UNIVERSIDADE DE LISBOA

FACULDADE DE LETRAS

QUALITY IN HUMAN POST-EDITING OF MACHINE-TRANSLATED TEXTS:

ERROR ANNOTATION AND LINGUISTIC SPECIFICATIONS FOR TACKLING REGISTER ERRORS

INGRID TESTA

Relatório de estágio orientado pela Professora Doutora Sara Mendes e coorientado pela Professora Doutora Helena Moniz, especialmente elaborado para a obtenção do grau de Mestre em TRADUÇÃO

2018

ABSTRACT

During the last decade, machine translation has played an important role in the translation market and has become an essential tool for speeding up the translation process and for reducing the time and costs needed. Nevertheless, the quality of the results obtained is not completely satisfactory, as it is considerably variable, depending on numerous factors. Given this, it is necessary to combine MT with human intervention, by post-editing the machine-translated texts, in order to reach high-quality translations.

This work aims at describing the MT process provided by Unbabel, a Portuguese start-up that combines MT with post-editing provided by online editors. The main objective of the study is to contribute to improving the quality of the translated text, by analyzing annotated translated texts, from English into Italian, to define linguistic specifications to improve the tools used at the start-up to aid human editors and annotators. The analysis of guidelines provided to the annotator to guide his/her editing process has also been developed, a task that contributed to improve the inter-annotator agreement, thus making the annotated data reliable. Accomplishing these goals allowed for the identification and the categorization of the most frequent errors in translated texts, namely errors whose resolution is bound to significantly improve the efficacy and quality of the translation. The data collected allowed us to identify register as the most frequent error category and also the one with the most impact on the quality of translations, and for these reasons this category is analyzed in more detail along the work. From the analysis of errors in this category, it was possible to define and implement a set of rules in the Smartcheck, a tool used at Unbabel to automatically detect errors in the target text produced by the MT system to guarantee a higher quality of the translated texts after post-edition.

Keywords: machine translation, post-edition, annotation, inter-annotator agreement, error analysis, register

RESUMO

Nas últimas décadas, a tradução automática tem sido uma importante área de investigação, no âmbito da qual os investigadores têm vindo a conseguir melhorias nos resultados, obtendo mesmo resultados positivos. Hoje em dia, a tradução automática desempenha um papel muito importante no mercado da tradução, devido ao número cada vez maior de textos para traduzir e aos curtos prazos estabelecidos, bem como à pressão constante para se reduzir os custos.

Embora a tradução automática seja usada cada vez com mais frequência, os resultados obtidos são variáveis e a qualidade das traduções nem sempre é satisfatória, dependendo dos paradigmas dos sistemas de tradução automática escolhidos, do domínio do texto a traduzir e da sintaxe e do léxico do texto de partida. Mais especificamente, os sistemas de tradução automática que foram desenvolvidos podem ser divididos entre sistemas baseados em conhecimento linguístico, sistemas orientados para os dados e sistemas híbridos, que combinam diferentes paradigmas. Recentemente, o paradigma neuronal tem tido uma aplicação muito expressiva, implicando mesmo a problematização da existência dos restantes paradigmas.

Sendo que a qualidade dos resultados de tradução automática depende de diferentes fatores, para a melhorar, é necessário que haja intervenção humana, através de processos de pré-edição ou de pós-edição.

Este trabalho parte das atividades desenvolvidas ao longo do estágio curricular na *start-up* Unbabel, concentrando-se especificamente na análise do processo de tradução automática, implementado na Unbabel, com vista a apresentar um contributo para melhorar a qualidade das traduções obtidas, em particular as traduções de inglês para italiano.

A Unbabel é uma *start-up* portuguesa que oferece serviços de tradução quase em tempo real, combinando tradução automática com uma comunidade de revisores que assegura a pós-edição dos mesmos. O *corpus* utilizado na realização deste trabalho é composto por traduções automáticas de inglês para italiano, pós-editadas por revisores humanos de *e-mails* de apoio ao cliente. O processo de anotação visa identificar e categorizar erros em textos traduzidos automaticamente, o que, no contexto da Unbabel,

é um processo feito por anotadores humanos. Analisou-se o processo de anotação e as ferramentas que permitem analisar e anotar os textos, o sistema que avalia a métrica de qualidade e as orientações que o anotador tem de seguir no processo de revisão. Este trabalho tornou possível identificar e categorizar os erros mais frequentes nos textos do nosso *corpus*.

Um outro objetivo do presente trabalho consiste em analisar as instâncias dos tipos de erro mais frequentes, para entender quais as causas desta frequência e estabelecer generalizações que permitam elaborar regras suscetíveis de ser implementadas na ferramenta usada na Unbabel, para apoiar o trabalho dos editores e anotadores humanos com notificações automáticas. Em particular, o nosso trabalho foca-se em erros da categoria do registo, o mais frequente nos textos anotados considerados. Mais especificamente, o nosso estudo consiste em definir um conjunto de regras para melhorar a cobertura do Smartcheck, uma ferramenta usada na Unbabel para detetar automaticamente erros em textos traduzidos no âmbito dos fenómenos relacionados com a expressão de registo, para garantir melhores resultados depois do processo de pós-edição.

O trabalho apresentado está dividido em oito capítulos. No primeiro capítulo, apresenta-se o objeto de estudo do trabalho, a metodologia usada na sua realização e a organização deste relatório. No segundo capítulo, apresenta-se uma panorâmica teórica sobre a área da tradução automática, sublinhando as características e as finalidades destes sistemas. Apresenta-se uma breve história da tradução automática, desde o surgimento desta área até hoje, bem como os diferentes paradigmas dos sistemas de tradução automática. No terceiro capítulo, apresenta-se a entidade de acolhimento do estágio que serviu de ponto de partida para este trabalho, a *start-up* portuguesa Unbabel. Explica-se o processo de tradução utilizado na empresa e as fases que o compõem, descrevendo-se detalhadamente os processos de pós-edição e de anotação humanas. São apresentadas também algumas informações sobre as ferramentas usadas na empresa para apoiar o processo de tradução, o Smartcheck e o Turbo Tagger. No quarto capítulo, apresenta-se o processo de anotação desenvolvido na Unbabel, como funciona e as orientações que o anotador deve seguir, descrevendo-se também alguns aspetos que podem ser melhorados. No quinto capítulo problematiza-se a questão do acordo entre anotadores, descrevendo-se a sua importância para medir a homogeneidade entre

anotadores e, consequentemente, a fiabilidade de usar os dados de anotação para medir a eficácia e a qualidade dos sistemas de tradução automática. No sexto capítulo, identificam-se os erros mais frequentes por categoria de erro e destaca-se a categoria de registo, a mais frequente e com repercussões evidentes na fluência e na qualidade da tradução, por representar a voz e a imagem do cliente. Apresenta-se uma descrição de um conjunto de regras que pode ser implementado na ferramenta Smartcheck, com vista a diminuir a frequência do erro e aumentar a qualidade dos textos de chegada. Procede-se ainda à verificação do correto funcionamento das regras implementadas, apresentando-se exemplos ilustrativos do desempenho do Smartcheck, na sua versão de teste, com dados relevantes. No último capítulo deste trabalho, apresentam-se as conclusões e o trabalho futuro perspetivado com base neste projeto.

Em conclusão, o objetivo do presente trabalho visa contribuir para a melhoria da qualidade dos textos traduzidos na entidade de acolhimento do estágio. Concretamente este trabalho constitui um contributo tangível para o aumento da precisão do processo de anotação humana e para a extensão da cobertura das ferramentas de apoio ao editor e ao anotador humanos usados na *start-up* Unbabel.

Palavras-chaves: tradução automática, pós-edição, anotação, acordo entre anotadores, análise de erros, registo

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my supervisor, Sara Mendes. Her patience and professionality in giving me advices and suggestions have played an important role throughout my academic journey, and without her help this work would not have been possible.

My gratitude goes also to my internship tutor, Helena Moniz. She has been a guidance in this new adventure and she constantly motivated me to work harder, and she helped me at a professional level, as well as at a personal level.

I would like to thank Unbabel, for the opportunity they gave me to work and grow with them, but a special thank goes to the Quality Team, which contributed in making this such a great experience.

A very big thanks to my family, especially to my parents and my sister who were always by my side, even if physically far from me. They have always been very supportive and understanding, and I would like to thank them one more time for the opportunity they gave me to live and experience this magic city, Lisbon.

Finally, a big thanks to all my friends, in Italy, in Lisbon and around the world, who make this Portuguese experience probably one of the best of my life so far.

CONTENTS

1. INTRODUCTION

The aim of this work is to analyze the translation process and the post-edited texts provided by Unbabel. We focused on giving proposals to improve the quality of the MT, starting from the problematization of specific phenomena. Unbabel is a Portuguese start-up, which hosted my internship, from September 2016 till June 2017. It offers translation services, combining MT with crowd post-edition. This process is done on an online platform and post-editors are not necessarily professional translators, but people who are fluent in English and native speakers of the target language. This approach, make it possible to increase the amount of translation produced, reducing the time and the costs. Therefore, Unbabel relies on MT for its translation process, alongside human post-edition.

In the last decades, MT has been an important area of research and through efforts, as the evolution not always was constant, researchers succeeded in obtaining remarkable improvements and positive results.

Nowadays, MT has become an important element in the translation process, as it increased the amount of translation and it decreased the time needed to produce it, as well as the costs. Even though MT is being increasingly used and reliable, the quality of the translations obtained are still variable and not totally satisfactory in many cases, heavily depending also on the paradigm of the MT system that is being used. Obtaining high quality results is very important in order to reach a better translation and to control or to assess the quality of the results.

As the quality assessment in machine translation and in post-edited texts is a process that still has to be improved in the translation pipeline, we studied the quality assessment mechanism and applied some improvements, in order to reach satisfactory translation texts, mostly in what register is concerned.

1.1. OBJECTIVES

The main objective of this study consists in analyzing the already existent translation process provided by Unbabel and improving the quality of the results in the translated

text, focusing on translation work form English into Italian. The *corpus* of texts that we collected are Help Centre e-mails of the language pair that we are taking into consideration.

One of the first aspects that we are going to study in our project is the process of error annotation, in the language pair English – Italian, and the tool that allows the annotators to analyze and annotate their texts. We will also tackle the framework that ascribes the quality metric to a translation and the guidelines the annotators have to follow.

The error annotation analyses the translation results, and, in this work, we are going to take in consideration both the results of the MT process and the results edited by humans. This process allowed us to identify and categorize the most common errors for the type of texts we are studying.

By the identification and categorization of error patterns, we were able to identify the information the system needs to integrate, in order to achieve the tangible results of improving the effectiveness and the quality of the system in this part of the translation process. Thanks to the identification and categorization of error patterns, it was also possible to suggest clearer guidelines for the definition of specific criteria the annotator has to follow, in order to achieve a high level of consistency and agreement with other annotators, working with the same language pair, and even comparing intra-annotator agreement. The inter and intra-annotator agreement shows how reliable the data are.

Another objective of this study is the definition of the most frequent errors in consideration, understanding the causes of this frequency and recognizing the typology of the error, in order to elaborate rules to automatically detect the errors and provide a warning to the post-editor. With this we aim at helping the translators and consequently at improving the post-edition process at Unbabel.

Our work points out that the register errors are the most frequently identified in the annotation process and the process of reducing the frequency of this error is a state-of-the-art one, it means that not only this topic is crucial within Unbabel, but it has also been tackled in recent literature.

As we will demonstrate along our work, we succeeded in achieve the objective of defining a set of rules that will be implemented in the system, in order to improve the

Smartcheck, a tool used at Unbabel that automatically detects errors in the texts that have been machine-translated. In our case, we focused only on the register errors. In this work, is also defined a set of rules that was not possible to implement, due to technical limitations, but that are linguistically presented and analyzed. This tool provides the editor with information regarding the register that has to be used in the target text, in order to guarantee better results after the post-edition process.

1.2. METHODOLOGY

The starting point of the present work is the historical and theoretical perspective of the machine translation, its importance in the translation market and how the machine translation process can be useful for humans, as it is faster, and it reduces the costs.

This study allowed us to identify the advantages and disadvantages of MT and, with the aim of analyzing the errors in machine translation at Unbabel, were also taken into consideration previous published works regarding MT systems and post-edition.

Concerning the detection of the errors in machine translation for our work, we collected a *corpus* of texts machine translated, from English to Italian, post-edited by humans, and finally annotated by us. After the process of annotation, the errors were categorized and were defined *criteria* in order to increase the inter-annotator. This process has been defined from the analysis of the annotated errors, from the divergencies and convergences among the annotators, with the aim to improve and increase the agreement.

The data were then studied again to outline some improvements in the quality of the translations, and the most frequent error, the register, was analyzed, in order to find repeated patterns, to allow the implementation of certain rules to automatically detect register errors in the post-editing stage. When this was not possible, because of technological limitations of the tool, we outlined linguistic generalizations that can be used to define formal rules in the future.

1.3. ORGANIZATION

The work presented here is organized as follows. In chapter two, we discuss the scientific domain of machine translation, pointing out its characteristics and its goals. An historical overview is provided, from the early stages of research until nowadays. Paradigms of MT systems are described as well in this chapter. This allows us to

understand the advantages and disadvantages of the MT system used at Unbabel, how it works and how we can improve it.

In chapter three, we describe the entity that hosted the internship, the Portuguese start-up Unbabel, explaining how the translation process is performed within the company, its pipeline and how the post-editing and the annotation process work. We also provide information about the tools used at Unbabel during the translation process, namely the Smartcheck and the semantic parser.

Chapter four presents the error annotation process, how it works, the instructions that the annotators have to follow, the guidelines, and what can be improved.

In chapter five, it is shown the inter-annotator agreement, i.e. the definition of some criteria the annotator has to follow in order to achieve a high level of consistency in the annotation process. It brings the first improvements to the translation process provided by Unbabel, by assessing the reliability of the data.

In chapter six, we introduce the error categorization used at Unbabel to annotate the data and we highlight the most common and frequent errors based on a data-driven approach. We then focus on the category of register, the most frequent one and also the category that has more impact on the fluency and quality of the translation, as it represents the voice and image of the client. We also provide possible solutions to address specific issues related to this error category and we implement, in the Smartcheck, a set of rules that decrease the frequency of this type of error and at the same time increase the quality of the target texts.

The final chapter of this thesis is dedicated to some conclusions and to the presentation of future work.

2. MACHINE TRANSLATION: THEORETICAL OVERVIEW

In this chapter, we present a brief historical and theoretical overview of machine translation. In the first section, definitions of machine translation and its functions are given. In the second section, we provide a historical overview of scientific and technological developments in machine translation, from the early stages until the present days. In the third section, main paradigms of MT systems are described; especially the ones based on linguistic knowledge, like *rule-based machine translation systems* and the direct, transfer and *interlingua* approaches. In this section, systems based on data (*corpus-based machine translation systems*) are also presented, that can be divided into two different types: the ones based on statistics (*statistical-based machine translation system*) and the ones based on examples (*example-based machine translation systems*). In the last two sections, *neuronal systems* and *hybrid systems* are presented.

2.1. WHAT IS MACHINE TRANSLATION?

According to Dorr, Jordan and Benoit (1999), machine translation (MT) is an automated translation, it is the process by which computer software is used to translate a text from one natural language to another.

This means that MT is focused on obtaining a target language text from a source language text by means of automatic techniques (Costa-Jussá, Fonollosa, 2014). In this process, there is no intervention of human translators. The source text is exclusively processed by computer systems. This characteristic is the main distinction between MT systems and computer-aided translation, in which the intervention of the human translator is crucial. In the latter, human translators do the translation work, while being aided by language resources and tools, such as dictionaries, translation memories, and glossaries.

Results obtained with MT processes are variable and depend on different factors, such as the genre and domain of the source text, the aim of the text, and the syntax and the lexicon. Most of the time, the generated text is a “raw” translation: its quality is poor. Therefore, in order to achieve a better level of quality in translated texts, human

intervention is needed, either by pre-editing or post-editing the source text or the target text, respectively. MT systems generate the first version of a translation, which has to be edited by a human to produce a high-quality translation. This edition is crucial to avoid some linguistic problems, such as ambiguity, either lexical or structural that can be generated by the MT systems. Furthermore, MT systems can be also used with the aim of creating a rough version of the target text and not only of producing a high-quality translation; thus, enabling access to the meaning of the source text. MT plays a crucial role in the contemporary society, in particular, because of political and social reasons: society is currently characterized by a multicultural environment, as we can see for example in Europe, multilingual for nature, in which translation is fundamental in human interaction, as machine translation makes the communication between people easier. Its importance has also grown thanks to the expansion of the Internet, the most used communication tool in the world, in which translation is a connecting process among people who speak different languages.

2.2. MACHINE TRANSLATION: HISTORICAL OVERVIEW

Machine translation is a field that investigates the development of computer programs that are used to translate text or speech from a language to another. The first efforts to develop a software that was able to translate are dated from the mid of the 20th century. Since the beginning, researchers were focused on the translation of technical texts, because there were fewer differences between language productions in different languages from a cultural and linguistic perspective, than, for example, in literary texts. The demand for translation was very high, but the results were not satisfactory. After a rough translation performed by the software, a human post-edition was required, which is expensive and time-consuming.

The first attempts to achieve full automated translation began in 1949, after the Second World War, when Warren Weaver created a memorandum that helped to catch the attention of the researchers in the field of MT in the United States. In this memorandum, he explained the importance of achieving automatic translation of scientific and technical texts, and he proposed methods to solve ambiguity, a well-known linguistic issue in natural language texts. From then till the mid-1960s, the developments made by the researchers led to high expectations and optimism. Thanks to the creation of large bilingual dictionaries and glossaries, and the developments in

computation and in formal linguistics, great improvements in quality were possible. By that time research groups had been established in many countries throughout the world, as the result of these achievements and the enthusiasm they created within the MT community. This first period of work in the field of MT is dominated by the "direct approach", presented in the 2.3.1. section, which means that a word-for-word translation was performed, by means of the use of bilingual dictionaries, without any type of linguistic analysis, and the results were obviously not satisfactory. Due to the huge investment and effort in the field, in 1964, the American National Science Foundation set up a committee, the Automatic Language Processing Advisory Committee (ALPAC) to examine the developments achieved and the opportunities created. In this 1966 report, ALPAC considered that MT was slow, less accurate and much more expensive than human translation and stated that there was no way of progress and so no need of further investments in this area. The results produced by MT systems were considered poor in terms of quality. The ALPAC recommend, instead, the development of machine aids for translators, such as automatic dictionaries (ALPAC report, 1966). The ALPAC report brought the research in MT to a virtual end, due to the fact that in the former decade the expectations were too high, considering that the obtained results were not good enough. From that moment on, researchers focused more on other fields, like computational linguistics and artificial intelligence (AI).

Despite this report and its finding, research did not stop completely. In particular, in the 1960s, some groups in the USA and in the Soviet Union were still working on MT, especially in the translation of technical and scientific documents from English to Russian and vice versa. In the 1970s, there was a high demand of MT for different reasons. In Canada, for example, there was an important demand in term of translation of official documents from English to French and vice versa.

In 1965 the TAUM project (Traduction Automatique de l'Université de Montréal), was put in place at the University of Montréal. It accomplished two major achievements: the Q-system formalism for manipulating linguistic strings and trees and the MÉTÉO system that was used for translating weather forecasts (Hutchins: 2010), from French into English, and which can be considered as the first completely automatic translation system.

In 1970, another operational MT system was launched: SYSTRAN, developed by Peter Toma. Its first version provided translations for the language pair English-Russian and was used by the USAF Foreign Technology Division and by NATO. In 1976 it was purchased by the Commission of the European Communities, in the English-French version and was later extended to new language pairs. The main rivals of SYSTRAN were LOGOS, at first, and METAL at a later stage. LOGOS appeared in 1972. It was a system for translating aircraft manuals from English to Vietnamese, by using contextual clues, which allowed to deduce meanings. METAL appeared at the end of the 1980s and it provided translations for the language pair German-English.

In the end of the 1980s, MT was used in different countries and the language pairs covered by this type of systems were growing. There was the need of a multilingual transfer system that met the need of the European Communities to have translations in all the languages of the Community. 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, so the EUROTRA project was launched with the aim of achieving complete and satisfactory translation in all the languages supported by the project, that at the time were 9.

In the end of the 1980s, a team at Carnegie-Mellon University developed the KANT system, a knowledge-based MT system that used lexicon, grammar and semantic resources (Nyberg, Mitamura and Carbonell: 1997). During this period, translators were not satisfied with the quality of the results of MT systems: they wanted to be in control of processes and of translation assisted tools.

For this reason, in the 1990s, new methods were introduced in the MT field. These new systems were no more *rule-based approaches* based on linguistic rules, but *corpus-based approaches*: deducting rules from corpora. In Japan the first MT systems based on examples (*example-based MT systems*) were created, as we will see in section 2.3.

