POLITENESS CONTROL AS A DOMAIN ADAPTATION PROBLEM IN NMT: fine-tuning vs. multi-register models for Castilian Spanish

Author: Celia Soler Uguet

Advisors: Nora Aranberri Monasterio

hap/lap

Hizkuntzaren Azterketa eta Prozesamendua

Language Analysis and Processing

Final Thesis

13th June 2022

Departments: Computer Systems and Languages, Computational Architectures and Technologies, Computational Science and Artificial Intelligence, Basque Language and Communication, Communications Engineer.

Laburpena

Itzultzaile automatiko neuronalen (IAN) itzulpen-proposamenak gizakiek ekoitzitakoen kalitate-mailara hurbiltzen doazela ikusita, uste dugu iritsi dela unea gramatika-zehaztasunetik haratago doazen hizkuntza-alderdiei erreparatzen hasteko. Ikerketa honetan aztertu ditugu kortesia-maila adierazten duten elementuak kontrolatzeko bi hurbilpen, domeinu-egokitze arazo gisa ulertuta, ingelesetik Espainiako gaztelaniara itzultzen duen IAN bat garatzeko. Emaitzek erakusten dute, Sennrich et al. (2016) lanean proposatutakoari jarraiki, sistema eleaniztun bat egokiagoa dela erregistro eta esaldi-mota desberdinetan kalitate orokor eta doitasun berdintsua mantentze aldera. Izan ere, oinarri-lerro baten erregistro-doiketa saiakerek ahazte katastrofikoa dakarte eta, ondorioz, sistema horien kalitate okerragoa.

Gako-hitzak: hizkuntza, erregistroa, kortesia, IAN, HP, azalgarritasuna, IA ebaluazioa, eleaniztasuna, Espainiako gaztelania

Abstract

Given that Neural Machine Translation (NMT) systems have been proven to generate translations with a quality that is being regarded as close to that of a human, we believe it is the time to start paying attention to aspects of language that go beyond grammatical accuracy. In our research, we explore different approaches towards training a NMT system from English to Castilian Spanish with politeness control of the output and deal with the task as a domain adaptation problem. Our results show that training a multi-register model to deal with different registers following Sennrich et al.'s approach (2016) is the best option when trying to find a balance between overall performance across different registers and types of segments as well as accuracy for producing the right honorific, while fine-tuning a baseline towards each specific register suffers from catastrophic forgetting, thus leading to a worse overall performance of such engines.

Key words in English: language, register, politeness, NMT, NLP, explicability, MT evaluation, multilingual, Castilian Spanish

Table of contents

List of Figures

List of Tables

1 INTRODUCTION

As Vanmassenhove et al. suggest (2021), now that Neural Machine translation (NMT) systems have reached a quality that is (arguably) close to that of human translations, it is time to start paying attention to aspects of language that go beyond grammatical accuracy.

In that sense, one of the main troubles that NMT has been facing so far has to do with register or the use of politeness, which can be crucial for tasks such as subtitling translation (Etchegoyhen et al., 2014). Moreover, deviations from what is expected in the use of politeness may give rise to social misunderstandings in the communication and although this might seem like a petty problem in some languages or cultures, it can become extremely critical for cultures such as the Japanese one, where the concept of *place* (place where one belongs or place where one stands) is fundamental to instances of politeness in this language (Haugh, 2005).

One of the largest source of communication nowadays is instant messaging (IM). WhatsApp, which is one the biggest IM applications of Western Europe, has over 2 billion users in 180 countries. Some studies claimed that in 2017, over 29 million messages were sent per minute through this application (Smith, 2018). However, even if the IM technology is so developed, there is still a barrier to communication for people having different native-languages (Yang and Hua-Yi, 2010).

On their study on Machine Translation (MT) in IM, Tekwa (2018) arrives at the conclusion that machine-translated IM could improve the willingness to communicate (WTC) of beginner FL learners and increase their opportunities to communicate (OTC).

Even if these IM applications are most commonly used by individuals, there is also an increasing number of companies or small business which are moving away from emails and that communicate with their customers throughout these apps. Therefore, we believe that adding a translating tool to an IM application could be of help for both non-native speakers who need to communicate with others in a different language that they may know but which they do not master; as well as for companies which might need to make use of it to expand their business to other countries. Moreover, going back to the importance of politeness in communication, it might be desirable to have an engine that uses specific forms of politeness depending on the person that the user is

communicating with, rather than an engine which is not consistent in the use of forms in languages with honorifics (lexical forms to express respect).

Having this motivation in mind, the main objective of this project is to explore different options towards training a NMT system with politeness control of the output. We focus on the translation of English to Castilian Spanish, which is a language that uses different honorifics depending on the person that one is talking to or the situation that one finds themselves in.

However, to do so, we believe it is important to take the linguistic theory as a starting point. Therefore, in the first pages of the project, we revise some linguistic concepts such as *linguistic variation*, *register* and *politeness*; revisit the concept of *human quality* and what it means in MT evaluation and revise some State-of-the-art (SOTA) approaches towards the task at hand. We then present an approach to be used for politeness classification of a Castilian Spanish parallel corpora and train NMT models following two different domain adaptation approaches. Finally, we carry out an in-depth evaluation of the resulting systems using automatic metrics such as COMET (Rei et al., 2020), and human evaluation, following the guidelines proposed by Marie, Fujita, and Rubino for improving the scientific credibility in the research of MT (2021).

2 LITERATURE REVIEW

2.1 SOME LINGUISTIC CONCEPTS

2.1.1 The notion of linguistic variation

The concept of *language* used in linguistics represents a practical way of reunifying the different means of communication used inside human groups. This reunification is the result of putting together great sets of components and rules that have things in common. If we use the concept of *sets*, we are implicitly referring to homogeneity. However, language is not an inert and simple object that is transferred and passively accepted, but an object that exists because it is used (Álvarez, 2006). Saussure (1915) established the distinction between *langue* and *parole*, which are commonly translated into English as *language* and *speech* respectively, although not without some danger of ambiguity. The concept of *language* is for him an abstract and homogeneous system, which is incorporated by the speakers and which is subjacent to the exercise of speech. On the other hand, *speech* is not homogenous and it is marked by diversity and differences among speakers. Sociolinguistics (the scientific study of social variation), however, suggests that *languages* have their stable spaces –characterized by homogeneity– and unstable spaces –characterized by heterogeneity and variation–. Therefore, linguistic variation would correspond in Saussarian terms to a *speech* phenomenon, rather than a *language* one.

The beginning of study in the variation of language was marked by a research carried by William Labov (1986), where he researched the relationship between linguistic variation and social stratification in New York, focusing on English from Afro-American speakers. This study contributed to the development of sociolinguistics. This new kind of research would try to describe the specificity of the real use of languages. From Labov's perspective, the different ways of speaking a language are neither random, nor the result of stylistic choices of speakers. These variation events are ruled by internal and external factors.

Moreover, all languages present variation phenomena in their use (Álvarez, 2006). Therefore, speaking of *the* French language, *the* German language or *the* Spanish language is in itself a considerable generalization and abstraction. There exist as many

different *ways of using speech* (languages) as many different collectivities who use language, and for the sake of rigor, one could even say that there are as many forms of expression as persons who make use of a language.

In this sense, according to Álvarez (2006), there are three linguistic varieties that correspond to three different axes:

- Diatopic variation: corresponding to the geographic axis,
- Diastratic variation: corresponding to the social axis,
- Diacronic variation: corresponding to the time axis.

All these variations show, as pointed out by Calvet (1987) that every speaker, even those who would consider themselves monolingual, are plurilingual or *pluridialectal* to some extent, since they have a wide range of abilities within the same set of linguistic rules. The actual use of this individual *pluridialectism* would activate a range of dialectal abilities and what Hymes defined as communication competence (1971). The notion of communication competence introduces the important role that the situation plays in the origin of the situational variation, which leads to a selection of style of language according to the specific situation, what Coseriu (1958) denoted as *diaphasic variation*.

Thus, as we can see, languages are not a simple and homogeneous phenomenon, but an abstract generalization of ways of using speech. The situation plays an extremely important role in the selection of words, leading to these inner variations inside of what we call *languages*. Therefore, we believe that such variation phenomena should not be overlooked or fall into oblivion when dealing with synthetic language in tasks such machine translation or language generation.

2.1.2 The notion of register

The concept of *register* is also central to Halliday´s theory of Systemic Functional Linguistics (SFL) (Lukin et al., 2011). SFL was developed during the 1960 and studied the language through meaning (i.e. its function). Halliday was concerned with the writer's purpose for writing a sentence, rather than the sentence itself (Matthiessen & Halliday, 1997). His model of language analyzed texts in four ways: context, semantics, lexico-grammar and phonology. Without going into much detail about Halliday's model of language, we would like to draw the reader's attention to the first element of such list: the context. For Halliday, context is integral to the process of creating meaning

(1997). The author also makes the distinction between two types of contexts within an act of speech:

1- The context of culture

2- The context of situation

While *the context of culture* has often been referred to as *genres*, *the context of situation* is what has been technically referred to as *register*.

Halliday´s idea of register is composed of three dimensions: field, mode and tenor. The field refers to the area in which the linguistic activity is operating (specialized vs. non-specialized discourse); the mode has to do with the mean in which communication is taking place (written vs. oral); and the tenor has to do with the relationship between the speakers (relatives vs. workmates) (Carrera, 2014). It might seem as a straight definition; however, the concepts of *mode* and *tenor* present their own complications.

Regarding the definition of mode, it is normally said that there exist two modes: written and oral. However, this definition excludes those oral discourses that are much alike a written discourse, e.g. a dissertation; and, as well, written discourses that look much alike an oral discourse, e.g. instant messaging writing. Silva-Corvalán (2001) presents a much more detailed description of the different types of modes that can exist in her book *Sociolingüística y pragmática del español* as it can be observed in Figure 1.

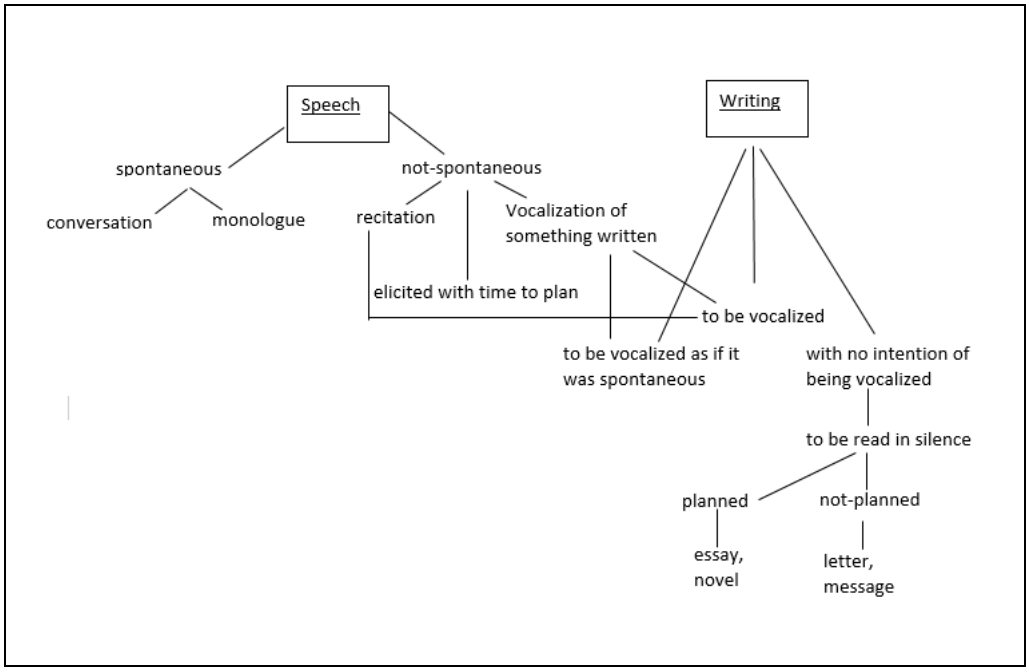

Figure 1: Subcategories of mode (translated from Silva-Corvalán (2001, p. 22))

Regarding the tenor, it can be linked to the relationship between the speakers in the linguistic exchange and their respective intentions. It comprises from the degree of formality or informality, to the speaker´s intentions or the value given to the linguistic exchange by the different participants of the conversation (Carrera, 2014).

However, as claimed by Halliday (2002 [1977], p. 58):

> "…we shall not expect to be able to show that the options embodied in one or another particular sentence are determined by the field, tenor and mode of the situation. The principle is that each of these elements in the semiotic structure of the situation activates the corresponding component in the semantic system, creating in the process a semantic configuration, a group of favored and foregrounded options from the total meaning potential that is typically associated with the situation type in question. This semantic configuration is what we understand by the register."

What Halliday is trying to explain is that all the processes or signs coming from each of the elements (*field, mode* and *tenor*) creates what he denoted as *semantic configuration* (e.g. some forms that will tend to appear more in some situations that in others), and this is what we understand by a register. In other words, these three register variables are used to explain people's intuitive understanding that individuals use different resources and different parts from the system of language (Matthiessen & Halliday 1997) (i.e., different words) depending on the situation.

2.1.3 Register studies

As Lukin et al. (2011, p. 206) put it:

> "At the moment, register studies are still divided between approaches which prioritize complex models of language but can´t quantitatively test their hypotheses yet, and those which prioritize sampling strategies and automated coding and therefore can only handle more parsimonious models of language, although computation register profiles based on co-selection of lexico-grammatical categories are gaining ground."

What we can extract from this quote is that being register such a complicated topic, there are several approaches towards its study: those that are complex and whose results are difficult to interpret; and those which might be over-simplistic. There is also research on register using computation approaches which select lexico-grammatical features in order to find those signs that the field, mode and tenor leave in the language depending on the register.

Briz (2010) also takes a deeper look into diaphasic variation and register in Castilian Spanish and points out that we should analyze variation as a global, gradual and hierarchic fact. His main point is that variation is dynamic. For instance, when somebody is speaking colloquially, dialectal characteristics (geographic axis) emerge, and so do sociolectal characteristics (social axis) such as sex, age or genre.

2.1.4 Register and politeness

"Politeness is essentially a matter of taking into account the feelings of others as to how they should be interactionally treated, including behaving in a manner that demonstrates appropriate concern for interactors' social status and their social relationship" (Brown, 2015, p. 1). In that sense, we could consider politeness as one of the aspects that conform the tenor, and therefore, which has an impact on the register.

Since politeness is crucial to the construction and maintenance of social relationships, politeness in communication goes to the very heart of social life and interaction. This phenomenon has called the interest from theorists in a wide range of social sciences. Brown and Levinson (1987) studied the concept of politeness across cultures and discovered that there were parallels in the construction of polite utterances across widely differing languages and cultures, and they argued that universal principles underlie the construction of polite utterances. For them, there exist at least three social factors that guide the choices of the speakers: (1) one tends to be more polite to social superiors (the superior is less polite to an inferior); (2) one tends to be more polite to people one does not know (politeness is symmetrically exchanged), and (3) in any culture there are norms and values affecting the degree of imposition or unwelcomeness of an utterance, and one tends to be more polite for more serious impositions.

Linking this to research in MT, we believe that now that NMT systems have reached a quality that could be considered close to that of human translations (Vanmassenhove et al., 2019), focusing in other aspects that go beyond grammatical accuracy should be a major interest for the community. One of these is register, a concept that we have been trying to describe throughout the section, focusing on the concept of tenor and, more precisely, on politeness. No matter how accurate the translator system might be, if the engine is not able to use the correct register or polite term, there might be misunderstandings in the communication, which can be extremely impactful in certain

cultures such as the Japanese one, where the concept of *place* is fundamental in social interactions (Haugh, 2005).

2.1.5 Characterization of the informal and formal registers

In his study on registers, Briz (2010) gives a definition of colloquial and formal register, or what he denotes as the prototype of colloquial (+colloquial) and prototype of formal (+formal).

On the one hand, the prototype of colloquial register presents the following features:

+ Equality between interlocutors

+ Experiential relationship of proximity: context and experience are shared between interlocutors

+ Familiar interactional frame

+ Daily life thematic: non-specialized topics

+ Interpersonal purpose

+ Informal tone

According to this definition, *colloquial* is a concept that describes a precise communicative situation that can be characterized by immediacy, social proximity and other aspects associated to the situation. The more or less presence of all these aspects determines different degrees of informality, thus creating peripheral registers. For instance, as pointed out by Briz, one can speak colloquially with a professor, even if there is a distance between the interlocutors and one can speak colloquially even if the topic is specialized.

