da baseball

49. Leather bag
Borsa di pelle

Finally, when there is more than one modifier of the head noun, the order in the translation is critical because it can be an inlaid modification structure and, therefore, a an adjective or a modifying noun modify another modifier, and not the head noun. In such sequences of constituents, dependency may be different from case to case, as we saw above in the section 5.3.1.3.

If we could intervene on the target text taking into account also the source text, the tool to help to solve error in word order would be a syntactic parser. By establishing the dependencies of all the modifiers, it would allow us to understand what could be the right order in Italian. Ideally, a syntactic parser could be used in English, to establish the dependency tree of the original phrase, a syntactic parser in Italian could do the same, and then the two trees could be compared and matched. However, the POS of constituents would not always be equivalent in English and Italian, because a modifying noun in English can correspond to an adjective or a prepositional phrase in Italian, for instance, introducing additional challenges. Since the Smartcheck takes into account only the target text, addressing word order in an automatic way is a more challenging task. As different combinations of modifiers are admitted in Italian, we cannot have the Smartcheck highlight a phrase every time a sequence of modifying constituents occurs. Otherwise the errors marked would be too many, and often not errors, and the editor would not pay attention to the Smartcheck indications (see sections 3.4.1 and 4.4 for the relevance of precision in the Smartcheck performance). We counted in the *corpus* the number of occurrences in English of the problematic combinations of constituents presented in section 5.3.1.3., and the number of times an error is annotated in the translation of these structures into Italian. In doing so, we were aiming to establish the percentage of incorrect translations of the given combinations, i.e. the correlation between certain structures in the source text and the generation of word order errors. However, the process did not work for several reasons:

- First of all, the parser labels named entities as proper nouns and excluded them from the count. However, we were able to repeat the process including named entities (with the label PROPN) in the sequence.
- Second, many sequences were not correctly analyzed by the parser and the POS of the constituents in the expression was not correctly identified.
- Third, no occurrences were found for the sequences N+N+ADJ+N, N+ADJ+ADJ+N, ADJ+ADJ+N+ADJ+N, N+N+ADJ+N+N, PROPN+ADJ+PROPN, and NOUN+ADJ+PROPN even if some errors in the translation of the sequences into Italian were marked having these sequences of constituents in the source text.
- Finally, the expressions identified are sometimes only a part of the error annotated. This can be due to the fact that the parser did not correctly analyze the constituent in English. For example, the sentence "many gorgeous art nouveau buildings" was not recognized as the sequence of ADJ+ADJ+N+N+N, but the sequence "gorgeous art nouveau" was classified as a sequence of ADJ+ADJ+N and the word "building" was not recognized as the head noun of the phrase.

However, even if the process did not give the expected results, and we cannot thus use its results to formulate any hypothesis, we were able to count the occurrences in the table below, which we present as merely indicative.

Correlation between the occurrence of specific POS sequences in English and annotated errors in Italian		
Sequence	Number of occurrences identified in English	Number of errors annotated in Italian
N+ADJ+N	4	0/4
ADJ+N+N	41	6/41
ADJ+ADJ+N+N	3	0/3
PROPN+ADJ+N	9	4/9
PROPN+N+ADJ+N	1	1/1
ADJ+PROPN+PROPN	6	0/6

Table 3.

It is important to point out that in 9 cases the sequence recognized by the parser was part of the error annotated. They were not counted in the table above, because, as they do not completely correspond to the actual sequence, they would lead to wrong conclusions. Even if it was not possible to count the percentage of errors in the translation of specific sequences of constituents of the source language, the experiment helped us identify the most common sequences of modifiers in English in the *corpus* analyzed, namely N+ADJ+N, ADJ+N+N, ADJ+ADJ+N+N, and PROPN+ADJ+N. Although they are common, it is difficult to generalize a rule to generate the correct translation for them, because there is more than one possible dependency relation between the constituents involved, and, therefore, it was not possible to come up with a rule to automatize the post-editing of such sequences. Additionally, as the sequences are frequent, having the Smartcheck highlight all of their occurrences would be counterproductive since, as we saw, only a in small percentage of sequences an error was actually annotated.

The identification of the sequences in which the greatest amount of errors was annotated and the analysis of the sequences that are not admitted in Italian, can help us create rules for the Smartcheck to intervene with more precision in the target text. From the analysis presented in section 5.2, we realize that only the following are possible sequences of constituents in Italian[10]:

- N+ADJ^{+}
- N+PP^{+}
- N+ADJ*+PP*+ADJ*

Therefore, every time a sequence of adjectives, prepositional phrases and nouns occurs before the head noun in Italian, the Smartcheck should highlight it and provide a warning to the editor. In order to make the task easier, the wrong sequences the Smartcheck has to identify are:

- ADJ^{+}+ ADJ+N
- PP^{+}+N
- ADJ^{+}+PP^{+}+N
- ADJ^{+}+N+N

With regard to the sequence N+N, the expressions accepted in Italian being very reduced in number and crystallized, such as “indirizzo email”, they should be included in a

[10] Please note that regular expressions were used to present rules and generalizations.

glossary or in the lexical resource, while the Smartcheck should highlight all the other cases in which the sequence occurs in Italian.

As for named entities, we already saw in section 5.3.1, they occur in almost half of the noun modification errors annotated. Due to the high number and to the fact that they often have an idiosyncratic behavior, the solution can be having the Smartcheck highlight them and ask the editor to check the translation. As we will see in the agreement chapter, this strategy will also help in addressing errors related to agreement. In order to reduce the number of alarms given by the Smartcheck, we propose that only the cases in which the named entity is preceded or follow by a modifier (adjective or PP) should be highlighted.

We will now list some rules that can be implemented in the Smartcheck, and some that can be integrated in the MT system in the future or in any tool that checks both the source and the target text. Please note the implementation of the rules for MT presupposes, apart from more tools, that the classification of adjectives is available or provided by some tool or resource. Additionally, the implementation presupposes that the system is able to recognize whether a modifying noun has to be translated into an adjective or into a PP (see rule 2 in section 5.3.2.2), according to the information included in the lexical resource.

5.3.2.1 Rules for the Smartcheck

Rule 1

If a noun or a PP precede the head noun, ask the editor to check the sequence with the following message: “Word_Order: Controllare l’ordine degli elementi nella frase.” (Word_Order: Check the order of the elements in the sentence.)

Rule 2

If a noun or a sequence of nouns follow the head noun, ask the editor to check the translation of the structure with the following message: “Word-Order: Controllare la categoria sintattica degli elementi della frase.” (Word_Order: Check the part of speech of the elements in the sentence.)

Rule 3

If one of the sequences listed below are detected, ask the editor to check the order of the words in the sentence with the following message: “Word_Order: Controllare l’ordine degli elementi nella frase.” (Word_Order: Check the order of the elements in the sentence.)

$\mathrm{N+ADJ^{+}+N;}$

$\mathrm{ADJ^{+}+N+N;}$

$\mathrm{ADJ\ +ADJ^{+}+\ N+N^{+}.}$

Rule 4

When a named entity occurs in the target text and is preceded or followed by an adjective or a PP that modifies it, highlight the sequence and ask the editor to check the order of the constituents with the following message “Word_Order: Controllare l’ordine degli elementi nella frase.” (Word_Order: Check the order of the elements in the sentence.)

5.3.2.2 Rules Applied to MT

Rule 1

If there is an adjective modifying a noun in English and the adjective is a quality adjective, then the order in the target language should be noun adjective.

$\mathrm{ADJ_Q+N} \rightarrow \mathrm{N + ADJ_Q}$

Rule 2

If there is a noun preceding another noun in English, and the first noun modifies the second, invert the order and convert the noun into an adjective phrase or a PP.

$\mathrm{N1+^{modifies}N2} \rightarrow \mathrm{N2+(ADJP|PP_{N1})}$

In this chapter we analyzed the errors belonging to the “word order” category and generalized the most common. This way, we were able to provide a solution to some frequent errors. In chapter 6, we will take into account errors belonging to the second category analyzed in this study, namely “agreement.

6 AGREEMENT

6.1 INTRODUCTION

Agreement, in linguistics, is the morphosyntactic covariation of two or more words in a sentence. The word form and the morphemes change in order to agree in a particular feature with the word they are related to and to bear similar information. Grammatical agreement is related to morphological, syntactic and semantic aspects. The term can be used to describe the covariation of different word pairs:

- the subject and the verb;
- the determiner and the noun;
- the adjective and the noun;
- sequences of verbs;
- the pronoun and its antecedent.

The features of agreement are gender, number, person, and case. “By gender is meant a grammatical classification of nouns, pronouns, or other words in the noun phrase, according to certain meaning-related distinctions, especially a distinction related to the sex of the referent” (Quirk et al.: 1985, 315). The gender feature refers to the feminine, masculine, or neutral aspect of a constituent. The number feature takes into account the singularity or the multiplicity of the constituent. The values of this feature can be singular, plural or dual. Person agreement refers to the selection of the verb person, first, second, or third.

According to Wunderlich (2013:2), nouns and pronouns are the controllers of an agreement relation because they bear relevant information that determines the covariation. For example, nouns include information about the gender and about the number, pronouns carry information about the number and the person, and in some cases also about the gender, such as in “his” and “her”, in English. Gender classification is semantically based because feminine nouns denote females, and masculine nouns denote males. However, when the entity is inanimate or abstract, such as “moon”, “sun”, “bridge”, “mountain”, the distinction is not semantic but grammatical. The other elements, that do not carry relevant information for the agreement, are the targets, or controllees, for example determiners and adjectives. Within the noun phrase, for the same feature, the value of the noun and that of the related elements have to correspond.

In the second section of this chapter we will present the way agreement works in English and Italian. In the third we will analyze how it is solved in MT. In the fourth section we will take into account the errors annotated and will present possible solutions to address the issue.

6.2 AGREEMENT IN ENGLISH AND ITALIAN

6.2.1 Gender in English and Italian

With regard to agreement, English and Italian present different mechanisms. This is due to the fact that Italian has a richer inflectional morphology than English, and that gender, number and person are morphologically marked in different ways in the two languages.

In English, the gender distinction is not always marked in the inflection of a word. For instance, it is not marked in determiners, such as “the”, “a” and “an”, that have the same form for the feminine and the masculine. In determiners such as “that”, “this”, “those”, “these”, “my”, “your”, “our”, “their”, the gender is not morphologically marked, even if the number (in “that”, “this”, “those”, “these”) and the person (in “my”, “your”, “our”, “their”) are marked. In the pronouns “his” and “her” gender and person are marked.

Gender is also not morphologically marked in adjectives, in English, as these do not change their form depending on the gender of the noun they modify (see example 1). The same happens with past participles (see example 2).

1. The <u>old</u> lady
   The <u>old</u> man
2. The young woman was <u>found</u> guilty.
   The young man was <u>found</u> guilty.

Pronouns express natural gender distinction in the 3rd person with the forms “he”, “himself”, “his” for the masculine and the forms “she”, “herself”, “hers” for the feminine. Additionally, the relative pronouns “whom” and “which” distinguish between animate and inanimate entities.

In English nouns, however, gender distinctions are more complex. There are no morphologically-marked gender distinctions in nouns and they are not classified grammatically, even if some of them are classified semantically, according to this feature. In the taxonomy of nouns, first of all, we can distinguish between animate and inanimate nouns. The former are those that denote people, animals and living beings, while the latter refer to entities that are not alive. Among animate nouns, some are morphologically marked for gender, i.e. a morpheme changes in order to express the gender. Some are morphologically unmarked, i.e. the word form changes completely in order to express the gender distinction, which means that the distinction is lexically-based and not morphological. Let us look at some examples of these two cases in example 3 and 4.

3. Nouns morphologically marked for gender:
prince – princess,
waiter – waitress,
widower – widow

4. Noun morphologically unmarked for gender:
father – mother,
boy – girl,
brother – sister

As we can see in 3, the suffix "–ess" is added in the first two examples, to mark the feminine gender. In the third example, the suffix "–er" is added to mark the masculine form. On the contrary, in 4, the masculine and feminine forms are not related to each other and are completely different.

A number of nouns, such as "artist", "friend", "student", "doctor", "cook", can be used for both men and women, having a dual gender. The class is increasing due to the need not to sexually connote some words that refer to professions traditionally done by men or women, for instance "fisher" instead "fisherman".

Some nouns can be considered animate when a personal relationship with the referent is highlighted, and inanimate when no relation is present or when the class of the referent is taken into account (see examples 5 and 6). This determines the selection of the pronoun referring to the noun, that is masculine or feminine if the noun is considered animate or an individual, and neutral if the noun is inanimate or a class.

5. The baby needs all the comfort of its environment.
Her baby was born yesterday. His name is John.

6. The cat is a small animal; it can live in apartments.
My cat is white. She is lovely.

In Italian gender is expressed in all determiners. Articles have a masculine and a feminine form, e.g. “il”, “la”, “i”, “gli”, “le”. In possessives and demonstratives, gender is also morphologically marked, e.g. “mio”, “mia”, “tuo”, “tua”, “questo”, “questa”, “quegli”, “quelle”. Since determiners overtly express gender, when there is a contraction between the preposition and the article the contracted form expresses gender too, e.g. “della” (“di” + “la”), “agli” (“a” + “gli”), “nello” (“in” + “lo”).