The first example of this new approach was a system called Candide, developed in 1989 by a group at IBM. This system used word correlations between the source and target languages to output a translation of a given source sentence. The *example-based approach* was developed in the same period. The system extracts from a database of

corpora, equivalent phrases that have already been aligned by a statistical or rule-based method. A deeper analysis is presented in the section 2.3.1.

This period, the 1990s, is also marked by the higher usage of Internet. This, of course, was going to have an influence also on MT systems. There was the presence of new MT software products specialized in the translation of web pages and e-mails. Translation software for personal computers was also made available. According to Hutchins (2010), the first example of this kind of software is the French SYSTRAN. After that, other free online MT services were also developed, such as Babelfish, on the AltaVista site, that offers SYSTRAN versions to translate French, German and Spanish into and from English. The quality of the online services was often poor, but it was enough to get the general meaning of the text, and therefore enough for people that only wanted a rough translation of a given source text.

From the first years of 2000, we can observe a large use of *statistical-based MT systems*, due to the large number of available corpora, online and free tools for the text alignment, but we can also verify a large use of hybrid systems that try to combine parts of *rule-based MT systems* with parts of the *corpus-based MT systems*, as detailed in the next section.

2.3. PARADIGMS OF MACHINE TRANSLATION SYSTEMS

Among all MT systems we can make a general distinction between the ones that are knowledge-based and the ones that are data-driven. This classification allows us to understand the type of resources the translation process uses. A third type of MT system is considered, that of hybrids systems that combine at least two MT paradigms, and also a fourth type, the most recent, the neuronal paradigm. In the next sections all these paradigms are introduced and then analyzed in general.

2.3.1. RULE-BASED MT SYSTEMS

In *rule-based MT systems* (RBMT), the method used for translation is based on rules derived from grammatical rules and linguistic principles, such as morphological, syntagmatic and syntactic principles. The aim of this system was to produce high-quality translations, converting the source language structures into target language structures, but, at the same time, it was very costly, and it requested extensive manual work.

In RBMT, we can distinguish three different approaches in which the translation process can be performed: direct, transfer and *interlingua*. They can be distinguished on the basis of the analysis that is involved in the translation process in these specific systems.

The direct approach can be defined as the "first generation" of MT systems. It was adopted by most early MT systems, like Texas's METAL and Montreal's TAUM (Slocum: 1985), around the 1950s, till around 1990s. In this approach, the system translates words of the source text, directly into words of the target text, without any intermediate stages. The analysis of the source language is limited to a basic morphological analysis, which identifies word endings and reduces inflected forms to their uninflected forms, i.e. the basic information needed to produce a target text. This leads to a frequent mistranslation at the lexical level and inappropriate syntactic structures. The resources used are generally limited to a bilingual dictionary, providing target language word equivalences (Hutchins: 1978). Some local reordering rules of the text are present in these systems, in order to give more acceptable target language output. These are systems that require a minimum of linguistic, consequently, the resolution of some problems is very difficult, such as lexical ambiguity and inappropriate syntax structures.

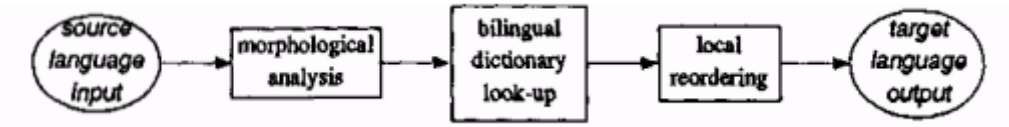

Figure 1 – Direct approach. (From Hutchins and Sommers: 1992)

This approach, despite the lack of sophisticated linguistic information that allows for a correct translation process, can give positive results when the two languages can be considered "close", as the ambiguity and order problems are reduced to the minimum. On the other side, as we explained, this approach has several problems, i.e. lack of linguistic information and difficulties in solving the ambiguity. This means that this approach can only produce acceptable results when relying on a post-editing process.

The transfer approach is based on a deep analysis of the source text. The main objective of these systems is to obtain a target text that is correct at a syntactic level, transforming the representations of the source text into syntactic proper representations of the target text. The representations are language-specific: the source language intermediate representation is specific to a particular language, as is the target language intermediate representation (Hutchins and Somers: 1992). In this approach, a translation process involves three different stages: the analysis of the source language, a syntactic and a semantic transfer, and the synthesis and creation of the target language.

After the first phase of analysis of the source text, there is the transfer stage. In this central stage, there are some mapping rules between the source and the target language, which operate from the "surface" of the target and source text till the deeper structures and representations. Each phase of the process uses specific dictionaries: a dictionary of the source language for the analysis phase, a bilingual dictionary for the transfer phase, and a dictionary of the target language for the creation of the text in the generation phase.

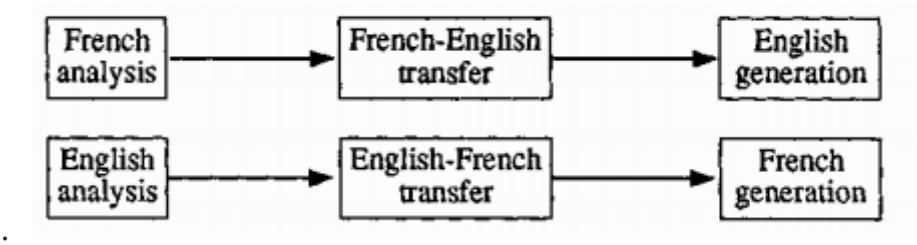

Figure 2 – Transfer approach. (From Hutchins and Sommers: 1992)

The translation performed by these systems are acceptable translations, because these systems can solve some ambiguity issues of the text, based on the first phase of syntactic analysis, in which is possible to recognize the lexical category of the words of the source text. On the other side, they use complex rules that vary according to the language pairs used, or sometimes the rules are not complete enough to give all the requested information.

The *interlingua* approach is basically aiming at the creation of "meaning" representations common to more than one language, to generate the target text translation. In this approach, the translation process is thus a two-stage process: from the

source language to the *interlingua* and from the *interlingua* into the target language. The *interlingua* is an abstract representation of the language, it includes all information necessary to the generation of the target text (Hutchins and Somers: 1992). This abstract representation is suitable for two or more languages and this is, therefore, an advantage of this approach, but the concrete definition of this abstract representation can create some difficulties.

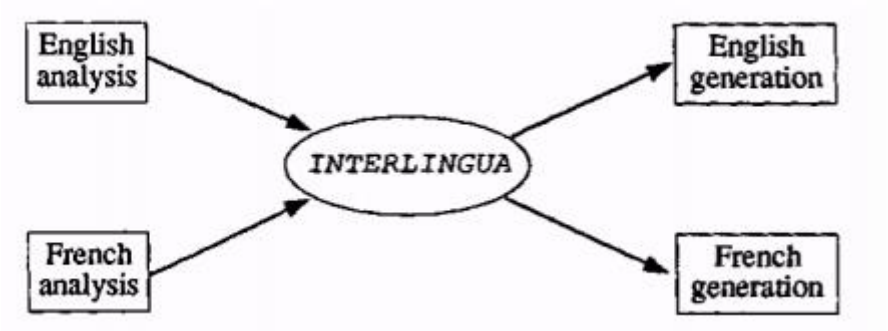

Figure 3 – Interlingua approach. (From Hutchins and Sommers: 1992)

After having described the three approaches, we can resume them in one triangle, the Vauquois Triangle (1968):

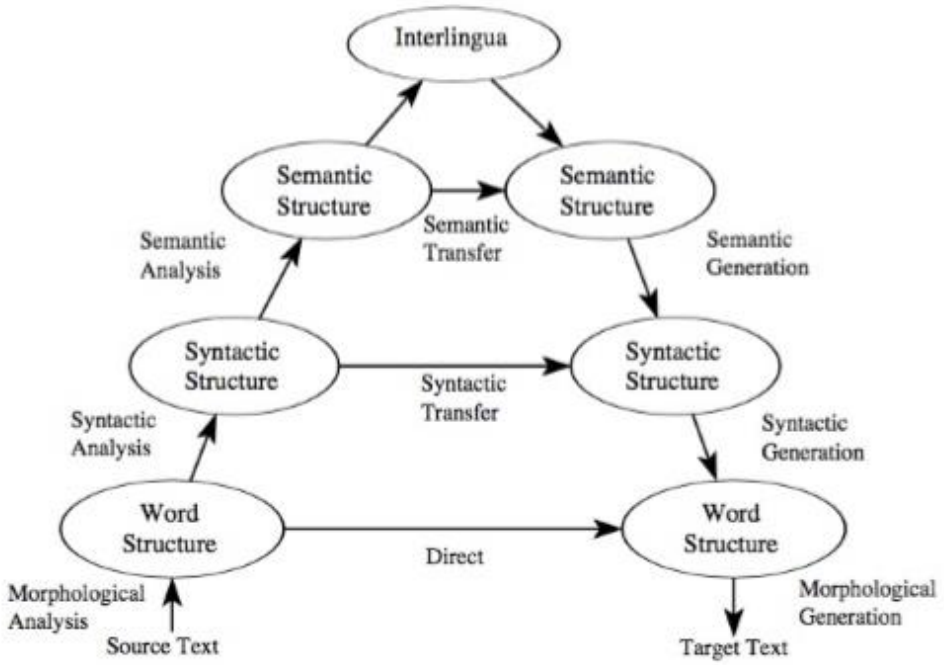

Figure 2 – Simplified version of the Vauquois Triangle. (From Dorr et al., 1999)

The figure above illustrates the three different architectures that can be used in an MT process. In general, the type of translation process depends on the level of analysis. On the left side of the triangle is represented the analysis of the source text and on the right side is represented the generation of the target text. The base of the triangle represents the first approach, the direct one, the transfer is only at a lexical level, a word of the source text is replaced by one of the target text, is a word-to-word translation. In the central part of the triangle are represented the systems that make a deep analysis of the structures of the target text, at a semantic and syntactic level. After this step, the information is transferred for the generation of the target text. At the top of the triangle is represented the *interlingua* approach, a deep analysis of the source and target language is given, providing an abstract representation, in order to generate the translated text.

2.3.2. CORPUS-BASED MT SYSTEMS

In the 1990s, after de decadence of the RBTM dominance, the *corpus-based approach* started to be predominant, built upon faster-running computers and the availability of large bilingual corpora. The corpora are constituted by parallel-translated texts, either bilingual or multilingual. In this CBMT approach, we can distinguish *statistical MT systems* and *example-based MT systems*.

The first example of statistical methods was developed within the Candide project, developed by researchers at IBM in 1988. The project was based on a corpus of French and English Canadian parliamentary debates. Statistical MT systems use corpora and pure probabilistic calculations to produce translations for such reports.

The translation processes performed by the statistical method can be described following three important phases: alignment, calculation of the correspondences, and reordering. In the phase of alignment, sentences, words and sequences of words are aligned in order to achieve correspondences. After this first phase of alignment, correspondences among words are calculated, by applying algorithms and probabilistic calculations. In the last phase, the reordering of the words is done, in order to obtain a more accurate and fluent translation.

The second approach that appears in this period is the *example-based MT* (EBMT) *approach*, first proposed by the Japanese Makoto Nagao, in 1981, although his project was only implemented towards the end of the decade. The main idea of this new

approach was to find correspondences among words, with the aim of achieving the best option between the source language and the target language, by using texts that were already translated by other translators. This system takes examples, i.e. fragments of sentences, from dictionaries and pairs that set lexical equivalences, to create a bilingual corpus. This approach is divided into three different phases: correspondence, alignment and recombination. In the first phase, examples are selected and extracted from the corpus. After the selection of the examples, the system finds the correspondences and also stores the examples that are useful for the translation. In the second phase, the alignment phase, the phrases in the source and target texts from the parallel corpus are aligned. In the last phase, the system recombines and reorders all segments into translation units.

2.3.3. HYBRID SYSTEMS

After a long period in which the two different approaches mentioned above, RBMT and CBMT were used, some MT researchers developed hybrid systems to further improve the performance of MT systems.

Hybrid approaches attempt to combine characteristics of both *corpus-based machine translation systems* and the *rule-based machine translation systems*, to produce better quality translation, combining 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.

Hybrid systems can be either guided by RBTM, in which corpus information is integrated into a rule-based architecture, or by CBMT, in which linguistic rules are integrated into a corpus-based architecture.

Hybrid systems, through the combination of the already mentioned systems, RBMT and CBMT, aim at extracting the best features of each approach, allowing the exploration and the improving of both systems.

2.3.4. NEURAL MACHINE TRANSLATION SYSTEM

Neural machine translation systems (NMTS) are a recent approach of the traditional statistical machine translation (SMT) that takes inspiration from the neuronal system of the human brain. According to Yonghui, Schuster, Chen, V. Le, Norouzi (2016) the

strength of NMTS lies in its ability to learn directly, in an end-to-end fashion, the mapping between an input text and its associated output text.

The machine translation system used at Unbabel is a NMTS and, at the time the corpus used in this work was translated, the MT system was Moses. Moses started to be used at Unbabel in September 2016. Before Google Translator was used.

Moses is an open-source SMT system trained on parallel data of two different languages in which each sentence in one language is aligned with its equivalent in the other language. It is composed by a training pipeline and a decoder. In the pipeline all the stages of the translation process are included: tokenization of the text (dividing it into smaller parts called tokens), word alignment, the creation of a language model, and tuning (the definition of criteria for the selection of the best possible translation). The decoder identifies the sentence with the highest score, according to the translation model, and selects it as the translation of the input text.

2.4. SUMMARY

The main objective of this chapter is to give a theoretical overview of machine translation, in order to outline the characteristics of different MT systems and to identify the differences and strengths of each paradigm. This allows us to define the machine translation system used at Unbabel and to describe its pipeline, as we see in chapter 3.

3. UNBABEL'S PIPELINE

In this chapter, we present the Unbabel's Pipeline, i.e. the workflow of the Portuguese start-up. By doing so, we explain how the translation process is done, in a step-wise perspective. In the final section, we also present the Natural Language Processing (NLP) tools that are used at Unbabel in this complex pipeline.

Unbabel, a Portuguese start-up headquartered in the USA and backed by Y Combinator, combines human editing and machine translation into an online translation platform. It offers translation services involving several language pairs and relies on a community of 50,000 editors, which work online on the company platform. Currently covering 28 languages, Unbabel has a growing list of supported languages.

Knowing that the 75% of the world does not speak English, which represents a big problem for companies that want to have an international presence, Unbabel presents itself as a start-up that can connect people from different languages, breaking down language barriers. Translation can be used, for instance, to improve the performance of a customer service team: by answering thousands of tickets per day, or by translating FAQs and knowledge centers. Unbabel is targeting also distinct contents, such as product descriptions, blogs, video subtitles, user reviews, and other UGC (User-Generated Content) and documents.

Unbabel adopts a crowd translation model, which involves multiple translators for a single translation project[1]. The process used at Unbabel consists of dividing texts into small chunks and distributing them to translators. This allows for cost and time reductions.

Unbabel does not work only with professional translators, but also with bilingual speakers. Combining bilinguals with map-reduced distributed methods allow for the company to be faster in delivering translations and in reducing the costs.

[1] See Yamamoto, Aikawa, Isahara, (2012).

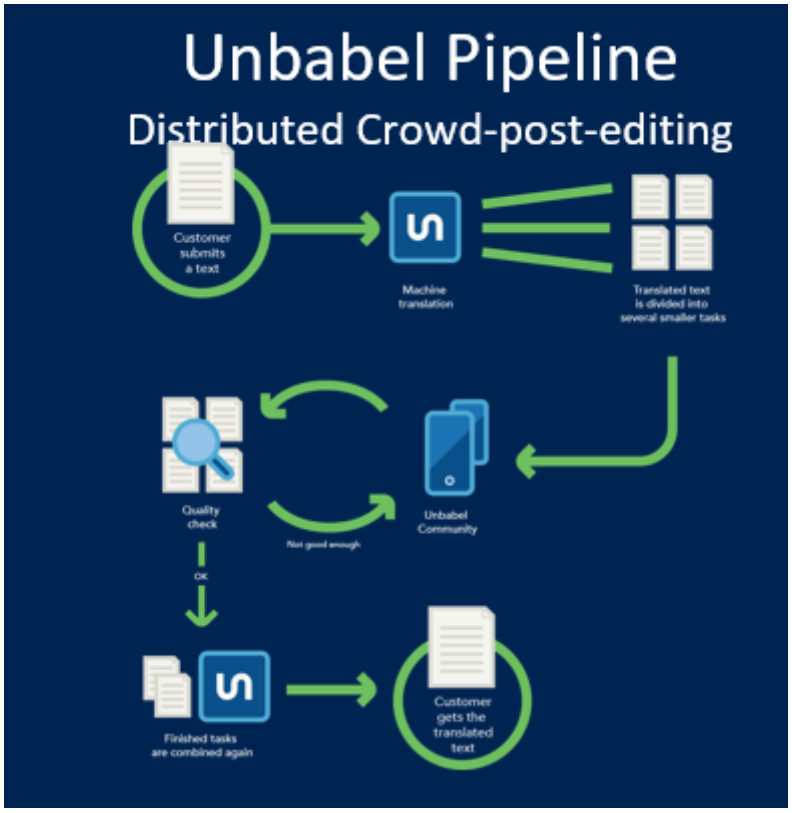

Figure 5 – Unbabel's Pipeline

Figure 5 represents Unbabel's pipeline, the process that is followed to produce satisfying translations in an efficient way. Firstly, a customer submits a text, it is analyzed, a step where a range of factors that will influence the process in the pipeline is detected and determined. This can include customer glossaries, style and register guides. The content of these texts can be, for example, customer service e-mail on a platform like Salesforce, Zendesk or Freshdesk.

We will now zoom in and present a more detailed analysis of each step that is presented in Figure 5.

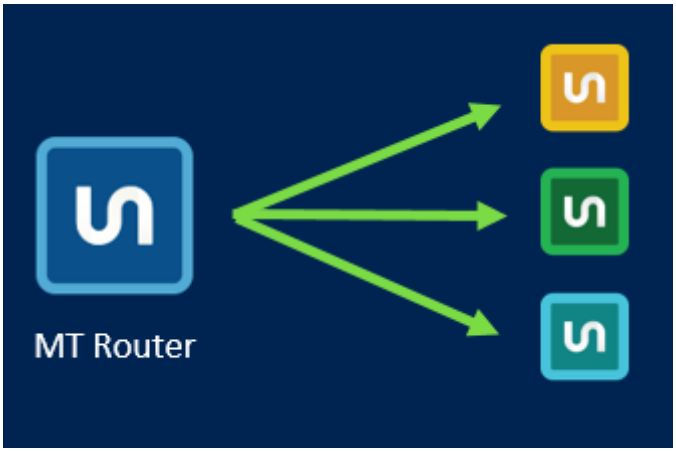

Figure 6 – Machine Translation step

Once all the markups are removed, the glossary words are extracted, and translation memories are kept, the text is sent to the neural machine translation system, adapted by content type and sometimes by client. The machine produces a first translation and then the obtained text is post-processed. In the post-edition process, the editors are assisted by automatic NLP tools, which detect or highlight potential errors, so the editors may correct them. These tools include a spellcheck and a Smartcheck.

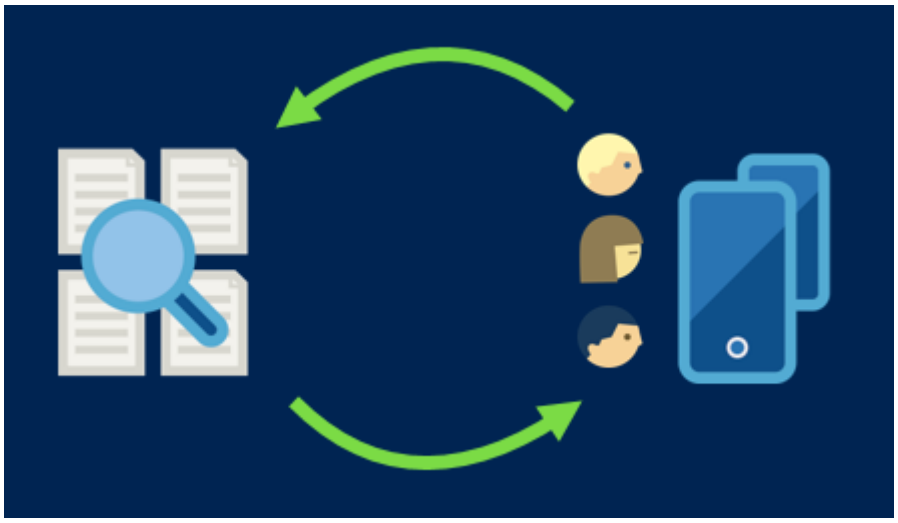

Figure 7 – Map step

The third step illustrated in Figure 5, represents the second step of the Unbabel pipeline. Here, the source and target text are divided into chunks, made available on the platform for editors and then distributed for the translators of the community to edit them. In this platform, both the source and target text are shown. The human translators will also find client instructions, register and style guidelines, along with warnings or suggestions provided by the Smartcheck. Once a chunk has been edited, quality is automatically checked, through means of a Quality Estimation algorithm, in order to determine whether the segment needs to be edited again or if it is ready to be delivered to the client. In the first case, when the quality is poor, the previously edited chunk is sent to the platform again and ascribed to a senior editor.