On the other hand, the prototype of formal register presents the following features:

- Equality between interlocutors

- Experiential relationship of proximity: context and experience are shared between interlocutors

- Familiar interactional frame

- Daily life thematic: non-specialized topics

- Interpersonal purpose

- Informal tone

2.1.6 Politeness in Castilian Spanish: *tú* vs. *usted*

Following this characterization of what is colloquial (+colloquial) and what is formal (+formal) by Briz (2010), we take look into one of the aspects that can mark this distinction in Spanish: the use of different forms of address (honorifics) to express politeness, while English only has the form *you* and express politeness using other phenomena such as indirect speech or impersonalization (Fukushima & Iwata, 1985).

La Nueva gramática de la lengua Española (NGLE) (2009) defines the different forms of addressing someone as the pronominal varieties that are chosen depending on the social relationship between interlocutors of the speech act. As described by Fernández, (2003) in peninsular Spanish (which we will refer to as *Castilian Spanish* from now on, given that is the common term in NMT), *tú* tends to be the form used in situations where interlocutors are close to a certain extent; meanwhile *usted* tends to be the form used to show respect and distance. It is worth to highlight the fact that our research focuses on this specific diatopic variety of Spanish because this is the dialect that is spoken in the area where the study is taking place, and thus, we believe that the results of a potential human evaluation can be more reliable if the study deals with this variety of Spanish.

At this point, it is time for us to characterize two different registers in Castilian Spanish by referring to the use of *tú* and *usted* and linking them to the different concepts we have been discussing so far:

Firstly, the use of these two forms is directly related to the concept of *politeness* described in Section 2.1.4. Without deepening too much into diachronic aspects of language (u*sted* is not being used in Castilian Spanish as often as it was some generations ago (Verdeguer, 2017)), when trying to speak to somebody in a polite manner, the speaker tends to use *usted,* because as we pointed out in the previous paragraph, it shows respect and distance from the other person.

Secondly, going back to Briz's theory on registers, even if one can use *usted* while speaking colloquially and the other way around, we can argue that the form *tú* would be the most common form in +colloquial speech acts, and the form *usted* would be the most common form in +formal speech acts. Therefore, these two aspects represent the two prototypes from Briz: a +colloquial one (using *tú*), and a +formal one (using *usted*).

Finally, if we try to link these forms to Halliday's idea of register, we can easily start to notice that the two words at hand are indeed signs left by the *tenor* (relationship between the speakers). We can therefore say that in texts where *tú* appears, this word is a sign of informality (*informal tenor* or *informal tone* as denoted by Briz); and in texts where *usted* appears, such word is a sign of formality (*formal tenor* or *formal tone*).

A summary of the characteristics of the two registers from Castilian Spanish according to the use of politeness and tenor and how they relate to the prototypes from Briz can be found in Table 1.

	INFORMAL REGISTER	**FORMAL REGISTER**
POLITENESS	less polite	more polite
BRIZ REGISTERS	colloquial register	formal register
TENOR	informal tenor	formal tenor
SIGNS	*tú*	*usted*

Table 1: Characterization of two registers in Spanish

Although through this research we will be referring to these configurations (informal and formal) as registers, we are characterizing them with respect to only two forms (*tú* and *usted*) and one aspect (politeness). As explained throughout the previous sections, register is a complex and multifactorial concept, and thus, although being aware of how simplistic this characterization might be from a linguistic perspective; we believe that it might be a good starting point for the study of register in NMT.

2.2 OTHER RELATED CONCEPTS

In the following lines we present some concepts that, while not being directly connected to MT or register, they have some connection with the and will be referred to later in the study.

2.2.1 Register in audiovisual translation: prefabricated orality

In her study on audiovisual translation, Baños (2009) focuses on the research of what is defined as *prefabricated orality*. Audiovisual texts are characterized not only by the unusual mode of the linguistic discourse, but also by the way in which they are

produced and received (Salvador, 1989). In that sense, Baños claims that these texts have been written but should be read as if they had actually not. That is where the term *prefabricated orality* comes into play. This kind of texts presents characteristics from spontaneous speech (which it tries to imitate) and from other unique characteristics that come from writing. Therefore, when referring to a translated audiovisual text, both the source and the translations of audiovisual texts present these characteristics (Chaume, 2001).

Going back to the section 2.1.2 where the concept of *mode* was explained; we can start to notice that the dichotomy *written* vs. *oral* is not as simple as it might seem, and this should be taken into account when characterizing a particular register.

2.2.2 Pro-drop languages

Spanish is what is denoted as a *pro-drop language* (Świątek, 2012). The term *pro-drop* stems from Noam Chomsky's Lectures on Government and Binding (1981) as a cluster of properties of which *null subject* was one. According to this parameter, languages like Italian and Spanish are classified as pro-drop languages, while English or French are not. This pro-drop parameter (also referred to as *null subject*) is specific to certain languages which allow subject pronouns to be omitted because the person is implicitly present in the verb. These personal pronouns tend to be used for emphasis in some cases such as expressing that somebody did an action and it was actually that person. An example of this phenomenon can be found in Table 2:

Source
Speaker 1: *-Hoy he ido a comer a casa de mi abuela. ¿Qué habéis hecho* ***vosotros****?* Speaker 2: *-****Yo*** *he salido a correr.* ***Mi hermano*** *se ha ido a Barcelona.*
Translation
Speaker 1: Today **I** went to grandma's for lunch. What did **you guys** do? Speaker 2: **I** went running. **My brother** was off to Barcelona for the day.

Table 2: Examples of sentences where the subject is omitted in Spanish

Since the first speaker is explaining what he did that day without any other reference, he avoids using the pronoun *yo* (*I*). However, since the second speaker is actually talking about what he and another person did, he uses *yo* to remark that it was actually him who did it and not his brother.

This phenomenon is of interest for the research of register, since there are times where the forms *tú* and *usted* are omitted and the register is implicit in the verb form used, i.e. is marked grammatically rather than lexically.

2.2.3 Instant messaging and Machine Translation

Information and communication technologies (ITC) impose a great technologic paradigm (Castells, 1999), which is characterized by the capacity of penetrating in all areas of human activity and processing of knowledge, information and communication. According to Sánchez (2015), this activity has raised some concern amongst teachers and linguists, there being those really open to it, to those that consider that this phenomenon is impoverishing language to a great extent.

McLuhan (1964) said that societies have always been molded by the means of communication they use, rather than the content of the communication itself; and, according to Sánchez (2005), the mean is actually the message: it acts upon senses, drastically changing the human way of perceiving and leading the change on society and culture.

Yang & Hua-Yi (2010, p. 1) already stated the important role that instant messaging played in our society around 10 years ago:

> 'Along with the rapid development of Internet technologies, the instant messaging has become nowadays an important medium for a huge number of people to communicate with friends, family, and even colleagues while working. People who come from different corners of the world speak different native-languages. Even if the instant messaging technology is so developed, there is a barrier to communication for people having different native-languages.'

Moreover, on his research on machine translation, IM and foreign language learning (FLL), Tekwa (2018) arrives at the conclusion that machine-translated IM could improve the willingness to communicate (WTC) of beginner FL learners as well as increasing their opportunities to communicate (OTC). He defines OTC as the 'ability to use a specific tool or take advantage of a specific platform to communicate in a foreign language' (2018, p. 7).

Even if FL learning is not among the topics of this research, we believe that the conclusions drawn from Tekwa's work are in fact essential to understand the usefulness of MT in Instant Messaging for non-native speakers or FL learners. Given that our research is focused on NMT, we believe it could be of interest to think of a real-word

application of the systems we are creating. Having this motivation in mind, many decisions that will need to be made throughout the research will be geared towards this final use case. Therefore, given the importance of IM (also referred to as *chat messaging*) in nowadays social interactions, we believe that NMT could help many people with communicating in a language that they might know but do not speak fluently with either their friends, work colleagues or clients.

2.3 REGISTER AND POLITENESS IN NMT

With the appearance of Neural Machine Translation (NMT) on the scene, MT has achieved a much higher quality than a couple of years ago. However, as it has been exposed throughout the first section of this research, languages are complex phenomena and cannot be reduced to a completely homogeneous set of linguistic rules. In this sense, register presents itself as one of the aspects of languages that can be explained as the context of a situation and which creates variations within the same language. We have also gone through the concept of politeness as one of the aspects that conform the tenor (one of the three dimensions that conform the register in Halliday's theory). However, the task of creating a NMT engines for specific levels of politeness has not been a priority so far.

Even when we can find research that provides different alternatives towards achieving a control of the NMT output such as using terminology constraints (Dinu et.al, 2019) or constrained decoding (Post and Vilar, 2018), these are not specifically designed for politeness control. To the best of our knowledge, one of the only research which focuses on the exact same task that we are dealing with is that of Sennrich et al. (2016) who address it as a domain adaptation problem. In their research, the authors use a preliminary step for classifying a corpus into different registers by using parsing, and use them as domains for training a multi-register engine. However, there have been other attempts of dealing with the task of register or politeness classification using different techniques.

2.3.1 Classifying text based on politeness

In the next section, we succinctly introduce some techniques that have been used for classifying text according to politeness. This is the starting point for creating in-domain

data-sets divided by levels of politeness and tackle the task of training an NMT with a focus on politeness as a domain adaptation problem.

Danescu-Niculescu-Mizil et al. (2013) present a classifier to predict the level of politeness in a sentence. For the task, they create a 10,000-long annotated corpus with requests from the Wikipedia community of editors (Wiki) and the Stack Exchange question-answer community (SE) and label each request using ratings. In this research, the authors compare two classifiers: a bag of words approach (BOW) and a linguistically informed classifier (Ling.). They train the classifiers using a Support Vector Machine (SVM) and the unigram features (BOW), plus the linguistically annotated features (Ling.) Their results show an absolute improvement of 3-4% of the Ling. classifier over the BOW model.

Another approach is presented by Aubakirova and Bansal (2016). In their paper, the authors present an interpretable neutral network for predicting politeness. Their models are based on a simple convolutional neural network directly on raw text, avoiding any manual identification of complex sentiment or syntactic features. As an experimental setup, they make use of the two datasets released in the aforementioned paper (Danescu-Niculescu-Mizil et al., 2013). Their experiments show that without using any manually defined, theory-inspired linguistic features, a simple Convolutional Neural Network (CNN) model (85.8% accuracy on the Wiki test-set) performs better than the feature-based methods (Ling. model from Danescu-Niculescu-Mizil achieved 82.6% on the Wiki test-set).

Moreover, Niu and Bansal (2018) present a politeness classifier trained using a bi-directional LSTM followed by a convolutional layer (LSTM-CNN) in order to capture long-distance relationships in the sentence as well as windowed filter based features. The classifier outputs probabilities over two labels, namely *Polite* and *Rude*. Their model achieves comparable in-domain accuracy (85.0% accuracy) to the results reported for the Wiki dataset in Danescu-Niculescu-Mizil et al. (2013) (82.6% accuracy) and Aubakirova and Bansal (2016) (85.8%).

As we can notice by the previous aforementioned approaches, researchers have proposed to address the task of politeness classification by training regression models or binary classifiers. However, most of these are supervised approaches which require a labelled dataset for training. Moreover, all of them focus on the English language, while

resources with labelled data with respect to politeness in other languages are still scarce. However, politeness is not marked in the same way in each language. For instance, while English does not make a difference in the form of addressing the person and includes lexical forms of expressing gratitude (*please, thanks*) or grammatical constructions such as modal verbs (*Could* you please call me when you get back?), Spanish has two distinctive forms of address which might be easier to identify with a parsing approach than with a binary classifier.

That is why, in their research for creating NMT systems with a control in politeness in English→German, Sennrich et al. (2016) perform a study to control honorifics via *side constraints*. They claim that, for languages without these honorifics such as English, the task of predicting the appropriate register can be complicated. However, they prove that by marking up the source side of the training data with a feature that encodes the use of honorifics on the target side, it is possible to control the honorifics produced at test time. They claim that when training the NMT engine, the correct feature is extracted from the sentence pair as described in the following lines; while at test time, the side constraint is provided by a user who selects the desired level of politeness of the translation. They add side constraints as special tokens at the end of the source text (<*T*> for formal instances and <*V*> for informal instances). The attentional encoder-decoder framework is then able to learn to pay attention to the side constraints, without generating them in the target.

To do so, the authors automatically annotate politeness on a sentence level with rules based on a morpho-syntactic annotation by ParZu (Sennrich et al., 2013). Sentences containing imperative verbs are labelled as informal, while sentences containing an informal or polite pronoun from a list of pronouns are labelled with the corresponding class. Since some pronouns are ambiguous –polite pronouns (*Sie*) are distinguished from 3rd person plural forms (*sie*) by the capitalization, but are ambiguous in sentence-initial position–, in this sentence-initial position, the authors consider them polite pronouns if the English source side contains the form *you* or *your*. For *Ihr* and *ihr* (plural *you*), ParZu is used for carrying out morphological annotation to distinguish between the informal 2nd personal plural nominative, the 3rd person singular dative, and the possessive; for possessive pronouns, they distinguish them between polite forms and 3rd person forms by their capitalization. If a sentence matches rules for both classes, the authors label it as *informal* and all sentences without a match are considered neutral.

2.3.2 Domain adaptation in NMT

High quality domain specific machine translation systems (MT) are in high demand whereas general purpose MT has limited applications. What is more, general purpose translation systems usually perform poorly and hence it is important to develop translation systems for specific domains (Koehn and Knowles, 2017). Domain adaptation in Neural Machine Translation involves leveraging out-of-domain corpora to improve in-domain translations (J. Kirkpatrick et al., 2017).

Chu and Wang's (2017) analyze different techniques for domain adaptation in real word scenarios, and divide these techniques into two categories: data centric and model centric. Data centric approaches focus on the data being used, meanwhile model centric approaches focus on NMT models that are specialized for domain adaptation. In the case of NMT, data centric approaches include the use of monolingual corpora or out-of-domain parallel corpora and the generation of synthetic parallel corpora; and model centric approaches include the use of cost weighting, regularization or fine-tuning (training-objective centric), the use of deep fusion or a domain discriminator (architecture centric) or the use of ensembling (decoding centric).

In the following lines we offer a description of two of the approaches that we considered for our research: a widely used technique such as fine-tuning and the *multi-register* approach that Sennrich et al. proposed for politeness control in NMT (which was introduced in the previous section).

2.3.2.1 Fine-tuning approach

The first approach is considered model centric, or more precisely, training-objective centric. In this method, an NMT system is trained on a resource rich out-of-domain corpus until convergence, and then its parameters are fine-tuned on a resource poor in-domain corpus. An overview of how fine-tuning is performed can be found in Figure 2. One of the main drawbacks from these systems is what is called *catastrophic forgetting.* This is a phenomenon, whereby after a model has been trained on task A and then retrained on task B, it forgets much of what it originally learned on task A (Kell, 2018). In other words, the new model achieves a better performance in task B while worsening the performance in task A.

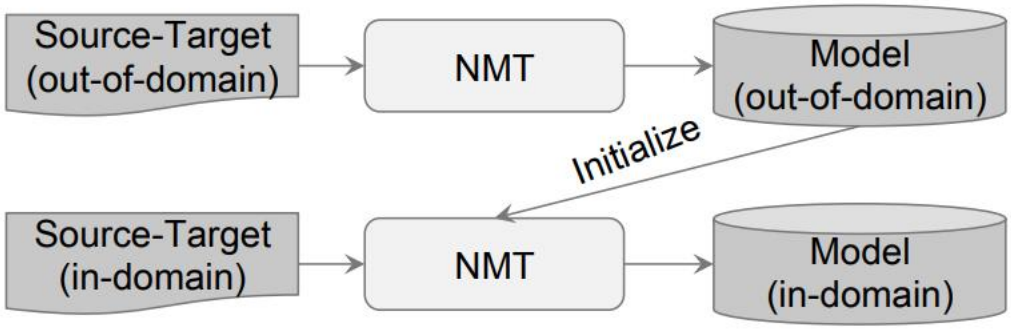

Figure 2: Outline of the fine-tuning approach (extracted from Chu et al. (2018, p. 386))

Different approaches have been proposed for tackling this problem, such as combining multi-domain and fine-tuning methods or using regularization techniques such as elastic weight consolidation (Kell, 2018).