In adjectives, gender is morphologically marked and agrees with that of the noun the adjective modifies. The majority of adjectives in Italian have a masculine form ending in “–o” and a feminine ending in “–a”. Adjectives ending in “–e” have the same form for the masculine and the feminine, e.g. “instabile”, “eccezionale”, “enorme”. Past participles can express gender in the same way adjectives do, when they are used in predicative position as we will explain in section 6.2.4.

With regard to nouns, a distinction is made between real gender, i.e. the gender motivated by the sex of the referent, and grammatical gender, when the referent does not have a sex because it is inanimate. In animate nouns, gender is morphologically marked by the ending (“-o” for the masculine and “–a” for the feminine) and some suffixes, such as “–tore” for the masculine, and “–essa” and “-trice” for the feminine. There are some nouns that have a single form for the masculine and the feminine. In this case gender is made apparent by the form of the determiner or the adjective(s). This happens with some nouns ending in “–e”, such as “preside”, nouns with the suffixes “–ista”, such as “dentista”, with the suffix “–iatra”, such as “odontoiatra”, and some nouns ending in “–a”, such as “atleta” and “collega”.

With regard to inanimate nouns, they are usually divided into masculine and feminine according to a classification of entities depending on their nature. Of course, there are many exceptions, but the taxonomy is an attempt to describe grammatical gender in Italian. In the table, the determiner was added in order to make the gender visible.

Categories of masculine and feminine nouns in Italian (Serianni, 1989: 106-109)	
Categories of masculine nouns	Categories of feminine nouns
Trees: il melo, il pero, il ciliegio.	Fruits: la mela, la pera, la banana. Exceptions: tropical fruit: il kiwi, il mango.
Metals and chemical elements: l’oro, l’azoto, il piombo.	Cities, islands, regions, continents: la Roma, la Corsica, la Lombardia, l’Oceania.
Cardinal points: il nord, il sud.	Military functions: la guardia, la sentinella, la pattuglia, la vedetta.
Months, weekdays: il febbraio, il sabato. Exception: la domenica.	Sciences and abstract disciplines: la matematica, la sociologia.
Seas, mountains, rivers, lakes: il Mar Mediterraneo, il Monviso, il Po, il Garda.	
Wine brands or types: il moscato, il prosecco, il lambrusco.	

Table 1.

For the nouns that do not belong to any of these classes, it is possible to establish whether they are feminine or masculine, according to the morphology.

Nouns endings in Italian (Serianni, 1989: 110-111)	
Masculine nouns	Feminine nouns
Nouns ending in -o: lo zaino, il pennarello. Exception: la eco.	Nouns ending in –a: la casa, la sedia. Exceptions: il tema, il problema, il cinema, il dramma.
Nouns from foreign languages: il bar, lo scotch.	Nouns ending in –i: la crisi, la tesi, l’analisi.
Nouns with the suffix –tore: l’acceleratore, l’evidenziatore.	Nouns ending in –tà e –tù: la società, la schiavitù.
	Nouns with the suffixes –trice, -tite, and –zione: la lavatrice, la dermatite, l’attenzione.

Table 2.

Nouns ending in “–e”, without the “e” being part of a suffix, can be either masculine or feminine, as illustrated below.

7. Masculine: il bicchiere, il dolore, il cognome,
   Feminine: la gente, la fame, la chiave.

6.2.2 Number in English and Italian

As we said above, number in grammar has to do with the number of entities that are denoted by the noun phrase. In the majority of languages, number can be singular, when only one entity is denoted, or plural, when more than one entity is denoted. Some languages also have dual number, that is used when a pair of entities is denoted. In such languages, such as ancient Greek, Sanskrit and Slovenian, there is a particular morpheme marking dual number in nouns and pronouns and, in some cases, there is also a dual form in verbal inflection.

In English, number can be singular, when one entity is denoted, or plural, when more than one entity is denoted. The dual number existed in Old English and there are still some traces of it in modern English, in expressions like “both”, “either”, and “neither”. Demonstratives express number in their morphology, e.g. “this” and “that” for the singular, “these” and “those” for the plural. As for adjectives, these do not express number in their form, as do not participles either. The articles “a” and “an” are only singular, while for the plural other words or expressions like “some” or “a few” are used.

Nouns morphologically express the number with the suffixes “–s” or “–es”. This happens in both animate and inanimate nouns. Some nouns are morphologically unmarked because they have an irregular form for the plural, such as “mice” (plural of “mouse”), “children” (plural of “child”), “people” (plural of “person”). There are nouns that only have a singular form. Such is the case of uncountable nouns like “butter” or “salt”. On the other side, there are nouns that only have a plural form, such as “trousers” or “scissors”. Additionally, there are some nouns that are described as collective, because they denote a group that can be referred to as the sum of its parts or as a whole. Some examples of collective nouns are “police”, “cast”, “class”, “family”, and “staff”. The verb is usually plural when the group is considered as a collection of individuals, and singular when the focus is on the collectivity.

8. Police has recently arrested the thief.
Police have blocked the street after the explosion.

In Italian, number is expressed morphologically in determiners, adjectives, pronouns and nouns. Articles, demonstratives, and possessives have a specific form for the plural, e.g. “i”, “le”, “questi”, “queste”, “miei”. Adjectives and past participles in predicative position also have a plural form. In nouns, the last vowel changes to mark the plural and becomes a “i” for the masculine, such as in “amici” (plural of “amico”), and “e” for the feminine, such as in “matite” (plural of “matita”). There are some nouns that have the same form for the singular and the plural, such as “città” and “virtù”, or have an irregular form, such as “uomini” (plural of “uomo”) and “uova” (plural of “uovo”). Like in English, some nouns are only used in the singular, e.g. “fame”, “pazienza”, and some are only used in the plural, e.g. “occhiali”, “manette”.

Foreign words that are not translated into Italian do not have a different form for the plural. Therefore, the singular form is used both in the singular and in the plural, while the determiner and the verb agree with the number of the noun even if it does not overtly show the value of number.

9. How do websites use cookies?
Come i siti web usano i cookie?
10. I have two computers.
Ho due computer.

6.2.3 Person in English and Italian

Verbs have three persons both in English and in Italian. With regard to verb conjugation, in English only the third person singular of the present is morphologically marked with the morpheme “–s” or “–es”, while other forms are identical. In other tenses, no form is marked. The expression of the person feature is achieved by the presence of the subject, either a personal pronoun or a noun, since in English the subject of a sentence must always be overtly produced. In Italian, each person has a different form, in all the tenses.

6.2.4 Agreement in Different Word Pairs in English and Italian

As we mentioned above, agreement can involve different word pairs in a sentence. In the beginning of this chapter we listed some of these pairs. In this section we will analyze some relevant cases for the study presented in this work.

Subject – verb agreement: the verb must agree in person and number with the subject (see example 11). As mentioned above, while in English the verb form only changes in the third person singular in the present tense, in Italian the verb form is different for each person and number in all the tenses. In consequence, in Italian, it is not necessary to overtly express the subject of the verb as the information can be derived from the verb form and, therefore, in these cases, the element determining the form of the verb is not visible (see example 12).

11. Mary and John live in Lisbon.
    Mary e John vivono a Lisbona.
12. We have a dog.
    Abbiamo un cane.

NP agreement: in the noun phrase, the modifiers (13), the determiners (14 and 15), the possessives (16), and the demonstratives (17) must agree in gender and number with the noun. In these cases, the agreement in Italian is more complex than in English, because the right morphological form has to be selected.

13. The red shirt is in the wardrobe.
    La camicia rossa è nell’armadio.
14. The chair is in the bedroom.
    La sedia è in camera da letto.
15. A letter arrived for you.
    È arrivata una lettera per te.
16. His party is on Saturday.
    La sua festa è sabato.
17. I bought these apples.
    Ho comprato queste mele.

Anaphora – antecedent: the anaphoric constituent must agree with its antecedent. When the antecedent is a noun phrase, the pronoun in the anaphora must agree in gender and number with it (see example 18. Please note that the personal pronoun was used in Italian to show the agreement, but could be omitted.) The adjectival constituents depending on the anaphora must agree too with the antecedent (see example 19). This is the case even if the pronoun in the anaphora is not expressed because it can be omitted, as it happens in Italian, the same being true for the verb form (see example 19).

18. John’s youngest nephew is Bob. He is four.
Il nipote più piccolo di John è Bob. Egli ha quattro anni.
19. Mary is Ann’s daughter. She is blond and tall.
Mary è la figlia di Ann. È bionda e alta.

Among the pronouns, we can mention the case in which the anaphoric element is a relative pronoun, because it has some particularities. In English, the form of the relative pronouns changes depending on the antecedent being an animate or inanimate noun and agrees in gender and number with it (see examples 20 and 21). In Italian we can distinguish relative pronouns that do not have a gender and number, such as “che” and “cui” (see examples 20 and 21), and relative pronouns that have a gender and number such as “il quale”, “la quale”, “i quali”, “le quali”, and, therefore, agree in these features with their antecedent (see example 22).

20. The books that are on the table are mine.
I libri che sono sul tavolo sono miei.
21. The man who lives in that house is very old.
L’uomo che vive in quella casa è molto anziano.
22. Those who finished the test can leave the classroom.
Coloro i quali hanno finito l’esame possono uscire dall’aula.

Noun - past participle agreement: with regard to the past participle in Italian, we have to analyze it more thoroughly. The past participle can vary in gender and number as an adjective, as we saw in the previous section. However, it varies only in some cases. When it is used as an adjective in a noun phrase, for instance, it varies and agrees in gender and number with the noun it modifies (see example 23). When it is used as a part of a verb form, agreement restrictions vary depending on whether the auxiliary is the verb “essere” (to be) (example 24), whether the past participle is used in a passive construction (25), or whether the auxiliary is the verb “avere” (to have) (26,27,28). In the former contexts, the past participle agrees with the subject, in the latter it does not and the form used is the masculine singular (26). However, there are two cases in which the past participle form varies: when there is an anaphora (see example 27), and when the object is a clitic pronoun that precedes the verb (see example 27 and 28).

23. I soldi spesi non erano miei.
[The money spent was not mine]
24. Maria è andata a scuola.
[Maria is gone to school]
25. La mela è stata mangiata dal bambino.
[The apple is been eaten by the child]
26. Anna aveva mangiato tutte le mele.
[Anna had eaten all the apples]
27. Le mele, le ho mangiate tutte io.
[The apples, them had eaten all I]
28. Ci avevano visti.
[Us had (3rd person plural) seen]

6.3 AGREEMENT IN MT

Agreement is a complex issue in MT and a frequent source of error. Errors can occur both in the analysis of the source text or in the generation of the target text. In the former, the system, in case of error, does not extract relevant information about the gender, the number or the person in the source text and is therefore unable to provide crucial information to the module which generates the translation. In the latter, although the system extracts the correct information regarding the relevant agreement features in the source text, it is unable to generate the correct output in the target language. This can be due to shortcomings of the generation module, for instance when the system does not have information to solve the problem that is given to it, as it happens when a word is rare and the system does not have information on whether it is masculine or feminine, for example, or when the system cannot generate the correct inflected form. Let us consider some specific difficulties for MT systems related to agreement.

The first difficulty we want to mention is the fact that a word can be feminine in one language and masculine in another, or singular in one and plural in another. The system must have this information in a lexical resource. However, rich lexical resources are expensive and require time and effort to be always updated and complete. Additionally, lexicon is open and constantly changing, making it difficult to achieve completeness and accuracy in lexical resources.

The second difficulty amounts to the fact that the source and target languages can have contrasting morphological systems, as in the case of Italian and English, since Italian has a richer inflectional morphology than English. When the source language has a richer morphology, the source text contains more information than what is needed for the generation of the target text and, therefore, the system is able to translate correctly. On the contrary, when the target language has a richer inflectional morphology than the source language, the system does not find in the source text the information needed for the generation of the target text. The system must recognize the form in the source text and be able to select the correct one in the target language with the correct values for Gender, Number, and Person features, which is often not the case.

Another difficulty consists of assessing the correct dependency between constituents in long or complex sentences. The structure of a noun phrase can be ambiguous due to the position of the constituents. This happens when a word can agree with more than one word co-occurring with it in a sentence, for example when an adjective can modify both the head noun and another modifier (see examples 29 and 30). This situation is more common when the adjective is not morphologically marked in the source language, for example.

29. Old books shelves
Scaffali di vecchi libri
Vecchi scaffali di libri

30. A precious ring box
Una scatola di un prezioso anello
Una preziosa scatola per anelli

As we can see from the examples 29 and 30 above, the adjectives “old” and “precious” can refer both to “books” and “shelves”, in 29, and to “ring” and “box”, in 30. While in English the adjective is not morphologically marked, in Italian it has to agree with one of the two words, so it has to be morphologically marked for gender. This means that the English structure is always ambiguous, while in the translation into Italian the ambiguity needs to be solved. The SMT system has to necessarily disambiguate the NP and does not have enough information in the source text to do so. Even if the first option seems the more obvious in the two cases, the second one cannot be excluded, because the meaning of the sentence depends on the context, and it is one of the possible readings of the English structure.