Figure 8 – Agglutination step

In the third step, the combination of all chunks is done and then the text is sent to the client. Sometimes, the text is sent to a senior editor before being sent to the customer, to double-check the edited text and to correct possible inconsistencies, and thus improve fluency.

Quality at Unbabel is growing constantly, due to the constant feedback between improving the algorithms of the system and learning from the data and the results obtained. Following these two features, a way to improve the quality of the system is through the annotation process, that we present and analyze in the next chapter. The annotations are performed on a weekly basis on works already automatic translated and edited. The annotators, linguistic experts, analyze and annotate errors, to understand what is working in the pipeline and what still can be improved, in order to reach better quality standards.

The translation process just presented is really fast. This is because the intervention of the human editor only takes place in a phase of post-edition and Unbabel can rely on a large number of human editors, who can process large volumes of texts in a short period of time. The translation is made by the machine and not by the human, and this means a reduction of time and costs. But, human intervention is fundamental to guarantee a high quality.

3.1. TOOLS TO DETECT ERRORS: SMARTCHECK AND DEPENDENCY PARSER

At Unbabel, apart from the MT system that is the core of the translation process, there are other NLP tools that are used to improve the quality of the translation, also helping to speed up the translation and the post-editing process. We are referring to the Smartcheck and to a dependency and syntactic parser developed by Martins et al. (2013).

3.1.1. SMARTCHECK

With the help of researchers in Natural Language Processing (NLP) and other field specialists, Unbabel was able to develop tools like Smartcheck, which provides alerts and suggestions to the community of editors to aid with proofreading. This tool helps to reduce the time needed to accomplish a post-edition task and to make the process faster, while reducing errors. The Smartcheck helps translators during the post-edition process, not only pointing out possible errors but also offering helpful hints to correct the problem.

The Smartcheck can show warnings or errors. In the first case, the word or expression is underlined in green and the editor can decide whether to introduce or not any changes in the translation. In the second case, the word or expression is underlined in red and the editor has to read the message shown by the Smartcheck and he/she can decide whether to mark the error or to ignore it, before submitting the translated text.

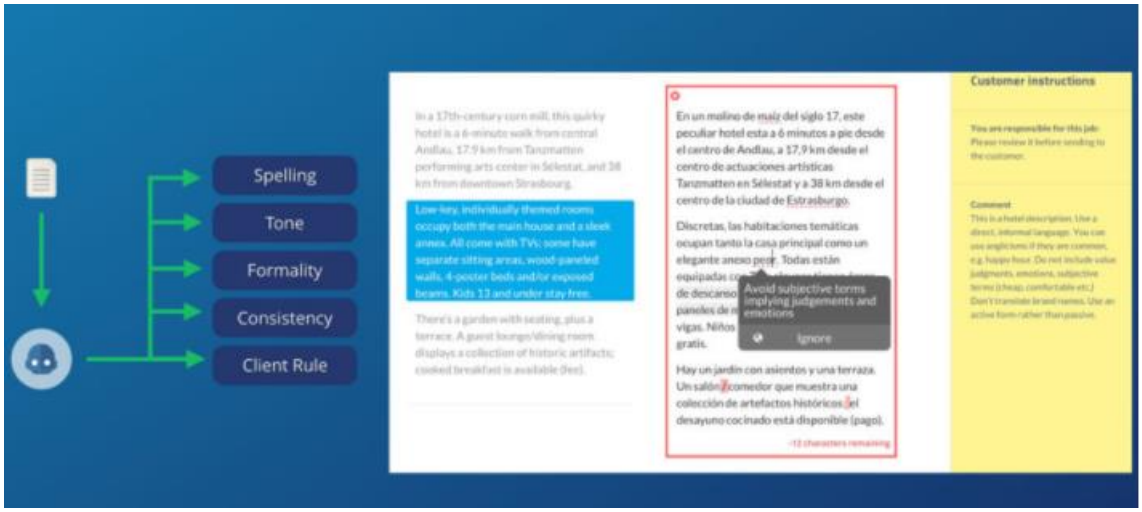

Figure 9 – Example of a Smartcheck suggestion.

The Smartcheck includes a battery of tests in order to identify and tag different issues that may occur in a text. However, not all the checks are available for all languages. Those available for Italian, already tackled by Comparin (2017) 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 misspelt 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".

Register: checks if the register used in the text is correct and if it is coherent to the one set by the client.

The Smartcheck does not automatically edit the text; rather it provides only warnings or suggestions. It is the human editor who has to take the final decision, in order to improve the quality of the translation. The tool should show only relevant warnings or suggestions, because it takes time for the editor to go through all the suggestions, thus, too many warnings or suggestions would result in a slower process instead of a faster one.

3.1.2. DEPENDENCY PARSER.

A dependency parser is a syntactic analyzer that provides information regarding the structure of a sentence.

A parser is, therefore, an important tool in the process of automatically establishing the correct syntactic dependency between constituents occurring in a sentence and to provide part-of-speech (POS) tagging of each word. It is a very powerful tool to solve both syntactical and lexical ambiguity issues, depending on the relation between constituents and the meaning of a constituent depending on the POS.

The parser used at Unbabel was developed by Martins, Almeida and Smith, (2013).

> "The parser is fast, accurate, direct nonprojective, with third order features. Our approach uses AD3, an accelerated dual decomposition algorithm, which we extend to handle specialized head automata and sequential head bigram models. Experiments in fourteen languages yield parsing speeds competitive to projective parsers, with state-of-the-art accuracies for the largest datasets" (Martins, A., Almeida, M., Smith, N.: 2013).

The parser is used to analyze data in order to disclose more specific information to the Smartcheck, with the aim to improve the precision of the corrections. The parser supports all the morpho-syntactic information needed to process the Smartcheck rules.

The information provided by the parser is given for each word, its POS and values for specific features (for example, number, gender, person, mood, tense, verb form). A dependency tree representing the syntactic structure of the sentence is also provided.

3.2. SUMMARY

The presentation of the Unbabel's workflow allows us to understand how the translation process is performed at the start-up and the NLP tools that are used. This analysis makes possible the evaluation of the translation process and it focuses on what can still be improved, in order to increase the translation quality. This improvement is possible also thanks to the annotation process of the target texts after the first human post-edition, as it is explained in the next chapter.

4. ANNOTATION

Error annotation is a process that aims at identifying and categorizing errors in machine translated texts. In particular, we analyze the error annotation performed at Unbabel, which is not an automatic process, but it is made by humans. Error annotation can be performed either by one annotator or by more annotators, as it is presented in chapter 5.

This chapter focuses on the role of the annotator in the Unbabel community, but also on the tool, which allows the annotators to analyze and annotate the texts. The tool assigns a quality score according to the number of errors and its severity. The annotators must follow instructions during the annotation process: general guidelines of annotation and specific customer instructions. After this first section, which is more descriptive, we are going to present some proposals on how to improve both the general annotation guidelines and the more specific instruction given to annotators working on Italian.

4.1. ANNOTATION TOOL

The next objective of the European Union is to have a Union free of barriers, in particular, language barriers, to achieve a free flow of ideas, commerce, and people. Nowadays, 27 official languages are spoken in EU and many of them are not supported by machine translation technology, due to the fact that these languages are considered as minority languages because of historical events, political issues or just because they have entered in the EU in the last years. This can bring to a variable translation quality so that experts, in order to assess the quality of the machine translation systems used, created a specific project: QT21. The Quality Translation 21 is a machine translation project that aims at bringing down all language barriers and improving the quality of translation. Another goal of the project is to enhance statistical and machine-learning based translation models, to improve the evaluation and continuous learning from mistakes, guided by a systematic analysis of quality barriers, informed by human translators. The QT21 project developed a framework, the Multidimensional Quality Metrics (MQM).

MQM is a comprehensive framework for developing MT quality assessment metrics; it defines a typology of language issues to identify specific problems and to

underline the strengths and weaknesses of a translation. “The MQM framework does not provide a translation quality metric, but rather a framework for defining task-specific translation metrics” (Lommel, 2015). Some aspects of the quality of a translation are assessed and categorized in this framework, for example, the accuracy, fluency, and verity.

The study of the MQM framework helped us to analyze the annotation process performed by Unbabel and allowed us to understand if the tool used was adequate, in order to reach a high quality in the translated texts.

The annotation process is possible thanks to a tool developed by Unbabel that is used to assess the quality of texts. The tool shows two blocks of text, the source text, on the left, and the target text, on the right, as well as the annotation area, with the glossary terms highlighted.

The top bar shows the number of the job, composed by the source text in English and the target text in Italian, that is being annotated, and the number of jobs that still have to be annotated, the title of the batch, composed by all the jobs, usually 25, that have to be annotated on a weekly basis: only a batch per week is available. The QT21 score and the register, more specifically which register should be used, formal or informal, as well as the client’s instructions are also shown.

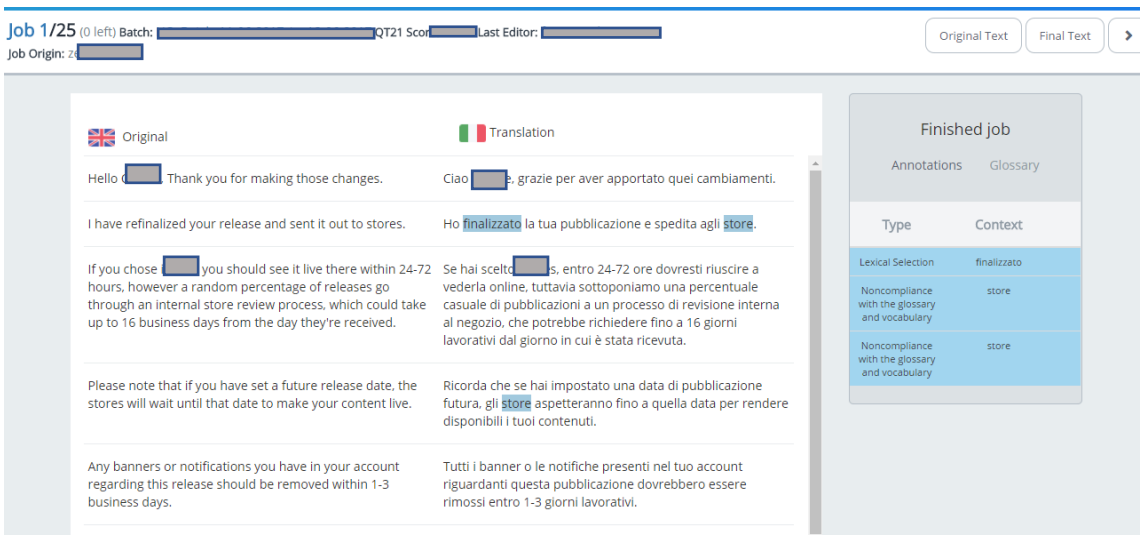

Figure 10 – The annotation process. The image is edited; we deleted all the client’s information due to privacy concerns.

Annotators are asked to identify errors in the target text, the text translated by the editor, and to classify them according to the taxonomy of errors, presented in subsection 4.2.1., shown on the panel on the right, once a word, or a group of words, is selected.

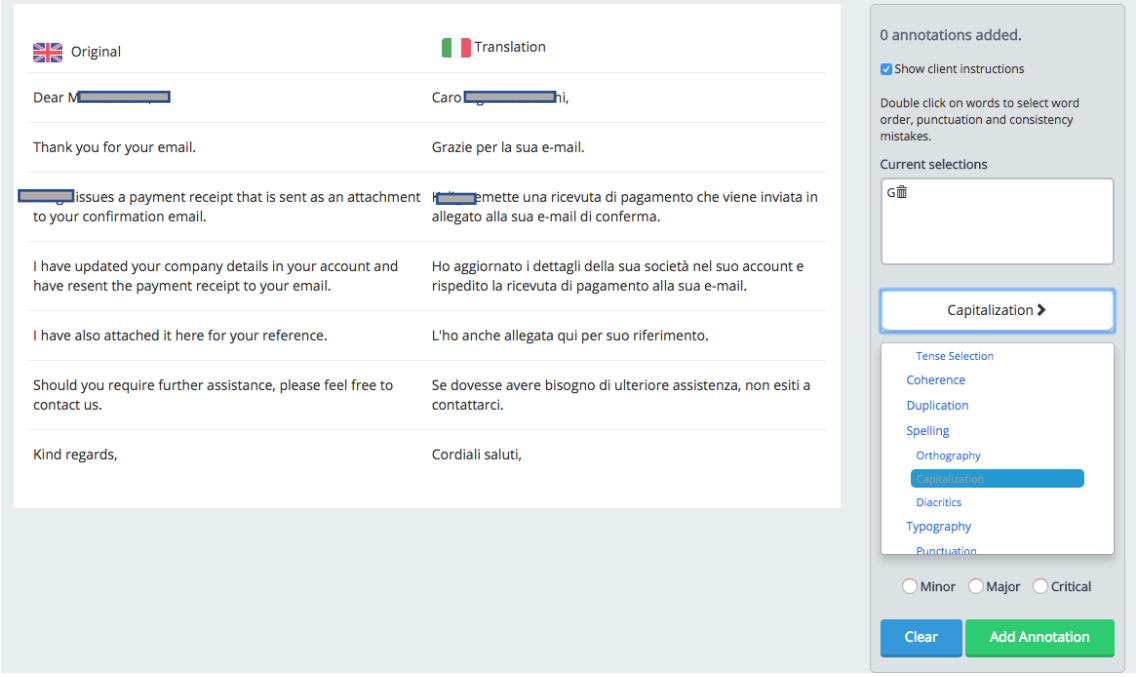

Figure 11 – Error identification. The image is edited; we deleted all the client's information due to privacy concerns.

Once the annotation process is finished, the annotator is asked to assess the fluency of the translated text on a scale from 1 to 5, where 1 refers to a very low fluency and 5 refers to a very high fluency.

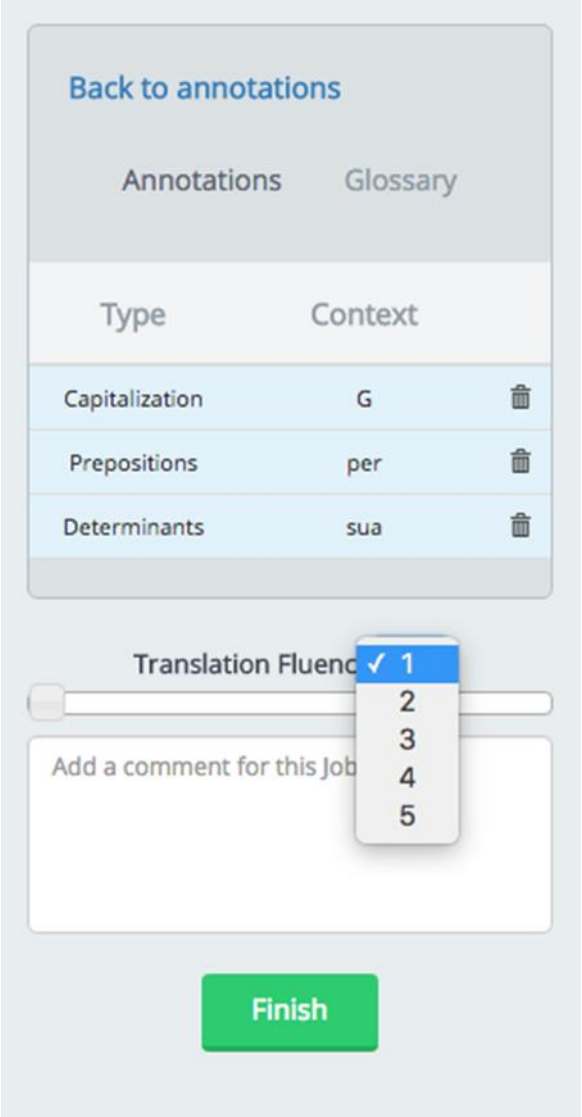

Figure 12 – Fluency Assessment scale.

The minimum unit available for annotation is a word, and the maximum is the whole expression or sentence. However, a whitespace can be selected when there are extra spaces, when the error is a punctuation error, and when a word is missing.

4.2. GUIDELINES

The role of the annotator is to analyze, identify, and categorize errors in a text. Thanks to the annotation we can determine the average quality of a translation or of a group of translations. The annotator always has to follow the instructions of the client, for example, requests for a certain register or stylistic guides.

Annotators must follow some directives during the annotation process, so that parameters for selecting error categories in the typology and assessing the severity of the errors are used uniformly. In this section, both the general guidelines, along with the typology of errors used at Unbabel, that are suitable for all language pairs, and the guidelines for Italian, specifically, are presented.

4.2.1. TYPOLOGY

In this section is presented the typology of errors used at Unbabel for the annotation process. The error types are divided into different categories: accuracy, fluency, style, terminology, language variety, named entities, formatting and encoding.

ACCURACY: errors that are related to the translation of the meaning in the target language.

Mistranslation: an incorrect translation of the word, or expression in the target language.

- **Overly literal**: direct translation, literal translation of idiomatic expression, sentences and structures.
- **False friend**: wrong translation of a word in the target language that looks and/or sounds similar to the word in the source language, but with a different meaning.
- **Should not have been translated**: content that does not have a translation in the target language and that does not have to be translated.
- **Lexical selection**: terms translated incorrectly in the target language.

Omission: omitted words, sentences or even paragraphs in the target text.

Untranslated: content is not translated in the target text.

Addition: insertions of contents in the target text.

FLUENCY: errors that affect the quality of a text; if a text is readable and well-written.

Inconsistency

- **Word selection**: translation of a same content differently throughout the target text.
- **Tense selection**: temporal cohesion throughout the target text is not corrected.

Coherence: the text is not clear and consistent, difficult to be understood.

Duplication: repetition of the same content in the target text.

Spelling

- **Orthography**: wrong orthography.
- **Capitalization**: wrong use of capital and small letters.
- **Diacritics**: wrong use or missing symbols.

Typography

- **Punctuation**: wrong use or missing punctuation.
- **Unpaired Quote Marks and Bracket**: one of the quote marks or brackets is missing.
- **Whitespace**: addition or omission of whitespaces.
- **Inconsistency in character use**: especially added for Chinese or Japanese, when the characters that are used are inconsistent.

GRAMMAR

Function words

- **Prepositions**: wrong use or missing preposition.
- **Conjunctions**: wrong use or missing conjunction.
- **Determiners**: wrong use or missing determiner.

Word form

- **Part-of-speech**: wrong use of the word category in the target language.
- **Agreement**: inconsistency in number and person between words.
- **Tense/Mood/Aspect**: wrong use of tense, mood and aspect of a verb.
- **Word order**: wrong word order of the target language.
- **Sentence structure**: wrong sentence structure in the target language.

STYLE

- **Register**: use of a wrong register, informal instead of formal, or vice versa.
- **Inconsistent register**: incoherent use of a register, presence of both registers throughout the target text.
- **Repetitive style**: repetition of expressions or words.
- **Awkward style**: presence of unnaturalness in a sentence throughout the target text. Used when the error does not fit in any other category.

TERMINOLOGY: error in the use of the terminology.

Noncompliance with client or company style guide: translation does not follow the given directives.

Noncompliance with the glossary and vocabulary: translation does not follow the glossary.

WRONG LANGUAGE VARIETY: wrong use of language variety: added to differentiate the European Portuguese from the Brazilian Portuguese, the European Spanish from the Latin American Spanish, and the British English from the American English.

NAMED ENTITIES: wrong translation of names, products, and organizations.

- **Person**: wrong translation of a person's name
- **Organization**: wrong translation of an organization's name.
- **Location**: wrong translation of a geographical name.
- **Function**: wrong translation of a person's position or charge.
- **Product**: wrong translation of a product's name.
- **Amount**: wrong use of a unit of measure.
- **Time**: wrong use of the time format.

FORMATTING AND ENCODING: errors in the layout of the text.

4.2.2. SEVERITY

Once the annotator, as defined the type of error he/she has identified, he also has to decide on the severity of this error, i.e. that he/she must decide whether the error is minor, major or critical.

• **Minor error**: errors that do not introduce a big loss of meaning and do not produce misunderstanding nor confusion. This kind of error leads to a loss in the quality of the target text and to a loss in the clarity and fluency of the target text. They can be, for example, punctuation errors, capitalization errors, and repetitions.