The first approach, denoted by the authors as *mixed fine-tuning* suggests training an NMT model on out-of-domain data until convergence and then resuming training the NMT model from step 1 on a mix of in-domain and out-of-domain data until convergence. Therefore, an out-of-domain development set is first used for training the out-of-domain NMT models, then a mix of in-domain and out-of-domain development set are used for mixed fine-tuning. An overview of this method can be found in Figure 3.

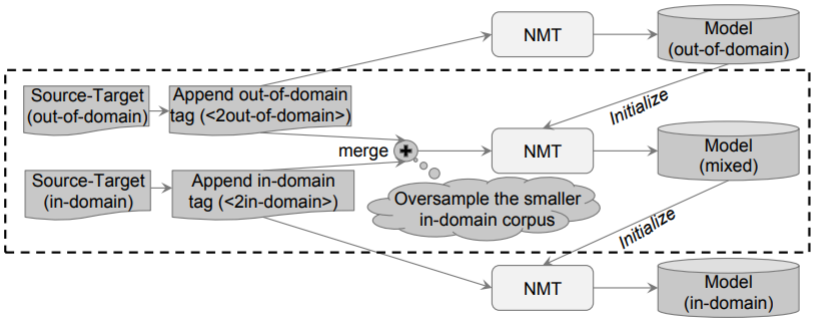

Figure 3: Outline of the mixed fine-tuning approach (extracted from Chu et al. (2018, p. 387))

The second one, EWC, regularizes the parameters which are relevant to the first task by adding a penalty term to the loss function (Kell, 2018). This method was shown to mitigate the degradation on the second task after 20,000 iterations, although to a lesser

degree than other interpolation models. However, the computational and runtime requirements were significantly less, which meant that this NMT system could be deployed in situations where these are constrained.

2.3.2.2 Multi-domain approach

The second approach was first introduced by Sennrich et al. (2016) and has been used for other tasks such as multilingual NMT (Johnson et al., 2017). This method uses the placement of tags in the training for helping the decoder at translation time.

This method has been used by Google for training a single Neural Machine Translation model to translate between languages, (Johnson et al., 2017). Instead of applying changes to a model architecture from a standard NMT system, it introduces an artificial token at the beginning of the input sentence to specify the required target language. Using a shared wordpiece vocabulary, their approach enables Multilingual NMT systems using a single model. In WMT'14 benchmarks, a single multilingual model achieved comparable performance for English→French and surpassed state-of-the-art results for English→ German.

Sennrich et al. (2016) perform English→German experiments on OpenSubtitles (Tiedemann, 2012), a parallel corpus of movie subtitles, since politeness is considered an open problem for automatic subtitle translation (Etchegoyhen et al., 2014). They train an attentional encoder-decoder NMT system using Groundhog[1] (Bahdanau et al., 2015; Jean et al., 2015) and use BPE to represent the texts with a fixed vocabulary of subword units with size 90,000. The training corpus consists of 5.58 million sentence pairs, out of which they label 0.48 million sentence pairs as polite, and 1.09 million as informal, while the rest is left out unlabeled.

For training, the source side is annotated with the politeness feature as described in Section 2.3.1. Note that there are only two values for the politeness feature, and neutral sentences are left unmarked. Their intuition is that in this way, the NMT model would ignore side constraints when they are irrelevant. To ensure that the NMT model does not overly rely on the side constraints, and that performance does not degrade when no side constraint is provided at test time, only a subset of the labelled training instances is marked with a politeness feature at training time. They set the probability that a labelled training instance is marked, α, to 0.5 in their experiments. To ensure that the NMT

model learns to ignore side constraints when they are irrelevant, and does not overproduce address pronouns when side constraints are active, they also mark neutral sentences with a random politeness feature with probability α. They claim that keeping the mark-up probability α constant for all sentences in the training set prevents the introduction of unwanted biases. They also re-mark the training set for each epoch of training. The model is trained for approximately 9 epochs.

At test time, they test translation without side constraints and translations that are constrained to be polite or informal, and take that as a baseline. In another oracle experiment, they use the politeness label of the reference to determine the side constraint, which simulates a setting in which a user controls the desired politeness. In that case, BLEU is strongly affected by the choice in politeness: the oracle experiment showed an improvement of 3.2 BLEU over the baseline (20.7 →23.9). They also analyze those sentences with a strongly marked politeness (such as the sentence *You foolish boy*) and find out that the NMT is translated with the informal pronoun, regardless of the side constraint.

They conclude that MT should not only produce semantically accurate translations, but should also consider pragmatic aspects, such as socially appropriate forms of address. And, although their paper focuses on controlling politeness, they claim that side constraints could be applied to a wide range of phenomena such as tense, gender or number.

2.3.3 MT evaluation

2.3.3.1 Human-quality of MT systems. Are we nearly there yet?

With the improvement of MT quality, there has been an increasing interest in comparing MT output with human translations (HT) and there has been research claiming that the quality of MT (specifically of NMT) has achieved human parity. For instance, on their paper, Wu et al. (2016) claim that their NMT system approaches the accuracy achieved by average bilingual human translators (on some test sets) and Hassan et al. (2018) affirm that the translation quality is at human parity when compared to professional human translators.

[1] github.com/sebastien-j/LV_groundhog

However, on their research (Läubli et al., 2018) test Hassan's claim following alternative evaluation protocols, contrasting the evaluation of single sentences and entire documents. Their experiments show that raters presented a stronger preference for human translations when evaluating at the level of documents, as compared to an evaluation of single, isolated sentences.

More claims to Hassan et al.'s research were made by Toral et al. (2018) on their study where they reassessed the claim that MT has reached human parity by considering more variables that were not taken into account in the original study and by comparing the judgements of professional translators against those of non-experts. One of their findings is that, between professional translators, there is a higher inter-annotator agreement between the experts' annotations and a better discrimination between human and machine translation output.

Other interesting aspects of language beyond grammatical accuracy were analyzed by Vanmassenhove et al. (2020). On their research, they analyze the linguistic richness (on a lexical and morphological level) of translations created by different data-driven MT paradigms (SMT and NMT) for two language pairs (EN←→FR and EN←→ES). By using traditional metrics from L2 learning such as MTLD (McCarthy, 2005), Yule's I (Yule, 2014) and TTR (Oakes & Ji, 2012) for calculating lexical richness, and others such as Shannon Entropy (Shannon, 1948) and Simpson's Diversity Index (Simpson, 1949) for grammatical diversity/ On this study, they draw three conclusions: (i) all the metrics indicate that the original training data has more lexical and morphological diversity compared to translations produced by the MT systems; (ii) that there is a strong indication that Transformer models outperform Statistical Machine Translation (SMT) systems in terms of lexical and morphological richness and (iii) that the MT systems have a stronger negative impact (in terms of diversity and richness) on the morphologically richer languages.

Finally, in another recent study, Bizzoni et al. (2020) analyze translationese patterns in translation, interpreting, and machine translation outputs. Their research leads to the conclusion that machine translation shows traces of translationese, but does not reproduce the patterns found in human translation. They, however, emphasize that while they find evident differences between translationese of human and machine translations, their results are complex to analyze. They also highlight the fact that there are still many independent patterns from machine translationese that are still unknown to the

community and that a further understanding of them could help improving machine translation implementation.

All in all, although there has been an increasing number of claims about the quality of MT and its parity with HT, we can remark that there are reasons to believe that this parity is very specific and limited to the type of evaluation that has been traditionally carried out by the community. For instance: whether the evaluation was carried out at sentence or document-level or whether experts or non-experts were used for evaluation. Moreover, as the aforementioned studies suggest, MT tends to produce less rich language (measured by lexical richness) and there are still many characteristics of the machine translationese that are still unknown.

Therefore, the claim that MT has achieved HT quality should not be taken for granted and more research should be carried out on this matter. In this way, we will be able to know what NMT systems are still lacking and how we can tackle these missing milestones.

2.3.3.2 Considerations for carrying out a sound MT evaluation

The importance of the evaluation phase for claiming that MT has reached human quality has already been introduced in the previous section. We also introduced how those claims where dubious due to a number of factors.

Marie, Fujita and Rubino (2021) introduce some other practices that might lead to dubious conclusions in MT evaluation and present them in their paper which represented the first large-scale meta-evaluation of machine translation (MT). In this study, they annotated MT evaluations conducted in 769 research papers published from 2010 to 2020 and arrive at the conclusion that practices for automatic MT evaluation have dramatically changed during the past decade and follow concerning trends:

- An increasing number of MT evaluations exclusively rely on differences between BLEU scores to draw conclusions, without performing any kind of statistical significance testing nor human evaluation, while at least 108 metrics claiming to be better than BLEU have been proposed.
- MT evaluations in recent papers tend to copy and compare automatic metric scores from previous work to claim the superiority of a method or an algorithm without confirming neither exactly the same training, validating, and testing data have been used nor the metric scores are comparable.

- Furthermore, tools for reporting standardize metric scores are still far from being widely adopted by the MT community.

They highlight how, even if most of these pitfalls are well-known by the MT community, most of MT publications that they analyzed were affected by at least one of these. That is why, they propose a set of guidelines to be followed in future studies:

- An MT evaluation may not exclusively rely on BLEU, other automatic metrics that better correlate with human judgements may be used in addition or in lieu of BLUE.
- Statistical significance testing may be performed on automatic metric scores to ensure that the difference between two scores, whatever its amplitude, is not coincidental.
- Automatic metric scores copied from previous work may not be compared. If inevitable, copied scores may only be compared with scores computed in exactly the same way, through tools guaranteeing this comparability, while providing all the necessary information to reproduce them.
- Comparisons between MT systems through their metric scores may be performed to demonstrate the superiority of a method or an algorithm only if the systems have been trained, validated, and tested with exactly the same pre-processed data, unless the proposed method or algorithm is indeed dependent on a particular dataset or pre-processing.

That is why, when it comes to the step of evaluating MT models, researchers should try not to fall into any of the pitfalls that Marie, Fujita and Rubino remark and try to follow their guidelines in order to arrive to a sound and trust-worthy conclusion.

2.4 RATIONALE AND RESEARCH QUESTIONS

Having set up the basic concepts for our study, we will try to answer the following questions:

- Is it possible to control politeness in the output of EN>ESES NMT systems (being politeness characterized by the use of *tú* and *usted*)?
- If so, dealing with the task as a domain adaptation problem, which is the best approach to use?
- What is the impact that politeness features have in the evaluation of NMT systems?

Moreover, politeness has not been a common research topic in this area, especially when it comes to the language pair at hand. Therefore, we believe that our study could be of interest for the MT community, since not only does it offer a way of classifying Castilian Spanish text with respect to the use of politeness using traditional NLP techniques, but also compares two approaches using different parameters and training techniques, which have not previously been compared for this task and this language pair.

3 METHODOLOGY

In this section, we go over the methods that were used for the research: Firstly, we focus on the data-set used for training and explain the reasons behind this choice, the methodology used for dividing the data-set, carrying out an analysis of the resulting subset and doing some final cleaning; secondly we describe the NMT systems that were trained and the methods, and finally, we dwell over the steps that we followed for the evaluation as well as the tools that we used.

3.1 DATA-SET

3.1.1 Selecting the data-set

The first step of the research was to select which data-set was going to be used throughout the project. As it was already mentioned in the Section 2.3.1, there are no resources with differentiation in politeness for Spanish available online. Therefore, we decided to select one of the multiple available open-source corpora and create a pipeline to divide its content according to the occurrence of more or less polite forms. We based our classification in the two registers that were characterized in section 2.1.5, where we explained that Castilian Spanish tends to use *tú* for situations where there is some closeness between interlocutors or some relationship of familiarity (what we denoted as *informal register*); meanwhile *usted* tends to be used for situations where the speaker wants to show respect to the other person, because there is some kind of distance between them and the addressee (what we denoted as *formal register*).

We wanted our engines to be the most consistent possible when it comes to register, i.e. to comply with the characteristics of the domain. While there is going to be changes in the politeness (informal vs. formal), the rest of characteristics of the text outputted by the engines, such as the field or the mode shall be given by the characteristics of the corpus at hand. Therefore, we had to take into account the final aim of the engines in order to choose the appropriate corpus for our task.

According to an article from the Journal 'Clarin' (2011), a normal person uses around 500 to 1000 terms in their daily lives, meanwhile youngsters use around a 25% (a bit more than 240 words). This can be crucial information when choosing our training corpus. Our engine was going to be used for IM; therefore, even if the speaker wants to

express politeness by using *usted*, we could expect the field not to be too technical and sophisticated.

We also needed to take into account the characteristics of instant messaging speech: it is written, but presents characteristics that are closer to an oral mode, which is really similar to the prefabricated orality of audiovisual content (see Section 2.2.1).

Therefore, taking into account the lower sophistication of the corpus and the oral characteristics of the final use of the future engines, we considered the OpenSubtitles corpus (Tiedemann, 2012) to be the most suitable corpus for our task at hand.

OpenSubtitles is a collection of parallel corpora extracted from *Opensubtitles.org* which contains a collection of user contributed subtitles in various languages for movies and TV programs. The release includes a total of 1689 bi-text spanning 2.6 billion sentences across 60 languages. The main drawback from this data-set is that due to scraping, it might contain segments with spelling mistakes, punctuation mistakes and parallel segments that are misaligned. Moreover, corpora are not divided into their different diatopic varieties, therefore the English←→Spanish corpus might contain instances from several dialects of the Spanish language. We hypothesized that once we classified the corpus according to forms of politeness, most of the instances containing *tú* forms would contain mostly Castilian Spanish, since it is one of the only Spanish-speaking countries where this form is broadly used. However, some problems regarding the dialect being used could arise with segments containing *usted* forms, since this form tends to be used in many dialects of the language much more frequently. Nevertheless, we believed that the advantages of this corpus over others which contained more formal and technical terminology such as Europarl (Koehn, 2005) (which contained only Castilian Spanish) outweighed this disadvantage.

3.1.2 Dividing the data-set into subsets by politeness

Given that Castilian Spanish is a language that marks politeness by the use of different honorifics and pronouns (mainly), we decided to use a rule-based approach rather than a model-based one (training a classifier) in order to split the dataset according to the use politeness, since the identification of politeness can be carried out more easily by using exact-matching and parsing. For that, we followed the two-step approach from Sennrich et al.'s work (2016): (1) Searching for lexical forms in the whole corpus, and (2) searching for grammatical forms in a subsection of the corpora with none of the lexical

forms searched on the previous step to search for those segments where the subject was elided.

In the first step, we searched for all appearances of any lexical form that belonged to the paradigm of *tú* or *usted*. This can be done using a simple case-insensitive regex and Python[2]. All the forms are presented in Table 3.

INFORMAL LEXICAL FORMS	FORMAL LEXICAL FORMS
tú, tu, tus, contigo, tuyo, tuyos, tuya, tuyas, ti, te, vosotros, vosotras, vuestro, vuestros, vuestras, vuestros	usted, ustedes, le, les, su, sus, se, suyo, suyos, suya, suyas

Table 3: Lexical forms in Castilian Spanish for the different levels of politeness

As it can be seen, we made no distinction between plural and singular forms, which is something to be analyzed in further research, since we wanted to start by focusing only in politeness distinction. In this way, having tested if it is possible to control the politeness of the output, we could move on to explore the same approach with other linguistic aspects such as number or gender in future research.

This lexical search could be enough for some non-pro-drop languages (see Section 2.2.2 on the notion of pro-drop languages). However, this is not the case for Spanish, which tends to omit the subject in many occasions.

Therefore, if we reduced our division to the appearance of these lexical forms, the language produced by the engine would sound quite unnatural, since it might over-generate the subject when it actually should not. In order to avoid this problem, we divided the rest of sentences with no explicit honorifics by parsing the verbs. For that, we decided to make use of Stanza (Qi et al., 2020), which is a natural language analysis package which offers pre-trained neural models in 70 languages that can be used for text analytics. However, we noticed that this package was behaving quite strangely with formal forms in Spanish and was not identifying some of them, so we filtered those segments with Spacy[3], which is another natural language processing toolkit which offers pre-trained language models for parsing text.

[2] https:/www.python.org

[3] https://spacy.io

In that sense, the *informal you* (*tú*) in Spanish is conjugated using the second person, meanwhile the *formal you* (*usted*) is conjugated using the third person. Therefore, if the Spanish sentence contained a verb conjugated in the second person, we classified it as informal; if the Spanish sentence contained a verb conjugated in the third person, we classified it as formal, and if there was no verb or there was a verb but it was conjugated using a different person from the other two, we classified it as neutral.