Another difficulty is the fact that named entities in the two languages can have different genders. In English, for example, named entities do not have a gender when they refer to an inanimate entity (when they refer to an animate entity they express the gender), while in Italian they do and it is not always apparent what it is by taking into account the word form, specially when it ends with a consonant or a vowel that is not characteristic of the feminine or the masculine, i.e. when it is not an "–a" or an "–o".

31. The Big Ben
    Il Big Ben
32. The famous Buckingham Palace
    Il famoso Buckingham Palace

Even if there is a classification for some entities (see table 1 and 2), such as countries, geographical nouns and topographic nouns, there are a lot of exceptions. In those cases in which the named entity includes a common noun referring to it (both when it is in the target language, such as in "Piazza Maggiore", and when it is not translated, such as in "Buckingham Palace"), the common noun's gender can determine the gender of the entity (see 32, the gender of "palazzo" determines the gender of "Buckingham Palace"). Additionally, named entities in a foreign language are not always translated or used in Italian in the same way. For instance, "The Shard" is sometimes used in Italian with the English determiner, while when it is used with the Italian determiner, the masculine form is used. "Stonehenge" is usually used without a determiner. There are no general rules applying to all cases, they are by nature idiosyncratic.

Another difficulty, particularly at Unbabel due to the specifications of the translation process followed, is the segmentation of the texts translated, i.e. the division of the text in paragraphs or sentences before the MT is done and during the post-edition. Even if segmentation makes the process faster and easier for the editor, it is possible that two or more words that must agree are in two different segments. Therefore, it is sometimes impossible for the Editor to assess what the constituent must agree with, without having access to the complete source text. There are cases in which the segmentation is correct, but it is still impossible to establish the agreement, as it happens, for instance, in lists.

6.4 AGREEMENT ERRORS IN THE *CORPUS*

There are many agreement errors annotated in the *corpus*. On one side, the errors from this category are apparent, and the editors usually correct them without spending too much time, as in the majority of the cases the correction of agreement errors only involves changing a character and checking contiguous constituents. On the other side, agreement errors are common and, if the Editor happens not to notice one of them, they cause the text to be considered a sloppy translation. This is the reason why agreement errors are considered severe even if they do not prevent to understand the text. Given all this, it is useful to automatize the agreement in post-edition as much as possible, because it is a common and very visible error, that can have a major impact on the translation quality. In the annotated texts, there were cases in which both number and gender were not correct. Since it is not possible to mark the two errors in the annotation tool, only the most relevant one in terms of severity and of interest for this study was marked.

Agreement errors classification	
Gender agreement errors	137
Number agreement errors	19
Person agreement errors	3
Total	159

Table 3.

Among the 159 agreement errors identified in the annotation task, only approximately 1/3 was detected by the Smartcheck, as presented in Table 4.

Agreement errors detected by the Smartcheck		
Category	Total number	Number of errors detected by the Smartcheck
Agreement errors	159	51
Gender agreement errors	137	51
Number agreement errors	19	0
Person agreement errors	3	0

Table 4.

As we can see in the table, all the agreement errors detected were gender agreement errors. All the errors detected by the Smartcheck were errors in the gender agreement between the determiner (article) and the head noun in a given NP. Some of them were repeated, because a few frequent words were recurrently incorrectly translated, for instance the English word "petition" was translated correctly as "petizione" but the determiners specifying it were all masculine, while the word is feminine in Italian. The Smartcheck detected all the errors related to the word "petition" and some other simple cases. It did not detect, however, any gender agreement errors between a noun and a past participle in predicative position, like those in 33 and 34.

33. Your petition is saved.
*Il tuo petizione viene salvato.
La tua petizione viene salvata.

34. The library is recognized.
*La libreria è riconosciuto.
La libreria è riconosciuta.

6.4.1 Gender Agreement Errors in the *Corpus*

Gender agreement errors were divided into subcategories, in order to study possible solutions to address them.

Gender agreement errors in the *corpus*	
Number of unique errors	81
Number of gender agreement errors involving constituents in the NP	61
Number of gender agreement errors involving constituents in the VP	20
Total	137

Table 5.

As in the word order errors reported and analyzed in Chapter 5, among the gender agreement errors observed some were repeated. We found 81 unique errors, while the remaining 56 were repeated and were mainly errors involving the agreement with the words "petizione" and "applicazione". The 81 unique errors were divided into those that occurred within constituents of the NP and those that involved constituents in the VP. The agreement errors that involved constituents in the NP occurred in modifiers such as adjectives and past participles used as adjectives, in determiners, possessives and demonstratives. Those involving constituents of the VP occurred in adjectives and past participles.

6.4.1.1 Gender Agreement Errors Involving Constituents in the NP

Specifiers and modifiers in the NP can occur in a prenominal and postnominal position in the target language, as we already saw in chapter 5. In this study, it is helpful to use this distinction in order to analyze the errors and understand what could be the contribution of the Smartcheck in trying to automatically solve the issue in the post-editing stage.

Gender agreement errors in different linear order positions	
Errors in prenominal position	50
Errors in postnominal position	11
Total	61

Table 6.

As we can see from Table 6, the majority of errors are observed in prenominal position. This can be explained by the fact that the majority of the gender agreement errors in this subcategory occur between the article and the noun. Another factor to take into account is that the word order in the target language is not always the correct one, and, therefore, the two errors co-existed.

Gender agreement errors involving different constituents in the NP	
Determiners	33
Determiner + adjective	14
Adjective	12
Quantifier	2
Total	61

Table 7.

The majority of the errors annotated amount to agreement between the determiner and the head noun (see example 35). However, as we saw, the majority of these errors are detected by the Smartcheck. We also have to highlight the fact that 20 of these 33 errors occurred in NP’s headed by a named entity (36). As we said before, named entities are a challenge in MT, because it is not possible to list them all in a lexical resource and because they sometimes show an idiosyncratic linguistic behavior, distinct from that of common nouns, in particular with regard to properties such as gender.

35. The defeat
*Il sconfitta
La sconfitta
36. The Chain Bridge
*La Ponte delle Catene
Il Ponte delle Catene.

14 errors involved both the determiner (usually the article) and the adjective preceding the head noun.

37. A nice picnic
*Una bella picnic
Un bel picnic

12 errors involved only the adjective co-occurring with the head noun. In this subcategory, two errors occurred in prenominal position and 10 in postnominal position. As expected, the 10 postnominal errors in Table 6 correspond to those concerning an adjective, since articles, demonstratives, possessives, and quantifiers in Italian always precede the noun. In three cases there was also a named entity involved. There were eight cases in which the error occurred in an adjective (see example 38), and four in which it occurred in a past participle used as an adjective (39).

38. You will then visit the palace, famous for its architecture.
*Visiterai quindi il palazzo, famosa per l’architettura.
Visiterai quindi il palazzo, famoso per la sua architettura.
39. A hand-picked collection
*Una collezione raccolto a mano
Una collezione raccolta a mano

Among the errors involving constituents in the NP, there were eight errors occurring in a phrase where an adjective ending in “–e” was present.

40. After a spectacular day spent surrounded by nature
*Dopo una giornata spettacolare passato circondato dalla natura
Dopo una giornata spettacolare passata circondato dalla natura

In conclusion, the most problematic cases for gender agreement in MT are those in which a named entity is the head noun of the noun phrase (see example 41) and when a word not morphologically marked occurs in the NP in the target text (see example 42).

41. The Sistine Chapel
*Il cappella Sistina
La Cappella Sistina
42. A pleasant boat tour
*Una piacevole giro in barca
Un piacevole giro in barca

6.4.1.2 Gender Agreement Errors Involving Constituents in the VP

As we said above, some constituents of the VP must agree in gender with the subject expressed in the NP, i.e. adjectives in predicative position (see example 43) and past participles (44).

43. This rose is red.
Questa rosa è rossa.
44. The book is titled “Pinocchio.”
Il libro è intitolato “Pinocchio”.

Gender agreement errors involving different constituents in the VP	
Adjective	6
Past Participle	14
Total	20

Table 8.

As we can see from the table above, the majority of the errors involving constituents in the VP occurred in the agreement of past participles. In the 14 errors, only one occurred in a past participle of a complex active verb form (example 45), the other 13 errors occurred in past participles in passive constructions (46). In these sentences, the past participle follows the verb "to be", that can occur in different tenses and moods.

45. Cards which we hand-picked
*Carte che abbiamo raccolte a mano
Carte che abbiamo raccolto a mano

46. The magnificent church was designed
*La magnifica chiesa è stato progettato
La magnifica chiesa è stata progettata

There was only one case (example 47), among those annotated, in which the verb preceding the past participle was the verb "viene", from the verb "venire" (to come). The verb "venire", as well as the verb "andare" (to go), is sometimes used as the auxiliary verb in passive constructions.

47. Your request is saved.
*La tua richiesta viene salvato.
La tua richiesta viene salvata.

The other errors occurred in an adjective following the copula verb heading the VP. The verb was in different tenses and in three cases there was an adverb between the verb and the adjective.

48. His behavior is disorderly.
*Il suo comportamento è disordinata.
Il suo comportamento è disordinato.

6.4.1.3 Possible solutions

Gender agreement errors can be avoided if the syntactic dependency among the constituents in a phrase is correctly identified. An accurate parser is able to identify the correct syntactic dependency between the constituents both in cases in which all the modifiers modify the head noun, and in cases of inlaid modification (see examples 35, 36, and 37 in chapter 5). After assessing the dependency between two constituents, the tool should check whether the value of the feature Gender in the modifier agrees with that of the modified constituent. The ideal situation would be having a parser in English establishing the tree, and a parser in Italian reproducing it, before checking the value of the feature. Since in this study, for the reasons and motivations already discussed, we focus on the target language, we will exclude this possibility and concentrate only on the information we can acquire from the target text, immediately after MT. The dependency can be analyzed directly in the target text and a dependency tree can be drawn. The parser used at Unbabel correctly identifies the value for the feature Gender in the majority of the cases. When an incorrect sentence or phrase was analyzed by the parser, it was able to correctly identify the value for the feature Gender for the separate constituents. For example, when the phrase “la bella palazzo” (the beautiful palace) was analyzed, the parser recognized “la” and “bella” are feminine, while “palazzo” is masculine. The parser always identifies correctly the gender of the elements when the error involves a noun and an adjective or a past participle, while it does not always do so, when a noun is preceded by a determiner with the wrong gender. For example, when the wrong phrase “le concerti” (the concerts, in which the determiner is feminine and the noun is masculine) is analyzed, the noun is identified as feminine instead of masculine. The performance of the parser in the analysis of an incorrect sentence or phrase is crucial in understanding which solutions are possible to implement and, on the contrary, which solutions require different strategies to be implemented. When an adjective ending in “–e” occurs, the parser does not provide any information about the gender, but only about the number. With regard to foreign words used in Italian, there are some cases in which the parser only identifies the word as a noun, as for “cookie”, in some other cases it correctly identifies its gender, as for “feedback”, while, in some other cases, it does not, as for “picnic” (the information provided by the parser is that the word is feminine, while it is masculine in Italian). With regard to named entities, the parser recognizes them and classifies them as proper nouns, without providing any information about their gender.

Taking into account all the aforementioned considerations, with regard to gender agreement errors that occur in frequent words, such as “petizione” (petition) and “applicazione” (application), the parser correctly identifies the gender of the two words. Even so, a more accurate and updated lexical resource can be used, in order to have gender information regarding the words that are frequent in gender agreement errors and are used by several clients that may be not correctly analyzed by the parser. If the information is present in the lexical resource, the value for the feature Gender can be checked for agreement with that of the specifiers and modifiers. With regard to rare words used in Italian in particular texts, a solution to address this can be extracting a list of such words depending on the client and having the Smartcheck highlight the word and ask the Editor to check the agreement. Some false positive cases could be present, but since the words are not frequent, we do not expect these false alarms to affect the post-editing in a relevant way.

With regard to adjectives ending in "–e" in Italian, several errors were detected in sequences of modifiers that included one of such adjectives. Since these adjectives can be both feminine and masculine, the parser does not provide any information about their gender and the form that should be used. Therefore, the solution can be having the Smartcheck highlight the cases in which such an adjective occurs in sequences of modifiers, because it is likely that the modifiers following it do not have the correct value for the feature Gender.

As we already mentioned, the parser classifies named entities as proper nouns, but does not analyze the feature Gender. Therefore, the only solution to address errors involving this type of nouns would be highlighting them when they occur in Italian. A list can be added to the lexical resource, with information about the gender of the named entities that represent the most regular and systematic cases, such as topographical named entities. The list can be then updated for specific clients or kinds of texts.

The classification of masculine and feminine nouns provided in section 6.2.1 helps us to come up with some rules for the Smartcheck. From Table 2 we can understand that nouns ending in "–tore" are always masculine, while nouns ending in "–tù", "–tà", "–trice", and "–tite" are always feminine. The following rules can account for the errors occurring in phrases headed by such nouns.

Rule 5

If a noun ending in "–tore" occurs in the target text, check if its specifiers and modifiers are masculine.