• **Major error**: errors that lead to a lack of meaning, the comprehension of the text results more difficult. This type of error can change the meaning of the target text. For example, lexical selection, agreement, noncompliance with glossary, etc.

• **Critical error**: errors that lead to a complete lack of meaning, making impossible the comprehension of the target text. This type of error can also affect the company's reputation or may carry health, safety, legal or financial implications. It has a very negative impact on the client's opinion towards the product.

An annotator can only choose one category to associate to each selected segment in the text. Once the annotation is performed, the quality of the translation can be measured, thanks to the MQM: the higher the score, the better the quality of the text, considering 95% professional quality.

4.2.3. ITALIAN GUIDELINES

At Unbabel, contrarily to editors' guidelines, the annotation guidelines are for all the languages. The work of an editor consists in editing a text, that was translated by the machine translation engine. Jobs generally have less than 100 words, so that they are easy and fast to edit. Once this process is concluded, the annotator is the person that reviews these translations to perform the annotation process that we described in the previous section. Editors work directly on texts translated by the machine; annotators, instead, control the quality of reviewed texts by editors. Both editors and annotators, in distinct ways, contribute to increase the quality of the outputs. However, we can try to give some proposals to help the annotator in this process, so that he can categorize and analyze the errors in the most correct and linear way.

4.3. IMPROVEMENTS TO THE GUIDELINES

During the annotation task we performed, we were also trying to improve the guidelines, especially the guidelines for Italian, so that certain errors, that may lead to a slight agreement[2] between annotators, can be categorized without any problems, this way allowing the annotators to reach an almost perfect agreement and an objectiveness, which is not always easy, due to the fact that annotators are human and the interpretation of an error can be different from one annotator to another. Reach an almost perfect agreement was possible thanks to the definition of decision trees by the two annotators for Italian working at the company at the time this work was developed. The decision trees are analyzed further in the next chapter.

Table 1 resumes the most common errors in general, how they are typically categorized, and the associated severity degree. It also displays the instructions that annotators have to follow, in order to take always the same decision when trying to categorize an error region. This is not a mandatory table, i.e. when the meaning of the sentence of the target text is changed from the one of the source text, the severity can vary. This table was made by Italian linguists in order to improve the quality of the translated texts and to improve the inter-annotator agreement, as we will observe in the next chapter.

[2] According to Landis and Koch, 1977, is a specific term presented in section 5.1

MINOR	MAJOR	CRITICAL
Accentuation missing	Agreement	Translation does not make sense
Punctuation errors (such as misplaced commas)	Wrong grammatical subject	Word selection that may have a negative influence on the reader towards a certain product
Double spaces	Tense/Mood/Aspect	Different meaning of the source text that may lead to legal, health or economic repercussions
Use of the decimal point instead of a comma	Coherence issues	The meaning of the source text is changed
Misplaced commas	Wrong word order	Wrong geographical references[3]
Hyphens missing	Wrong function word	
Repetition of the same term in the same sentence	Noncompliance with the glossary[4]	
	Register	

Table 1 – Example of instructions for annotators.

[3]There are cases in which it is major, for example, if the named entity is not completely changed.
[4]The error is critical when the word is completely different from the term in the glossary. However, when it is only an error of capitalization or a missing preposition or determiner, the error is major.

Observations:

1. Please use the critical severity only when it is really necessary. Critical errors affect the quality score significantly, so we need to be careful in its use.

2. Please be aware of the meaning of the source text. When the translation changes the meaning of the source text, mark it as critical.

Using our annotation experience we outlined some suggestions to improve the guidelines for Italian. These are integrated and formalized in the decisional trees used to improve the inter-annotator agreement, which we present and analyze in Chapter 5.

In particular, we focused on the severity of the errors within *register*. Errors associated with register are considered major errors, because they modify the way a customer addresses its audience, and sometimes it can result in an inadequate way or even show a lack of respect, with a negative impression linked ot the voice and the image of a company.

We analyze some examples that present a part of the source text (a) and its translation in the target text, after the first human edition (b), which is marked as an error by the annotator, and then we present a third sentence (c) that is the form supposed to be, the correct translation.

In these first two examples, the register in the instructions provided by the customer is set to formal:

(1a) Hi there, …

(1b) *Ciao, …

(1c) Buongiorno, …

(2a) I hope to hear from you soon.

(2b) *Spero di sentirti presto.

(2c) Spero di sentirla presto.

In Italian, there are a lot of English words that are currently used, and people are now getting used to them. In fact, in some translations, we can find English words that

are not translated, because they are transparent to the target public. But sometimes, in certain contexts, they have a different meaning, and thus they should be considered and categorized like *untranslated* errors. Moreover, they have to be considered like major errors, because they bring to a lack of meaning: the target text is not clear, and it leads to some difficulties in understanding it.

(3a) Ticket

(3b) Ticket* - it can be ambiguous with other meanings, for example, it can be interpreted as the ticket for a show, for the cinema, which should be translated as "Biglietto", or it can also be interpreted as the ticket in the hospital, a fee that people have to pay when they are visited by a doctor.

(3c) Richiesta di assistenza

Another improvement that we can bring to the guidelines is in the category of *prepositions*. The problem has to do with verbal valency. In Italian, there are a lot of verbs that require a specific preposition, according to the meaning of the sentence and according to the text. These valency errors are categorized as minor errors.

(4a) Thanks for your e-mail.

(4b) *Grazie per la tua e-mail.

(4c) Grazie dell'e-mail.

The categories of punctuation and capitalization are also taken into account in this analysis. Even though they are considered minor errors, they are very important for the understanding of the target text by the target public. They are analyzed together because one depends directly on the other.

(5) Lists:

(5a)

- Fill it with your first name
- Write your e-mail address

(5b)

- Inserisci il tuo nome*
- Scrivi la tua e-mail*

(5c)

- inserisci il tuo nome;
- scrivi la tua e-mail**;**

It results that, at the end of every sentence, the annotator has to mark a punctuation sign, and at the beginning of the sentence he/she has to mark a *capitalization* error. This is because, in Italian, after every element of the list, a punctuation mark is required, and it is usually a semicolon or a comma, and the following element of the list has to be written with the a low-case.

4.4. SUMMARY

This chapter focused on the role of the annotator, on the way he works, the process of annotation he has to follow, and on the rules, he has to apply, the guidelines. We also provided some suggestions to improve and implement the guidelines, both for Italian and the general guidelines, so that accuracy and fluency of the translated text, as well as its quality, can be improved.

5. INTER-ANNOTATOR AGREEMENT

This chapter presents how the inter-annotator agreement works and how important it is to measure homogeneity among annotators, and thus compare the effectiveness and quality. As defined by Nowak & Rüger (2011), the inter-annotator agreement describes the degree of consensus and homogeneity in judgments among annotators and it is used as a measure, showing that the data are reliable.

This chapter is divided into two sections. They present both the inter-annotator agreement among annotators for Italian, but in the first one it is calculated before the definition of specific guidelines that the annotator has to follow, and in the second section, the inter-annotator agreement is calculated after the definition of these guidelines. The inter-annotator agreement is calculated in terms of types of errors, severity of errors and both aspects together, as we see in the next sections. The aim of this chapter is to underline the importance of the guidelines and inter-raters agreement, in order to better assess the data and the quality.

5.1. FIRST PHASE: ANNOTATION BEFORE THE DECISIONAL TREES

After a thorough reading of the annotation guidelines and a clarification of some doubts with the help of other annotators and linguists, we started a training stage, annotating batches of translated texts on our own, gaining experience, in order to begin the process of annotation.

This was a crucial period because it allowed us to face usability issues of the annotation system and to define criteria used along the annotation process.

We then annotated a batch of data per week, from the 22nd January to the 26th February of 2017. The annotated batches of data were also annotated by another annotator, for Italian with the same linguistic background as ours, and then compared, so we could calculate the inter-annotator agreement in terms of types of errors, severity of errors, and both aspects together.

For this first analysis that was made, the two annotators did not speak to each other and they did not discuss hypothetical criteria to use during the annotation process. This

allowed us to underline the differences between two annotators, thus showing how the human component and subjectivity is difficult to manage and, for this reason, the importance of having proper and specific guidelines that help annotators during the annotation process. Defining such guidelines amounts to trying to detect in the more objective way possible all the errors, and to classify and evaluate them homogeneously with the goal of obtaining an almost exact agreement.

To analyze the annotated data and compare the work conducted by the two annotators, as we can see in the table below, we considered the number of jobs that were accomplished by the two annotators, every week, as well as the number of words annotated.

	ALL JOBS	TOTAL OF WORDS
22/01/2017	20	2053
29/01/2017	20	1978
05/02/2017	20	1977
12/02/2017	20	1991
19/02/2017	20	1929
26/02/2017	20	2145
TOTALS	120	12073

Table 2 – Jobs and words annotated per week

To evaluate the level of the inter-annotator agreement between the two annotators, we rely on a specific coefficient, the *kappa coefficient (K)* and on the table proposed by Landis & Koch (1977) to evaluate the K value we obtain.

The kappa coefficient (K) measures pairwise agreement among a set of coders, making category judgments, and it takes into account the possibility of the agreement occurring by chance. As defined by Carletta (1996), K is calculated as follow:

$$\kappa = \frac{p_a - p_\varepsilon}{1 - p_\varepsilon}$$

where P(a) is the proportion of times that the coders agree and P(s) is the proportion of times that we would expect them to agree by chance. The calculation is based on the difference between how much agreement is actually present ("observed" agreement), compared to how much agreement would be expected to be present by chance alone ("expected" agreement).

Along our work, we chose to consider the Cohen's kappa, because it is used as a measure of agreement between two coders. In our case, the two coders considered, to measure the inter-agreement, are two chosen annotators for Italian.

Kappa values range on a scale from -1 to 1, where 1 is perfect agreement, 0 is exactly the agreement that would be expected by chance, and negative values indicate less agreement than chance.

Landis & Koch (1977) provided guidelines for the interpretation of the kappa values:

< 0 →Less than chance agreement

0.01–0.20 → Slight agreement

0.21– 0.40 → Fair agreement

0.41–0.60 → Moderate agreement

0.61–0.80 → Substantial agreement

0.81–0.99 → Almost perfect agreement

All the statistics were possible due to the use of a website, http://dfreelon.org/utils/recalfront/recal3/ , that shows all the details of the statistics.

The inter-annotator agreement that we expect from this first part of the analysis is a *slight* or *fair* agreement (Landis and Koch: 1977), as the two annotators did not have a training session together, they did not speak about possible common criteria to use during the annotation process. The level of inter-annotator agreement expected is very

poor, as a reasonable level of agreement, in which annotators take almost the same decisions in term of annotation, starts from a *moderate* agreement (Landis and Koch: 1977).

In the next sections, the annotated batches of data, which were also annotated by another annotator, are presented, and the inter-annotator agreement is calculated and analyzed in terms of types of errors, severity of errors, and both aspects together.

5.1.1. TYPE OF ERRORS

	# errors
# cases	318
avg pairwise agreement	68.239%
avg pairwise Cohen's kappa	0.329

Table 3 – Agreement on the type of errors.

In this overall analysis of the data annotated, from the 22nd January 2017 to the 26th February 2017, the average pairwise agreement, the percentage of pairwise agreement among a set of coders, and the value for Cohen's kappa, that is 0.329, meaning that the inter-agreement between the two annotators is a *fair agreement*. It shows that the inter-annotator agreement between the two annotators is not that good, and this is due to some particular differences in the recognition of the errors, that we identify and discuss below.

To better analyze the aforementioned differences and try to overcome them, in the table below we compare the work of the two annotators per week in the period considered in this analysis, to identify systematic and/or regular differences, in order to determine criteria that have to be taken into account by all the involved annotators.

Weekly:

	# errors	
	Annotator 1	Annotator 2
22-01-17	92	77
29-01-17	91	91
05-02-17	73	83
12-02-17	57	71
19-02-17	79	104
26-02-17	79	82
TOTAL	471	508

Table 4 – Number of errors annotated by the two annotators

The table above makes apparent that the Annotator 2 identified more errors than the Annotator 1, even though these differences are not stable throughout the period considered, thus indicating that certain factors, still to be identified, play a role in this contrastive behavior of the two annotators. There is a big difference in the fifth weeks, but then we can also see a perfect agreement between them in the second week. During the other weeks, we find out that there is not such a big difference between them. We wanted to show the data per week, in order to underline that, even after a long period of annotation, there are still differences between the two annotators, this means that only an individual training session is not enough in terms of homogeneity and consistency among annotators. We have not absolute data to rely on, but from this sample, it results that Annotator 2 is stricter, and Annotator 1 is more permissive in annotating.

We can now say, after a deep analysis of the types of errors, that the differences between the two annotators reside in three different categories: *preposition*, *punctuation*, and *capitalization*.

Concerning the category of *preposition*, we observed that while Annotator 1 annotated 41 errors involving a preposition, Annotator 2 only annotated 14. Further looking at the data allowed us to realize that the main difference between the judgments of the two annotators is related to a particular construction:

(6a) Thanks for your e-mail

(6b) Grazie **dell'**e-mail – Annotator 1

(6c) Grazie **per** l'e-mail – Annotator 2

This difference is due to the fact that, in spoken Italian it is acceptable to say "grazie per + noun" (6c), but not in written language. The standard for written Italian states that the correct construction in this type of example is "grazie di + noun" (6b). (Dizionario Treccani). This is, nonetheless a minor error, as the content of the message is not affected by the wrong use of the preposition.

The second error category for which significant contrasts were observed is *punctuation*. The Annotator 1 reported 43 errors, whereas the Annotator 2 reported just 14. This contrast depends on the fact that sometimes punctuation is related to style. In particular, we can find some differences in the way editors used the punctuation in the lists and in the way, annotators identify the errors, as we can see in the examples below.

The third category involved in contrasting annotation is *capitalization*. Depending on the type of punctuation used, the capitalization changes. The first annotator identified 12 errors, and the second annotator just 3.

(7a) fill it with your first name

(7b) **I**nserisci il tuo nome**;** – Annotator 1

(7c) inserisci il tuo nome – Annotator 2

(8a) write your e-mail address

(8b) **S**crivi il tuo indirizzo e-mail**;** – Annotator 1

(8c) scrivi il tuo indirizzo e-mail – Annotator 2

Concerning the analysis made about the types of errors, it results that there is a tendency for Annotator 1 to identify less errors than Annotator 2, independently from the categorization of the errors. This means that the Annotator 1 is more permissive than Annotator 2 in identifying and categorizing the errors. From this analysis, it also results that the major contrast in the types of errors annotated between the two annotators links to "minor" errors, where the impact in the transmission of the message is null or reduced.

In the next section we present the same process of analysis and of calculation of the inter-annotator agreement, but by taking into account only the severity of errors.

5.1.2. SEVERITY OF ERRORS.

In terms of severity of the errors, i.e. their impact on the quality of the output, we compared 318 cases considered in the previous section. The table 5 presents the details concerning the three different degrees of severity of the annotated errors, more specifically: minor, major and critical. The severity of the error has a big impact on the MQM of the translation, and for that reason the annotators have to agree on the severity ascribed to each error annotated, in order to address the error in the same way.

	Minor	Major	Critical
# cases	318		
avg pairwise agreement	78,931%	78,302%	97,17%
avg pairwise Cohen's kappa	0.307	0.238	0.484

Table 5 – Agreement of the severity of errors

Table 5 shows that critical errors achieve a Cohen's kappa of 0.484, which means that there is a *moderate agreement,* according to the table proposed by Landis & Koch (1977). The table also shows that for the other two degrees of severity, there is much less consistency between the annotators, 0.307 for minor and 0.231 for major, meaning that there is a *fair agreement* both for minor and for major errors.

As for the previously analyzed data, we are going to present the differences between the annotators per week, in order to better analyze the behavior of the two annotators.

Weekly:

	Annotator 1	Annotator 2	Annotator 1	Annotator 2	Annotator 1	Annotator 2
	Minor	Minor	Major	Major	Critical	Critical
22-01-17	79	39	12	38	1	0
29-01-17	80	36	12	54	1	1
05-02-17	66	25	6	58	1	0
12-02-17	48	37	5	35	1	2
19-02-17	64	35	13	67	2	2
26-02-17	56	19	21	61	2	2
TOTAL	393	191	69	313	8	7

Table 6 – Number of errors annotated by the two annotators

From the table above, we may observe that there is a very big difference between the way the two annotators consider the severity of the errors, in particular there is a big difference between minor and major errors. Regarding the critical errors, there is almost a perfect agreement between the two annotators (Ann1 – 8/ Ann2 – 7).

We can observe that Annotator 1 classifies the majority of errors as minor, whereas Annotator 2 considers most annotated errors as majors. This discrepancy is caused by differences in the way the two annotators considered the severity of some errors, and the most indicative example is in the category *register*. For Annotator 1, *register* is generally considered as a minor error, whereas for Annotator 2 it is a major error. This contrast in the interpretation of the gravity of the error, on its possible consequences, and it

underlines the importance of having a training period before starting the process of annotation, so that decisions made by annotators are consistent, particularly in the assessment of the severity of annotated errors. Not doing so, it can lead to a very low inter-annotator agreement and sometimes different decisions in the annotation process can determine different measures of quality, and consequently to a lower MQM.

5.1.3. TYPE OF ERRORS AND SEVERITY

Following this first analysis, in which the results were not so satisfactory, due to the fact that the inter-annotator agreement was very low, we are going to study the inter-annotator agreement per type of error and severity together, in order to have a more precise idea of the real agreement between the two annotators, so action can be taken to improve it.

	type of error and severity
# cases	318
avg pairwise agreement	81,237%
avg pairwise Cohen's kappa	0.127

Table 7 – Agreement of the type of errors and severity together

The table shows that annotation decisions considering the combination of the two features achieves a Cohen's kappa of 0.127, i.e. there is a slight agreement between the two annotators, according to Landis & Koch (1977). The lowest agreement above its due to the contrast mentioned above.

This was the litmus test of the work of the two annotators, it shows us the contrast between the two annotators, their behavior in the annotation process, underlying that even after a long individual training, the decisions taken are still not homogeneous. As a consequence of these results, we will analyze the same categories further in our work, to examine if, after a discussion between the two annotators on the general criteria that

have to be used during the annotation process, the inter-annotator agreement has improved or not.

As Table 8 shows, the differences between the two annotators are presented, but in this case, we are going to consider only two weeks, the most representative ones: the ones that demonstrate the presence of a very low inter-annotator agreement, they present more data on which annotators do not agree, and that show that the Cohen's kappa coefficient is always almost quite the same value.

Weekly:

22-01-17	type of error and severity
# cases	53
avg pairwise agreement	81,761%
avg pairwise Cohen's kappa	0.083

Table 8 – Agreement of the type of errors and severity together

12-02-17	type of error and severity
# cases	53
avg pairwise agreement	81,761%
avg pairwise Cohen's kappa	0.085

Table 9 – Agreement of the type of errors and severity together

The inter-annotator agreement of the two weeks is very low: it achieves a Cohen's kappa of 0.083 and 0.085, for the week 22-01-17 and for the week 12-02-17, respectively, values which correspond to a *slight agreement* (Landis and Koch: 1977).

We can notice that these two weeks are not in the same month or in sequence, they were chosen because they result to have almost the same agreement, and to underline that, without common criteria on the decisions that have to be taken in the annotation process, the factor of time is not relevant. What is important is to have common guidelines for the annotation, in order to improve the homogeneity among annotators, thus the inter-annotator agreement.

5.2. SECOND PHASE: AFTER THE GUIDELINES FOR THE ANNOTATION

After this first period of training on how to annotate, we went through a process of determination of criteria that should be taken into account when annotating. All the in-house annotators participated, so that everyone agreed with the criteria, independently of the language of the annotator. This means that following such criteria leads not only to a better inter-annotator agreement between annotators of the same language, but also to a better inter-annotator agreement between all the annotators working for Unbabel, i.e. this process contributes to the consistency of the annotation process. In particular, we worked with other four annotators, an Italian, a Spanish, a Portuguese and a German annotator. This work allowed us to define particular criteria that can be applied to all these languages, besides the specific criteria for Italian. These guidelines, in which are explained and listed criteria on how annotators have to annotate certain errors in the annotation process, in order to reach homogeneous and consistent decisions, are important also due to the fact that the annotation process has some limits. For example, there are cases in which one annotator can choose to define the error between two categories:

When a formal word is used, and the required register is informal, we could choose between the category of *lexical selection* and the category of *register*:

(9a) Dear X

(9b) Egregio X

The word “egregio” is a formal word, used in an informal context. In this case, the annotator should mark this word as a *register* error instead of a *lexical selection* error.

Another example can be the difficulty in deciding which error to address a certain word when it contains more than one error:

(10a) You have to send an e-mail to the customer service

(10b) devi inviare un’e-mail al servizio clienti

If the register is set to formal and the verb is at the beginning of the sentence, after a full stop, we can find two errors in the verb “devi”, the informal register and the lack of the capital letter.