However, there was a main issue with this approach. We could be certain that all second person verb forms were indeed forms of *tú*; but third person verb forms in Spanish are ambiguous and can belong to either *tú* or to third person pronouns such as *él*, *ella*, *ellos* or *ellas*. The same happens with lexical forms such as *su*, *suyo*, *suya*, etc. (see Table 4)

SOURCE	TRANSLATION
Su *casa es muy bonita.*	**Your/His/Her/Their** house is wonderful.
Le gusta *la cerveza.*	**He/She** likes beer.
*¿**Le gusta** la cerveza?*	Do **you** like beer?/ Does **he/she** like beer?

Table 4: Ambiguous use of pronouns and possessives in Spanish

To this regard, Stanza does have a parser for polite and impolite forms, which labels terms according to their politeness degree. However, it is not useful when trying to disambiguate between pronouns, possessives or verbs. Therefore, we carried out disambiguation in or own, given that by looking at a Spanish sentence in a parallel corpus, one of the ways of telling whether a verb in third person is indeed referring to a third person or is used as a polite form, is by searching for *you*, *your* or *yours* in the original English segments (Sennrich et al., 2016). Although it is not a perfect approach, we believed that it would filter out most of the segments that did not have any second person forms and classify them as neutral.

Having followed this two-step approach for diving OpenSubtitles (not in its entirety) we ended up with the following number of segments in each subset (see Table 5), which we believed were a fair amount for training different NMT engines.

FORMAL	INFORMAL	NEUTRAL
2,553,392 segments	5,658,102 segments	4,536,955 segments

Table 5: Number of segments after classification

3.1.3 Analysis of the resulting subsets

Given that we did not have the correct labels for each segment (i.e., our task was unsupervised), we tested our classifying approach by extracting 100 segments for each register (50 segments extracted using regex and 50 segments extracted using parsing) and manually checked whether the segments matched the intended class. The pie charts containing the percentage of correctly classified segments for each level of politeness can be found in Figure 4.

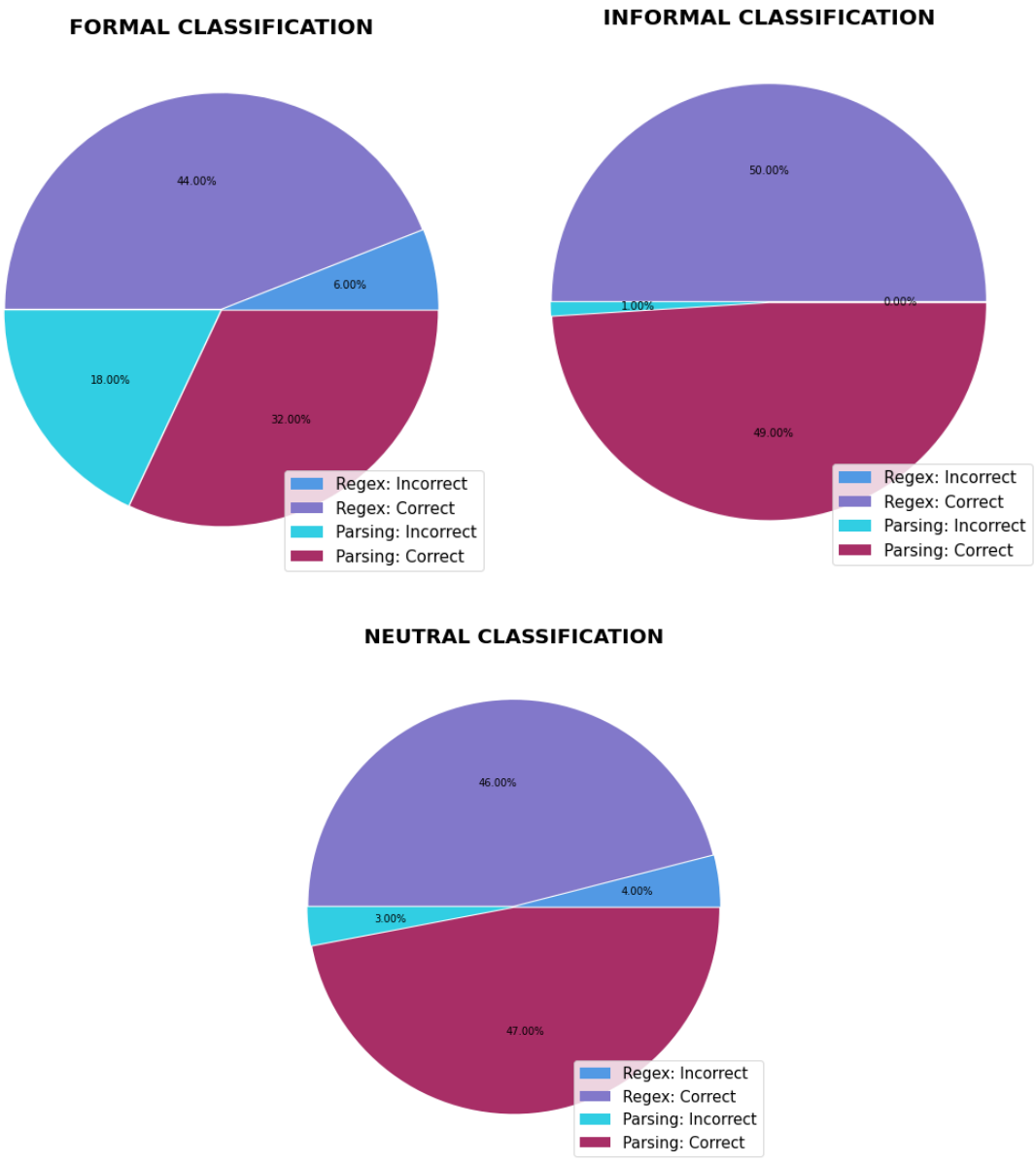

Figure 4: Percentage of correctly and incorrectly classified segments per subset

In Tables 25 to 30 in the Appendix section, we also include some examples of correctly and incorrectly assigned segments.

As it can be seen in Figure 4, the subset of informal segments containing lexical forms was completely filtered, and all the segments from the extracted subset presented the intended register. In the case of informal segments containing extracted by parsing, 1% of the total should have been given a different label. In the case of the neutral classification, 7% of the segments were misclassified: 4% of them extracted using regex and 3% using parsing. However, the subset that achieved worst results was the formal subset. A total of 24% of the segments were given the wrong label, which amounted to a 6% of segments extracted with regex and a 18% of segments extracted with parsing. These results show that according to the subset of segments that we analyzed, almost a quarter of the formal data contained false positives (i.e., segments which were either informal or had no verb or no actual forms *you* in the source).

When carrying out a rather qualitative analysis of the results in the formal subset, we found that the number of segments where there appeared a verb in the second person form (which should be labelled as *informal*) amounted to a 9% of the total (5% segments extracted with regex and 4% of the segments extracted with parsing). We suspected that, to a great extent these misplacements might have been due to incorrect parsing of Stanza or Spacy. The rest of incorrect segments (15% of the total) seemed to be due to the fact that there was a *you* in the original English sentence but it was actually not related to the grammatical form that appeared in the Spanish sentence. An example of this type of segments would be the following:

- SOURCE: I think **you** got it cheap.
- TARGET: *Creo que es barato*. (‘I believe this is cheap’)

In this example, while there was a *you* in the source sentence and a verb conjugated in a third person form in the target sentence (*es*), the Spanish translation had no forms of address. Therefore, this segment should have been given the label *neutral*.

To avoid these false positives, a more complex approach such as word-to-word alignments could have been applied. However, we assumed that the amount of segments that were filtered into the formal corpus and that should have actually belonged to the informal or neutral subset should not affect the quality of the final engine to a great extent. What is more, we also believed that this difference in the strictness of filtering between the different subsets could be interesting in the next steps of the research for

comparing the engines trained with the informal subset to the engines trained with the formal subset.

We also found other problems in the informal and neutral subsets, where some false positives seemed to be due to misalignments or segments of suspicious quality. An example of this would be the following parallel segment which made it to the neutral corpus:

- SOURCE: THAT'S A COP KILLER.
- TARGET: *¡Quítense!* ('Go away!')

In this case, when dealing with misalignments of the source and target segments, there are different options. According to Khayrallah and Koehn (2018) the quality of the data we use for training an NMT system is extremely important for the system's performance. To avoid this, Bane and Zaretskaya (2021) proposed different methods for filtering out segments of dubious quality before training. They analyzed four different scoring methods to approach this task: (i) marian-scorer[4] (Junczys-Dowmunt et al., 2018) (ii) LASER[5] (Schwenk, H. & Douze, M., 2017), (iii) MUSE[6] (Conneau et al., 2017) and (iv) XLM-R[7] (Conneau et al., 2019). Their study concluded that marian-scorer and MUSE were the approaches with a higher correlation with human annotators and produced the best results when using their filtered data to train an NMT engine. Being marian-scorer faster and less computationally costly, we decide to use this tool for filtering our data.

3.1.4 Filtering out bad-quality segments

In order to use marian-scorer, we would need to use an already-existing NMT system at hand to generate the scores. For this, we made use of the open-source NMT model available from Helsinki-NLP for EN>ES[8] (BLEU 54.9 on the Tatoeba-test.) (Tiedemann, 2020).

The marian toolkit is a free Neural Machine Translation framework written in pure C++. It is mainly being developed by the Microsoft Translator team and provides fast

[4] https://marian-nmt.github.io

[5] https://github.com/yannvgn/laserembeddings

[6] https://github.com/facebookresearch/MUSE

[7] https://huggingface.co/xlm-roberta-base

multi-GPU training and GPU/CPU translation using NMT architectures such as deep RNNs and Transformers. Marian offers its own scorer which calculates negative log likelihood of a segment with respect to a model, i.e. the probability that the translation was done by such model. We made use of this scorer for scoring the segments at hand and set a threshold score. If segments were beneath that score, such segment is less likely to have been produced by the NMT system, so we could assume that it is of less quality; while if segments were above that score, such segments were more likely to have been produced by the NMT system, and therefore, could be considered to have a more desirable quality.

Choosing a threshold for filtering can be a rather complicated task, since the score is language-dependent and model-dependent, which means that segments resembling the original training-data get higher scores. In that sense, Tiedemann's models were trained using available corpora from OPUS[9], however the author does not specify which particular corpus was used, so we could assume that several domains were present in the training data of such models. Moreover, we did not want to reduce the number of our segments to a great extent, because the original number of segments was a bit scarce. However, to the best of our knowledge, there were no guidelines available so as to which was a good threshold score to be used for filtering out data using marian-scorer. Therefore, we did some experiments and chose a threshold that, while not too tight, it would leave us with a fair amount of segments to train our engines, while getting rid of really dubious segments. We chose -6.5 as our filtering threshold which preserved around an 80% of our corpus. The final number of segments after filtering can be found in Table 6.

FORMAL	INFORMAL	NEUTRAL
1,821,381 segments	4,453,708 segments	3,670,602 segments

Table 6: Number of segments for each subset after filtering

3.1.5 Word clouds of the subsets

Having filtered our segments, we decided to make us of word clouds for visualizing the most common words appearing in each subset of the original corpus by using the

[8] https://huggingface.co/Helsinki-NLP/opus-mt-en-es

[9] http://opus.nlpl.eu

WordCloud[10] package from Python. We removed the stop-words using NLTK[11] and transform them into lower case. The world cloud for each division of the corpus can be found in Figures 7 to 9 in the Appendices section. In these figures, we can visualize the type of terms that are in abundance in each subset, being *usted* the most common term in the formal subset; *si* (if), *ti* and *tú* among the most common the informal subset, and *bien* (well), *si* (if) or *asi* (like this) the most common in the *neutral* subset. These word clouds provide a visual evidence of the successful divison of the data-set using our approach.

3.2 NMT SYSTEMS

We now present the different NMT systems that we decided to train and explore for our research. Two approaches towards politeness control in MT were implemented: A fine-tuning approach (FINE) and a multi-domain –or multi-politeness– approach (MULTI). We also implemented these two approaches in two different ways resulting in a total of four approaches.

For training, we made us of the Fairseq toolkit[12] (Ott et al., 2019), since on the Findings of WMT 2021 report (Akhbardeh et al., 2021), Fairseq appeared to be the most used framework (6 times) followed by Marian and OpenNMT[13] (Klein et al., 2017) (3 times each).

For carrying out tokenization and Byte-per-encoding (BPE) segmentation, we made use of Moses[14] and Subword-NMT[15] (Sennrich et al., 2015).

3.2.1 Fine-tuning approach

For the fine-tuning approaches, we first trained a baseline model using 3 million segments containing a balanced mix of segments from the formal, the informal and the neutral subsets. The total number of segments used for training, validation and test can be found in Table 7.

[10] https://amueller.github.io/word_cloud/

[11] https://www.nltk.org

[12] https://github.com/pytorch/fairseq

[13] OpenNMT - Open-Source Neural Machine Translation

[14] https://github.com/moses-smt/mosesdecoder

[15] https://github.com/rsennrich/subword-nmt

TRAINING SIZE	VALIDATION SIZE	TEST SIZE
2,996,000	2,000	2,000

Table 7: Size of training, validation and test-sets for the baseline engines

We trained a joint BPE vocabulary with size 32,000 and applied it to the sets, and when applying binarization with Fairseq, we created separate dictionaries for the source and the target language. Then, we trained a system based on the Transformer architecture (Vaswani et al., 2017) using Adam as an optimizer, a learning rate of 5e-4, dropout of 0.3, label-smoothing of 0.1 and trained it for 50 epochs. While the first engine was trained with an early-stopping of 10 validation runs (FINE_loose), the second one was trained with a tighter early-stopping of 5 validation runs (FINE_strict).

Then, we carried out the fine-tuning of each baseline to each degree of politeness. For that, 700,000 segments were extracted from the informal and formal subsets and were used to fine-tune the baseline engine using the last epoch of each training. We split the data into training, validation and test and ended up with the following number of segments (see Table 8):

TRAINING SIZE	VALIDATION SIZE	TEST SIZE
696,000	2,000	2,000

Table 8: Size of training, validation and test-sets for the fine-tuned engines

We re-used the BPE code from the baseline engines but, as pointed out by Subword-NMT best practices[16], the vocabulary for each engine was extracted and passed when applying the BPE with a vocabulary threshold of 50 so that the script only produces symbols which also appeared in the vocabulary above this frequency. According to these authors, for languages that share an alphabet, learning BPE on the concatenation of the involved languages increases the consistency of segmentation, and reduces the problem of inserting/deleting characters when copying/transliterating names. Moreover, applying a vocabulary prevents words from being segmented in a way that was seen only in the other language.

The engines were fine-tuned using the same parameters from the baseline and for 10 epochs. For the FINE_loose engine, we applied no early-stopping, while for the

FINE_strict engine we applied an early-stopping of 2. This difference was meant for analyzing whether training with a looser early-stopping value would help the final performance of the fine-tuned engines.

3.2.2 Multi-register approach

For the multi-register approach, we trained two engines. For this approach, we wanted to train engines that could handle three directions, which we denote as English → Informal Spanish, English → Formal Spanish and English→Neutral Spanish. For the first approach, we trained the multilingual system treating each register as if they were completely different languages, and used their respective subsets for training. We denote the third register as *neutral* as a way of indicating that the neutral subset was the one used for training that direction. Moreover, the second approach followed Sennrich et al.'s approach (2016), where a number of segments from the neutral subsets were added to each of the informal and formal corpora and vice-versa to avoid bias (MULTI_Sen). We believed that by training both systems we could get an idea of the impact that adding data from the different subsets could have in these multi-register models.

On their end, Sennrich et al. prepend a token to each segment to signal the level of politeness that appears on the target. However, for our research, we made use of Fairseq's implementation for training a multilingual system, which handles this procedure internally.

3.2.2.1 Training of MULTI_Own

For the MULTI_Own engine, we extracted 1,5 million segments from each of the subsets (formal, informal and neutral) amounting to a total of 4,5 million segments. We split each group of data into train, validation and test and ended up with the following number of segments per each direction (see Table 9):

TRAINING SIZE	VALIDATION SIZE	TEST SIZE
1,498,600	700	700

Table 9: Size of training, validation and test-sets for the multi-politeness approaches

[16] https://github.com/rsennrich/subword-nmt

We trained a joint BPE vocabulary of size 32,000, and following the same procedure as in the fine-tuned engines, we passed a vocabulary filter when applying the BPE code. The English vocabulary was trained using the English data from all three training-data, while the other three were extracted from each particular subset (i.e., a vocabulary for the formal subset, a vocabulary for the informal one and a vocabulary for the neutral one). During binarization, we created a common dictionary for English, and separate dictionaries for each of the registers. Finally, for training, we made use of the Transformer architecture for multilingual translation from Fairseq and used shared encoder-embeddings as well as the following parameters: Adam as an optimizer, learning rate of 5e-4, label-smoothing of 0.1 and dropout of 0.3. The model was trained for 50 epochs with early stopping of 5 validation runs.