$\mathrm{SPR}^{*} + \mathrm{N}_{-\mathrm{tore}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\mathrm{masc}} + \mathrm{N}_{-\mathrm{tore}} + \mathrm{MOD}^{*}{}_{\mathrm{masc}}$

Rule 6

If a noun ending in "–tà", "–tù", "–trice", or "–tite" occurs in the target text, check if its specifiers and modifiers are feminine.

$$\mathrm{SPR}^{*} + \mathrm{N}_{-\text{tà}|-\text{tù}|-\text{trice}|-\text{tite}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\text{fem}} + \mathrm{N}_{-\text{tà}} + \mathrm{MOD}^{*}{}_{\text{fem}}$$

With regard to foreign words, they are masculine in the majority of the cases but, as we already mentioned, the parser does not identify whether a given noun is a foreign word or not, and it does not always identify its gender. However, no Italian noun ends in a consonant, therefore it is possible to say that if a noun ends in a consonant, it is a foreign word. Therefore, the rule for the Smartcheck can be the following[11].

Rule7

If a noun ending in a consonant occurs in the target text, check if its specifiers and modifiers are masculine.

$$\mathrm{SPR}^{*} + \mathrm{N}_{-\text{consonant}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\text{masc}} + \mathrm{N}_{-\text{consonant}} + \mathrm{MOD}^{*}{}_{\text{masc}}$$

6.4.2 Number Agreement Errors in the *Corpus*

Number agreement errors involving different constituents	
Determiner + noun	7
Adjective in a partitive genitive	1
Constituents in the VP	9
Foreign words	2
Total	19

Table 9.

Among the number agreement errors annotated in the *corpus* we can distinguish different classes depending on the constituent presenting the error. In seven cases, the error occurred between the determiner and the head noun (see example 49). There were some

[11] Please note that the rule presented does not cover all the cases in which a foreign word occurs, since there are foreign words ending in a vowel, such as "cookie". However, the rule accounts for all the cases in which the foreign word ends in a consonant and, therefore, significantly reduces the number of errors.

cases in which the Smartcheck detected the error, but the percentage is smaller than in gender agreement. In one case there was an adjective occurring between the determiner and the head noun in the source text (see example 50). In two cases the head noun of the phrase was a named entity.

49. The mystery of its pasts
*Il misteri del suo passato
I misteri del suo passato

50. The misty streets of London
*Il strade nebbiose di Londra
Le strade nebbiose di Londra

One error occurred between the adjective and the noun, in particular it involved adjective modifying the noun in a partitive genitive (example 51).

51. One of the world's most unique cities.
*Una delle città più unica al mondo.
Una delle città più uniche al mondo.

In 9 cases the error was annotated in a constituent of the VP. Among these errors we could distinguish two involving a past participle related to the subject of the main verb (52 and 53). One error involved an adjective following the verb "to be" in the VP (54), and six the main verb and its auxiliary. Among these six cases in which the error occurred in the main verb, in three cases the main verb was preceded by the modal verb "potere" (can) (see example 55). In three cases the constituent in which the error occurred is the past participle following the verb "to be" in the passive construction (56). The errors involving constituents in the VP occurred in verb forms both in the main and in the subordinate clause. We can also point out that the subordinate was a relative clause (example 57). Considering this, we can hypothesize that the generation of the error could be due to the presence of the relative pronoun "che", that, as we saw above, is not morphologically marked for number. The errors annotated involved both active and passive constructions. In the former the errors occurred in the auxiliary verb "avere" or in the main verb "potere". In the latter they occurred in the past participle. Two errors occurred in a sentence that had a quantifier, namely "nessuno" and "chiunque" (see examples 55 and 56). In the example in 55, the quantifier "nessun" modifies two head nouns coordinated with the conjunction "or". This fact is relevant since, while in English, when the quantifiers "no" or "any" modify two nouns the subject is considered plural and, therefore, the verb is plural, while in Italian the verb must be singular. In the example in 56, the quantifier “chiunque” is singular and therefore not only the main verb (“abbia”) but also the past participle in the complex infinitive verb form (“essere fatto”) and the noun depending on it (“prigioniero”) must be singular.

52. The Orchestra is leading the performance, accompanied by professional ballet dancers.
*L’Orchestra sta eseguendo la performance, accompagnati da ballerini professionisti.
L’Orchestra sta eseguendo la performance, accompagnata da ballerini professionisti.

53. Exposed partially open-air, guests will be able to see the city.
*Esposto parzialmente a cielo aperto, gli ospiti potranno vedere la città.
Esposti parzialmente a cielo aperto, gli ospiti potranno vedere la città.

54. The prices are not exactly low.
*I prezzi non sono esattamente basso.
I prezzi non sono esattamente bassi.

55. No refunds or re-bookings can be provided.
*Nessun rimborso o riprenotazioni possono essere forniti.
Nessun rimborso o riprenotazione può essere fornito.

56. Anyone unlucky enough to be taken prisoner
*Chiunque abbia la sfortuna di essere fatti prigionieri
Chiunque abbia la sfortuna di essere fatto prigioniero

57. In the war rooms that have been untouched
*Nelle stanze di guerra che sono stati intatto
Nelle stanze di guerra che sono state intatte

As we can see from the analysis presented in the previous paragraph, certain structures are more critical than others. Among such structures we can mention in particular those including a modal verb, a passive construction, a relative pronoun that is not morphologically marked in Italian, and a quantifier.

Finally, among the 19 number errors annotated, in two cases the error involved a foreign word. The suffix “–s” marking the number in English was kept in Italian, and this introduces an error since, as we saw in the section on number in Italian, foreign words in Italian do not vary in form between singular and plural.

6.4.2.1 Possible Solutions

By using the syntactic parser, we saw that the system correctly identifies the value of Number of the modifier and the noun in the NP. Therefore, in the cases in which the syntactic dependency is not ambiguous, the Smartcheck can automatically detect the error after the parser analysis, when the two values do not match.

When the error occurs in an adjective, a past participle or a quantifier in the VP, it is important that the parser correctly analyzes the dependency. When it does, then it is possible to automatically check if the value for the feature Number of the adjective, past participle, or quantifier corresponds to that of the NP it refers to. The parser correctly analyzed the dependency in both the cases annotated in the *corpus*, with the past participle “accompagnati” modifying a singular noun, in the example 52, and with the past participle “esposto” and the plural subject “gli ospiti” and, in example 53.

With regard to the quantifiers “nessuno” and “chiunque”, as we already mentioned they are always followed by a singular verb form. Therefore, a rule regarding these constituents can be made for the Smartcheck:

Rule 8

If the quantifier “nessuno” or “chiunque” are part of the subject of a sentence, then the head verb form of the sentence must be singular.

In the parser for Italian the feature Number is not provided for named entities, therefore it is difficult to check the agreement in constituents involving proper nouns automatically. The only solution, as we saw for gender agreement, would be having the Smartcheck highlight the named entity and its specifiers and modifiers for the Editor to check them.

With regard to foreign words, since they are not included in the Italian lexicon, they should be highlighted in the Smartcheck with a message asking the Editor to pay attention to the form of the word. Since in Italian there are no words ending in “–s”, if the Smartcheck highlighted all the words with this suffix found in the target text, apart from named entities (that are tagged as PROPN by the parser, and are correctly identified in the majority of cases), only foreign nouns in the plural form, which are incorrect in Italian, will be listed.

Finally, the superlative structure in Italian in the sentence “una delle città più uniche al mondo” is not recognized by the parser, because the plural of the word “città” is the same as the singular. If another word is used, the parser correctly analyzes the constituents and the value of the feature Number can be checked automatically. With regard to the irregular plural forms, a possible solution can be just checking if the lexical resource includes the cases in which the singular and the plural forms are the same. In such cases, since the form can be both singular and plural, it is ambiguous and the best solution is to have the Smartcheck highlight the modifiers of the noun, in order for the editor to check the value of the feature Number.

6.4.3 Person Agreement in the *Corpus*

Person agreement errors	
Number of person agreement errors in a coordinate clause	2
Number of person agreement errors in a subordinate clause	1
Total	3

Table 10.

Three errors of person agreement were annotated in the *corpus*. As we saw earlier in this work, this kind of errors occurs in the agreement between the subject and the verb, and between the subject and the pronoun. The errors annotated involve the pair subject – verb and occurred twice in a coordinate clause (58 and 59) and once in a subordinate clause, more precisely in a relative clause (60).

58. We review these reports daily and can easily evaluate them.
*Esaminiamo queste relazioni quotidianamente e può facilmente valutarle.
Esaminiamo queste relazioni quotidianamente e possiamo facilmente valutarle.

59. I went ahead and reported the issue.
*Sono andato avanti e ha riferito il problema.
Sono andato avanti e ho riferito il problema.

60. We apologize for the issue you are experiencing.

*Ci scusiamo per il problema che si riscontrano.

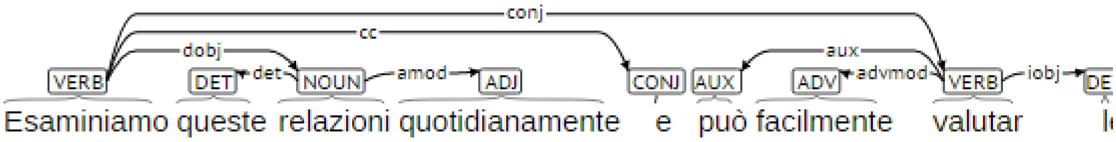
Ci scusiamo per il problema che riscontri.

As we can see from the three errors listed above, in 58 and 59, the third person was used instead of the first, and in 60 the third person was used instead of the second.

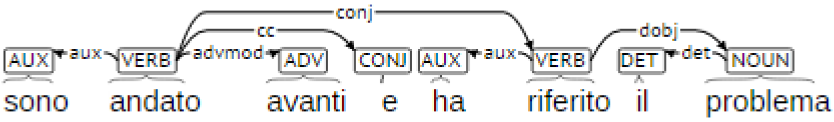
6.4.3.1 Possible Solutions

When the error occurs in a coordinate clause, when the syntactic structure is not complex and the coordination occurs between two main verbs, the parser identifies correctly the dependency and the number of the subject and the verb in the majority of the cases. The syntactic trees of the sentence in 58 and 59 are presented below:

(58)



(59)



Since the coordination cannot be acceptable between verbs with different forms, the solution can be checking if the value for the feature Person of the two verbs and auxiliaries is the same for the two or more constituents.

With regard to the relative clause, the solution could be checking if the value of the feature Number in the verb of the relative is the same as that of the antecedent, that is identified by the parser.

In chapter 6, we considered the errors belonging to the “agreement” category. After analyzing them, we were able to generalize the most common and critical errors. In some cases, we were able to provide a solution to some frequent errors. In chapter 7, we will take into account the third category of errors we focused on in this study, i.e. “tense/mood/aspect”.

7 TENSE/MOOD/ASPECT

7.1 INTRODUCTION

The category “Tense/mood/aspect” includes errors regarding the selection of these three features of the verb form. It is a broad category that takes into account verbs in both main and dependent clauses. Apart from tense, mood and aspect, inflectional features of the verb include also person and number, that were addressed in chapter 6, as a subcase of agreement errors.

According to Quirk et al (1985:96), a general distinction can be made among full verbs, modal auxiliary verbs, and primary verbs. The first are also called lexical verbs, because they bear semantic meaning. Full verbs can only function as main verbs (see example 1). Modal verbs act as auxiliaries (see example 2). Examples of modal verbs in English are “can”, “should”, and “must”. Their contribution regards the expression of modality of the action denoted by the main verb, because they provide information about volition, probability, and obligation. The verbs “do”, “have” and “be” are primary verbs, and these act both as main verbs and auxiliaries, depending on the way they are used in a sentence (see examples 3 and 4).[12]

1. I usually <u>go</u> to work by car.
2. You <u>should</u> study more.
3. <u>Do</u> you usually wake up early?
4. I <u>do</u> my homework every day.

Additionally, verb forms can be described as finite and nonfinite. The former have person and number features, and they agree with the subject (see example 5). The latter do not have those features and, as such, they cannot head a finite verb phrase, they are always dependent on a main verb (see example 6).

5. Yesterday I <u>went</u> to the theater.
   Ieri <u>sono andato</u> a teatro.
6. I like <u>swimming</u>.
   Mi piace <u>nuotare</u>.

[12] This distinction accounts for the majority of the cases, even if there can be exceptions, for instance semi-auxiliary verbs used in periphrastic constructions such as “be about to”, “used to”, etc.

As we can see in the examples 5 and 6, “went” and “sono andato” are the main verbs of the sentences in 5, while “like” and “piace” are the main verbs in 6, as “swimming” and “nuotare” are nonfinite verb forms.

The verb form is considered simple when it consists of only one word (like in example 7), and complex when it consists of two or more words (like in 8, 9,10, and 11). Modal auxiliaries (8), perfective (9), progressive (10), or passive constructions (11) correspond to complex verb forms.

7. See you tomorrow!
Ci vediamo domani!
8. Can I open the window?
Posso aprire la finestra?
9. I had studied Chinese for three years, before moving to China.
Avevo studiato cinese per tre anni, prima di trasferirmi in Cina.
10. Mary is listening to music.
Mary sta ascoltando la musica.
11. All the work was done by Ann.
Tutto il lavoro è stato fatto da Ann.