These examples show in a clear way the importance of the guidelines, in order to make the data reliable.

The in-house annotators agreed upon the following general criteria considering them valid for all the languages:

1. **Sentence structure/prepositions/conjunctions**

- When the sentence structure in the target language is not correct:

- if the sentence could be corrected by adding simply a preposition or a conjunction, then mark, respectively, “Prepositions” and “Conjunctions”;
- if the sentence cannot be corrected by simply adding a preposition or a conjunction, then mark “Sentence structure”.

2. **Pronouns/Prepositions/Conjunctions**

- When a pronoun, preposition, or conjunction is missing, then mark “Function words”, “Prepositions”, or “Conjunctions”, respectively.

3. **Tense/Mood/Aspect vs Agreement**

- When the person of the verb is not correct, then mark “Agreement”.

4. **Lexical Selection**

- If a word has different meanings in the target language depending on the context, and the meaning used is not correct for the context, then mark “Lexical selection”.

  Example: the English word “support” can mean, in Portuguese, both “suporte” and “ajuda”, if “suporte” is used instead of “ajuda”, mark it as “Lexical selection”.

5. **Register**

- If the register used is wrong in the entire text, then mark “Register”, and its severity is major. If the register is wrong in only one or two sentences, then mark “Inconsistent register”.
- If an informal word is used and the register required is formal, or vice versa, then mark “Register” and NOT lexical selection.

Concerning the specific criteria for the Italian, we arrived at the following:

1. Translation of the second person into Italian

In personal pronouns:

- When the register required is formal and the pronoun "tu” is used → register
- When the register required is informal and the pronoun "lei" is used → register
- When the pronoun "voi” is used and when the impersonal construction is used → word form

In possessives pronouns:

- When the register required is formal and the possessive "tuo” is used → register
- When the register required is informal and the possessive "suo” is used → register
- When the possessive "vostro” is used → word form

2. Punctuation and capitalization in lists

- When there is not a strong punctuation sign at the end of every sentence → punctuation
- When there is a capital letter at the beginning of the sentence → capitalization

Example:

- **F**ill it with your first name**;**
- **W**rite your e-mail address**;**

Once these criteria were established, both annotators annotated another batch and the results were compared. As in the previous work, the inter-annotator agreement was calculated, and the results analyzed, to check if it remained the same or if it has improved.

The new batch that was annotated by the two annotators, corresponding to data from the 23rd April to the 30th April 2017, contains 20 jobs and a total of 1649 words.

The analysis made in the next section is specular to the work just presented: the inter-annotator agreement is calculated in terms of type of errors, severity of errors and both aspect together, focusing on the results obtained after the definition of the general guidelines and the ones specific for Italian.

5.2.1. TYPE OF ERRORS

Considering only the type of error, analyzing 53 cases, we find out that the two annotators reached a Cohen's kappa of 0.712, which corresponds to a *substantial agreement* (Landis and Koch: 1977) between the two annotators.

	# errors
# cases	53
avg pairwise agreement	88,925%
avg pairwise Cohen's kappa	0.712

Table 10 – Agreement of the type of errors

	Annotator 1	Annotator 2
3-05-17	114	112

Table 11 – Number of errors annotated by the two annotators

As the table shows, we can notice a very little difference between the number of errors of the two annotators. This sample shows a big difference between the behavior of the two annotators in the first weeks of annotation and their behavior in this specific sampled week, after the definition of the guidelines. Annotator 1 annotated 114 errors, Annotator 2, 112 errors instead. The difference in annotating between the two annotators is minimum, thus the inter-annotator agreement, as we expect, is higher than the one calculated in the first weeks.

5.2.2. SEVERITY OF ERRORS

In terms of severity of the errors, i.e. their impact on the quality of the output, we compared 318 cases considered in the previous section. The table 5 presents the details concerning the three different degrees of severity of the annotated errors, more specifically: minor, major and critical. The severity of the error has a big impact on the MQM of the translation, and for that reason the annotators have to agree on the severity ascribed to each error annotated, in order to address the error in the same way.

	Minor	Major	Critical
# cases	53		
avg pairwise agreement	83,019%	81.132%	100%
avg pairwise Cohen's kappa	0.616	0.646	1

Table 12 – Agreement of the severity of errors

As it is represented in Table 13, the numerical differences between the two annotators are minimal. There is only a single difference in the minor category and another one in the major category. As we can see, in the critical category there are no errors, both annotators agreed with the absence of errors and the decision of not annotating any error, thus they took the same decisions, starting from the guidelines they have to follow. This is the litmus of the high inter-annotator agreement that we obtained and that we were expecting.

	Annotator 1	Annotator 2	Annotator 1	Annotator 2	Annotator 1	Annotator 2
	Minor	Minor	Major	Major	Critical	Critical
3-05-17	26	25	88	87	0	0

Table 13 – Number of errors annotated by the two annotators

5.2.3. TYPE AND SEVERITY OF ERRORS

The last analysis performed regards the type of error and severity, together. As we can see from the table below, the definition of clear criteria, such as those presented in section 5.2, allowed us to reach a high inter-annotator agreement. In fact, the Cohen's kappa of this category is of 0.686, which means that there is a *substantial agreement* (Landis & Koch, 1977) between the two Italian annotators.

	type of error and severity
# cases	53
avg pairwise agreement	87,516%
avg pairwise Cohen's kappa	0.804

Table 14 – Agreement of the type of errors and severity together

5.3. SUMMARY

This chapter focused on the inter-annotator agreement among annotators. It results that before having a definition of criteria to guide annotation decisions, during the training process, there was only a *slight agreement* (Landis and Koch: 1977) between the annotators. In this chapter, it is underlined how important the decision trees are: they allow the annotators to follow the same criteria while annotating, in order for decisions to be consistent during the error annotation process; and they lead to a *substantial agreement*, and sometimes to a *perfect agreement* between the two annotators, so that also the MQM of the translation keep on improving. That means that the guidelines for the annotation have a key role in the categorization of the errors during the annotation process and they result in a considerable homogeneity and consistency in the decisions made by all the annotators, trying to make the annotation process the more objective as possible, limiting the subjectivity of the human component, so that the annotated data turn out to be reliable.

6. ERROR ANALYSIS

In this chapter, we are going to study the most frequent error categories used during the process of annotation. Section 6.1 will be devoted to the analysis of the top 6 types of most frequent errors. We will then focus on the most frequent type of error, the category *register*. We will explain why it is the most frequent type of error found in our data and describe the impact that errors in this category have on the fluency and quality of the translation.

In section 6.3, we will describe the tools used at Unbabel to help translators detect register errors, how they work and how these tools can be improved, by creating additional rules.

Section 6.4 presents the deployment of the set of rules that was established under the scope of this work. Non-deployed rules will also be presented in this section, along with a description of the limitations that prevent or hinder their integration in the system.

6.1. MOST FREQUENT ERROR CAEGORIES

Once the annotation process was concluded, we conducted an analysis of the data. We investigated all the batches annotated, from January 22nd, 2017 till February 26th, 2017, and from April 23rd, 2017 to April 30th, 2017. In total we analyzed 174 jobs and 14222 words. In table 15, we list the top 6 types of error identified in the annotated texts, so that we can have an overall view of the most frequent errors.

	MINOR	MAJOR	CRITICAL	TOTAL
All error types	518	137	8	663
Register	186	46	0	232
Preposition	98	7	1	105
Punctuation	73	4	0	73
Capitalization	53	0	0	53
Whitespace	38	0	0	38
Lexical selection	15	15	1	31

Table 25 – Top 6 most frequent types of error in the annotated texts and their distribution per severity level

In this sample, the most common error is the one regarding the *register* category. We can also underline that this is one of the errors with the greatest impact on the quality and fluency of the translation, as it can result in the disappointment of a client, particularly when he has provided a style guide, which often includes specific indications regarding *register*. Errors in this category are considered major/critical both by annotators and the client, because they can result in a lack of respect in addressing costumers. The second most frequent type of error is the *preposition* category. There are many errors involving verbal valency in particular, as we mentioned in previous chapters. After these, *punctuation* and *capitalization* errors, which are, actually, mutually dependent, appear on the list. *Punctuation* errors are, sometimes, also a question of style, especially in lists, as the punctuation can vary, depending on the person:

Example:

(11a) To create an account, you have to:

- insert your full name**;**
- insert your email address**;**
- insert your phone number**;**

(11b) To create an account, you have to:

- insert your full name
- insert your email address
- insert your phone number

In example (11a), we can see that the person decided to put a semicolon mark at the end of each sentence and to start the next one with a low case letter. In the second example (11b), instead, there are no punctuation marks besides the colon introducing the list of items needed to create an account.

Capitalization, as mentioned before, is often a direct consequence of the former category, as if we decide to use a strong punctuation mark we must start the following sentence with an initial capital letter, but if we decide to use a weak punctuation mark, we must use an initial lower-case letter. The fifth most frequent error category is *whitespace*: this is an error inserted by machine translation which often goes unnoticed by the editor and it is easy to solve through automatic processes. The last most frequent error type in Table 15 is *lexical selection*. Errors in this category are frequent because in English, as in all natural languages, there are many polysemic words that can lead to ambiguities or mistranslations. This is particularly problematic when it involves the choice of the correct equivalent in the target language, Italian in our case, which can be problematic either for the MT system or for the editors, or for both.

We will now proceed by analyzing the most frequent errors before and after the definition of the guidelines for the annotation (section 5.2.), in order to see whether the categories chosen by the annotators and the severity associated to the errors are the same or if differences are made apparent after the specification of the aforementioned guidelines.

In Table 16, we present the batches annotated from January 22nd, 2017 till February 26th, 2017. In total we analyzed 150 jobs and 12573 words.

	MINOR	MAJOR	CRITICAL	TOTAL
All error types	489	78	8	575
Register	186	0	0	186
Preposition	89	5	1	95
Punctuation	64	4	0	68
Capitalization	48	0	0	48
Whitespace	38	0	0	38
Lexical selection	15	13	1	29

Table 16 – Top 6 most frequent types of error in texts annotated between January 22nd, 2017 and February 26th, 2017 and their distribution per severity level

Data in Tables 16 and 17 show that the categories of error are essentially the same, although the annotator, in the batches before the definition of the guidelines, marked almost all errors as minor, except for the *lexical selection* category, in which we can find almost the same number of errors annotated as minor and major.

We will now look into the results that we obtained from the annotation performed between April 23rd, 2017 and April 30th, 2017. This batch contains 24 jobs and a total of 1649 words and it was annotated after the definition of the annotation guidelines (section 5.2.).

	MINOR	MAJOR	CRITICAL	TOTAL
All error types	29	59	0	88
Register	0	46	0	46
Preposition	8	2	0	10
Punctuation	9	0	0	9
Capitalization	5	0	0	5
Orthography	4	0	0	4
Lexical selection	0	2	0	2

Table 17 – Top 6 most frequent types of error in texts annotated between April 23rd, 2017 and April 30th, 2017 and their distribution per severity level

Comparing Table 16 and 17 makes apparent that the categories of error and their relative frequency in the table are almost the same. The only exception is the substitution of "whitespace" errors by "orthography" errors as the fourth most common error type annotated. An important thing that changes is the severity associated to the annotated errors, particularly the severity of register errors. This is naturally consistent with what we said in former chapters, when we presented the guidelines for annotation, where the severity classification of register errors was pointed out as the most common inconsistency between trained and untrained annotators.

We decided to focus on the category of *register*, not only because it is the most frequent error, but also because it is the error with the major impact on the quality of translations, as discussed in the next section.

6.2. REGISTER

As the error category of register is the most frequent in the annotation data and not only does it have a significant impact on the final quality of the translation and on the client's perception of its quality, but it can also lead to omissions of signs of respect or to infringements regarding good manners, we decided to study it in detail and to make a deeper analysis on how this category is treated throughout the translation process at Unbabel. Upon this analysis, we are going to implement some heuristics for the automatic detection of register errors, in order to reduce its frequency in translation outputs at Unbabel.

Register touches different aspects of grammar, and this is why it is difficult to accurately encode this phenomenon in a natural language processing system. Register is materialized in the selection and use of certain expressions, some of which are linked to language variation in Italian. As all natural languages, Italian is continuously changing (Proudfoot & Cardo: 2005). The fact that some of these changes-in-progress become register marks makes it very challenging to categorize all the syntactic and morphological features needed for assuring full coverage of register-related rules. In this work we focus our analysis on some of the register-related expressions, identifying generalizations to formulated rules, some of which are formalized and implemented in the system.

In our study, we are going to analyze three major categories of grammatical phenomena, which are involved in the expression of register in Italian: pronouns, tense/mood/aspect and lexicon. We will focus on pronouns and tense/mood/aspect as these categories are particularly important and clear-cut in differentiating between the two registers (formal vs. informal), as they correspond to closed morpho-syntactical categories and to systematic grammatical phenomena. Thus, observations regarding these categories are bound to be generalized and consequently to be covered by a set of finite rules. Concerning lexical phenomena, these involve open categories, and for this reason are more difficult to be generalized and described by a set of finite rules. In fact, dealing with lexical phenomena in the context of the translation workflow at Unbabel would necessary involve encoding rich information in lexical resources, which is not within the short-term priorities of the company. Considering this along with the fact that variation on the lexis choice often implies also a variation on the grade of formality

(Giordano & Voghera: 2002), in this work we are going to concentrate only on some common and recurring *formulae*.

6.3. TOOLS TO TACKLE REGISTER

Other factors that we have to take into consideration in this kind of work are the tools and information available in the system at Unbabel, in order to deal with the grammatical phenomena involved in the expression of register and in its modelling.

Regarding the tools used at Unbabel to help editors in the post-edition of machine translated outputs, particularly in detecting whether the register is correct and consistent with the indications of the client, there are two main tools being used at the company: the Smartcheck and the Turbo Tagger, a dependency parser. In the following sections we briefly describe these tools, the way they work and the information they work with.

6.3.1. SMARTCHECK

As described in the first chapter of this work, the Smartcheck is a tool developed at Unbabel to check format, grammar and style in the texts translated on the company's platform.

The Smartcheck is a checker, which means that it does not automatically edit the text, but only provides warnings or suggestions to the human editor, who is the responsible for deciding what to do regarding these warnings and suggestions.

In this chapter, we are going to analyze the performance of this tool as a checker of the register of a text, i.e. an automatic process to verify if the register used in a translated text is correct and consistent with the one set by the client.

Besides considering the aforementioned automatically-generated messages, the editor has to thoroughly go through the translated text because the Smartcheck, at the time the internship leading to this work took place, did not incorporate morpho-syntactic context rules[5], its action being limited to word spotting actions. This is why the dependency parser described in section 6.3.2. is also needed, as described below.

[5] These information, now, are already incorporated. As they were implemented after the internship, they are not taken into account in our work.

6.3.2. TURBO TAGGER

The turbo tagger is a dependency parser that provides information regarding the structure of a sentence. It is an important tool in the process of automatically establishing the correct syntactic dependency between constituents occurring in a sentence. In this phase, the parser is used only for the first part of the parsing process, i.e. POS tagging: only morpho-syntactical information is taken into account, and not dependency information. The results obtained with the parser are also useful to understand why the Smartcheck does not detect certain errors: sometimes the POS tag is wrongly ascribed by the tagger which hinders the recognition of certain errors by the Smartcheck; other times the problem is strictly related to limitations of the technology implemented in the Smartcheck.

When a sentence is analyzed by the parser, it provides information on the base form of each word occurring in it, its POS, the value for specific features of the word (for example, number, gender, person, mood, tense or verb form), and a dependency tree representing the syntactic structure of the sentence is also provided. This was useful during the phase of creation of new rules to be implemented in the Smartcheck to check the register, as it allows the definition of more robust rules, by avoiding rules that over generate and thus cover phenomena of different nature than those being modelled by a given rule. The results defined and implemented in the Smartcheck are presented in section 6.4, and the rules that were not implement in section 6.5.

6.4. DEPLOYED RULES

To characterize the errors of register identified in the annotation process, we analyzed the specific occurrences of these errors. In a data driven process, in which we considered our examples as a starting point, we conducted a linguistic analysis to identify generalizations that were the base of our work in the definition of rules to be implemented in the Smartcheck tool for it to tackle register errors. In doing so, we aimed at reaching better results in the annotation process, by accomplishing that all the suggestions given by the Smartcheck turn out to be correct, so translators can save time and be more efficient. After the creation of the rules, they were tested in a process of staging of the tool to verify whether the phenomena modelled by the rules are in fact recognized by the Smartcheck. This procedure has been put in place during the month of September 2017.

The process of staging consists in writing the Italian sentence that we want to analyze in the translated text box, selecting the target language, in our case Italian, and writing the correspondent English sentence in the source text box, and selecting the source language, in our case English (Figure 13). After that, we have to select which checks we want Smartcheck to perform, namely the tone[6] (Figure 14), and choose whether a formal or an informal register was to be used in the translation (Figure 13). Further, in Fig. 16, we have a sample and an explanation of the staging process when the Smartcheck detects an error.

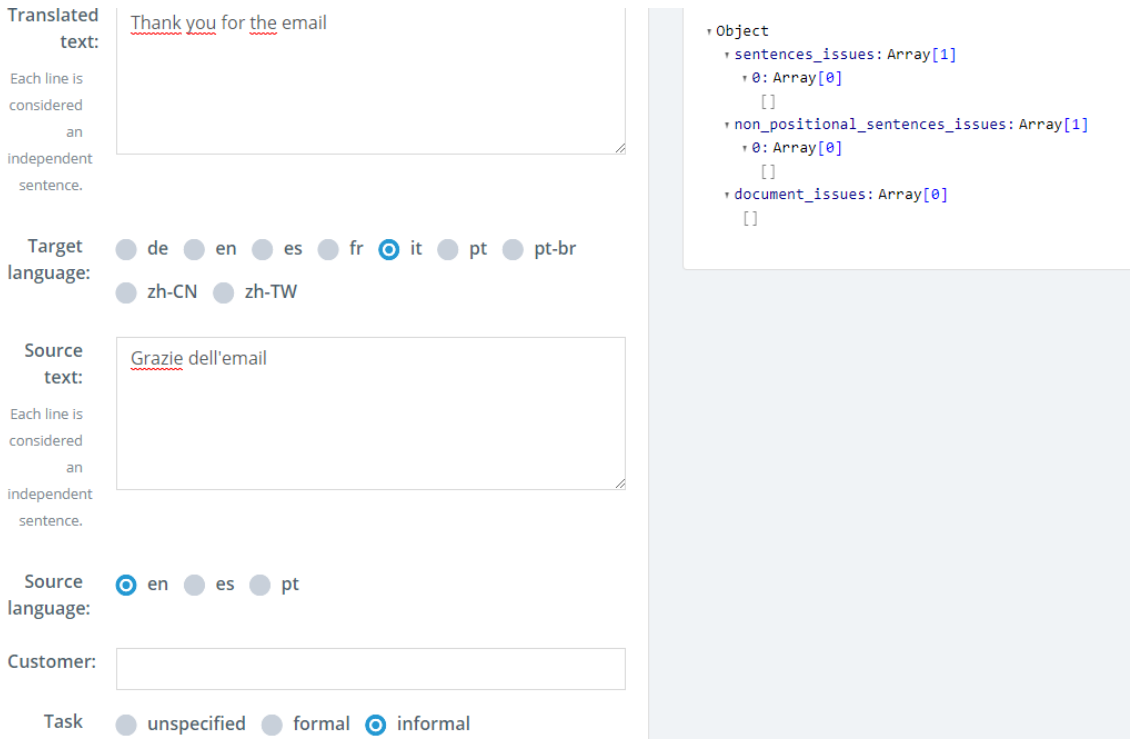

Figure 13 – a screen of the staging tool for evaluating implemented rules

[6] The verifications associated to register, in the interface of the tool, are selected via the option “tone”, and not “register”. This has to do with implementation issues and not to linguistic knowledge of such structures.