3.2.2.2 Training of MULTI_Sen

For the second engine (MULTI_Sen), reusing the training sets from the MULTI_Own engine, we combined the different subset, i.e., a number of sentences from the neutral subset were added to both the informal and the formal training data and vice-versa. In this way, we tried to imitate the way in which Sennrich et al. set the probability of an instance being marked to 0.5. They claimed that in this way, biases could be reduced.

In Table 10 we present the number of segments from each classification that we used for training each direction of the engine.

Language direction	English → Informal ES	English → Formal ES	English → Neutral ES
Number of segments from each subset	750,000 informal segments 325,000 neutral segments	1,000,000 formal segments 75,000 neutral segments	750,000 informal segments 750,000 formal segments 750,000 neutral segments
TOTAL	1,075,000 segments	1,075,000 segments	2,250,000 segments

Table 10: Number of segments from each subset included in each direction of the MULTI_Sen system

For the English → Formal Spanish direction, given that the formal subset was not completely filtered (contained around a 1/4 of false positives (see Section 3.1.3)), we decided to add a higher amount of segments coming from that subset, and less segments from the neutral classification. In that way, if according to the tests that we carried out, the wrongly classified segments amounted to a quarter of the 1M formal segments, then, around 250,000 segments out of the 1M should have been classified as informal or neutral. Therefore, we could just add 75,000 neutral segments more and end up with roughly the same amount of non-formal segments as in the En→Informal Es direction (325,0000).

For training, we followed the same procedure as the MULTI_Own engine and trained a joint BPE code, applied it passing a vocabulary filter, created separated dictionaries for English and the respective directions, and finally, trained the engines using the same parameters and the same model.

A summary of the different characteristics of each engine that were used for training can be found in Table 11 and Table 12.

	FINE_loose (baseline)	FINE_loose (fine-tuned engines)	FINE_strict (baseline)	FINE_loose (fine-tuned engines)
early-stopping	10	Not applied	5	2
Architecture	transformer_wmt_en_de			

Table 11: Summary of the characteristics of the FINE systems

	MULTI_Sen	MULTI_Own
Number of segments per direction:	**1,075,000 EN> Informal ES:** 750,000 segments from informal subset + 350,000 segments from the neutral subset **1,075,00 EN> Formal ES:** 1,000,000 segments from formal subset + 75,000 from neutral subset **2,250,000 EN > Neutral ES:** 750,000 segments of each subset	1,500,000 per direction
early-stopping	5	
Architecture	multilingual_transformer_iwslt_de_en	

Table 12: Summary of the characteristics of the MULTI systems

3.3 TOOLS FOR ANALYSIS

Having trained the models, the next step involved was to test how well they performed. In this section, we present the methods used for the evaluation of the engines, which can be essentially divided into two steps: an automatic and a human evaluation.

When generating the translations that were used for testing, we used the last checkpoint from each engine, as well as a beam search of 5 and a batch size of 128.

3.3.1 Automatic evaluation

For the automatic evaluation, we made use of MT-Telescope[17], which is a toolkit for comparative analysis of MT systems that provides a number of tools that add rigor and depth to MT evaluation. It gives easy access to MT evaluation metrics such as COMET (Rei et al., 2020), sacreBLEU (Post, 2018), Prism (Thompson & Post, 2020) or chr-F (Popović, 2015) as well as statistical tests with bootstrap resampling (Koehn et al., 2004).

Our automatic evaluation was two-fold: we first tested the engines over their specific test-sets, and then over a common test-set to all the engines. The reason behind this distinction was that, while the specific test-sets would give us an overall idea of how well each of the systems was performing on their specific tests-sets, they were not suitable for comparing the engines with each other, since these test-sets contained sentences extracted from the same distribution of each engine (i.e., from the same subset). Therefore, we believed that creating a common test-set and using it for the evaluation of each engine would lead to much more sound conclusions.

However, the task of creating a common test-set was not simple, since we needed to find a balance between every distribution. For that, we extracted 200 segments from each of the following specific test-sets: 600 segments from the FINE corpora (200 from the informal, the formal and the baseline test-sets respectively), and 600 from the MULTI corpora (200 from the informal, the formal and the neutral register respectively). Therefore, the final test-set contained 1,200 segments. We believed that in this way, all forms of politeness would be represented as well as sentences not containing verbs and forms of address, and therefore, all engines could be tested in equal conditions.

For each test-set the system-level sacreBLEU, chr-F and COMETINHO (light-weight version of COMET) scores were extracted using the reference segments. Moreover, when carrying out the evaluation on the common set, t-tests were performed by using bootstrap resampling with default parameters (re-samples of 0.5 and 300 iterations).

MT Telescope also provides an error comparison tool, therefore, we included these analyses in our research, since they could provide more details about the performance of each engine.

3.3.2 Human evaluation

Given that automatic metrics are dependent on the reference segments and their original quality, we decided to perform a human evaluation of each of the engines. This human evaluation had three steps: a general-quality assessment where 30 evaluators contributed to scoring segments in terms of general-quality, a register-specific assessment where we wanted to calculate the accuracy of each system for producing the desired register (*politeness test*), and finally a succinct evaluation of how each engine behaved with regards to segments with a clearly marked register (*opposite test*).

3.3.2.1 General-quality assessment

The first step of the human evaluation had the following goals:

- Gaining an overall estimation of the quality of each system without taking into account the politeness constraints.
- Assessing the quality of each system with regards to different types of segments.
- Getting an overall idea of the impact that politeness features had in the task of evaluating NMT output.

For this task, we created a specific test-set which we denoted *Ling_test*, and which tried to cover different linguistic aspects of the Spanish language. The Ling_test consisted of 50 segments with the following types of instances (see Table 13):

[17] https://github.com/Unbabel/MT-Telescope

LINGUISTIC ASPECT COVERED	NUMBER OF SEGMENTS
Segments which contain the pronoun *you* as subject in the source (You_EN)	3
Segments where the pronoun *you* (*tú* and *usted*) should not be elided in the Spanish translation (You_ES)	5
Segments containing possessives of second person: *you, yours* (Possessives)	9
Segments containing imperatives (Imperatives)	6
Segments containing the construction preposition + personal pronoun *you* either in English or in the Spanish translation (Pronouns)	7
TOTAL NUMBER OF SEGMENTS SECOND PERSON FORMS (2PERSON)	**30**
Phrases containing no verb (No_verb)	10
Phrases containing no second person forms in the source (No_you)	10
TOTAL NUMBER OF SEGMENTS WITH NO SECOND PERSON FORMS AND NO VERBS (NO_FORMS)	**20**

Table 13: Number of segments from each linguistic phenomenon contained in the Ling_test

This test aimed at giving a coverage of the kind of segments that the engine would be seeing in a real-life scenario, being those produced by an individual about daily-live topics; or by a company communicating with their clients using terms such as *item, address, service or contact*. In this test-set, we divided the segments into two main categories: those with second person forms in the source (denoted as 2PERSON), and those with no second person forms or no verbs in the source (denoted as NO_FORMS). Among the first type of segments, we tried to represent different grammatical and syntactic phenomena with respect to the use of *tú* and *usted* forms such as the use of *you* as a subject, as a possessive or after a preposition. This is intended to cover the different forms that *tú* and *usted* can take in Spanish depending on their grammatical category (*tu, tuyo, ti, te,* etc.). On the other hand, among the NO_FORMS segments, phrases which contained no verb or which presented other grammatical persons different that

you (e.g. *I, he, she, we* and *they*) were included. We considered it was worth paying attention to this kind of segments in order to observe how each engine was behaving with their respective in-domain and out-of-domain sentences (i.e. whether those engines trained with only this kind of sentences would perform better than those which had not seen any of these segments and those which had seen only a small amount).

For further reference, we include the Ling_test in Table 31 in the Appendices Section.

Moreover, for obtaining objective results, we could not handle the task of evaluating the overall quality of our engines by ourselves. Therefore, in order to carry out the first-step of the human evaluation, we got in contact with volunteers with different backgrounds: linguistics, translation, philology or computer sciences. What all of them had in common was that they were fluent or at least native-like in Spanish and had some background in Natural Language Processing. A total of 30 people carried out the evaluation. The only information that the evaluators were given before the task was that they needed to evaluate some segments generated using MT. However, they were not aware of the fact that we were studying politeness control in MT and the fact that they were evaluating different models. We decided not to give them this information ahead in order to avoid biases when evaluating the segments and observe if the different use of politeness in the same test-set had an impact on their decisions.

In addition to that, they were given instructions to score each segment focusing on both adequacy and fluency using a scale from 1 to 5. The instructions that the annotators received can be found in Table 14.

Please evaluate the segments from the next page focusing on these two aspects:	
ADEQUACY: How much of the meaning is preserved?	**FLUENCY: Is the language in the output fluent?**
5: all meaning	5: flawless
4: most meaning	4: good
3: some meaning	3: non-native
2: little meaning	2: disfluent
1: none	1: incomprehensible

Table 14: Instructions given to linguists to score each segment in the evaluation

We also asked them to try to avoid using number 3 when it was not strictly necessary, in order to achieve more conclusive results. Moreover, we added a section for comments on the file, so that evaluators could provide any insights about the evaluation task.

We translated the Ling_test using each of the 12 engines (the three engines from each of the four approaches) and this resulted in a total of 600 translations.

Since we wanted to obtain a robust result with the evaluation, we decided to create 12 test-sets containing a mix of the translations generated by each engine. In this way, we ended up with 12 completely different test-sets containing segments generated by each engine. Moreover, being a quite subjective evaluation, we decided to have three evaluators per test-set. However, due to a lack of evaluators, six of them evaluated two test-sets instead of just one.

For obtaining the final human score, we averaged the scores for each segment given by each of the three evaluators and obtained a segment-score. We then averaged all the segment-scores for each model to obtain a final system-level score.

Moreover, in order to get an idea of the inter-annotator agreement, we calculated the Fleiss' Kappa for each test-set and averaged it to obtain the overall inter-annotator agreement for this research.

Finally, we also took a more detailed look into the scores given to the different types of segments in order to see if the performance of each engine degraded or improved with their respective in-domain and out-of-domain segments.

3.3.2.2 Register-specific assessment

With the previous test, we wanted to get some insights on the overall quality of each system, focusing as well on their overall quality for different types of segments. We then decided to carry out another test for calculating the accuracy of each model when producing their intended register and for observing whether the engines were over-producing honorifics in sentences with no second person forms, while trying to leave fluency and adequacy to one side. For this test, we made use of the same Ling_test scored by the evaluators and set a labelling approach for evaluation.

In the 2PERSON segments (segments with no second person forms), we used the following labelling:

- INFORMAL: Use of informal forms.
- FORMAL: Use of formal forms.
- ELLIPSIS: Segments where second person forms were erased or neutralized. An example of this could be:
    - SOURCE: Who did it? Was it you?
    - MT: *¿Quién lo hizo?*
- MIX: Mix of more than one form of politeness (both informal and formal). An example of this can be:
    - SOURCE: Can you check your agenda and let me know when you are free?
    - MT: *¿Podrás comprobar su agenda y hacerme saber cuándo serás libre?* (Can you (informal) check your (informal) agenda and let me know when you are (informal) free?)
- OTHERS: Those segments where there might be additions, or incorrect use of verb forms or incorrect choice of persons which are not directly linked to the use of more or less polite forms, but to other linguistic aspects. We do not go into much depth with this type of segments, but by getting a number of how many of these are produced by each engine, we can get an idea of how many suspicious outputs are produced by each engine. An example of this type of segments would be the following:
    - SOURCE: Click on the item you wish to purchase
    - MT: *Scock on the item you desea comprar*.

In the NO_FORMS segments, we used the following labels:

- NO ADDITIONS: Although the sentence might be more or less adequate or fluent, there are no additions with relation to politeness. Example:
    - SOURCE: Hey, there!
    - MT: *¡Oye, oye!*
    - An example of addition in this sentence could be: *¡Oye, tú!*
- ADDITIONS: The engine produces an output with extra honorifics or referrals to the addressee that might or might not influence the adequacy or fluency of the sentence. These are normally referred to as *hallucinations*. Example:

  - SOURCE: They were supposed to come today
  - MT: *Se suponía que iban a venir hoy, ¿no crees?* (‘They were supposed to come today, don’t you think?’)
- OTHERS: Following the same idea as in the 2PERSON segments, we gave this label to segments which did not fit into any of the other categories: Use of a different person than the one indicated in the source or untranslated segments, among others. An example of this could be the following:
  - SOURCE: Let's go together
  - MT: *Andando juntos.*

With this test, we expected to discover how accurate the systems were with respect to producing their intended politeness and to what extent the engines were overproducing honorifics, (i.e., are somewhat biased) for segments with no *you* and no verb in the source. For reference, we denoted this test as *politeness test*.

3.3.2.3 Assessment of sentences with a clearly marked register

In a final step of the testing, we decided to create another test with segments containing a marked politeness, i.e. sentences that should normally be translated with a certain register no matter the situation, since they contain clearly formal (such as sentences with honorifics such as *Mr.*, *Miss*, *Sir*) and clearly colloquial language (such as sentences containing swear words). One may think that this lacks interest for our research, since what we want to achieve is control over the NMT output. However, it might be desirable in some situations to have a system that keeps the ability of producing the opposite register when the source sentence is clearly marked as being informal or informal.

Therefore, with this test-set we expected to test whether our engines completely lost the ability of using the opposite register that they were intended to generate in this kind of situations. We denoted this test as *opposite test*, and the segments that it contains can be found in Table 32 in the Appendices Section for reference.

4 FINDINGS

In this section, we will go through the results that we found during the evaluation stage, which consisted of the following steps: the automatic evaluation of each engine in their specific test-sets as well as in a common test-set, and the human evaluation, which covered an overall assessment of quality of the engines, an assessment of the accuracy of each engine with regards to the use of politeness (what we denote as *politeness test*), and finally, a small evaluation of sentences where politeness is clearly marked in the source (*opposite test*).

We analyze 12 systems: each of the two fine-tuned engines from the FINE systems and their respective baselines, and the three language directions contained in the MULTI_Own and MULTI_Sen engines. Even if the latter are not technically engines *per se* but different language directions contained in one engine (i.e. they are multi-register systems), at some point we will might refer to them by the term *engine, system* or *direction* interchangeably to avoid confusion when comparing them to the fine-tuned models. Therefore, we refer to each engine by the approach which was used for training (FINE_strict, FINE_loose, MULTI_Own and MULTI_Sen) followed by their register (*inf* for informal, *frm* for formal, *neutral* for the neutral engines of the MULTI systems, and finally, *baseline* for the baselines of the FINE systems).

4.1 AUTOMATIC METRICS

We start by analyzing the results from the automatic metrics. We first take a look at the results of each engine on their specific test-sets, which contained sentences extracted from the same distribution of the data in which they were trained, and then, analyze the results from the common-test, which contained a set of sentences coming from each distribution (neutral, formal and informal sentences). It is worth mentioning that for all the automatic metrics, scores were multiplied by 100 for readability reasons and that best scores for each metric are presented in bold and second best are underlined.

4.1.1 Specific test-sets

Table 15 presents the automatic metrics extracted for each engine on their specific test-sets.

By taking a look at these results, we can observe that almost every engine obtained sacreBLEU scores over 30 points and that there are only two engines which obtained a score of 40 or close to it: the MULTI_Own_inf engine and the FINE_strict_inf engine. On the other hand, the FINE_loose_baseline achieves the worst scores across all metrics.

Moreover, in general, the informal engines achieve the highest scores for each approach, while the formal engines tend to achieve better scores than their respective baselines and neutral models (except for the FINE_strict engine, where the baseline outperforms the formal engine).

These results show that, when it comes to their respective registers, the informal engines perform better than the formal, the neutral and the baseline systems, which might be a sign of overfitting to the training-data.