In section 7.2 we will analyze the features we considered in this chapter. We will consider the concepts of “tense”, “mood”, and “aspect” of lexical verbs, and we will see how they are expressed in English and Italian. We we will also take into account modal verbs in English and Italian and their characteristics. We will not analyze auxiliaries in particular, since, with regard to tense, mood, and aspect, they behave like lexical verbs. In section 7.3 we will analyze the challenges in selecting the right tense, mood, and aspect in MT, while in section 7.4 we will analyze the errors annotated in the *corpus* and provide possible solutions to address them and improve the quality of the translation produced.

7.2 TENSE, MOOD, AND ASPECT IN ENGLISH AND ITALIAN

7.2.1 Mood

Mood expresses the way the communication between speakers is established and the status the speaker has in his/her own communication. Mood is related to the concept of modality. “Modality may be defined as the manner in which the meaning of a clause is qualified so as to reflect the speaker’s judgment of the likelihood of the proposition it expresses being true" (Quirk et al., 1985: 219). Mood has not only a syntactic value in the sentence, but also a pragmatic value, as we can see in the use of imperative to give an order. Moods can be distinguished into finite and nonfinite. Finite moods are those in which person and number features are expressed, such as in indicative or in imperative forms. Nonfinite moods are the so called nominal forms of the verb, such as infinitive, participle and gerund. They take on the mood of the corresponding finite verb. We have to note that nonfinite moods are not considered moods in all the languages, as we will see further below. In the examples below we can see the use of both finite (underlined in 12 and 14) and nonfinite moods (underlined in 13 and 15).

12. Singing, the girl left the room.
    Cantando, la ragazza uscì dalla stanza.
13. Singing, the girl left the room.
    Cantando, la ragazza uscì dalla stanza.
14. Once finished, the painting will be given to John as a present.
    Una volta finito, il dipinto sarà dato in regalo a John.
15. Once finished, the painting will be given to John as a present.
    Una volta finito, il dipinto sarà dato in regalo a John.

With regard to the mood of the verbs in 13 and 15, the gerund in 13 corresponds to "while she was singing", and "mentre cantava", in English and Italian respectively. The past participle in 15 corresponds to "it is finished" and "sarà finito".

A general characterization of the semantic contribution of the finite moods that are present both in English and Italian can be made. The indicative expresses a real and objective action (16), the subjunctive expresses an action that is neither real nor objective, but can be a wish (17) or a hypothesis (18), and the imperative expresses an order (19).

16. John goes to the university every day.
    John va in università tutti i giorni.
17. I wish I could help you.
    Magari ti potessi aiutare.
18. If I could, I would go on holiday.
    Se potessi, andrei in vacanza.
19. Open the window!
    Apri la finestra!

7.2.1.1 Mood in English

In English there are three moods, the indicative, the subjunctive, and the imperative. The infinitive, the participle, and the gerund are not considered moods in English, due to the fact that, as we already said, they do not provide in their form any information about the modality of the action they denote, but take on the modality of the main verb.

- Indicative: it marks the factual status of the predication. It is used to express facts and objective actions, both in main clauses (see examples 20, 22, and 23) and dependent clauses (example 21). The time when the action takes place can vary, being present (20), past (22), or future (23), but the action denoted by the verb is a fact.

  20. He is 30.
  21. I didn’t go to the park because it was raining.
  22. Yesterday we went to the theater.
  23. Next summer we will travel around Portugal.

- Subjunctive: this mood is used as mandatory (example 24), when it is introduced by verbs such as “decide”, “insist”, “order”, “request”, adjectives such as “advisable”, “desirable”, “imperative”, or nouns like “decision”, “order”, “requirement”, or “resolution”. It can also be formulaic in expressions like “Come what may”, “So be it!”, or “God save the king!”. The past subjunctive, that is also called “were-subjunctive”, has a hypothetical or unreal meaning and is used in clauses introduced by “if”, “as if”, “as though”, “though”, “wish”, and “suppose” (examples 25, and 26).

  24. I insist you come and visit me during the summer.
  25. I wish you were here.
  26. If I were you, I would try again.

- Imperative: it is used to express commands and request. The imperative only exists in the second person, singular or plural, while other verb forms are used when the order given involves a first or third person.

  27. Close the door!

7.2.1.2 Mood in Italian

In Italian, there are four finite moods, and three nonfinite moods. The latter are considered moods, even if they do not express any modality of the action, because, like the finite moods, they have tenses and can form a VP in subordinate clauses. They are the infinitive, the participle, and the gerund. They take on the modality of the verb they depend on, or the mood of the corresponding finite mood, as we saw in the previous section.

The finite moods are:

- Indicativo: it is used when the action denoted by the verb is real and objective. It can occur in the present (28), in the past (29) or in the future (30). The mood can be used in both main clauses (28, 29, and 30) and dependent clauses (31).

  28. Mario gioca a calcio.
  [Mario plays to soccer]
  29. Ieri siamo andati a teatro.
  [Yesterday are gone to theater]
  30. Chiamerò domani.
  [Will call tomorrow]
  31. Mario ha detto che è andato al cinema.
  [Mario has said that is gone to cinema]

- Congiuntivo: it is used when the action is not completely real and is not objective, it can be used to express a wish, a fear, a volition, a hypothesis. It is used mainly in dependent clauses, such as in cause clauses, and clauses of condition, concession, time, consequence, and comparison. In the examples below, it is used to express a wish (32), a hypothesis (33), and a concession (34).

  32. Spero tu stia bene.
  [Hope you are well]
  33. Credo che Anna sia uscita.
  [Believe that Anna is out]
  34. Nonostante piovesse, siamo andati in spiaggia.
  [Even if rained, are gone to beach]

- Condizionale: it is used when there is a condition that does not depend on the subject, that can be either real or not (example 35). It can be used in main clauses, for example in conditional sentences (example 36). It can indicate a future action in the past (example 37).

35. Se non dovessi leggere questo libro, leggerei il tuo.
[If not should read this book, would read yours]

36. Domani mi piacerebbe andare in spiaggia.
[Tomorrow me would like to go to beach]

37. Aveva detto che sarebbe tornato.
[Had said that would come back]

– Imperativo: it is used to express an order, a request, an invitation. It only exists in the second person, singular or plural. To express orders given to the first or third person or negative orders other verb forms are used.

38. Passami il libro!
[Pass me the book!]

39. Alzatevi!
[Stand up!]

7.2.2 Modal Verbs in English and Italian

Modal verbs are related to the concept of "modality" that we saw in section 7.2.1. Modal verbs can express permission, obligation, volition, possibility, necessity, and prediction. Modal verbs are defined as auxiliaries, because they express in their form mood, tense, person, and number, while the lexical verb occurs in its infinitive form. This happens in both affirmative and negative forms, without exceptions, as we can see in the following examples.

40. You may be wrong.
Potresti non aver ragione.

41. You shouldn't smoke.
Non dovresti fumare.

7.2.3 Modal Verbs in English

Modal verbs in English provide information on the certainty or probability of an action and express ability, permission, request, or offer.

They are “can”, “could”, “may”, “might”, “shall”, “should”, “will”, “would”, “must”. They are all followed by the base form of the lexical verb. In the negative form, while lexical verbs combine with the auxiliary “do” in the negative “don’t”, modal verbs add the particle “not” and can have a contracted form, e.g. “cannot”, “can’t”, “must not”, “mustn’t”, “should not”, “shouldn’t” (examples 42, and 43). Similarly, in questions, modal verbs do not combine with the auxiliary “do”, but precede the subject and the main verb (examples 44, and 45). This shows that modal verbs form a particular class of verbs with syntactic and specific morphological characteristics that determine the fact that they are considered a category of verbs in between lexical verbs and auxiliaries. The syntactic behavior of modal verbs is similar to that of auxiliaries, but they bear semantic meaning like lexical verbs.

42. I <u>couldn’t</u> find the place
43. You <u>mustn’t</u> smoke during pregnancy.
44. <u>Shall</u> I close the door?
45. <u>Can</u> I borrow your pen?

7.2.3.1 Modal Verbs in Italian

Modal verbs in Italian, as in English, are inflected and precede the base form of the main verb. Among modal verbs we can distinguish three, “volere”, “potere” and “dovere”, with three characteristics in common: they are followed by the base form without any preposition, they have the same subject as the base form, and, when there is an unstressed pronoun, it can precede or follow the modal. These verbs express volition (example 46), permission (47), and obligation (48). Other modal verbs are “preferire”, “solere”, “osare”, “desiderare”, and “sapere”.

46. <u>Voglio</u> farti un bel regalo!
[Want to give you a nice present]
47. <u>Posso</u> chiamarti domani?
[Can call you tomorrow]
48. <u>Devo</u> finire i compiti.
[Must finish the homework.]

7.2.4 Tense

The tense of the verb provides information regarding the time when the action expressed by the verb takes place. In order to correctly understand the concept of tense, we need to explain the distinction between “time” and “tense”. In some languages, such as English, two terms are used to denote the two concepts, while in Italian the term is only one (“tempo”). The concept of time is non-linguistic and can be divided into past, present and future. On the contrary, tense is a grammatical category that establishes a correspondence between the verb form and time. Therefore, the verb can denote a present (49), past (50), or future action (51) relative to the time of utterance. It is also possible that the time the tense refers to does not correspond to the actual time in which the action occurs, considered the time of utterance, like in the example 52.

49. I am going to the library.
Sto andando in biblioteca.
50. I was born in 1991.
Sono nato nel 1991.
51. Next year Ann will start school.
L’anno prossimo Ann comincerà la scuola.
52. In one week they will tell me if I passed the exam.
Tra una settimana mi diranno se ho passato l’esame.

In 49, 50, and 51 the verbs refer to a present, past, and future action, considered the time of utterance. In 52 the verbs “passed” and “ho passato” refer to an action that will take place in the future, but that will happen before another future action expressed by the verbs “will tell” and “diranno”. Therefore, the past tense is used to denote the sequence of events in the future.

The action denoted by the verb can be looked upon depending on the time of utterance, or on a referential level. In the first case, it establishes a relation with the time the speaker says or writes the sentence (53), in the second case a relation is established with the action expressed by another verb (54).

53. I will go to New York next year.
Andrò a New York l’anno prossimo.

54. Two years ago I said I was going to New York the following year.
Due anni fa dissi che sarei andato a New York l’anno seguente.

In both cases the verb expresses a future action, however in 53 the phrase “next year” establishes a relation with “now”, that is the time of utterance. In 54, the phrase “the following year” is related to the time in which the action “said” took place, i.e. “two years ago”.

Additionally, verb forms used to express the tense can be simple, when they consist of only one word (55), and complex, when they involve an auxiliary verb (56).

55. I live in Lisbon.
Vivo a Lisbona.
56. I have been to Paris twice.
Sono stato a Parigi due volte.

7.2.4.1 Tenses in English

English has three main categories that refer to the time when the action occurred: present, past, and future. The tenses can be expressed by simple verb forms, when there is only one verb in the verb form, and by complex verb forms, when they involve the use of an auxiliary. The tenses involving simple verb forms in the indicative are:

- Simple Present: it has a broad use. That is why it is referred to as the “non-past tense”. It is used for facts (57), for repetitive actions (58), for habits (59), and for actions taking place at the time of utterance (60).
  57. The Sun shines.
  58. I go to school every day.
  59. I usually get up at 07.30 a.m.
  60. Here comes the bride.
- Simple Past: it is used to refer to actions that occurred and finished in the past (61 and 62).
  61. Yesterday I cooked a delicious dinner.
  62. The football team lost the match.

The tenses involving complex verb forms in the indicative are:

- Present Perfect: it expresses a situation that began in the past and continues in, or still has an influence on the present time (63). It can also be used when an event has occurred once or several times in a period that precedes the time of utterance (64 and 65).

63. I have lived here since 1990.
64. I have been to England twice.
65. We have always known each other.

- Past Perfect: it expresses an action in the past that preceded another past action (66 and 67).

66. I had already closed the gate when the car arrived.
67. He had just finished the university when his sister got married.

- Future: it is expressed through the modal “will” (68), or the semi-auxiliary “be going to” (69). The former is used to express a future action that is certain, while the latter is used to convey a future action that is a prediction based on the evidence of a present situation.

68. Mary will start school next year.
69. The tree is going to fall.

The subjunctive mood, as we already mentioned in section 7.2.1.1, has a present and past tense. The former is used in formulaic expressions or to express orders (see example 24), while the latter is used in clauses that express hypothesis or unreal actions (see examples 25 and 26).

The infinitive and the gerund, like the indicative, have a simple and a complex form. The former is present (see examples 70 and 72), while the latter is perfect (examples 71 and 73). They have a referential value, so they establish a relation in time with the verb from the principal clause.

70. I hope to see you soon.
71. He pretended to have fallen from the chair.
72. I like listening to music.
73. He denied having eaten all the cake.

With regard to the participle, it has a present and past form. The former is the same form as present gerund (–ing form), but is used with an adjectival function (example 74), while the latter is used in complex verb forms (example 75) and as an adjective (example 76).

74. The singing bird was flying in the room.

75. I have studied.

76. The broken window is in the living room.

7.2.4.2 Tenses in Italian

In Italian tenses can also be simple or complex. They are simple when the verb is formed only one word, they are complex when an auxiliary is needed (“avere” or “essere”).

The indicative is the mood that has the most varied system of tenses in Italian. It has eight tenses:

- Present: it is used to express a present action (77), a habit (78), an action that is out of time (79), a future action that is real and objective (80). It can also be used instead of the past tense in narrations.