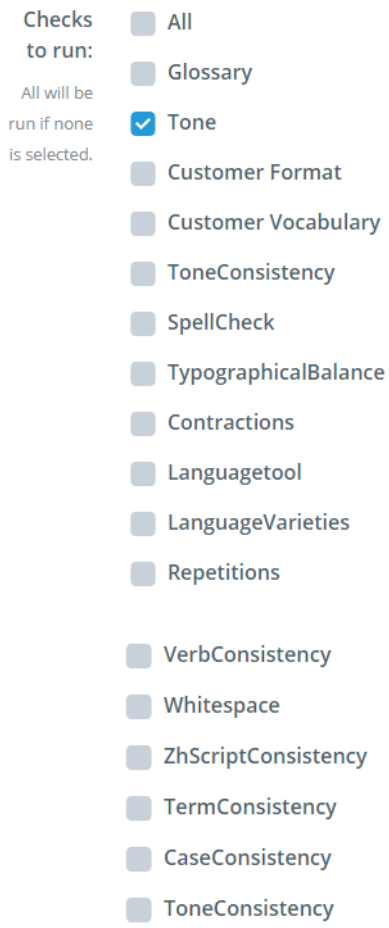

Figure 14 – Checks selection in the staging tool

Some of the rules described below are “case sensitive”, as the norm for the Italian formal register specifies that certain words are to be capitalized and as there is a clear indication by the company that only robust rules are to be implemented, i.e. only rules that do not cover phenomena of different nature and thus do not overgenerate. Guaranteeing that the rules defined do not overgenerate involves the use of a capital letter in some categories, such as personal pronouns, possessives, accusative and dative clitics and some *formulae* in the formal register. It is important to mention that, if the norm was systematically observed by users, the implemented rules would be sufficient to distinguish a formal register from an informal register in a clear-cut way, with the exception of the contexts in which the expressions covererd by the rules occur in the beginning of a sentence. Specific rules involving case-sensitivity are identified and described below. However, in some contexts, there is often a deviation from the

aforementioned norm, as it is quite common for Italian speakers to write personal pronouns, possessives, accusative and dative clitics with an initial low case letter, even when the register is formal. This fact naturally affects the coverage and performance of our rules, as we will explain in section 6.5. Such deviations to the norm makes the definition of non-overgenerating rules a complex process, as many forms and syntactic structures become ambiguous.

This section is organized depending on the type of register: a first part is dedicated to deployed rules covering specifications related to the informal register and the second part to the ones covering phenomena related to the expression of a formal register.

6.4.1. RULES COVERING SPECIFICATIONS RELATED TO THE INFORMAL REGISTER

Rule 1:

*If in a sentence in Italian as a TL the form **"tu"** occurs, then the register is informal*

The rule formulated above means that the Smartcheck will look for occurrences of the word "tu", the second person singular personal pronoun, that is specific for the informal register, and verify the association of the job to the informal register.

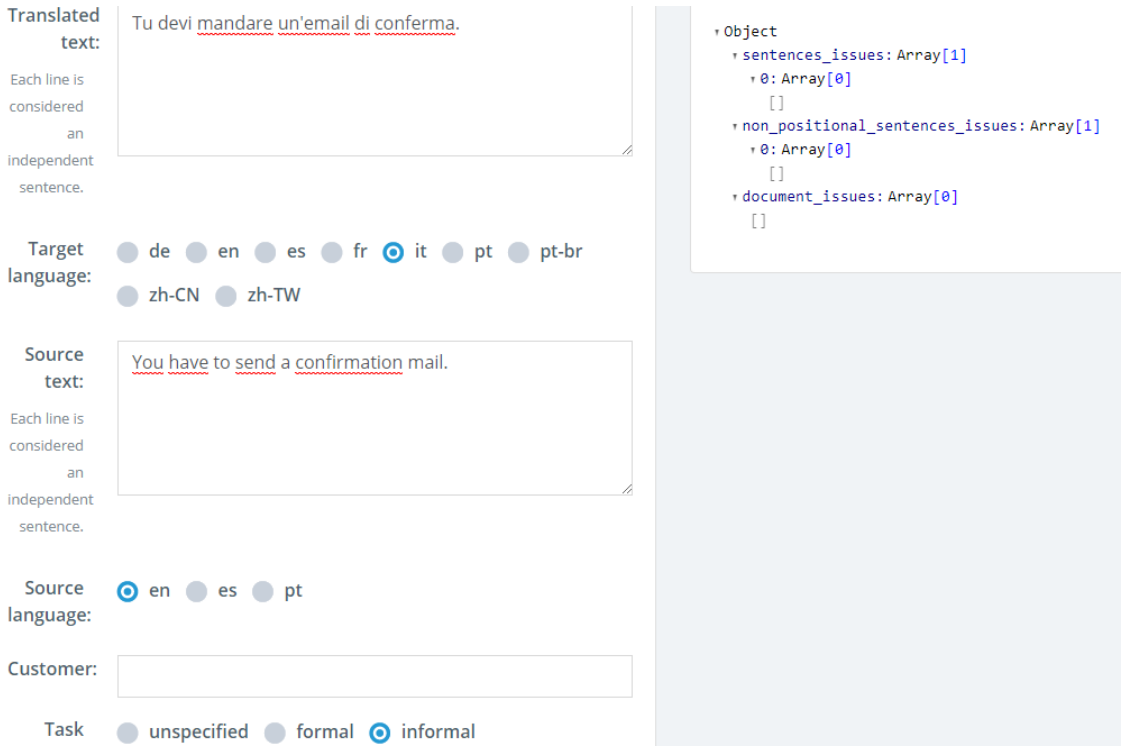

Figure 35 – Submission of a sentence in which the pronoun “tu” occurs to the staging tool when the register is set to “informal”

In Figure 15, we can also see the information output by the server, which allows us to identify if the Smartcheck is working correctly regarding the application of a given rule, and consequently, the recognition of a specific error. In Figure 15, no errors are marked. The server response is different when an error is detected, as we can see, for example, in Figure 16. In this figure, we can see that the server identifies the category of the error, in this case “tone”, the expression involved in the error, the personal pronoun “tu” in our example, its severity, namely whether it is an error or a warning, and a description of the error, “undesired register” in the case below.

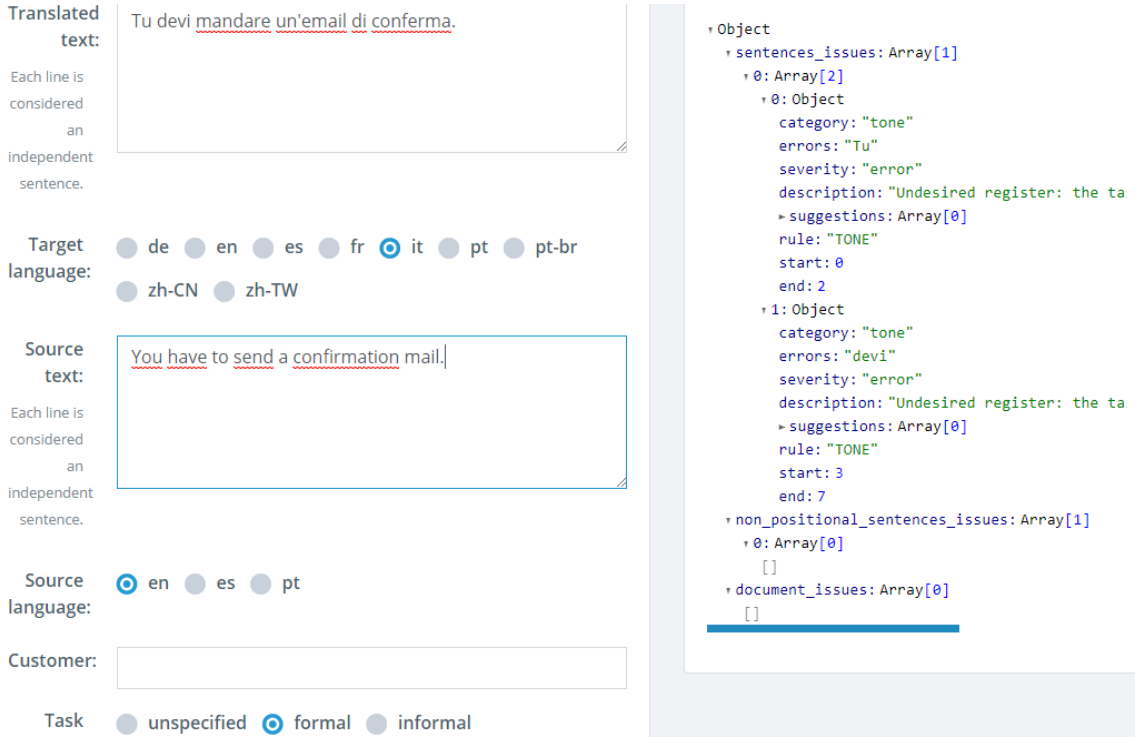

Figure 16 – Submission of a sentence in which the pronoun “tu” occurs to the staging tool when the register is set to “formal”

The two figures above, 15 and 16, allow us to verify that the Smartcheck correctly applies rule 1, at the right hand side of Figure 15 no errors are identified when we set the register to informal, but, when we set it to formal, as shown in Figure 16, we observe that the word “tu” is marked as a register error, as intended.

In the table below, we present the rules we created to identify errors involving the informal register. In this table, grammatical categories of the phenomena are provided, along with the Smartcheck rule, a short description of the rule, and whether it works correctly once it has been implemented in the Smartcheck, depending on the required register, formal or informal.

CATEGORY	SMARTCHECK RULE	DESCRIPTION	SMARTCHECK RESPONSE	
			If we set an informal register	If we set a formal register
POSSESIVES	*If in a sentence in Italian as TL the forms "tuo"/ "tuoi"/"tua"/"tue" occur, then the register is informal*	Possessives 2nd person singular and plural, masculine and feminine	No error is marked	*Tuo/tuoi/tua/tue* is recognized as an error
ACCUSATIVE CLITICS	*If in a sentence in Italian as TL the form "ti" occur, then the register is informal*	Accusative clitics in 2nd person singular	No error is marked	*Ti* is recognized as an error
DATIVE CLITICS	*If in a sentence in Italian as TL the forms "ti" / "a te"*[7] *occur, then the register is informal*	Dative clitics in 2nd person singular	No error is marked	*Ti/a te* is recognized as an error
VERBS	*If in a sentence in Italian as TL a verb, 2nd person singular occurs, then the register is informal*	Verbs in 2nd person singular	No error is marked	*Verb forms* are recognized as an error
GREETINGS	*If in a sentence in Italian as TL occurs the formula "Ciao", then the register is informal*	*Formula*	No error is marked	*Ciao* is recognized as an error
CLOSINGS	*If in a sentence in Italian as TL the formula "Ciao" occurs, then the register is informal*	*Formula*	No error is marked	*Ciao* is recognized as an error

Table 16 – Rules deployed in the Smartcheck to cover informal register errors

[7] Treated as a formula, it does not vary.

6.4.2. RULES COVERING SPECIFICATIONS RELATED TO THE FORMAL REGISTER

A process similar to the one described in the previous section was developed to create rules to cover formal register errors. In the table below, with the same format used in table 16 and described in the previous section, we present the rules for identifying formal register errors.

CATEGORY	SMARTCHECK RULE	DESCRIPTION	SMARTCHECK RESPONSE	
			If we set an informal register	If we set a formal register
PERSONAL PRONOUNS	*If in a sentence in Italian as TL the form "Lei" occurs, then the register is formal*	Personal pronouns in 3rd person singular, and with an initial capital letter	*Lei* is recognized as an error	No error is marked
POSSESSIVES	*If in a sentence in Italian as TL the forms "Suo"/ "Suoi"/ "Sua"/ "Sue" occur, then the register is formal*	Possessives in 3rd person singular and plural, masculine and feminine, and an initial capital letter	*Suo/Suoi/Sua/Sue* is recognized as an error	No error is marked
ACCUSATIVE CLITICS	*If in a sentence in Italian as TL the form "La" occurs, then the register is formal*	Accusative clitics in 3rd person singular, with an initial capital letter	*La*[8] is recognized as an error	No error is marked

[8-9] The forms "La" and "Le" are recognized by the Smartcheck, but they are ambiguous between the feminine singular determiner, at the beginning of a sentence, and the dative clitic in 3rd person singular and the feminine plural determiner. This results in an overgeneration of error marking in the informal register a specific case described in detail in section 6.5.

DATIVE CLITICS	*If in a sentence in Italian as TL the forms "Le"/ "a Lei" occur, then the register is formal*	Dative clitics in 3rd person singular, with an initial capital letter	*Le*[9]*/ a Lei* [10]is recognized as an error	No error is marked
GREETINGS	*If in a sentence in Italian as TL the formulae "Salve"/ "Gentile Signor"/ "Gentile Signore"/ "Gentile Signora"/ "Gentile Signorina"/ "Gentile Sig."/ "Gentile Sig.ra"/ "Gentile Sig.na"/ "Egregio Signore"/ "Egregio Signor"/ "Egregia Signora"/ "Egregia Signorina"/"Egregio Sig."/ "Egregia Sig.ra"/"Egregia Sig.na"/ "Signor, Signora"/ "Signorina"/ "Sig"./ "Sig.ra"/ " Sig.na", occur, then the register is formal*	*Formulae*	All the *formulae* considered are recognized as an error	No error is marked
CLOSINGS	*If in a sentence in Italian as TL the formulae "Cordiali saluti"/ "Arrivederci"/ "Arrivederla" occur, then the register is formal*	*Formulae*	All the *formulae* considered in this rule are recognized as an error	No error is marked

[10] Treated as a formula, it does not vary.

Table 17 – Rules deployed in the Smartcheck to cover formal register errors

6.5. NON-DEPLOYED RULES

Among all the new rules that were defined, there were some that despite their importance could not be included in the Smartcheck because they overgenerate, as the same form can have two different meanings and/or grammatical functions, and, for this reason, they do not systematically work as intended, as it will be described in a thorough way in this section. This means that not all the linguistic patterns identified in register errors were deployed at this stage, but their linguistic description has been done and is suitable to be used for further improvement in the future, particularly when richer linguistic information is included in the tools, especially morphosyntactic context information[11].

In this section we present some cases in which the linguistic pattern identified matches phenomena of different nature, which is problematic if they were to be implemented in the Smartcheck. In this case, we are not talking about lexical ambiguity, rather ambiguity involving functional words which despite their identical form have different syntactical and morphological functions.

We provide an analysis of these phenomena, trying to explain at a linguistic and technical level, when possible, why presently they are not unambiguously recognized by the Smartcheck. We will also define the information necessary to analyze and categorize the phenomena, i.e. morphological rules and contextual rules, which would accurately model the data in case such information was to be added to the tools.

1. *Pronoun "lei" that can occur in both registers*

The personal pronoun 'lei', in lower-case, is the most ambiguous case in terms of register binary decisions, because when a feminine person is at stake, the same form is used as a courtesy form to address a feminine interlocutor or as a pronoun to refer to a third person entity, singular and feminine.

[11] By the time this thesis was submitted, Unbabel was working on context dependent rules.

If a masculine person is referred to, it is quite easy to understand whether we are using a courtesy form, so a formal register, to address a masculine interlocutor or if we are referring to a third person entity, singular and masculine and hence the expression is unmarked in terms of register, as in this case the form of the pronouns is not the same (see 11a and 11b).

Formal: addressing Mr. Rossi, a second person interlocutor

(12a) È importante verificare che Lei sia connesso a Internet.

Unmarked: Luca, a third person entity

(12b) È importante verificare che lui sia connesso a internet.

In both examples we refer to masculine entities (see the terminations of the participles), but in (12a) we have the word "Lei" and hence a formal register. "Lei" is a courtesy form, as the unmarked 3rd person singular subject pronoun for the masculine is "lui" as in (12b).

If we talk about feminine entities, recognizing whether a formal register or an informal register is being used, is more complicated, due to the fact that there are no differences in the construction and in the forms used in the sentence.

Formal: addressing Mrs. Rossi, a second person interlocutor

(13a) È importante verificare che Lei sia connessa a Internet.

Unmarked: Anna, a third person entity

(13b) È importante verificare che lei sia connessa a Internet.

In this case, the only way to distinguish the formal "lei" from the informal "lei" is through the use of an initial capital letter when "lei" is a courtesy form and hence a mark of a formal register. Another way to distinguish the homonymous "lei" would be by checking the referent in the sentence, i.e. if we are considering a third element, different from the interlocutors (unmarked register), or if we are considering the interlocutor (formal register), i.e. using a 2nd person pronoun.

Considering all this, and the type of information that can be encoded in the Smartcheck, if we set the register to formal, there is not any problem regarding this phenomenon, as the Smartcheck does not mark any error, as shown in the figure below.

This is an accurate error marking, independently of “lei” being used as a 3rd person pronoun or a courtesy form.

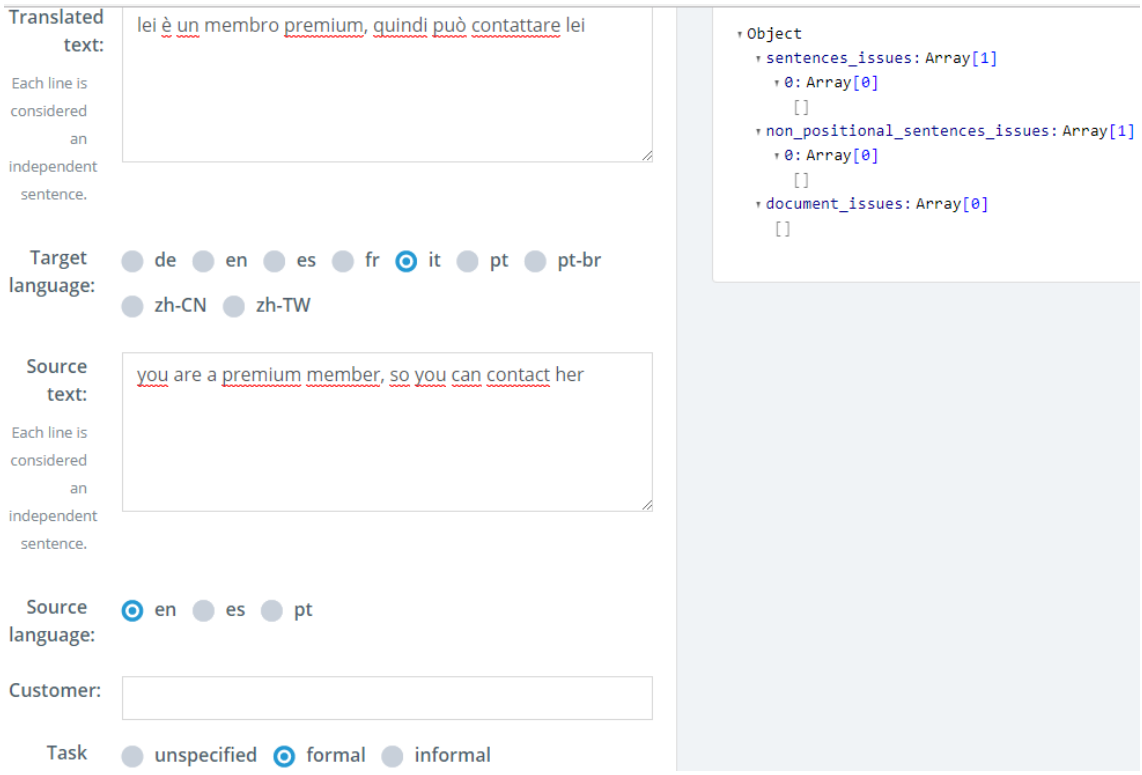

Figure 17– Submission of a sentence in which the pronoun “lei” occurs in subject position to the staging tool when the register is set to “formal”

If we set the register to informal, the ambiguity between the two functions of “lei” becomes a problem. The Smartcheck is not able to recognize whether it is a courtesy form subject or a third person pronoun, referring to a singular and feminine entity. When confronted with an example such as (14), the Smartcheck does not recognize “lei” as a courtesy form and hence does not mark it as a register error (see Fig. 18) when it should.