	SacreBLEU	COMETINHO	chr-F
FINE ENGINES			
FINE_strict_baseline	35.3	38.9	56.8
FINE_strict_inf	39.7	**47.5**	58.7
FINE_strict_frm	35.0	37.3	56.9
FINE_loose_baseline	29.5	12.9	49.6
FINE_loose_inf	39.4	45.6	58.4
FINE_loose_frm	35.4	36.8	57.3
MULTI ENGINES			
MULTI_Own_neutral	36.8	38.6	58.0
MULTI_Own_inf	**40.3**	46.6	**59.5**
MULTI_Own_frm	38.4	42.2	59.3
MULTI_Sen_neutral	30.3	25.5	53.0
MULTI_Sen_inf	32.8	30.0	54.1
MULTI_Sen_frm	31.8	27.8	55.1

Table 15: Automatic metrics for the specific test-sets

4.1.2 Common test-sets

Table 16 presents the results from the common test-set.

By taking a look at these results, we can observe that the MULTI_Sen_neutral engine achieves the best scores across all metrics, with a difference of +0.5 for SacreBLEU, of +0.1 for COMETINHO and of +0.3 for chr-F over the second-best engine, which is the FINE_loose_baseline. Worst scores are achieved by the MULTI_Own_neutral engine across all metrics.

Moreover, in this test-set, the difference between the scores of the informal and formal engines of each approach is not as marked as it was in the evaluation of the specific test-sets and is dependent on the automatic metric. However, there is indeed a clear distinction between the baseline engines from the FINE_strict and the FINE_loose systems, which achieve around +5.0 points for SacreBLEU, +8.0 points for COMETINHO and +4.0 points for chr-F, with respect to their informal and formal registers. The same applies to the MULTI_Sen_neutral engine, which achieves around +2.0 points for SacreBLEU, +3.0 points for COMETINHO, and +2.0 points for chr-F with respect to the informal and formal registers of the same approach.

This is an interesting result, since the baseline and the MULTI_Sen_neutral engines achieved some of the worst scores in their respective specific test-sets, which strengthens our idea of the fact that the informal engines might be in general over-fitted to their training data, while the baseline and neutral models might be better suited to deal with other kinds of data.

At this stage, given that we are evaluating the engines with respect to a common test-set, we carry out statistical testing using bootstrap re-sampling. These tests are carried out following a two-step approach:

- Firstly, each of the registers from the FINE_loose is compared with those from the FINE_strict approach, and each of the registers from the MULTI_Own with those from the MULTI_Sen approach. If results are statistically significant, we mark the score with an * in Table 16. In this sense, while there is not statistical significance between any of the registers from the FINE systems (except for the formal, where the FINE_loose_frm system achieves significantly better scores than the FINE_strict_frm), the results obtained by the MULTI_Sen_neutral and MULTI_Sen_informal are indeed statistically significant compared to their

respective registers in the MULTI_Own engine. This is a signal of how the MULTI_Sen_neutral and MULTI_Sen_formal engines are significantly better at dealing with all kinds of segments, compared to the MULTI_Own_neutral and MULTI_Own_formal respectively.

- Secondly, we compare the results from the MULTI_Sen approach with those from the FINE_loose approach, since they gave better results than the other two systems in the first-step, and observe that the MULTI_Sen informal and formal engines obtain results that are statistically significant with respect to those from the FINE_loose formal and informal engines (this is marked with a † in Table 16). However, this is not the case when comparing the FINE_loose_baseline and the MULTI_Sen_neutral engines. This might be a sign of the fact that, when fine-tuning a baseline towards the different registers, there is a bigger decrease in quality than when training a multi-register model from scratch with noise introduced in each direction.

Finally, what is also interesting from this evaluation is that the MULTI_Sen engines present a smaller variability in the results obtained by the three registers, specially across SacreBLEU, since the difference between the best scoring engine of the three (neutral) and the worst scoring engine (formal) is only 2.8 points, while for the rest of techniques there is a higher gap between their best scoring engine and their worst. This is a sign of how the MULTI_Sen approach achieves a rather similar quality among all its registers, while in the rest of approaches, there might be one or more registers where performance drops.

	SacreBLEU	COMETINHO	chr-F
FINE ENGINES			
FINE_strict_baseline	35.4	36.3	56.7
FINE_strict_inf	30.5	28.3	52.7
FINE_strict_frm	30.7	27.3	53.1
FINE_loose_baseline	36.0	38.0	57.2
FINE_loose_inf	31.3	29.4	53.2
FINE_loose_frm	31.5	28.9	53.9*****

MULTI ENGINES			
MULTI_Own_neutral	30.1	23.6	52.8
MULTI_Own_inf	32.3	30.8	55.0
MULTI_Own_frm	33.7	30.8	55.5
MULTI_Sen_neutral	**36.5***	**38.1***	**57.5***
MULTI_Sen_inf	34.1*†	35.1*†	55.8*†
MULT_Sen_frm	32.8†	30.1†	55.2†

Table 16: Automatic metrics for the common test-set

We also present the average performance of each approach (averaging the scores from the baseline/neutral, formal and informal registers) and present it in Table 17. As it can be seen, the engines from the MULTI_Sen approach achieve the best average scores for all metrics, with a difference of more than 2 points for each metric over the second-best approach (FINE_loose). On the other hand, the MULTI_Own engine presents the worst scores for SacreBLEU and COMETINHO. These results also show that the metrics are rather consistent when deciding which are the best performing engines.

	SacreBLEU	COMETINHO	Chr-F
FINE_strict	32.2	30.6	54.2
FINE_loose	32.9	32.1	54.8
MULTI_Own	32.0	28.4	54.4
MULTI_Sen	**34.8**	**34.4**	**56.3**

Table 17: Average performance of each approach

4.1.3 Error comparison

Given all the functionalities that MT Telescope has to offer, we decide to make use of the error comparison tool for analyzing the types of errors that each engine generates. In this way, we might be able to get a more in-depth idea of the performance of the engines beyond the automatic scores. This tool compares systems according to the percentage of segments falling into 4 different category buckets: residual errors (dark green), minor

errors (light green), major errors (soft orange) and critical errors (red). These can be found in Figure 5 and Figure 6 below.

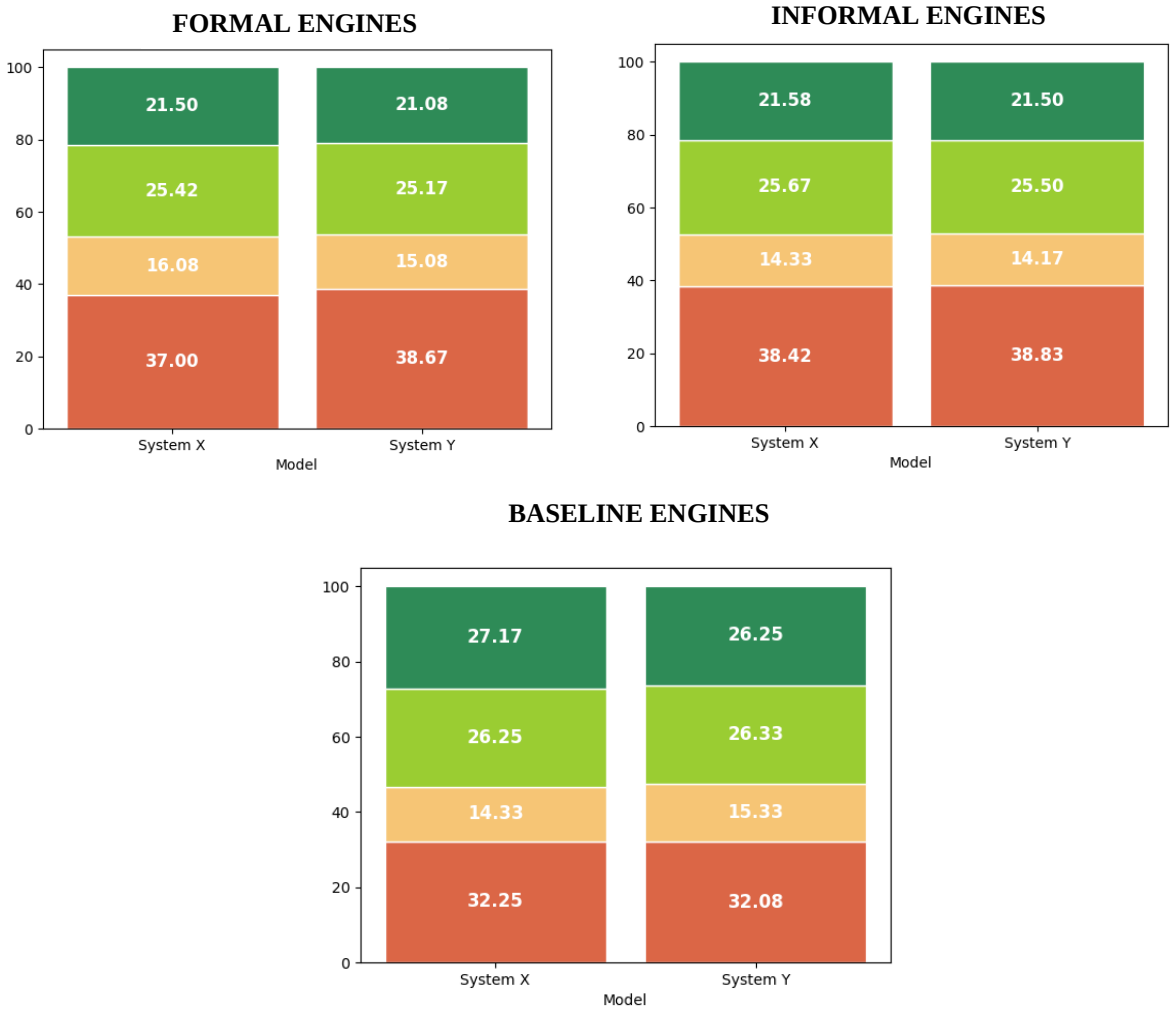

Figure 5: Error comparison of the fine-tuned engines. System X represents the FINE_loose systems, while System Y represents the FINE_strict systems.

As it can be observed in these figures, for the FINE informal and formal engines, in the case of critical errors, the four engines present around 38% segments of this type, while in the case of the baseline models, both engines present a much lower percentage of this kind of errors in comparison to the latter (around 6% fewer critical errors). This is in line with the better scores that the baseline models obtained in the previous stage of the evaluation. Moreover, the FINE_loose_baseline seems to present roughly the same percentage of critical errors as the FINE_strict_baseline engine.

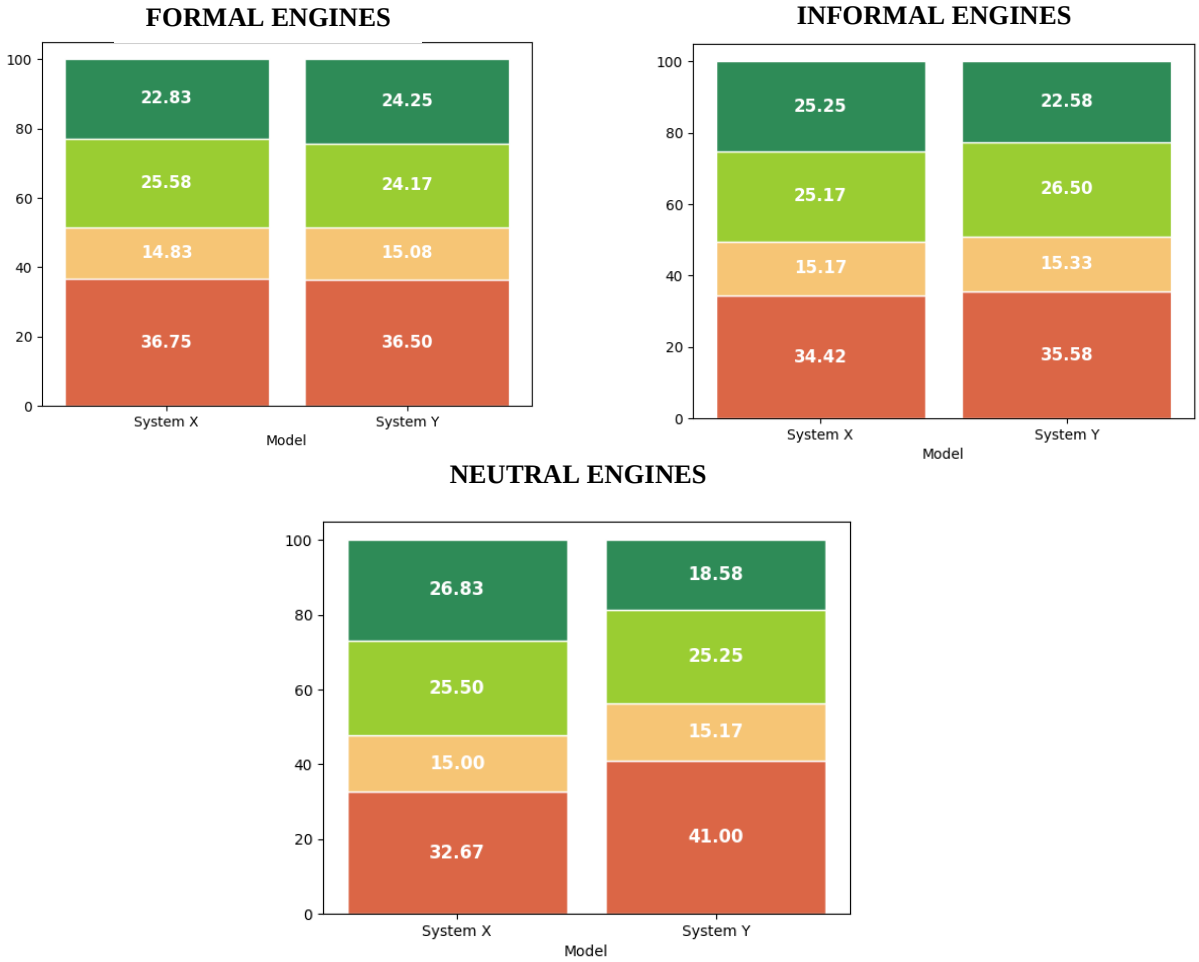

Figure 6: Error comparison of the multi-register engines. System X represents the MULTI_Sen systems, while System Y represents the MULTI_Own systems.

Regarding the MULTI engines, while there is not such a striking difference between the percentage of critical errors of the formal and informal registers (around 1% depending on the register), there is an important leap from the MULTI_Sen_neutral engine to the MULTI_Own_neutral engine, where the former presents 32.67% of critical errors, while the latter presents a 41%. This might be signaling some important phenomena occurring in the MULTI_Own_neutral that might have been penalized by the automatic tool as being critical errors with respect to a reference sentence. This is in line with the low scores that this engine received in the common test-set. However, we might need to take a look at the actual translations produced by the engine in order to get an idea of what problems this system is presenting. For that, we carry out a brief qualitative analysis of the output generated by this engine by using MT Telescope's segment-comparison tool. This tool plots segments into a 2D plot with regards to the scores that each system was given, thus facilitating the evaluation of those segments where each

system achieves a low score. With the help from this tool, we are able to observe some concerning outputs produced by the MULTI_Own_neutral engine, such as the following two:

- SOURCE: You are the one who knocked down.
- MULTI_Own_neutral: *You're the one who noqueado down* (Untranslated sentence.)
- SOURCE: You know, other than you like canoeing.
- MULTI_Own_neutral: *A parte de la canoa.* ('Apart from the canoe.')

These examples show that the MULTI_Own_neutral engine is indeed presenting some problems and leaving some segments as untranslated or erasing part of the sentence.

All in all, (except for the MULTI_Own_neutral engine) if we compare the four approaches (FINE_loose, FINE_strict, MULTI_Own and MULTI_Sen) both MULTI engines present around 2% less critical errors than the FINE engines in their respective registers, while the baseline engines and the MULTI_Sen_neutral systems present a similar percentage of these errors (~32%).

These results show that again, there is a higher decrease in performance when fine-tuning from a baseline, than when systems are trained from scratch as a multi-register model.

4.2 HUMAN EVALUATION

4.2.1 General-quality assessment

We now present the results from the human evaluation of each engine, where annotators had to score each sentence in a 1 to 5 scale for both adequacy and fluency (see Section 3.3.2.1). The average score for each metric and system resulting from the human evaluation is reported on Table 18.

As it can be observed, the FINE_loose_frm engine achieves the best scores for adequacy (4.62) while, in terms of fluency, the FINE_loose_baseline (4.54) engine achieves the best score. The FINE_strict_baseline and the MULTI_Sen_frm achieve the second-best scores (4.51 and 4.47 respectively).