77. Vivo a Milano.
[Live in Milan]

78. Frequento una scuola di musica.
[Attend a school of music]

79. La Terra è tonda.
[Earth is round]

80. Ci vediamo domani.
[Us see tomorrow]

- Imperfetto: it is used to express an action in the past that is not limited in time. It can be used to describe an action taking place while another action happened, in the past (see example 81). It is also used in descriptions (82), to talk about habits in the past (83), and for courtesy forms (84).

81. Stavo cucinando, quando hanno bussato alla porta.
[Was cooking when have knocked at the door]

82. Era un bel giorno di sole.
[Was a beautiful day of sun]

83. Andavamo tutti i giorni in spiaggia.
[Went every day to beach]

84. Volevo chiederti un piacere.
[Wanted to ask you a favor]

- Passato remoto: it is used when the action took place in the past and was completed. It does not establish a relation with the present.

  85. Dante morì nel 1321.
  [Dante died in 1321]

- Passato prossimo: the action has a relation with the present, but started or happened in the past (see example 86). It can be used for future actions preceding another future action (87).

  86. Ho lanciato il pallone contro la finestra. Ora è rotta.
  [Have thrown the ball against the window. Now is broken]

  87. Se tra due ore non se ne è andato, chiamo la polizia.
  [If in two hours is not gone, call the police]

  The difference among the three past tenses that we mentioned, i.e. the "imperfetto", the "passato remoto" and the "passato prossimo", is related to the aspect, not to a position in time. [13]

- Trapassato prossimo: it is used for an action that preceded another action in the past.

  88. Ero appena entrata nel supermercato, quando ho incontrato Anna.
  [Had just entered in the supermarket, when have seen Anna]

- Trapassato remoto: it is used as the "trapassato prossimo", but the use is nowadays only literary.

  89. Il castello di sabbia fu travolto dall'onda.
  [The castle of sand was devastated by the wave]

- Futuro semplice: it is used to indicate an action that takes place after the time of utterance (example 90). It is also used to mitigate a sentence, or to indicate an action that followed another action in the past (91). It is used to guess an action (92).

  90. Partirò domani.
  [Will leave tomorrow]

  91. Dopo aver vinto la guerra, nel 1345 il re dichiarerà la pace.
  [After having won the war, in 1345 the king will declare peace]

  92. Hai fatto molto esercizio, sarai stanco.
  [You did a lot of exercise, will be tired]

[13] We also need to point out that the use of the "passato remoto" in northern Italy is not common. The "passato prossimo" is used instead, both when the action is completed in the past, and when it is not.

- Futuro anteriore: it is used when the action precedes an action occurring in the future.

93. Prima di andare a dormire avrò letto tutto il capitolo.
[Before going to sleep will have read all the chapter]

In the other moods, the tenses have, in the majority of the cases, a referential value, so they establish a relation in time with the verb of the principal clause. This happens in the infinitive, in the gerund, and in the participle, as we saw in section 7.2.4.1 for English. The “congiuntivo” and “condizionale” have a referential value when they are in a subordinate clause, while they do not when the clause is independent. The “congiuntivo” has four tenses: “presente”, “imperfetto”, “passato”, and “trapassato”. When the “congiuntivo” occurs in a dependent clause, the verb establishes a relation with the verb of the main clause: simultaneity in the present (present “congiuntivo”, example 94) and in the past (“imperfetto” “congiuntivo”, example 95) or anteriority in the present (“passato” “congiuntivo”, example 96) or in the past (“trapassato” “congiuntivo”, example 97). As we can see from the examples below, the tense of the main verb in the main clause determines the selection of the “congiuntivo” tense: the present and the past are used when the main verb is present, the “imperfetto” and “trapassato” when the main verb is past.

94. Spero che tu stia bene.
[Hope that you are well]

95. Speravo che stessi bene.
[Hoped that were well]

96. Spero che tu abbia finito i compiti.
[Hope that you have finished the homework]

97. Speravo che tu avessi finito i compiti.
[Hoped that you had finished the homework]

When the “congiuntivo” is used in a main clause, the present and the past are used when the action denoted by the verb is possible (98), the “imperfetto” and “trapassato” when it is not possible anymore (99).

98. Che piova?
[That rains?]

99. Magari fosse vero!
[If only were true!]

The “condizionale” has two tenses: the present and the past. When it is used in dependent clauses, it establishes a relation with the main verb of the main clause. The present is used when the action is simultaneous or future (100), the past when it already occurred (101).

100. Penso che ti piacerebbe.
[Think that you would like it]

101. Penso che ti sarebbe piaciuto.
[Think that you would have liked it]

When it is used in main clauses, the present “condizionale” is used when the action is still possible (example 102), the past when it is not (103).

102. Partirei domani.
[Would leave tomorrow]

103. Avrei comprato quel libro.
[Would have bought that book]

7.2.5 ASPECT

The aspect of the verb provides information about the kind of action the verb denotes: progressive, prompt, repetitive, beginning or ending, complete or incomplete. The aspect is not deictic, i.e. it is not relative to the time of utterance. Aspect can be perfective, when the action is completed, or imperfective, when the action is described while happening, progressive, when the action is continuous in time, or when it is incomplete. There are languages, such as Russian, in which aspect is an explicit feature of the verb and there are two different verb forms for imperfective and perfective aspect. Both in English and Italian aspect is expressed in different ways, such as through tense selection, the semantics of the verb itself, or, in Italian, by the use of the suffix “–icchiare” (that denotes an action that is neither ended nor fully completed).

104. I was writing (imperfective)
Scrivevo (imperfective)

105. I wrote (perfective)
Scrissi (perfective)

106. I hit (perfective)
Colpii (perfective)

107. I run (imperfective)
Corro (imperfective)

108. Dormicchiare (imperfective)
[Sleep a little bit]

In the examples above, the aspect in 104 is imperfective, since the action is in progress. The aspect in 105 is perfective, since the verb denotes an action that is completed. The aspect is perfective in 106, because the action is punctual. It is imperfective in 107, because the action is a sequence of various movements. It is imperfective in 108, since the verb denotes an action that is beginning and not completed.

As we mentioned above, tense selection is not only a way to express the time the action denoted by the verb takes place, but also the aspect of the action. Therefore, the concept of "aspect" sometimes overlaps with that of "tense", specially in languages in which the aspect is not morphologically expressed in the verb form. For example, the contrast between "was writing" (104) and "wrote" (105) does not amount to the time these verb forms refer to (and therefore to their tenses), because they are both past, but to their aspect. The form "was writing" denotes a progressive action, while "wrote" denotes an action that was completed. Aspect is imperfective in the first case and perfective in the second. Progressive aspect can be expressed by verb forms in different tenses: present (109), past (110), and future (111).

109. I <u>am studying</u> right now.

110. He <u>was listening</u> to music, when the telephone rang.

111. I <u>will be playing</u> the piano tomorrow night at the concert.

There are also structures and expressions that add imperfective aspectual meaning to the verb, such as "to be going to" (112), "start to" (113), and the Italian "stare + gerund" (114).

112. It is going to rain (imperfective)
Sta per piovere (imperfective)

113. He started to dance (perfective)
Cominciò a danzare (perfective)

114. Stavo dormendo (imperfective)
[Was sleeping]

7.3 TENSE/MOOD/ASPECT IN MT

The right selection of tense, mood, or aspect is a challenge in MT, specially when English and Italian are at stake, due to the following reasons:

a. First of all, the use of verbs is not the same in the two languages considered in this study. SMT systems tend to translate the verb directly into Italian, and this may generate errors in the selection of the features of the verb expressing tense, mood, and aspect. With regard to tense and aspect, for example, while in English the present progressive is quite common, the majority of the times it would correspond to a simple present in Italian (example 115).

115. The account I am seeing is not yours.
*L’account che sto vedendo non è il tuo.
L’account che vedo non è il tuo.

The same happens in the translation of the present perfect: it can be translated into Italian as a “passato prossimo” or as a simple present, depending on the kind of action it is expressing. In the former, it expresses an action that happened a number of times in the past (example 116), in the latter an action that started in the past and is still taking place (example 117). So, while in English the present perfect is used to express these two types of action, they are conveyed in Italian by different verb tenses, which naturally poses a problem to MT systems, as they have to opt for one or the other in the translation.

116. I have read a lot of books lately.
*Leggo molti libri recentemente.
Ho letto molti libri recentemente.

117. I have studied English for ten years.
*Ho studiato inglese per dieci anni.
Studio inglese da dieci anni.

The same happens in the selection of the right mood in the target language, specially when the “congiuntivo” has to be used. Several contexts in which the indicative is used in English correspond to cases in which the “congiuntivo” is used in Italian.

This is due to the fact that some verbs require the use of the “congiuntivo” in the dependent clause, to express wish (118), and possibility (119).

118. I hope you are fine.
*Spero che tu stai bene.
Spero che tu stia bene.

119. It was possible that it was my fault.
*Era possibile che era colpa mia.
Era possibile che fosse colpa mia.

b. In the second place, there are verb complexes that result in syntactic constructions that are more challenging for a SMT system, specially in coordination contexts. Complex tenses and verbal expressions such as “be aware” are examples of such cases.

120. I am aware of the problem and working to solve it.
*Sono a conoscenza del problema e cercando di risolverlo.
Sono a conoscenza del problema e sto cercando di risolverlo.

c. In complex sentences, syntactic dependency is not always apparent, specially when the subject is not repeated. There are cases in which a verb can depend on one verb or be coordinated to another verb depending on the first verb. This is specially problematic in English, due to the poor inflectional morphology system, which makes less overt marks of syntactic dependency available and thus makes it more difficult to identify the actual syntactic structure of the sentence.

121. We know that the users may disagree with our decision and apologize.
*Sappiamo che gli utenti possono essere in disaccordo con la nostra decisione e chiedere scusa.
Sappiamo che gli utenti possono essere in disaccordo con la nostra decisione e chiediamo scusa.

The sentence structure is ambiguous in English, since the coordination can be established either between “we know” and “we apologize”, or “users disagree” and “users apologize”. It is not possible to keep the ambiguous structure in the translation into Italian, since the tenses, and therefore the verb forms, in the translation of the verb “disagree" are morphologically different (“ci scusiamo” if the verb “scusare” is coordinated to “we know”, and “chiedere scusa” if it is coordinated to “disagree”). As it is impossible to maintain the ambiguous structure, choosing between one of these possibilities depending on the structure being considered should be made based on more semantic knowledge.

d. Segmentation is sometimes a problem for translation, specially at Unbabel. As we already mentioned, it consists in dividing the text into sentences or group of sentences that are then translated and sent to different editors for post-editing. An incorrect segmentation can make it difficult or impossible to understand the correct syntactic structure of a sentence and, without further context, it often makes it impossible for the editor to make an informed and correct decision on what the right tense, mood, or aspect is.

As we can notice, some of the problems mentioned above regard the source language analysis, like the understanding of the right syntactic dependency holding between constituents, while others lie in the target language generation, like the use of the correct tense due to target language linguistic specifications.

7.4 TENSE, MOOD, AND ASPECT ERRORS IN THE *CORPUS*

In the *corpus*, 101 errors belonging to the category "tense/mood/aspect" were annotated. It is important to remember that some errors involving the selection of the correct inflectional features of the verb were accounted for through other types of errors, such as POS or agreement. As we saw in the agreement chapter, one of the types of agreement considered regards the person of the verb. Therefore, even if the errors occur in similar situations as some of those included in the Tense/mood/aspect category and discussed in this chapter, they were addressed as agreement errors since, as we already mentioned, only one error could be marked for each constituent. In this section we will then analyze tense/mood/aspect errors and suggest solutions to reduce the number of errors produced by the MT system.

Tense/mood/aspect errors	
Number of unique errors	83
Number of tense errors	4
Number of mood errors	79
Number of aspect errors	0
Total	101

Table 1.

As we can see from in Table 1, from the total of 101 errors annotated as tense/mood/aspect errors, 83 were unique. The majority of the errors involved the selection of mood. Only four errors involved the selection of the correct verb tense. There were no errors observed regarding the selection of the aspect feature of the verb. The lack of aspect errors can be due to the fact that the category, as we said above, sometimes overlaps with that of tense (see section 7.2.3). The small amount of tense errors, in comparison to mood errors, can be explained by the fact that there were cases in which no error was marked because the translation of the verb was a possible one, even if another tense was more natural in Italian. For example, when the present continuous in English was translated into the form “stare + gerundio” in Italian, no error was marked. The translation is correct in some cases, even if the tense is not the most appropriate in others, as the use in Italian is limited to describing an action taking place at the time of utterance. This is a difference that is difficult to generalize, but that a human editor can easily make based on semantic knowledge. Therefore, the error was not considered relevant in this study and was not annotated.

7.4.1 Tense Errors

Among the four tense errors, two occurred in a coordinated verb. The previous verbs, in both cases, were correctly translated into a future (see example 122) and a “passato remoto” (see example 123).