(14) Secondo il database, **lei** è un membro premium, quindi ha accesso a questa opzione

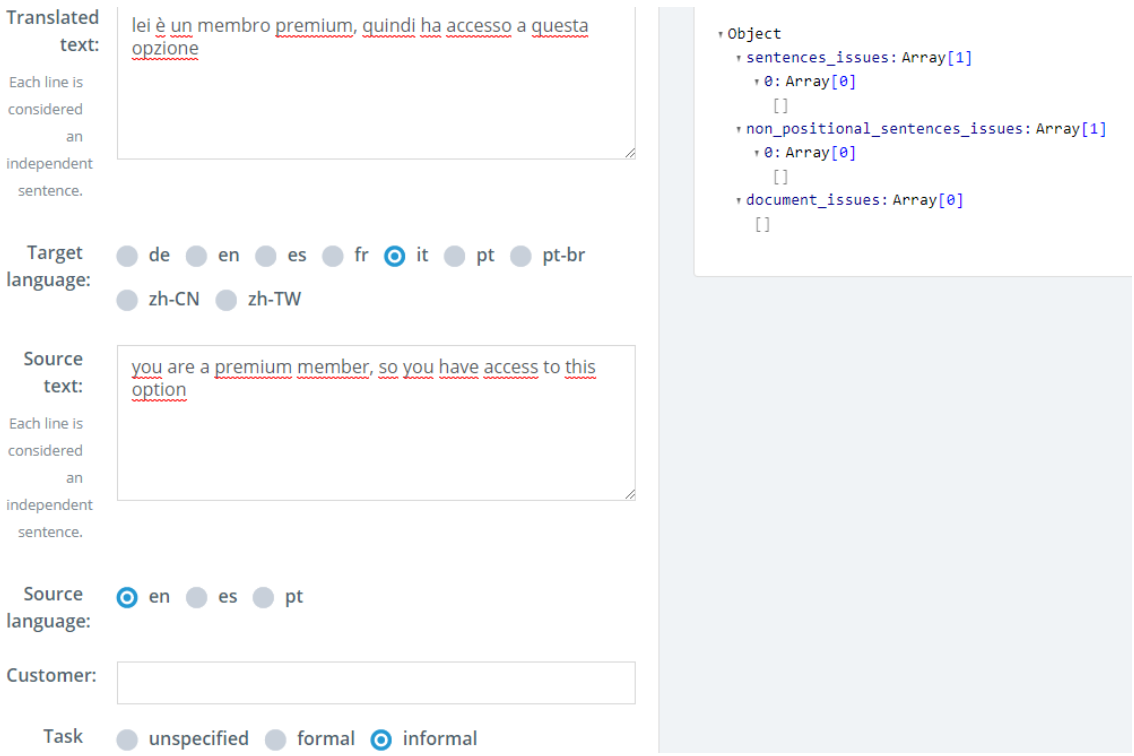

Figure 18– Submission of a sentence in which the pronoun “lei” occurs to the staging tool when the register is set to “informal”

Even if the Smartcheck does not recognize this kind of information, we analyzed the output of the parser, in order to check whether it corretly recognizes the syntactical fuctions of the constituents or not. As shown in Fig. 19, the parser does not provide any information regarding the gender associated to the pronoun “lei”. Thus, only using the analysis output by the parser, it is not possible to gather the information on whether it refers to a masculine or to a feminine entity. Given this, the information provided by the parser is not sufficient to resolve the ambiguity mentioned above.

1	lei	lei	PRON	Number=Sing|Person=3|PronType=Prs	4	nsubj
2	è	essere	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	4	cop
3	un	uno	DET	Definite=Ind|Gender=Masc|Number=Sing|PronType=Art	4	det
4	membro	membro	NOUN	Gender=Masc|Number=Sing	0	root
5	premium	premium	ADJ	Gender=Masc	4	amod
6	,	,	PUNCT	_	4	punct
7	quindi	quindi	ADV	_	8	advmod
8	ha	avere	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	4	acl:relcl
9	accesso	accesso	NOUN	Gender=Masc|Number=Sing	8	dobj
10	a	a	ADP	_	12	case
11	questa	questo	DET	Gender=Fem|Number=Sing|PronType=Dem	12	det
12	opzione	opzione	NOUN	Gender=Fem|Number=Sing	9	nmod

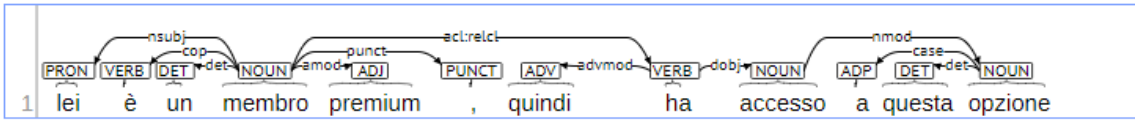

Figure 19 – POS tagging and dependency analysis of the sentence “lei è un membro premium, quindi ha accesso a questa opzione” output by the Turbo Tagger

The example we are now going to analyze, (15), differs from the example (14) only with regard to the syntactical function of the pronoun “lei”. In the former, “lei” is a subject, in the latter it is an object. The goal of contrasting these two examples is to verify whether the tool can recognize the two different syntactical functions and if they were treated correctly in terms of register marks.

(15) Per ulteriori informazioni può contattare **lei**

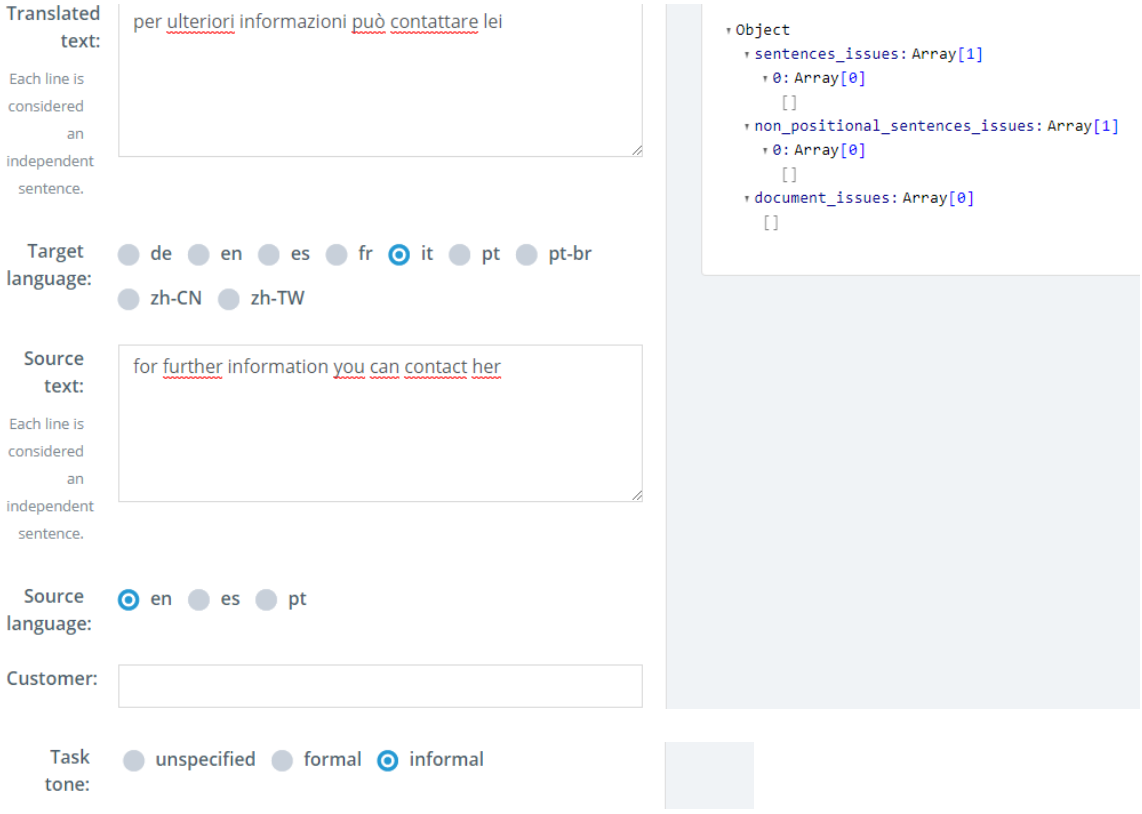

Figure 20 – Submission of a sentence in which the pronoun “lei” occurs in object position to the staging tool when the register is set to “formal”

1	per	per	ADP	_	3	case
2	ulteriori	ulteriore	ADJ	Number=Plur	3	amod
3	informazioni	informazione	NOUN	Gender=Fem|Number=Plur	5	nsubj
4	può	potere	AUX	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	5	aux
5	contattare	contattare	VERB	VerbForm=Inf	0	root
6	lei	lei	PRON	Number=Sing|Person=3|PronType=Prs	5	dobj

Download JSON

Figure 21– POS tagging and dependency analysis of the sentence "Per ulteriori informazioni può contattare lei" output by the Turbo Tagger

With examples (14) and (15), and Fig. 18-19 and Fig. 20-21, the problem is not the output of the parser, as it correctly tags all the words and their syntactical functions, but the fact that the Smartcheck only recognizes the word "lei" as a formal form when it is capitalized (see table 17), and that it cannot recognize whether it refers to a third entity or to an interlocutor. These are ambiguous forms that only a wider context can disambiguate when there is a deviation to the norm regarding the use of initial capital letter in a formal register and when "lei" is writen in lower case, or when it is at the beginning of a sentence and hence always with an initial capital letter that cannot be used as a mark of a formal register.

With the information currently being considered by the system, such rules cannot be implemented in the Smartcheck, but we did implement the warning below to be provided to the editor, in order to check the correct use of the register in these structures.

2. *Ambiguous forms 'la/le': depending on whether they are determiners or clitics, they can occur in both registers.*

The determiners 'la'/'le', in lower-case, are also ambiguous cases in terms of register binary decisions, as they can be either determiners (la = sing.; le = plu.) or clitics (in formal and unmarked register). In fact, as clitics, they can be used both to address a feminine interlocutor in a formal register, and to refer to a 3rd person feminine entity in an unmarked register.

Concerning the words “la” and “le”, it results that these can be both clitics or determiners. When they are determiners these forms are never marked in terms of register.

Clitics:

(16a) **la** ringrazio del feedback.

(16b) **le** inviamo un’e-mail di conferma.

Determiners:

(17a) deve inviare **la** conferma a questo indirizzo e-mail

(17b) qui potrà vedere tutte **le** recensioni riguardo la nostra azienda.

Considering the linguistic context in which these words occur allows us to clearly distinguish determiners from clitics: if we use “la” and “le” as determiners, these always precede a feminine name, singular in the former case and plural in the latter.

(18a) deve inviare **la** conferma a questo indirizzo e-mail → **la** + feminine singular name

(18b) qui potrà vedere tutte **le** recensioni riguardo la nostra azienda → **le** + feminine plural name

Instead, when they are used as clitics, they always precede or follow a verb.

(19a) **la** ringrazio del feedback. → **la** + verb

(19b) volevo ringraziar**la** del feedback → verb + **la**

(19c) **le** inviamo un’e-mail di conferma. → **le** + verb

For our purposes in this section, it is when “la” and “le” are clitics that these forms become relevant in terms of register information, as in this case there is an additional ambiguity, between whether “la” and “le” are referring to a feminine interlocutor in a courtesy form, or if they are referring to a third person feminine entity.

If the norm in Italian is respected, in this case, and without information from the context, the only way to disambiguate these forms is by taking into account whether these forms are capitalized or not:

(20a) **la** ringrazio del feedback → referring to a third person feminine entity

(20b) **La** ringrazio del feedback → addressing Mrs. Rossi

(21a) **le** inviamo un'e-mail di conferma → referring to a third person feminine entity

(21b) **Le** inviamo un'e-mail di conferma → addressing Mrs. Rossi

In terms of the performance of the tools used at Unbabel, concerning the formal register, there is not any problem involving "la" and "le", as they are never marked as an error in the Smartcheck, as shown in Figure 22.

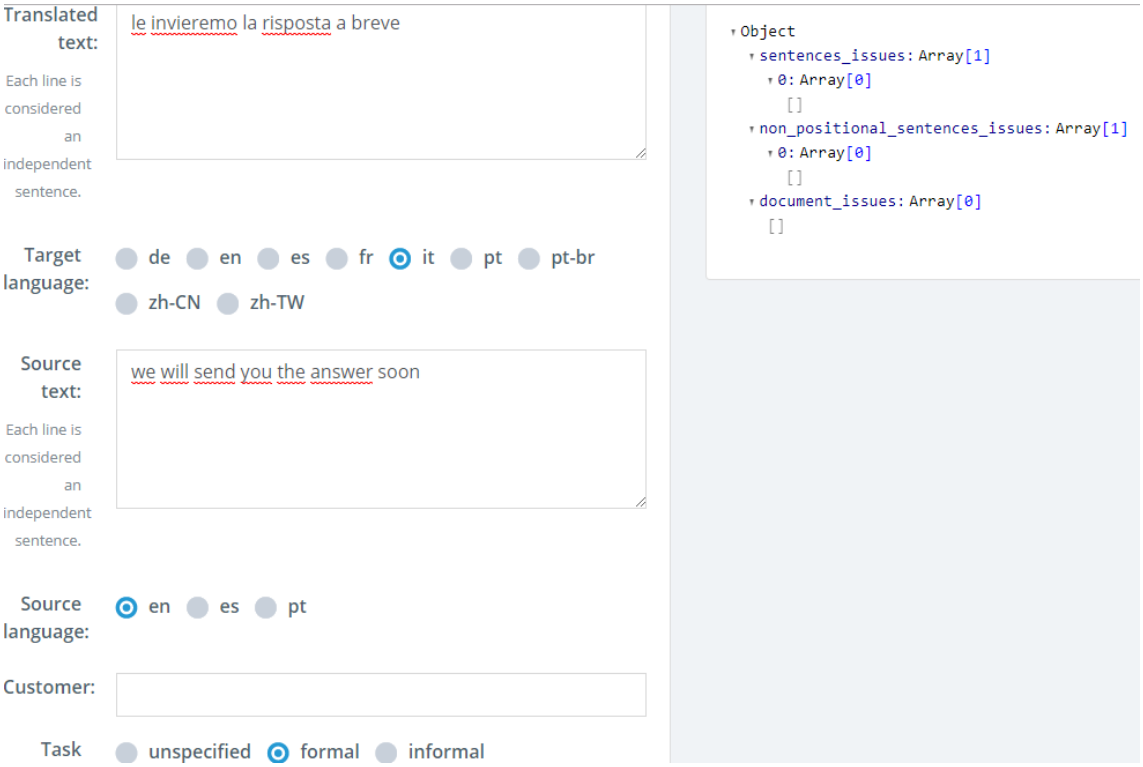

Figure 22 – Submission of a sentence in which the forms "le" and "la" occur to the staging tool when the register is set to "formal"

Concerning the informal register, a classification of the morpho-syntactical category of these forms would be necessary for the Smartcheck to be able to accurately mark register errors involving them, as they should be marked as errors when they are clitics

in an informal register and they should not be marked as errors when they are determinants.

(16) **le** invieremo **la** risposta a breve

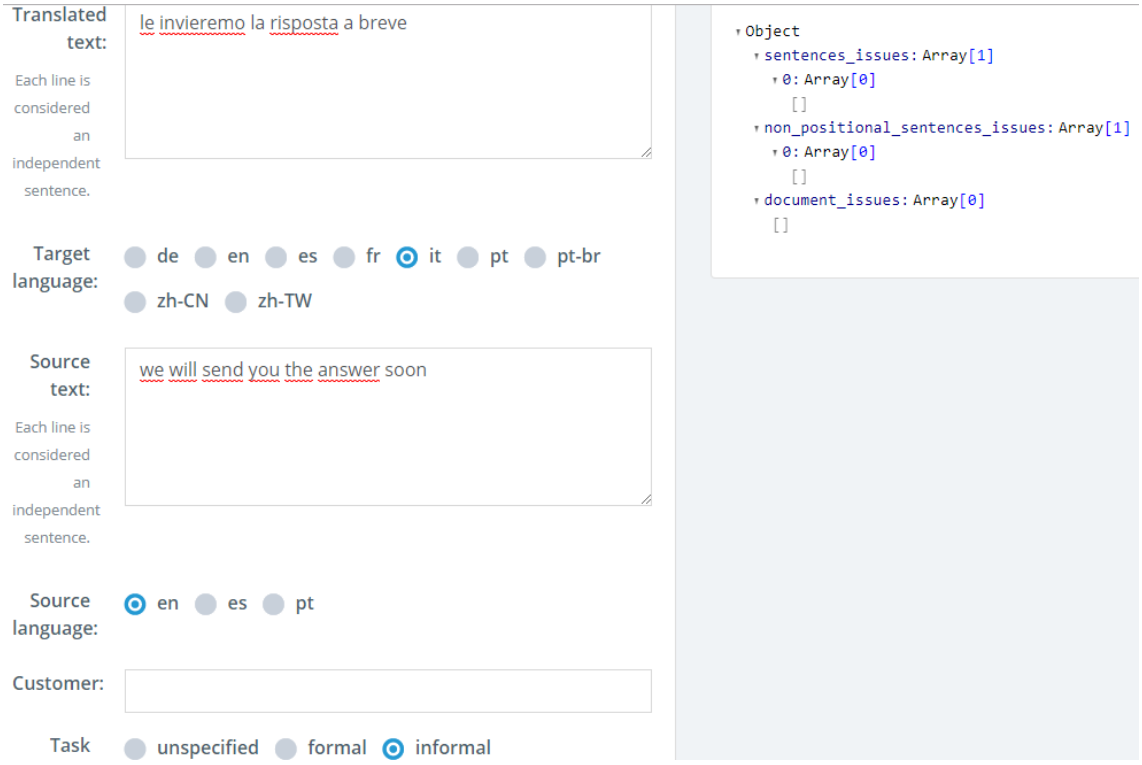

Figure 23– Submission of a sentence in which the forms “le” and “la” occur to the staging tool when the register is set to “informal”

1	le	il	DET	Definite=Def|Gender=Fem|Number=Plur|PronType=Art	2	det
2	invieremo	invieremo	NOUN	Gender=Masc|Number=Sing	0	root
3	la	il	DET	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	4	det
4	risposta	risposta	NOUN	Gender=Fem|Number=Sing	2	dobj
5	a	a	ADP	_	6	case
6	breve	breve	ADJ	Number=Sing	4	nmod

Figure 24 – POS tagging and dependency analysis of the sentence “Le invieremo la risposta a breve” output by the Turbo Tagger

The word “le” in (16) should have been detected as an error by the Smartcheck in Figure 23, when we set the register to informal, and was not. However, in contrast with what was the case for “lei” earlier in this section, the parser does not accurately tag “le” in (16) as a dative clitic, but as a determiner (see Fig.24) even if linguistic rules can easily solve this kind of ambiguity, namely by using the syntagmatic context, i.e. what precedes or follows the word we would like to disambiguate. Hence, in this case, not even by adding the information generated by the parser to the Smartcheck would allow these cases to be distinguished by the tool.

“la”, in (16), on the other hand, was not marked as an error by the Smartcheck, which is correct as it is a determiner in this example. We can also see, in Figure 24, that “la” is correctly tagged by the parser.

Considering all this, at the moment, a rule to mark errors involving the case of “la” and “le” cannot be implemented in the Smartcheck, but once again a warning can be provided to the editor, reminding him to be careful, in informal register contexts, and check whether the words “la” and “le” are clitics, and thus possibly errors in an informal register (if they are referring to a second person interlocutor), or if they are determiners and thus unmarked in terms of register.

3. Possessive ‘suo/sua/suoi/sue’ that can occur in both registers

The possessive ‘suo/sua/suoi/sue’, in lower-case, are ambiguous. They can be used as a courtesy form to refer to items belonging to an interlocutor, i.e. to a 2nd person, or when a third person, singular, masculine or feminine, is the possessor of something.

Once again there is no problem in a formal register, as the Smartcheck does not mark any error, independently of the possessives being a courtesy form related to a 2nd person interlocutor or used in relation to a third person entity, as we can see in Figure 25.

As the surface form of the Italian sentence is always the same whether we refer to a 2nd person possessor or to a 3rd person possessor, we are going to show only one example for the formal register (Fig. 25) and one for the informal register (Fig. 26), as the Smartcheck behaves in the same way. What changes is the original sentences in English:

- Your account is no longer valid – 2nd person possessor
- His/her account is no longer valid – 3rd person possessor

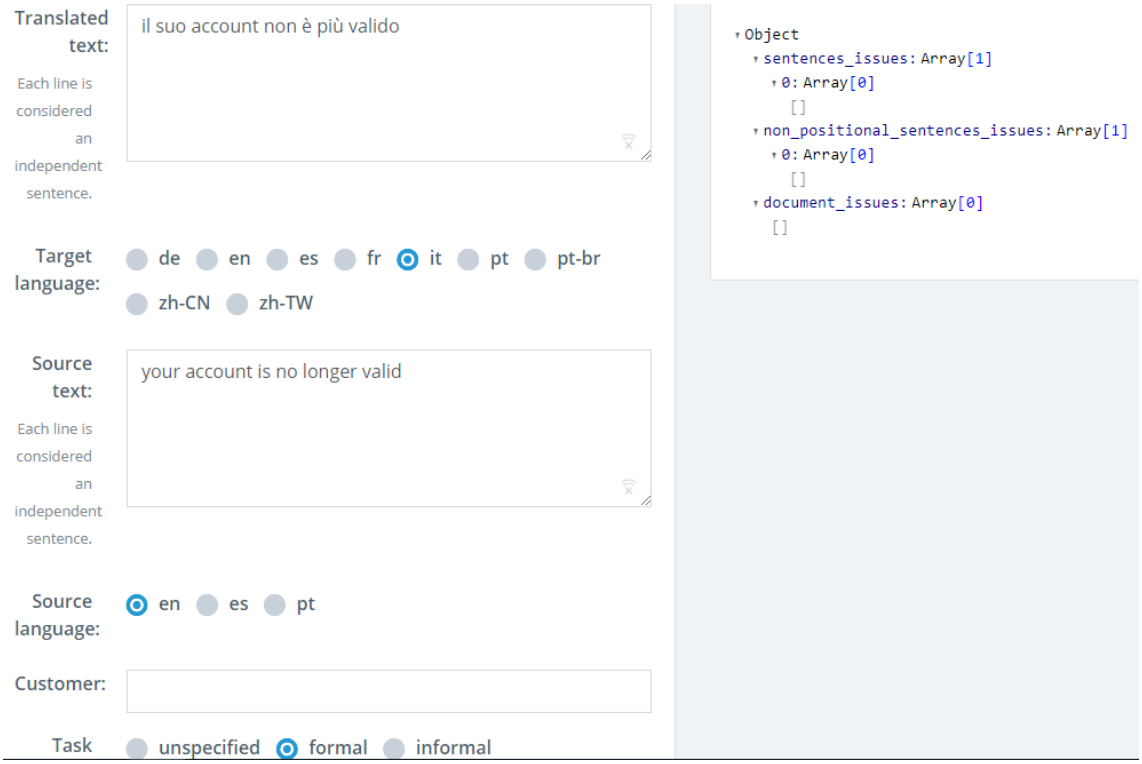

Figure 25 – Submission of a sentence in which the form “suo” occurs to the staging tool when the register is set to “formal”

Problems arise in informal register, in which the system should be able to automatically detect whether the possessives are a courtesy form, and thus a formal register is being used, or used in relation to a third person entity, in which case the possessive is unmarked in terms of register, as shown below.