When taking a look at the overall score (which is calculated as the average of the adequacy and fluency scores), the FINE_loose_baseline engine achieves the best score and is closely followed by the FINE_strict_baseline. This might be pointing to the fact that, while the FINE_loose_frm engine achieved the best adequacy scores, the fluency of the engine is more questionable and therefore, the overall quality of it is not as high as the one from the FINE_loose_baseline engine, although the difference is just 0.04 points.

Interestingly, all the informal engines achieved the worst results in adequacy except for the MULTI_Sen_inf engine (when comparing them to the formal and baseline/neutral engines from the same approach); while scores in fluency are less conclusive. This might be signaling that the MULTI_Sen_inf engine indeed benefited from the addition of sentences coming from the neutral and formal classification.

We also carried out t-tests following the two-step approach that we introduced in Section 4.1.2:

- Firstly, each of the registers from the FINE_loose are compared with those from the FINE_strict approach, and each of the registers from the MULTI_Own with those from the MULTI_Sen approach. If results are statistically significant with p-value < 0.1, we mark the score with * in Table 18. In that sense, the FINE_loose_frm engine achieves results that are significantly better than those of the FINE_strict_frm. However, there is no statistical significance between the any of the results from the MULTI_Own and the MULTI_Sen engines.
- Secondly, we compare the results from the FINE_loose approach with those from the MULTI_Sen approach, since they look consistently better than the MULTI_Own scores, and observe that again the FINE_loose_frm achieves results that are significantly better than the MULTI_Sen_frm (this is marked with a † in Table 18). This is signaling a clear preference of the annotators towards the FINE_loose_frm engine over the rest of engines.

Finally, it is also worth remarking that the average score for each engine is above 4 points, which in our measuring scale means all engines tend to preserve most of the meaning of the original sentence and have a good fluency, although not flawless (see Table 14 in Section 3.3.2.1).

	ADEQUACY	FLUENCY	OVERALL*
FINE ENGINES			
FINE_strict_baseline	4.51	4.45	4.48
FINE_strict_inf	4.05	4.32	4.18
FINE_strict_frm	4.18	4.14	4.16
FINE_loose_baseline	4.48	**4.54**	**4.51**
FINE_loose_inf	4.07	4.27	4.17
FINE_loose_frm	**4.62*†**	4.33	4.47
MULTI ENGINES			
MULTI_Own_neutral	4.21	4.16	4.18
MULTI_Own_inf	4.13	4.43	4.28
MULTI_Own_frm	4.47	4.35	4.41
MULTI_Sen_neutral	4.39	4.37	4.38
MULTI_Sen_inf	4.36	4.42	4.39
MULTI_Sen_frm	4.34	4.47	4.35

Table 18: Results of the human evaluation. The overall score is calculated as the mean of the adequacy and fluency scores of the engine.

Since it is also desirable to analyze the overall performance of each approach rather than of each separate engine, in Table 19, we present the average score for each approach for adequacy, fluency and overall performance. The score is obtained as the average of the informal, formal and neutral/baseline scores for each approach. To this regard, the MULTI_Sen approach achieves the best score for fluency and for overall performance, while the FINE_loose approach achieves the best score for adequacy and has an overall score that is just 0.01 points lower in overall performance. On the other hand, the FINE_strict systems receive the worst scores in all the metrics.

These results are interesting, since the MULTI_Sen approach did not achieve the highest score neither for adequacy, nor for fluency in each of the separate engines. However, when averaging the scores from each register (neutral, formal, informal), results are better than in the rest of engines, which shows that the three engines are

performing consistently well as a whole, while other approaches might present one or two engines that drop in performance for a specific register.

	ADECUACY	FLUENCY	OVERALL
FINE_strict systems	4.25	4.30	4.28
FINE_loose systems	**4.39**	4.38	4.38
MULTI_Own	4.26	4.31	4.29
MULTI_Sen	4.36	**4.42**	**4.39**

Table 19: Average score for each approach in human evaluation

4.2.1.1 Breakdown of types of segments in human evaluation

As it was already explained in Section 3.3.2.1, the Ling_test contained segments which represented different types of linguistic aspects of Spanish with the idea of carrying out a more in-depth evaluation of the engines. In Table 20, a breakdown of the average human scores that each type of segments (2PERSON vs. NO_FORMS) was given by the annotators can be found. Scores are calculated as the mean of adequacy and fluency for each engine. We also present the difference that there exists between the score of each engine for the two types of segments, which will indicate to what extent such engine is over-fitted to one type of segments over the other.

It is worth mentioning that we also tried to carry out a more fine-grained breakdown of the human evaluation per types of segments (with possessives, subjects, imperatives, etc.) but since the evaluation was not completely conclusive, we do not interpret the results in this section. However, for reference, we include the scores by type of segments in Table 33 in the Appendices.

Going back to the current evaluation, as it can be observed in Table 20, the MULTI_Own_inf engine achieves the best score for the 2PERSON segments and it is followed by the MULTI_Sen_neutral engine. Moreover, the FINE_strict_baseline achieves the best score in the NO_FORMS sentences and it is closely followed by the FINE_loose_baseline.

We compute t-testing once again following the two-step approach from previous sections. * marks that the results are statistically significant with p-value < 0.1 with respect to the same register when comparing FINE_loose with FINE_strict and

MULTI_Own with MULTI_Sen, while † marks that results are statistically significant when comparing the FINE_loose with the MULTI_Sen engines.

	MARKED	NO_FORMS	DIFFERENCE
FINE SYSTEMS			
FINE_strict_baseline	4.38	**4.59**	+0.21
FINE_strict_inf	4.41	3.79	-0.62
FINE_strict_frm	4.26	4.08	-0.18
FINE_loose_baseline	4.46	4.57†	+0.11
FINE_loose_inf	4.30	3.94	-0.36
FINE_loose_frm	4.55*****	4.37	-0.18
MULTI SYSTEMS			
MULTI_Own_neutral	4.16	4.23	+0.9
MULTI_Own_inf	**4.6**	3.87	-0.73
MULTI_Own_frm	4.45	4.47	+0.02
MULTI_Sen_neutral	4.51*	4.21	-0.30
MULTI_Sen_inf	4.44	4.33*†	-0.11
MULTI_Sen_frm	4.38	4.28	-0.10

Table 20: Breakdown of human evaluation by type of segment

In these results, we can observe the following phenomena for those segments where there are second person forms in the source (2PERSON):

- Regarding the neutral registers of the MULTI systems, the MULTI_Sen_neutral engine outperforms the MULTI_Own_neutral by almost 0.4 points, as well as the two registers from the same approach (MULTI_Sen_inf and the MULTI_Sen_frm). On the other hand, the MULTI_Own_neutral clearly drops in performance with respect to its informal and formal registers, and indeed, it achieves the worst performance of all the systems for this type of segments.
- The MULTI_Sen approach presents the smallest difference between its best and the worst system (0.13 points).

Regarding the segments with no verb and no *you* in the source (NO_FORMS), we can observe the following:

- The informal engines of those systems where the completely filtered corpus was used (MULTI_Own, FINE_loose and FINE_strict) achieve the worst scores, being all of them below 4.
- The formal engines achieve better results than the informal ones, and in the case of the MULTI approaches, the engines in this register (MULTI_Sen_frm and MULTI_Own_frm) outperform their respective neutral register, while the formal engines from the FINE approaches do not reach the scores of their respective baseline models.
- The MULTI_Sen engine also presents the smallest difference among the registers for this type of segments (0.12 points)

Finally, when taking a look at the difference in scores between the 2PERSON and the NO_FORMS, we can observe the following:

- There seems to be signs of catastrophic forgetting in the FINE systems, since all of the formal and informal engines clearly drop in performance when translating the NO_FORMS segments. However, this difference in performance seems to be smaller for the FINE_loose systems.
- The particular engine with the smallest drop in performance from one type of segments to the other is the MULTI_Own_frm. This engine presents an improvement of +0.02 from the 2PERSON to the NO_FORMS segments, being only trained with data coming from the formal classification. However, it is important to remember that in the formal subset, around 1/4 of the segments seemed to be false positives. Therefore, this result sheds some light as to how what would be a good ratio of noise in a training corpus so that the engine stays stable across all types of segments.
- The approach with smallest difference in performance across all of its registers is the MULTI_Sen approach. However, strikingly, the MULTI_Sen_neutral system drops in performance by 0.3 points for the NO_FORMS sentences. This might be signaling that this engine did not see enough neutral systems in order to achieve comparable results to the baseline systems from the FINE approaches.

Master HAP/LAP

All in all, we can extract the following conclusions from this analysis:

- The baselines from the FINE approaches seem to outperform the neutral systems from the MULTI engines in the NO_FORMS sentences, but not the MULTI_Sen_neutral system in the 2PERSON sentences. However, the MULTI_Sen approach seems to be more consistent when taking into account the performance of each engine across the two types of segments.
- When comparing the FINE systems with each other, the FINE_loose systems seems to be more consistent and presents a smaller difference between the results of each engine (0.15 in the 2PERSON sentences, and 0.63 in the NO_FORMS sentences).
- When comparing the MULTI systems with each other, the MULTI_Sen system seems to be more consistent and present a smaller difference between each engine than the MULTI_Own, while presenting better scores by the neutral engine in the 2PERSON segments. However, the formal and informal engines from the latter seem to achieve better scores in almost all cases. What this means is that the MULTI_Own formal and informal engines are probably over-fitted towards the 2PERSON segments, and thus achieve better results for this type of segments, while the MULTI_Sen presents worse results for these segments but a smaller drop in performance when dealing with NO_FORMS segments.
- The MULTI_Own_neutral engine did not benefit from having completely filtered data with sentences containing no verb and no *you*, since we observe that the difference with the MULTI_Sen_neutral in the NO_FORMS sentences is not significant enough (0.02) to compensate for the loss in quality in the 2PERSON sentences, where it obtained the worst results.

These results highlight the difficulty of finding a balance between the 2PERSON sentences and the NO_FORMS sentences at training-time. The engines that were trained with more strictly filtered data show better performance in the 2PERSON segments, while dropping in performance with the NO_FORMS segments to a larger or smaller degree. However, those engines trained with data coming from different subsets (even if unintentionally, such as the case of the formal engines) achieve slightly worse performance in the 2PERSON segments but do not experience such a sharp decrease in quality with the NO_FORMS segments. Therefore, in terms of overall quality, if we

assume that it is more desirable to have an engine (or set of engines) that achieves a consistent performance for all sentence-types and registers, then the MULTI_Sen approach presents itself as the most consistent option. However, if the priority is to use each engine for their specific type of segments, then it might be more desirable to use the FINE_loose approach and use each specific engine for each type of segments.

4.2.1.2 Inter-annotator agreement

In order to get an idea of how reliable the human evaluation that we carried out is, we calculate the inter-annotator agreement of each test-set using Fleiss' Kappa and present the results along with the average Kappa for the human evaluation of our research. Results can be found in Table 21.

As it can be observed, there is a clear difference in agreement between some test-sets and others: while some of them achieve an agreement that is close to 0.30 or even 0.35; three of them present an agreement that is lower than 0.20. However, when averaging all the scores, we achieve a score 0.25 for inter-annotator agreement. Out intuition behind this not-too-high score is that the number of classes to annotate (5) makes it more complicated to find an agreement between the annotators, especially in certain test-sets such as number 2, number 10 and number 12.

Test-set 1	0.29	Test-set 7	0.35
Test-set 2	0.17	Test-set 8	0.28
Test-set 3	0.24	Test-set 9	0.22
Test-set 4	0.29	Test-set 10	0.17
Test-set 5	0.24	Test-set 11	0.30
Test-set 6	0.29	Test-set 12	0.10
AVERAGE	0.25		

Table 21: Fleiss' Kappa per test-set

4.2.1.3 Comments from the evaluators

Even if, as mentioned in the previous sections, evaluators were not given any more information on the task at hand rather that they were evaluating a series of segments generated by MT, some of their comments shed some light about the task at hand.

A total of ten annotators added comments on their test, which amount to 1/3 of the annotators. Four of them commented that some segments seemed to use the wrong register. However, one of them did penalize these segments, meanwhile the other three did not (although they said that they had some doubts on what to do with these segments). Five other evaluators added some comments on the use of politeness and the forms *tú* and *usted,* and how one form was preferred over the other and how some thought the engine was changing from Castilian Spanish to other dialects of Spanish (when *usted* was used in a sentence that sounded rather informal). This shows that the evaluators were indeed paying attention to the honorifics and the register generated by the engine, and that they probably thought that the test-set contained segments generated by one engine rather than many. Two evaluators also mentioned the fact that the engine was translating those words with gender marks as being masculine. This shows that, for some of them, the gender generated by the engine was also important. Finally, two more annotators mentioned the fact that it was difficult to evaluate some segments without a given context.

What all of these comments prove is that evaluators did indeed pay attention to politeness and context when evaluating MT output, and that the evaluation of politeness can become a complicated task when no strict guidelines are provided with respect to the number of engines being evaluated or to the fact that the switch between *tú* and *usted* is expected.

4.2.2 Register-specific assessment

We now present the results of what we denoted as *politeness test* in our evaluation. Re-using the Ling_test from the previous section, we add labels to each translation generated by each engine following the classification explained in Section 3.3.2.2. We expect this test to offer some insights as to how accurate our engines are when producing the intended register in in the 2PERSON segments, as well as to whether the engines are over-generating *honorifics* when they should not. We present the results for the 2PERSON segments in Table 22.

2PERSON						
	FINE_strict			FINE_loose		
LABEL	Baseline	Informal	Formal	Baseline	Informal	Formal
INFORMAL	15	29	2	12	29	3
FORMAL	14	1	27	17	0	27
MIX	0	0	0	0	0	0
NEUTRALIZATION	1	0	1	1	0	0
OTHERS	0	0	0	0	1	0
TOTAL	30	30	30	30	30	30
ACCURACY*	-	96.7%	90%	-	96.7%	90%
	MULTI_Own			MULTI_Sen		
	Neutral	Informal	Formal	Neutral	Informal	Formal
INFORMAL	6	30	2	16	28	2
FORMAL	11	0	26	14	1	28
MIX	2	0	1	0	0	0
NEUTRALIZATION	6	0	1	0	0	0
OTHERS	5	0	0	0	1	0
TOTAL	30	30	30	30	30	30
ACCURACY	-	100%	96.7%	-	90.3%	90.3%

Table 22: Politeness test of 2PERSON segments. Accuracy is calculated as the percentage of segments where the intended register of the engine (formal or informal) was used.

For this type of segments. we can observe the following:

- Regarding the FINE engines, the FINE_strict_baseline seems to prefer forms of *tú* (50%) to *usted* (47%); meanwhile the FINE_loose_baseline seems to prefer *usted* (57%) to *tú* (40%). Both present one segment with neutralization. Moreover, when observing the fine-tuned engines, both informal engines present 96.7% of accuracy and both formal engines present 90% of accuracy.
- Regarding the MULTI engines, we start by taking a look at the neutral engines. In that sense, while the MULTI_Sen_neutral engine presents a preference for

informal (53%) over formal (47%) and no other type of phenomenon such a mixed of politeness or neutralization, the MULTI_Own_neutral presents a preference for formal (~37%) over informal (~20%). However, the latter presents other types of phenomena: 2 *mix*, 6 *neutralizations* and 5 *others*. This might be signaling an inconsistent way of dealing with the different types of segments. Moreover, taking a look at the informal engines, the MULTI_Own_inf engine presents an accuracy of 100%, while the MULTI_Sen_inf seems to be less consistent and presents 90.3% of accuracy. Regarding the formal engines, the MULTI_Own_frm presents an accuracy of 96.7% while the MULTI_Sen_frm engine presents 90.3%. The former presents as well other kinds of phenomena such as 2 *informal*, 1 *mix* sentence and 1 *neutralization*; while the latter presents only 2 *informal* segments, which might be signaling a more consistent performance in this engine.

Interestingly, there were several segments produced by the MULTI_Own_neutral engine which were labelled as *neutralization*, and although this engine presented other phenomena such as mixed registers and many segments with the label *others*, we believe that this neutralization phenomenon should be further analyzed. Some examples of neutralizations along with the kind of neutralization technique that was used in each case can be found in Table 23.