122. Your guide will tell you about the city’s origins, show you a palace, and take you to the Ponte Vecchio.
   *La vostra guida vi racconterà le origini della città, vi mostrerà un palazzo, e vi porta al Ponte Vecchio.
   La vostra guida vi racconterà le origini della città, vi mostrerà un palazzo e vi porterà al Ponte Vecchio.
123. He will explain how the city begged, borrowed, and stole almost all of them
   *Vi spiegherà come la città pregò, chiese in prestito e ha rubato quasi tutti.
   Vi spiegherà come la città pregò, chiese in prestito e rubò quasi tutti.

In 122, the verb forms “will tell” and “show” were correctly translated into a verb form in the future. The verb “take” was translated into a simple present. In example 123, the verbs “begged” and “borrowed” were correctly translated into two simple past verb forms, while “stole” was translated into a “passato prossimo”.

Another tense error occurred in the translation of the gerund after the expression "thank you for". The verb was correctly translated into an infinitive, but the tense should have been the past.

124. Thank you for getting in touch with us.
*Grazie di entrare in contatto con noi.
Grazie di essere entrato in contatto con noi.

The last tense error involved the selection of the right form of future in the translation of "you will then stop". The verb form used in the translation, "sarà quindi fermarsi", does not exist in Italian. With regard to the error in the translation of the future tense, the solution cannot be studied in this work, because we would have to take into account the steps of the generation of the target text. As we already said, we will only intervene in this study on the target text after it is produced.

7.4.1.1 Possible solutions

With regard to the error in the verb after the expression "thank you for", we mentioned above that the verb form tense in Italian should be in the past. The present can be used in some general expressions, but the action denoted by the verb following the expression "thank you for" (that is a gerund in English and an infinitive in Italian) is usually completed and, therefore, past. A rule could be added to the Smartcheck:

Rule 9

If the expression "grazie" is followed by a preposition ("di") and a verb in the present infinitive, then the verb has to be changed into a past infinitive.

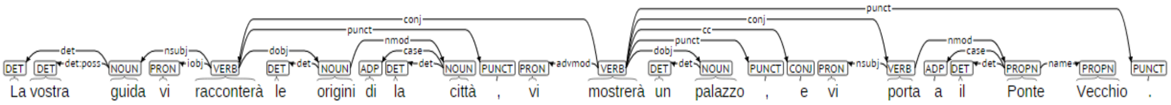
Grazie + di + V (present infinitive) → grazie + di + V (past infinitive)

125. Thank you for writing.
*Grazie di scrivere.
Grazie di aver scritto.

126. Thank you for being so kind.
Grazie di essere così gentile.
Grazie di essere stato così gentile.

One of the exceptions mentioned is the sentence in 126. The second Italian sentence is not incorrect, but the meaning is not the same as in the English sentence, since the verb “being” does not refer to a past situation and, therefore, could denote an action that is a fact and is not located in time. This means that some false positive of the rule 9 will be produced, but the frequency of occurrence will be reduced.

With regard to the errors involving tense in coordinated verbs, since the subject is not expressed in the sentence, the structure can be ambiguous. In the cases annotated, the parser correctly analyzes the structure of the sentence in Italian only in the case in 122, as we can see in the syntactic tree below. In 123, the parser does not recognize the coordination between the three verbs “pregò”, “chiese”, and “ha rubato” in the Italian sentence (it is an incorrect sentence, which probably has an impact on the performance of the parser.)


punct
conj
conj
punct
cc
nmod
punct
nmod
det
nsubj
dobj
case
dobj
case
det:poss
iobj
det
det
advmod
det
nsubj
det
name
DET DET NOUN PRON VERB DET NOUN ADP DET NOUN PUNCT PRON VERB DET NOUN PUNCT CONJ PRON VERB ADP DET PROPN PROPN PUNCT
La vostra guida vi racconterà le origini di la città , vi mostrerà un palazzo , e vi porta a il Ponte Vecchio .

In the case in 122, since the parser recognizes the structure, it is possible to adopt the same strategy presented in section 6.4.3.1 for the feature Person, i.e. have the tool check if the value for the feature Tense is the same in the coordinated verbs. However, when the coordination occurs between verbs of dependent clauses, the parser does not analyze correctly the structure in Italian, and therefore, it is not possible to provide a rule for the Smartcheck, that only takes into account the target text. In the future, when it is possible to use both the English parser and the Italian, since in English it establishes the dependency between the coordinated verbs and is able to recognize that the subject is the same for both verbs, it will be possible to force the tense to be the same in the two verbs in the target text, if there are no other instructions.

In conclusion, the errors involving the tense of the verb form are not common and given the small number of errors annotated no valid generalizations can be outlined. Since the SMT generally selects the correct tense in the translation from English into Italian, solutions were proposed regarding syntactic structures that can be problematic in general in the language pair considered, namely the coordination of two or more VPs. This issue regarding coordination is complex, since it involves also other error types, as we will see in the next section.

7.4.2 Mood Errors

Many mood errors were annotated in the *corpus*. We divided them into the following types:

Mood errors in the *corpus*	
Number of errors occurring in a main clause	39
Number of errors occurring in a coordinate clause	24
Number of errors occurring in a subordinate clause	16
Total	79

Table 2.

From Table 2, we can see that the majority of the errors occurred in a main clause, while 24 in a coordinate clause and 16 in a subordinate. This is mainly due to the kind of errors, as we will see in the next table. Among the number of errors occurring in the main clause, 7 times they occurred in an asyndeton, which is a figure of speech in which sentences are linked through punctuation instead of conjunctions. In this case, a period was used, instead of the conjunction “and”, hence, syntactically, two independent sentences were generated. In 13 of the 24 cases in which the error occurred in a coordinate clause, the verb was translated correctly in the main clauses. In the remaining cases (11), mood was already wrongly selected in the main clause too.

Distribution of mood errors per target mood type	
Number of cases in which the indicative should have been used	20
Number of cases in which the “congiuntivo” should have been used	10
Number of cases in which the “condizionale” should have been used	0
Number of cases in which the imperative should have been used	43
Number of cases in which the infinitive should have been used	5
Number of cases in which the gerund should have been used	1
Number of cases in which the participle should have been used	0

Table 3.

In Table 3 we can see the distribution of the mood errors considering the mood that should have been used. We categorized the errors in this way due to the high number of errors and to the fact that we are focusing on the target text. In 43 cases the verb mood selected should have been the imperative. In most cases, the infinitive was used instead of the imperative. This was the most common mistake in the number of tense/mood/aspect errors. The error occurred both in main clauses and in coordinate clauses. The lack of a morphological suffix in the imperative in English makes it more difficult for the MT system to distinguish between an infinitive and an imperative. Even if nonfinite sentences are less common than finite sentences in the kind of texts we annotated (emails from Help Centers and tourism texts), both in English and in Italian, we cannot exclude the possibility of having nonfinite sentences. Therefore, we cannot assume that every base form is an imperative. In 42 cases the infinitive was used instead of the imperative, while in only one case the simple present form of the verb was used in the second person instead of the imperative. In 10 cases the “congiuntivo” should have been used instead of the indicative verb form. As we already mentioned, there are several cases in Italian in which the “congiuntivo” is used while the indicative is used in English. In 20 cases, the indicative should have been used. Among these errors, 15 times the infinitive was used instead (in 9 cases, the infinitive was used instead of the future form of the verb, in 6 instead of the simple present form of the verb). As we already mentioned in the tense section, the future poses more issues than the other tenses of the indicative and is not always translated correctly. In one case the “congiuntivo” was used instead of the indicative in a coordinated sentence. In four cases the past participle was used instead of a past indicative form (simple past or “passato prossimo”). In one case the past participle was used instead of the gerund.

If we combine data from Table 2 and Table 3, we can notice that, if we consider only the errors in main clauses, five times the error occurred in a sentence with a modal auxiliary verb (the infinitive should have been used after a modal) and after a verb followed by a preposition (“cercare di” and “esitare a”). In six cases the indicative should have been selected instead of the infinitive. In the other 29 cases the imperative should have been used instead of the infinitive.

With regard to the errors in coordinated sentences, 15 times the imperative should have been used instead of the infinitive. In the remaining 9 cases, the indicative should have been used instead of the infinitive (8 cases) or instead of the past participle (one case).

In the 5 cases in which the error occurred in a subordinate clause, the indicative should have been used instead of the infinitive in 4 cases and the past participle in one case.

As we can see from the categorization presented above, there is a wide range of different errors occurring in the selection of the mood performed by the MT system. Therefore, we will suggest a solution only for the most common, as the remaining data available does not allow for valid generalizations. We will use, nonetheless, the other data available from the categorization for future work.

7.4.2.1 Possible Solutions

The parser in English correctly analyzes the verbs in the sentences in the source text. It recognizes the imperative forms and the dependency when a base form is coordinated with another verb. However, in our study it is not possible to use the English parser in order to create an equivalent sentence structure in Italian, as we saw in the previous chapters. The Italian parser alone does not recognize the errors in the majority of the cases. Therefore, we have to consider the different types of errors in order to solve them.

With regard to the errors in main clauses, a rule should be added to the Smartcheck for the verbs following a modal.

Rule 10

If a verb is preceded by a modal verb, then the verb should be an infinitive.

Modal + V → Mood(V) = infinitive

With regard to the errors in coordinate clauses, as we already saw in the tense errors, the coordinated verb has to have the same value for the mood feature as the verb it is coordinated to. Additionally, it is impossible in Italian for an infinitive to be coordinated to a verb that is not an infinitive. However, as we saw in section 7.4.1.1, the parser in Italian does not always correctly analyze the syntactic structure when the coordination occurs in dependent clauses. Therefore, a rule cannot be implemented in the Smartcheck until it is able to take into account also the source text.

The verbs in subordinate clauses are more difficult to correct, because they are related to specific selection restrictions which depend on the lexical items occurring in the sentence, as well as on particular constructions, such as the selection of the indicative instead of the "congiuntivo", for example. Additionally, they include cases in which the selection depends on the semantics of the verb, for example when the verb expresses a wish or a doubt. However, the verbs in which most errors occurred can be added to the Smartcheck with information regarding the mood they select. For example, the verbs "sembrare" (to seem), "sperare" (to hope), "pensare" (to think), "dispiacersi" (to be sorry), and the expressions "essere contento" (be happy), "fare in modo" (to ensure), and "fare tutto quanto" (to do everything to) are followed by the "congiuntivo" mood. Therefore, information about the selection restrictions in terms of mood these verbs introduce could be added to the lexical resource for the verbs mentioned, as well as for others that may be identified in the future.

8 CONCLUSIONS AND FUTURE WORK

The objective of this thesis consisted in improving the quality of the texts that were translated from English into Italian by the MT system considered and then post-edited by human editors. In order to do so, an error annotation of a *corpus* was performed, and possible solutions to the most frequent and systematic errors identified were provided. In this chapter, we will present our conclusions and possible developments of this work.

8.1 CONCLUSIONS

The annotation of the *corpus* and the error typology used in the task allowed us to identify the most frequent errors in the translated texts both after MT and after the first post-edition. The *corpus* consisted of Help Center tickets and tourism texts. The number of errors in each category in the two steps was calculated and the results were compared. The analysis of the errors and their distribution allowed us to select the issues to address first. This choice was based on the type of error, on the kind of tools available at Unbabel to solve the issues involved, on the impact the category of errors has on quality, and on the possibility to tackle the issues in an automatic way. Consequently, we identified three categories of errors that had an impact on the quality of the results, and were frequent and systematic. We considered in this work errors belonging to the categories “word order”, “agreement”, and “tense/mood/aspect”. The thorough analysis of the errors allowed us to identify patterns of errors and the constituents in which they occurred. When it was possible to outline a generalization characterizing the phenomena, and come up with a solution to address the majority of the cases, a rule was provided to be added to the tool that automatically detects errors in the target text, the Smartcheck. When the tools used at Unbabel did not allow to address the issue, possible strategies to obtain improvements were presented but no rules were provided.

Due to the fact that Unbabel is currently focused on other improvements in the translation process, it was not possible to implement the rules in the Smartcheck and check the results. This would allow us to see if rules generate false positives in the detection of errors and if the number of warnings and suggestions provided to the editor is too high and, therefore, counterproductive. Additionally, the results of the improved post-editing can be annotated and the number of errors corrected when more rules are added to the Smartcheck can be calculated. This way, we would be able to see the actual results of the work done, in a real post-editing situation.

Although it was not possible to provide a solution for all categories of errors, the data helped to identify the most critical errors, in terms of impact on quality and frequency. It was possible to understand which are the categories that should be addressed in order to further improve the quality of the target text. Two of such categories are “determiners” and “sentence structure”. A brief analysis of these two error categories was provided due to the high number of such errors and to the impact they have on the translation. In the former category, “determiners”, the high occurrence of errors makes the human post-editing time-consuming and increases the possibility that errors pass unnoticed. In the other category, “sentence structure”, the type of error makes it impossible for the human editor to understand the target sentence without reading the source text and, in the majority of cases, it is necessary to re-write the sentence. The analysis provided will help the elaboration of a rule to automatically detect this type of errors, when the resources at Unbabel will allow it.

The technology at Unbabel allowed us to intervene on the target text and on the detection of errors performed by the Smartcheck. It was not possible to integrate information in the MT system, since the translation was done by the Google MT system. Additionally, it was not possible to automatize the post-editing process, i.e. it was not possible to automatically correct the errors in the target text. When an automatic post-editing tool is available at Unbabel, the rules will be adapted to the task by providing a rule regarding the step the tool should take in post-edition, instead of providing a warning message to the editor.