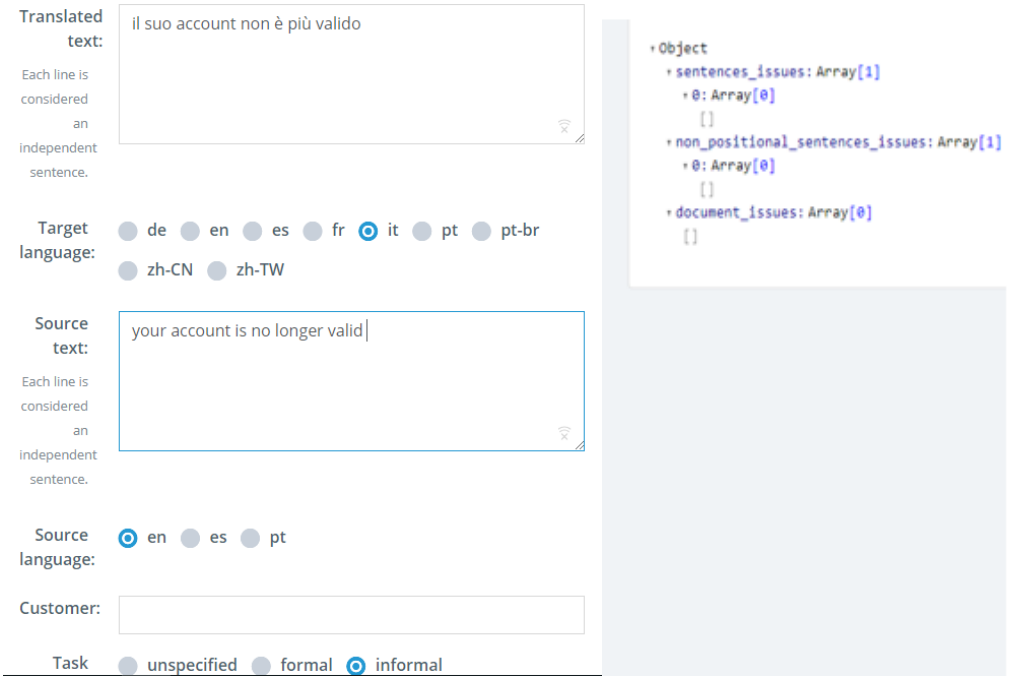

Figure 26 – Submission of a sentence in which the form “suo” occurs to the staging tool when the register is set to “informal”

1	il	il	DET	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	3	det
2	suo	suo	DET	Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs	3	det:poss
3	account	account	NOUN	_	7	nsubj
4	non	non	ADV	PronType=Neg	7	neg
5	è	essere	VERB	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	7	cop
6	più	più	ADV	_	7	advmod
7	valido	valido	ADJ	Gender=Masc|Number=Sing	0	root

Figure 27– POS tagging and dependency analysis of the sentence “Il suo account non è più valido” output by the Turbo Tagger

As made apparent in Fig. 26 and Fig. 27, the problem here is not linked to the parser, because it correctly provides an analysis of these structures, but the Smartcheck that does not detect “suo” as an error when we set the register to informal and the possessive refers to a 2nd person and thus is a courtesy form, as it does not distinguish it from when it is used in relation to a third person possessor, in which case it should not be marked as an error.

Once again, besides the context, there is no information allowing us to distinguish if the possessive is formal or unmarked in terms of register: the only way to distinguish these is by using an initial capital letter, as shown in the bexamples below (see (22a) and (22b) and (23a) and (23b)). When the possessive is written with an initial capital letter, it must necessarily be a formal form. When it is not, we cannot know, without contextual information, as made apparent in the examples below (see ((22b), (23b) and (24a) and (25a)).

Formal: Mr. Rossi, an interlocutor

(22a) Il **suo** feedback è stato molto utile.

(22b) Il **Suo** feedback è stato molto utile.

(22c) **Your** feedback was very useful.

Formal: Mrs. Rossi, an interlocutor

(23a) Il **suo** feedback è stato molto utile.

(23b) Il **Suo** feedback è stato molto utile.

(23c) **Your** feedback was very useful.

Informal: Luca, a third person entity

(24a) Il **suo** feedback è stato molto utile.

(24b) **His** feedback was very useful.

Informal: Anna, a third person entity

(25a) Il **suo** feedback è stato molto utile.

(25b) **Her** feedback was very useful.

Considering all this, this rule cannot be implemented in the Smartcheck, but these possessives can be highlighted as warnings for the editor, telling him/her to pay special attention, in informal register contexts, to verify whether possessives are courtesy forms, and hence an error, or if they refer to a third person.

4. *3rd person singular verbs are ambiguous between a formal register and an unmarked register.*

(26a) **può** risolvere il problema mandandoci una e-mail
(26b) **you** can solve the problem by sending us an e-mail

Formal

(27a) **può** risolvere il problema mandandoci una e-mail
(27b) **he** can solve the problem by sending us an e-mail

Unmarked

If we set the register to formal, the Smartcheck does not mark these verbs as an error, whether they correspond to a courtesy form of the verb related to a 2nd person subject (26) or to a third person subject (27), which is unmarked in terms of register.

If we set the register to informal, we face the problem of recognizing whether the 3rd person singular verb is related to a 2nd person subject, and in this case should be marked as an error because it should be a 2nd person singular verb form, or if it is related to a third person subject, and hence not marked as an error, as this form is unmarked in terms of register.

There is no way to disambiguate it with tense/mood/aspect information, but only with knowledge regarding the subject provided by the context.

6.5.1. GRAMMATICAL ASPECTS OF THE UNMARKED REGISTER

This last section is dedicated to some grammatical aspects that are typical of the unmarked register, but that could not be implemented as Smartcheck rules, as the tool does not include morpho-syntactical context information.

They are nonetheless very important aspects in order to reach a good quality and fluent translation, and hence we present their linguistic description below, with the aim that it can be used at some point in the future to improve the automatic tools used at Unbabel to check and post edit translations.

1. Distinction of subjunctive and indicative moods

The distinction of subjunctive and indicative moods is very important for unmarked register. Nowadays, in certain linguistic contexts, which require the use of the subjunctive, this mood is being replaced by the use of the indicative, a deviation from the norm, but that is starting to become a generalized mark of informal and spoken language. Being a deviation from the standard, it is generally avoided in formal contexts. This substitution takes place only in some cases and it is a linguistic change that is still ongoing and far from being stabilized. This means that not all the cases in which a subjunctive appears are necessarily in a formal register. (Maiden, Robustelli: 2013) Reversely, when an indicative appears in the contexts listed below, it must be in an informal register, otherwise it should be marked as an error.

We are now going to provide a list of structures that are involved in this phenomenon:

- After **expressions of belief, opinion, mental impression, seeming, doubting that…**

(28a) **Credo** che l'ncontro *sia* (subjunctive) alle 18.00. (unmarked)

(28b) **Credo** che l'incontro *è* (indicative) alle 18.00. (informal)

(28c) I think the meeting is at 18.00.

(29a) **Penso** che si *debba* (subjunctive) prendere in considerazione questo fatto. (unmarked)

(29b) **Penso** che si *deve* (indicative) prendere in considerazione questo fatto. (informal)

(29c) I think one must take this fact into consideration.

(30a) **Spero** che la nostra collaborazione *possa* (subjunctive) continuare. (unmarked)

(30b) **Spero** che la nostra collaborazione *può* (indicative) continuare. (informal)

(30c) I hope our collaboration can continue.

- After **qualunque**:

(31a) **Qualunque** cosa *faccia* (subjunctive), non lo perdonerà. (unmarked)

(31b) **Qualunque** cosa *fa* (indicative), non lo perdonerà. (informal)

(31c). Whatever he does, he will never forgive him.

- After **negated relatives**

(32a) **Non** c'è nulla che mi *possa fermare* (subjunctive). (unmarked)

(32b) **Non** c'è nulla che mi *può fermare* (indicative). (informal)

(32c) There is nothing that can stop me.

- After **superlatives**

(33a) Mario è il ragazzo **più inteligente** che *conosca* (subjunctive). (unmarked)

(33b) Mario è il ragazzo **più untelligente** che *conosco* (indicative). (informal)

(33c) Mario is the most intelligent guy I know.

- Indirect questions

(34a) Mi chiedo **chi** lo *abbia invitato* (subjunctive). (unmarked)

(34b) Mi chiedo **chi** lo *ha invitato* (indicative). (informal)

(34c) I am wondering who invited him.

2. Distribution od "di + infinitive" and "che + indicative"

The expression **di + infinitive** is the norm and is thus unmarked register. The expression **che + indicative**, instead, is used only in informal and spoken contexts, as it is a deviation to the aforementioned norm.

(35a) Sono dispiaciuto **di** non *poter venire* (infinitive). (unmarked)

(35b) Sono dispiaciuto **che** non *posso venire* (indicative). (informal)

(35c) I am sorry I cannot come.

3. *Past counterfactual sentences: "congiuntivo trapassato + condizionale passato" versus "imperfetto indicativo + imperfetto indicativo"*

In past counterfactual sentences, the norm is to use the "congiuntivo trapassato + condizionale passato" (see 36b), which is thus unmarked in terms of register, respectively in protasis and apodosis. Concerning informal contexts and everyday spoken language, it is more frequent the use of the "imperfetto indicative + imperfetto indicativo", which is a deviation from the norm, (see 36b).

(36a) Se *fosse venuto* lo avrei visto. (unmarked)

(36b) Se *veniva* lo vedevo. (informal)

(36c) If he had come, I would have seen him.

6.6. SUMMARY

Using annotated data as a strategic point, this chapter reports the importance of the error category of *register*, both in terms of frequency and of its important effect on the quality and fluency of the translation, as well as on the perception and, hence, satisfaction of the client. From a description and analysis of the data, we focus on outlining strategies to reduce the frequency of this error, especially strategies that can be integrated with the tools used at Unbabel to assist human editors, the Smartcheck and the Turbo Tagger. Hence, our approach essentially involved the creation and implementation of rules in the Smartcheck to automatically detect register errors.

The linguistic patterns observed in register errors allowed for the formulation of rules that, in some cases, were deployed, i.e. they were implemented and tested in the Smartcheck, while others were not, as the linguistic specifications involved are not recognized either by the Smartcheck or the parser at the present stage.

In this second case, a description of the phenomena is provided (section 6.5) as well as a discussion on the reasons why the generalizations were not implemented. The main reason why these rules could not be implemented was their propensity to overgenerate given the information available in the two tools used at Unbabel. In this chapter it is made apparent that, for most cases, the limitation to the deployment of these rules is technological, and not linguistic, i.e. having more accurate linguistic information available in the tools would allow us to deal with phenomena such as these. This means that in future stages of development, in which richer and more accurate linguistic information is incorporated in the tools, the work presented in the final sections of this chapter can be straightforwardly added to the automatic quality checking tools.

7. CONCLUSIONS AND FUTURE WORK

7.1. CONCLUSIONS

The general objective of the present work is to contribute to improve the quality of translated texts within the Unbabel pipeline. We focused on texts translated from English into Italian. Aiming at improving the quality in the translations output by Unbabel, we performed an error annotation and compiled a *corpus* in which all the errors detected were analyzed, categorized, and associated to a severity level.

From our annotation experience, we created guidelines for the annotation, aiming at leading to consistent annotation decisions and, thus, to an improvement of inter-annotator agreement metrics. Doing so crucially contributes to the reliability of the data and to the homogeneity of error decisions among annotators.

We focused on register errors, not only because it results to be the most frequent error, but also because it has a great impact on the quality of translated texts. It affects the fluency and the accuracy of the translation and it represents the voice and image of the client.

The thorough analysis of the errors allowed us to identify patterns of errors, enabling the implementation of certain rules, in order to reduce the frequency of the error in the translated texts.

We defined a set of rules that, when it was possible, were implemented in the Smartcheck, the tool that automatically detects errors in target texts to aid human editors in their work. Once the rules for the register were listed, and added to the Smartcheck, a testing stage was applied. For testing the rules deployed in the Smartcheck, we had a process of staging for each rule, in which we analyzed all the expressions included in our rules and checked whether they were recognized by the tool as an error or not. When a problematic expression was not recognized, we analyzed the results to diagnose the source of the problem: it was not recognized by the Smartcheck or it was wrongly categorized by the parser?

Some generalizations in the expression of register in Italian could not be implemented in the Smartcheck, because they involve ambiguous expressions, which are problematic for the Smartcheck -- at its current development stage cannot deal with all the linguistic information needed to tackle the aforementioned issues. These generalizations were nonetheless included in this work, as the description is bound to be useful for the formulation of additional rules at future stages.

The analysis presented in this study focused on the concrete results that we obtained in the improvement of the translated texts quality, in the process performed by Unbabel, as for example the creation of guidelines specific for annotation that improved the inter-annotator agreement metrics, and in explaining next steps to tackle register issues.

7.2. FUTURE WORK

With this work, certain improvements to the quality of translation were achieved, but we aim to continue to work on these features, expanding it also to more domains.

The future work may be focused on expanding the implemented rules for the register in the Smartcheck, not only to texts translated by the machine, but maybe also to texts translated by humans with translation aid systems.

Concerning the context, we can expand these features, not only to helpcenter e-mails, as we did in our work, but also to scientific/technical texts or literary texts, in order to support and help translators in their translation process, with the aim of reducing the time needed to complete each translation.

We also believe that future improvements in the register can be implemented, by trying to develop the tools used at Unbabel: the Smartcheck and the dependency parser. In this way, more rules could be implemented and correctly recognized, so that the translator can save time during the process of post-edition.

As a consequence, another future work line of research could be testing the rules at a production level, i.e. the choice whether to introduce or not warnings about possible errors in the Smartcheck, and whether this information can be useful for the editors in terms of error detection, reducing the frequency of register errors, or if these warnings are too much for the editors, so that it turns out to be a waste of time.

8. BIBLIOGRAPHY

Allen, J. (2003). "Post-editing". In Somers, H. (ed.). *Computers and Translation: A Translator's Guide*. Amsterdam & Philadelphia: John Benjamins Publishing. pp. 297-317.

Allen, J. (2005). "What is post-editing?". *Translation Automation*. 4: 1-5.

(Available at www.geocities.com/mtpostediting/)

Buchicchio, M. (2017). *Português Controlado para a Tradução Automática: Português-Italiano.* MA dissertation, Faculdade de Letras da Universidade de Lisboa, Portugal.

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/MQMusage-guidelines.pdf).

Carletta, J. (1996). "Assessing agreement on classification tasks: the kappa statistic". *Computational linguistics*, Volume 22, Issue 2, pp. 249-254.

Comparin, L. (2016). *Quality in Machine Translation and human post-editing: error annotation and specifications*. MA dissertation, Faculdade de Letras da Universidade de Lisboa, Portugal.

Costa-Jussà, M. R., Fonollosa, J. A. R. (2015). "Last trends in hybrid machine translation and its applications". In Moore, R. K. (ed.). *Computer Speech and Language*, vol. 32, Issue 1. Amsterdam: Elsevier, pp. 3-10.

(Available at http://www.sciencedirect.com/science/article/pii/S0885230814001077)

Di Eugenio, B., Glass, M. (2004). "The kappa statistic: a second look". *Computational Linguistics*, Volume 30, Issue 1, pp. 95-101.

Dorr, B. J., Jordan, P. W., Benoit, W. (1999). "A Survey of Current Paradigms in Machine Translation". In Zelkowitz, M. V. (ed.). *Advances in Computers*, Vol. 49. Amsterdam: Elsevier, pp. 1-68.

Giordano, R., Voghera, M. (2002). “Verb system and verb usage in spoken and written Italian”. In *JADT 2002: 6es Journées internationales d’Analyse statistique des Données Textuelles.*

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

Hutchins, J., Sommers, L. (1992). *An introduction to machine translation***.** London: Academic Press.

(Available at http://www.hutchinsweb.me.uk/IntroMT-TOC.htm)

Hutchins, J. (2000). “Machine Translation”. In Ralstion, A., Reilly, E. D., Hemmendinger, D. (eds.). *Encyclopedia of Computer Science, 4th Edition*. New York, Grove’s Dictionaries, pp. 1059-1066.

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. (2002). “The state of machine translation in Europe and future prospects”. *HLT Central*.

(Available at http://hutchinsweb.me.uk/HLT-2002.pdf)

Hutchins, J. (2005). *History of machine translation in a nutshell*.

(Available at http://www.hutchinsweb.me.uk/Nutshell-2005.pdf.)

Hutchins, J. (2010). “Machine Translation: a concise history”. In *Journal of Translation Studies*, vol. 13, nos. 1-2. Special Issue: Chan Sin Wai (ed.). *The teaching of computer-aided translation*, Chinese University of Hong Kong

Hutchins, J. (2015) “Machine Translation: history of research and applications”. In *Routledge Encyclopedia of Translation Technology*, Chan Sin Wai (ed.). London: Routledge.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

Landis, J. R., Koch, G. G. (1977). "The measurement of observer agreement for categorical data". Vol. 33, pp. 159-174.

Lommel, A. (2015). *Multidimensional Quality Metrics MQM Definition*.

(Available at http://www.qt21.eu/mqm-definition/definition-2015-06-16.html).

Maiden, M., Robustelli, C. (2013). *A reference grammar of modern Italian.* Routledge.

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.

(Available at: https://www.cs.cmu.edu/~afm/Home_files/acl2013short.pdf).

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).

Proudfoot, A., Cardo, F. (2005). *Modern Italian Grammar. A practical guide.* Routledge.

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

Serianni, L. (2010). *Grammatica Italiana*. *Italiano comune e lingua letteraria.* Novara, Utet Università.

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 Carl, M., Way, A. (eds.). *Recent Advances in EBMT*. New York: Springer.

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)

Wu, Y., Schuster, M., Chen, Z., V. Le, Q., Norouzi, M. (2016). Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

(Available at https://arxiv.org/pdf/1609.08144.pdf)

Yamamoto, K., Aikawa, T., Isahara, H. (2012). The Impact of Crowdsourcing Post-editing with the Collaborative Translation Framework. In *Advances in Natural Language Processing. 8th International Conference on NLP*, Jap TAL 2012, Kanazawa, Japan. pp. 1-10.

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.

LOGOS: *Logos Machine Translation System*.

(Available at https://aclweb.org/anthology/A/A97/A97-2009.pdf).

QT21 Launch Pad

(Available at http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics)

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Klingner, J. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation.*

(Available at https://arxiv.org/abs/1609.08144).

Dictionaries

English

Cambridge dictionary (available at http://dictionary.cambridge.org).

Merriam Webster dictionary (available at http://www.merriam-webster.com/).

Italian

Accademia della crusca (available at www.accademiadellacrusca.it).

Dizionario Zanichelli (available at http://www.zanichelli.it/dizionari/).

Vocabolario Treccani (available at www.treccani.it/vocabolario/).

Websites

www.unbabel.com

blog.unbabel.com

languagetool.org

www.taus.net

https://www.taus.net/academy/best-practices/evaluate-best-practices/error-typology-guidelines

https://www.taus.net/knowledgebase/index.php?title=Category:Evaluate

https://www.taus.net/academy/best-practices/best-practices

http://www.lti.cs.cmu.edu/Research/Kant

http://www.bbc.co.uk/languages/italian/tutors/grammar/language_notes/formal_informal.shtml

http://www.locuta.com/eforme.htmlhttp://www.effectivelanguagelearning.com/free-language-lessons/italian

http://aulalingue.scuola.zanichelli.it/benvenuti/2017/03/30/forme-di-cortesia/