In error annotation, we considered target texts not only after MT, but also after the first human post-edition. With regard to the post-edited texts, we used the data collected in annotation only to calculate the percentage of errors that were corrected by the human editor, and to better understand the post-editing steps. Due to time and space constraints, we did not analyze more thoroughly the errors annotated in edited texts and, therefore, we could not state anything regarding whether the errors that still occurred after the first post-edition were errors that were not corrected, or new errors introduced by the editors. In the first case, the automatic detection of the Smartcheck will assist the editor in the task. In the second case, more rules should be provided to address the new issues.

The analysis presented in this study focused on a limited part of the data collected. Nevertheless, the results outlined contribute not only to improving the services offered by

Unbabel, but also to understanding and improving the assessment of the performance of MT systems and their results.

8.2 FUTURE WORK

We believe that quality improvements in MT from English into Italian can be obtained by continuing the work started in this thesis and expanding it to more domains. The future work may focus on the error categories that were not addressed in this study either because the errors were not systematic, or because the necessary tools were not available at the time this work was developed. Apart from the two categories "determiners" and "sentence structure" already mentioned in the previous section, other error types can be analyzed in order to provide a solution. With regard to errors that are related to the creative use of language and to semantic or contextual knowledge, more tools, such as a semantic parser, are needed to tackle the issues, and the adoption of such tools would have a great impact on quality.

In this work, as already said, we decided to annotate texts after MT and after the first post-edition. In future work, texts that are completely post-edited, i.e. that are reviewed by both the first editor and the senior, may be annotated. This way, the percentage of errors corrected in the text that is actually delivered to the client can be calculated and the post-editing process can be analyzed in all its steps.

We believe that the analysis of errors can be improved by calculating the number of times a certain syntactic structure in the source text[14] was not translated correctly in the target text. This can be done by extracting all the occurrences of a syntactic structure from the *corpus* and by calculating the number of times an error occurred in the translation of such a structure. This was done, during this study, for the word order errors. However, the results were not satisfactory and future work is needed in order to improve the accuracy in extracting the occurrences of the syntactic structure considered in the source text and thus have data that can constitute a base for analysis.

[14] The annotation data collected and presented in this work can be used to select the structures to be analyzed, focusing on those that are problematic in MT.

9 BIBLIOGRAPHY

Burchardt, A., Lommel, A. (2014) *Practical Guidelines for the Use of MQM in Scientific Research on Translation Quality*. (Available at http://www.qt21.eu/downloads/MQM-usage-guidelines.pdf).

Dorr, B., Jordan, P., Benoit, J. (1998) *A survey of Current Paradigms in Machine Translation.* (Available at http://www.umiacs.umd.edu/users/bonnie/Publications/newai98.pdf).

España-Bonet, C., Costa-Jussà, M. (2016) “Hybrid Machine Translation Overview” in *Hybrid Approaches,* Costa-Jussà, M., Rapp, R., Lambert, P., Eberle, K., Banchs, R., Babych, B. (eds). Switzerland: Springer.

Frederking, R., Brown, R. (1996) *The Pangloss-lite Machine Translation System*. Carnegie Mellon University. (Available at http://www.mt-archive.info/AMTA-1996-Frederking.pdf).

Hutchins, J. (1978) “Machine Translation and Machine Aided Translation” in *Journal of Documentation*, Vol. 34, No 2., pp. 119-159. UK: Emerald Group Publishing.

Hutchins, J. (1986) “Machine translation: past, present, future” in *Ellis Horwood Series in Computers and their Applications.* Chichester: Ellis Horwood. (Second chapter “Precursor and pioneers” available at http://www.hutchinsweb.me.uk/PPF-2.pdf).

Hutchins, J., Somers, H. (1992) *An Introduction to Machine Translation*. London: Academic Press. (Available at http://www.hutchinsweb.me.uk/IntroMT-TOC.htm).

Hutchins, J. (1994) “Research methods and system designs in machine translation: a ten-year review, 1984-1994” in *Machine Translation: Ten Years On*, 12-14 November 1994, Chapter 4, pp. 1-16. University of East Anglia. (Available at http://www.mt-archive.info/BCS-1994-Hutchins.pdf).

Hutchins, J. (2001) “Machine translation over fifty years” in *Histoire, Epistémologie, Langage.* Vol. 23 (1), pp. 7-31. (Available at http://hutchinsweb.me.uk/HEL-2001.pdf).

Hutchins, J. (2005) *History of Machine Translation in a Nutshell.* (Available at http://wwwhutchinsweb.me.uk/Nutshell-2005.pdf).

Hutchins, J. (2010) “Machine Translation: A Concise History” in *Journal of Translation Studies,* Vol. 13, Nos. 1-2 (2010). *Special issue: The teaching of computer-aided translation,* Chan Sin Wai (ed.). (Chinese University of Hong Kong, 2010) pp.29-70. (Available at http://www.hutchinsweb.me.uk/CUHK-2006.pdf).

Hutchins, J. (2015) “Chapter 6. Machine translation: History of Research and Applications” in *Routledge Encyclopedia of Translation Technology,* Chan Sin-wai (ed.). London: Routledge. (Available at http://www.hutchinsweb.me.uk/Routledge-2014.pdf).

Koehn, P. (2005) “Europarl: A parallel *corpus* for statistical machine translation.” In *MT Summit 2005*. University of Edinburgh. (Available at: http://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf).

Lommel, A. (2015) *Multidimensional Quality Metrics MQM Definition*. (Available at http://www.qt21.eu/mqm-definition/definition-2015-06-16.html).

Lopez, A. (2008) “Statistical Machine Translation” in *ACM Computing Surveys,* Vol. 40, No3, Article 8. University of Edinburgh. (Available at https://alopez.github.io/papers/survey.pdf).

Marneffe, M.C., Manning, C. (2008) *Stanford Typed Dependency Manual.* (Available at http://nlp.stanford.edu/software/dependencies_manual.pdf).

Martins, A., Almeida, M., Smith, N. (2013) “Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers.” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL),* Sofia, Bulgaria, August 2013. (Available at: https://www.cs.cmu.edu/~afm/Home_files/acl2013short.pdf).

Papineni, K., Roukos, S., Ward, T., Zhu, W. (2002) “BLEU: a Method for Automatic Evaluation of Machine Translation” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL),* Philadelphia, pp. 311-318. (Available at http://www.aclweb.org/anthology/P02-1040.pdf).

Silva, A., Moniz, H., Graça, J., Macedo, H. (2016) *Unbabel’s Error Curation Platform Based on QT21: Towards Quality Assurance in an Ecological Set up* (unpublished).

Slocum, J. (1985) “A Survey of Machine Translation: its History, Current Status, and Future Prospects” in *Machine Translation Systems*. Cambridge: Cambridge University Press.

Somers, H. (2003) “An Overview of EBMT” in *Recent Advances in EBMT,* Carl, M., Way, A. (eds.). New York: Springer.

Williams, P., Koehn, P. (2011) “Agreement Constraints for Statistical Machine Translation into German” in *Proceedings of the 6th Workshop on Statistical Machine Translation,* pp. 217–226. University of Edinburgh. (Available at http://www.aclweb.org/anthology/W11-2126).

Wundelich, D. (2013) “Grammatical Agreement” in *International Encyclopedia of Social and Behavioral Sciences 2nd Edition.* (Available at http://www.zas.gwz-berlin.de/fileadmin/mitarbeiter/wunderlich/AGREE_2013.pdf).

Reports and Launchpads

ALPAC report (1966) *Language and Machines. Computers in Translation and Linguistics*. (Available at http://www.mt-archive.info/ALPAC-1966.pdf).

EUROTRA (1990) *The European Community’s Research and Development Project on Machine Translation.* Office for Official Publications of the European Communities. Luxembourg.

Nyberg, E., Mitamura, T., Carbonell, J. (1997) *The KANT Machine Translation System: from R&D to Initial Deployment.* Carnegie Mellon University. (Available at: http://repository.cmu.edu/cgi/viewcontent.cgi?article=1337&context=compsci).

LOGOS: *Logos Machine Translation System.* (Available at https://aclweb.org/anthology/A/A97/A97-2009.pdf).

White, J. (1985) “Characteristics of the METAL Machine Translation System at Production Stage” in *Proceedings of the Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate University, Hamilton, August 1985. (Available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.5168&rep=rep1&type=pdf).

QT Launch Pad. (Available at https://www.w3.org/International/multilingualweb/2014-madrid/slides/genabith.pdf).

Grammars and dictionaries

English

Cambridge dictionary: http://dictionary.cambridge.org.

Greenbaum, S. (1996) *The Oxford English Grammar.* Oxford, UK: Oxford University Press.

Merriam Webster dictionary: http://www.merriam-webster.com/.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985) *A comprehensive Grammar of the English Language.* London and New York: Longman.

Italian

Accademia della crusca: www.accademiadellacrusca.it.

Dizionario Zanichelli: http://www.zanichelli.it/dizionari/.

Serianni, L. (1989) *Grammatica italiana*. Torino: UTET Università.

Vocabolario Treccani: www.treccani.it/vocabolario/.

Websites

Unbabel website: www.unbabel.com.

Unbabel blog: blog.unbabel.com.

Language Tool website: languagetool.org.

Moses website: www.statmt.org/moses.

TAUS website: www.taus.net.

Videos

Estelle, J. (2013) *Google I/O 2013 - Found in Translation: Going Global with the Translate API*. (Available at https://www.youtube.com/watch?time_continue=374&v=lkwkx8NO4CY).

Haddow, B. *Moses Introduction* University of Edinburgh, UK. (Available at https://www.taus.net/translate/mosescore/machine-translation-and-moses-tutorial#training-systems).

ANNEX

RULES FOR THE SMARTCHECK

Rule 1

If a noun or a PP precede the head noun, ask the editor to check the sequence with the following message: “Word_Order: Controllare l’ordine degli elementi nella frase.” (Word_Order: Check the order of the elements in the sentence.)

Rule 2

If a noun or a sequence of nouns follow the head noun, ask the editor to check the translation of the structure with the following message: “Word-Order: Controllare la categoria sintattica degli elementi della frase.” (Word_Order: Check the part of speech of the elements in the sentence.)

Rule 3

If one of the sequences listed below are detected, ask the editor to check the order of the words in the sentence with the following message: “Word_Order: Controllare l’ordine degli elementi nella frase.” (Word_Order: Check the order of the elements in the sentence.)

$\mathrm{N+ADJ^{+}+N}$;

$\mathrm{ADJ^{+}+N+N}$;

$\mathrm{ADJ+ADJ^{+}+N+N^{+}}$.

Rule 4

When a named entity occurs in the target text and is preceded or followed by an adjective or a PP that modifies it, highlight the sequence and ask the editor to check the order of the constituents with the following message “Word_Order: Controllare l’ordine degli elementi nella frase.” (Word_Order: Check the order of the elements in the sentence.)

Rule 5

If a noun ending in “–tore” occurs in the target text, check if its specifiers and modifiers are masculine.

$\mathrm{SPR^{*}+N_{-tore}+MOD^{*}} \rightarrow \mathrm{SPR^{*}{}_{masc}+N_{-tore}+MOD^{*}{}_{masc}}$

Rule 6

If a noun ending in "–tà", "–tù", "–trice", or "–tite" occurs in the target text, check if its specifiers and modifiers are feminine.

$\mathrm{SPR}^{*} + \mathrm{N}_{-\text{tà}|-\text{tù}|-\text{trice}|-\text{tite}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\text{fem}} + \mathrm{N}_{-\text{tà}} + \mathrm{MOD}^{*}{}_{\text{fem}}$

Rule7

If a noun ending in a consonant occurs in the target text, check if its specifiers and modifiers are masculine.

$\mathrm{SPR}^{*} + \mathrm{N}_{-\text{consonant}} + \mathrm{MOD}^{*} \rightarrow \mathrm{SPR}^{*}{}_{\text{masc}} + \mathrm{N}_{-\text{consonant}} + \mathrm{MOD}^{*}{}_{\text{masc}}$

Rule 8

If the quantifier "nessuno" or "chiunque" are part of the subject of a sentence, then the head verb form of the sentence must be singular.

Rule 9

If the expression "grazie" is followed by a preposition ("di") and a verb in the present infinitive, then the verb has to be changed into a past infinitive.

Grazie + di + V (present infinitive) → grazie + di + V (past infinitive)

Rule 10

If a verb is preceded by a modal verb, then the verb should be an infinitive.

Modal + V → Mood(V) = infinitive

RULES APPLIED TO MT

Rule 1

If there is an adjective modifying a noun in English and the adjective is a quality adjective, then the order in the target language should be noun adjective.

$\mathrm{ADJ}_{\mathrm{Q}}$+N → N + $\mathrm{ADJ}_{\mathrm{Q}}$

Rule 2

If there is a noun preceding another noun in English, and the first noun modifies the second, invert the order and convert the noun into an adjective phrase or a PP.

$$\mathrm{N1}+{}^{\mathrm{modifies}}\mathrm{N2} \rightarrow \mathrm{N2}+(\mathrm{ADJP}|\mathrm{PP}_{\mathrm{N1}})$$