SOURCE	Yesterday, we went out for a couple of drinks downtown. What about you guys?	How was your experience with us?	We have all these new items for you!
TARGET	*Ayer salimos a tomar un par de copas al centro.* (‘Yesterday, we went out for a couple of drinks downtown.’)	*¿Qué tal la experiencia con nosotros?* (‘How was your *the* experience with us?’)	*¡Tenemos todos estos artículos nuevos!* (‘We have all these new ítems!’)
NEUTRALIZATION TECHNIQUE	Ellipsis of the second part of the sentence where the *you* appears.	It gets rid of the possessive *your* and translates it as an indefinite article, which is actually correct and fluent in Spanish.	Ellipsis of the part of the sentence containing the *you*

Table 23: Some examples of neutralization techniques from the MULTI_Own_neutral engine

As it can be observed, this engine uses an indefinite article (*la*) instead of the possessive determiner (*su*/*tu*), and this neutralization is indeed a fluent (an even more desirable) translation in Spanish. In other cases, it also gets rid of the part of the sentence where the second person form appears, although this might lead to some lack of information in the message (see first column of Table 23).

Moreover, Table 24 presents the results from the test with the NO_FORMS segments. We do not calculate accuracy this time, since there is no correct register. Instead, we calculate the percentage of number of segments with and without additions as well as other phenomena that is covered by the label *others* (see Section 3.3.2.2).

NO_FORMS						
	FINE_strict			FINE_loose		
LABEL	Baseline	Informal	Formal	Baseline	Informal	Formal
NO ADDITIONS	19 (95%)	16 (80%)	16 (80%)	19 (95%)	12 (60%)	17 (85%)
ADDITIONS	0	4 (20%)	1 (5%)	0	7 (35%)	1 (5%)
OTHERS	1 (5%)	0	3 (15%)	1 (5%)	1 (5%)	2 (10%)
TOTAL	20	20	20	20	20	20
	MULTI_Own			MULTI_Sen		
	Neutral	Informal	Formal	Neutral	Informal	Formal
NO ADDITIONS	19 (95%)	10 (50%)	18 (90%)	19 (95%)	19 (95%)	19 (95%)
ADDITIONS	0	10 (50%)	0	0	1 (5%)	1 (5%)
OTHERS	1 (5%)	0	2 (10%)	1 (5%)	0	0
TOTAL	20	20	20	20	20	20

Table 24: Politeness test of NO_FORMS segments

In this occasion, we observe the following:

- Regarding the FINE engines, we observe the same phenomena in both baselines, where only a 95% of the segments have no additions and only a 5% were labelled as *others*. In the case of the fine-tuned engines, while the FINE_loose_inf engine tends to over-produce *honorifics* to a greater extent than the FINE_strict_inf engine, both formal engines produce roughly the same percentage of segments with no additions.

- Regarding the MULTI engines, taking a look at the neutral engines, we can notice how both engines present the same percentage of segments with no additions (95%) and of other phenomena (5% of *others*). However, when turning to the informal engines, MULTI_Own_inf presents a 50% of segments with additions, while the MULTI_Sen_inf presents just a 5%. This is in line with the results from previous tests, where the MULTI_Own_inf engine achieved worst results in this type of segments. This is probably pointing at a present bias in the MULTI_Own engine for this register, which was palliated in the MULTI_Sen engine by introducing segments coming from the neutral subset. Finally, taking a look at the formal engines, the MULTI_Sen engine also presents an improvement from the MULTI_Own in the sense that the former presents a 95% of segments with no additions, while the latter presents a 90% as well as a 10% with labelled as *others*.

What we can extract from this test is the fact that those segments which contained more strictly filtered data-sets (FINE_strict, FINE_loose and MULTI_Own) are indeed between 10% and 6% more accurate than the MULTI_Sen when producing the intended register. However, they tend to over-produce *honorifics* in those segments where no verb or no second forms appear. Moreover, even if we did not take into account whether these hallucinations have an impact on the final sentence, we believe that they shed some light about the way in which each engine generates translations. In that sense, the MULTI_Sen approach is the one that presents less sentence labelled as *others*, from which we can conclude that while being less accurate when choosing the right registers, these engines produce less segments with dubious output, such as untranslated segments.

Therefore, in line with the results from the general assessment of the quality of each engine, we can conclude that the MULTI_Sen approach presents less accurate results when choosing the correct register of the output sentence, it does not over-produce *honorifics* to the same extent than other engines, and produces less dubious outputs, which highlights the consistency of this model when dealing with the different types of segments.

4.2.3 Assessment of sentences with a clearly marked register

In this section we carry out a final test to analyze to what extent the engines have lost – or not– the capacity of discerning which forms to use, i.e. of producing the opposite register if the source text is clearly marked as having a formal or informal register.

Regarding the baseline engines from the FINE approaches, both seem to produce the expected forms for each type of segments. For instance, they use *usted* for sentences like *Everything is ready for you, Madam;* and *tú* for *Your hair looks bloody amazing in that photo*. The same happens with the MULTI_Sen_neutral engine, while the MULTI_Own_neutral engine presents some neutralized segments (such as *Todo listo, señor*. instead of *Todo está listo para usted, señor*.) and segments using the appropriate form.

Moving on to the formal engines, the FINE_loose_frm along with the FINE_strict_frm and the MULTI_Sen_frm engine use the opposite register for *It was sooooo nice to see you!*. Moreover, the FINE_strict_frm, MULTI_Sen_frm and MULTI_Own_frm do the same with *Seriously, just piss off!*, while the FINE_loose avoids using any form for this segment.

Finally, taking a look at the informal engines, we observe that the MULTI_Own_inf engine does not produce any sentence with the opposite register, while the other three do handle one or two segments using the expected formal form. Those are the following: *Dear Mr. Smith, please contact us if you have any doubts.* and *We are really looking forward to welcoming you, Miss Wright.*

What we can conclude from this evaluation is that, while none of the engines presents the ability of the baseline and neutral models to produce the correct register in these situations, it is clear that these segments present a challenge to the systems and lead to translations with mix forms of politeness in some cases. In this sense, the MULTI_Sen and the FINE_strict approaches are the engines showing a slightly higher tendency to produce the correct form for these cases. However, the test-set is quite limited and more work should be carry out in this regard to arrive at a clear conclusion.

5 CONCLUSIONS AND FUTURE WORK

Throughout our work we reviewed the important role that register and politeness play in communication and have highlighted the fact that computational approaches towards the study and generation of language should not overlook these inner variations of languages. We have also presented an approach –following Sennrich et al.'s contribution for German (2016)– for classifying Castilian Spanish segments with respect to the appearance of *tú* and *usted* (whether they were explicit or elided). We then used the separated subsets for exploring different fine-tuning and multi-register approaches towards training a NMT system from English to Castilian Spanish with politeness control of the output and prove that the task can be addressed as a domain adaptation problem.

In the evaluation stage, the MULTI_Sen approach appeared to be the most consistent across registers and types of segments, presenting the best overall performance in the automatic evaluation across all metrics (SacreBLEU, COMETINHO and chr-F) as well as in the human evaluation, where it achieved the best score for fluency and overall performance. In the *politeness test*, the system achieved a 90.3% of accuracy for both registers while not over-producing *honorifics* in 95% of the segments that were analyzed. Moreover, the FINE_loose approach achieved the second-best overall results in the automatic performance and the best adequacy score in the human evaluation, which might be an indicator that fine-tuning the baseline engine with a looser early-stopping rate could be beneficial for this task. However, we cannot exclude the fact that the baseline models were trained with different early-stopping values, and therefore, results might not be due only to the fine-tuning stage. Finally, the annotators evaluated the MULTI_Own approach with some of the best scores for its informal and formal engines in the 2PERSON segments and the neutral register presented some impressive neutralization phenomena. However, this engine did not reach the quality of the baselines from the FINE models.

Therefore, when trying to give an answer to the question of which the best approach to use for the task is, there are different factors to take into account. In terms of overall quality, if we assume that it is more desirable to have an engine (or set of engines) that

achieves a consistent performance for all types of sentences and registers, then the MULTI_Sen approach presents itself as the most consistent option. However, if the priority is to use each engine for its specific type of segments (domain), it might be more desirable to follow the FINE_loose approach and use each specific engine depending on the type of segment that is inputted at inference time. In an implementation stage, this could be done by parsing the source segment and searching for a verb or *you* in any of its forms. If any of the latter are found, the translation can be sent to one of the fine-tuned engines (depending on the desired degree of politeness), while if none of the latter are found, the baseline model can be used. However, it is not among the purposes of this study to offer a solution towards the implementation of these systems.

Finally, with our last research question we wanted to focus on the impact that politeness features have in the evaluation of NMT output. To this respect, although we did not carry out an extensive experiment, we found that some annotators shared their doubts regarding the evaluation of segments containing *tú* and *usted* within the same test-set. The comments left by the annotators shed some light so as to how politeness is important in an evaluation task and so as to how, when carrying out an evaluation of the overall quality of an NMT system with politeness control of the output, a more in-depth guideline than the one we provided for this research should be granted to the evaluators.

This is one of the limitations of our work, since the mix of registers contained in the segments to evaluate and the lack of more information could be one of the reasons for the rather low inter-annotator agreement (0.25) and for the fact that most of the results were not statistically significant.

Be it as it may, we believe that our study leaves the door open towards future work on the study of politeness in NMT. Other approaches than can be further explored are the use of terminology constraints or Factored Neural Machine Translation (FNMT), which is related to the human way of learning how to construct correct sentences. Factors normally refer to linguistic annotations at word level such as POS tags or gender, and could potentially be used for applying politeness constraints by prepending a factor which states whether a lemma is *polite*, *neutral* or *impolite* before training the model. Moreover, there is much room for improvement when choosing the right amount of segments for fine-tuning an engine for politeness control in order to palliate the problem of catastrophic forgetting. Given that the engine with less strictly filtered data achieved

the best results, exploring the mixed fine-tuning approach proposed by Chu & Wang (2017) with a similar ratio could be an interesting option for this task. Finally, in line with the study of IM in NMT, we believe there is also still much opportunities in the study of discourse-level NMT, which can benefit from the exploration of techniques for controlling not only politeness, but also number and gender.

REFERENCES

Álvarez González, A. (2006). *La variación lingüística y el léxico.* Hermosillo: Universidad de Sonora.

B, L. (1994). The lexical profile of second language writing: Does it change over time? *RELC journal, 25(2)*, 21-33.

B. Laufer, &. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics 16(3)*, 307-322.

Bane, F., & Zaretskaya, A. (2021). Selecting the best filtering method for NMT training. *Proceedings of the Translation Summit XVIII: Users and Provides Track* (pp. 89-97). Virtual: Association for Machine Translation in the Americas.

Baños Piñero, R. (2004). La oralidad prefabricada en los textos audiovisuales: estudio descriptivo-contrastivo de Friends y Siete Vidas. *Forum de Recerca. Servei de Comunicació i Publicacions*, 42.

Bizzoni, Y., Juzek, T.S.,España-Bonet, C., Dutta Chowdhury, J., van Genabith, J., & Teich, E. (2020). How Human is Machine Translationese? Comparing Human and Machine Translations of Text and Speech. *Proceedings of the 17th International Conference on Spoken Language Translation* (pp. 280–290). online: Association for Computational Linguistics.

Brown, P. & Levinson, S. (1987). *Politeness: Some universals in language usage (Vol. 4).* Cambridge: Cambridge University Press.

Brown, P. (2015). Politeness and language. In A. C. Atkinson, *The International Encyclopedia of the Social and Behavioural Sciences (2nd ed.)* (pp. 326-330). Elsevier.

Bywood, L., Etchegoyhen, T., Georgakopoulou, P., Fishel, M., Jiang, J., Loenhout, G. V., ... & Maucec, M. . (2014). Machine translation for subtitling: A large-scale evaluation. *LREC 2014, Ninth International Conference on Language Resources and Evaluation*, 46-53.

Chaume, F. V. (2001). La pretendida oralidad de los textos audiovisuales y sus implicaciones. *La traducción en los medios audiovisuales*, 77-88.

Chu, C. & Wang, R. (2018). A survey of domain adaptation techniques for neural machine translation. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1304-1319). Santa Fe: Association for Computational Linguistics.

Chu, C., Dabre R., & Kurohashi, S. (2017). An empirical comparison of domain adaptation methods for neural machine translation. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 385-391). Vancouver: Association for Computational Linguistics.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek G., Guzmán, F., Grave, E. Myle, O., Zettlemoyer, L. , Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440-8451). Online: Association for Computational Linguistics.

Conneau, A., Lample, G., Ranzato, M. A., Denoyer, L., & Jégou, H. (2017, October 11). *Word translation without parallel data.* Retrieved from arXiv: https://doi.org/10.48550/arXiv.1710.04087

Davies, M., & Davies, K. H. (2017). *A frequency dictionary of Spanish: Core vocabulary for learners.* Oxon: Routledge.

Dinu, G., Mathur, P., Federico, M., & Al-Onaizan, Y. (2019, June 3). Training neural machine translation to apply terminology constraints. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3063-3068). Florence: Association for Computational Linguistics. Retrieved from arXiv:1906.01105: 2019

F., Arkhangorodsky, A., Biesialska, M., Bojar, O., Chatterjee, R., Chaudhary, V., et al. (2021). *Findings of the 2021 Conference on Machine Translation (WMT21).* Online: Association for Computational Linguistics.

Fernández, J. C. (2016, julio 20). *Aproximación a la traducción translectal de un corpus audiovisual de películas hispanoamericanas.* Retrieved from Universidad de Valladolid.[Versión electrónica]: http://uvadoc. uva. es/handle/10324/7662

Fernández, M. (2003). Constitución del orden social y desasosiego: pronombres de segunda persona y fórmulas de tratamiento en español. *Ponencia plenaria en el coloquio Pronoms de, 2.* Paris: Centro Virtual Cervantes.

Fukushima, S., & Iwata, Y. (1985). Politeness in English. *Jalt Journal, 7(1)*, 1-14.

Goffman, E. (1967). The nature of deference and demeanor. In E. Goffman, *Interaction ritual. Essays on Face-to-Face behaviour* (pp. 78-123). New York: Pantheon Books.

Grice, P. H. (1975). The logic of conversation. In P. Cole, *Syntax and Semantics 3: Speech Acts* (pp. 41-58). Elsevier.

Halliday, M. A. (1978). Language as Social Semiotic: The Social Interpretation of Language and Meaning. *Language in Society*, 84-89.

Halliday, M. A. (2002 [1977]). Text as semantic choice in social contexts. *Linguistic Studies of Text and Discourse. Volume 2 in the Collected Works of M. A. K Halliday*, 23-81.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., ... & Zhou, M. (2018, March 15). *Achieving human parity on automatic chinese to english news translation.* Retrieved from arXiv: https://arxiv.org/abs/1803.05567

Haugh, M. (2005). The importance of "place" in Japanese politeness: Implications for cross-cultural and intercultural analyses. *Intercultural Pragmatics, vol. 2, no.1*, 41-68.

Johnson, M., Schuster,M., Le, Q., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., Dean, J. (2017). Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistic*, 339-351. Retrieved from arXiv.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., ... & Birch, A. (2018). Marian: Fast neural machine translation in C++. *Proceedings of {ACL} 2018, System Demonstrations* (pp. 116-221). Melbourne: Association for Computational Linguistics.

Kell, G. (2018). *Overcoming catastrophic forgetting in neural machine translation (Doctoral dissertation).* Cambridge: University of Cambridge.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences, 114(13)* (pp. 3521-3526). Vancouver: Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *Proceedings of ACL 2017* (pp. 67-72). Vancouver: Association for Computational Linguistics.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT symmit vol.5*, 79-86.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of machine translation summit x: papers* (pp. 79-86). Phuket: Association for Computational Linguistics.

Koehn, P., Khayrallah, H., Heafield, K., & Forcada, M.L. (2018). Findings of the wmt 2018 shared task on parallel corpus filtering. *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 726-739.

Kyle, K. (2014). Measuring lexical richness. *Translation: A multidisciplinary approach*, 96-115.

Läubli, S., Sennrich, R., & Volk, M. (2018). Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). Brussels: Association for Computational Linguistics.

Leech, G. (1983). *Principles of Pragmatics.* London: Longman.

Marie, B., Fujita, A., & Rubino, R. (2021). Scientific credibility of machine translation research: A meta-evaluation of 769 papers. . *arXiv preprint arXiv:2106.15195.*

Matthiessen, C. M. I. M., & Halliday, M. (1997). *Systemic functional grammar*. Amsterdam and London: Benjamins & Whurr.

McCarthy, P. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD) (Doctoral dissertation).* Memphis: The University of Memphis.

McLuhan, M. (1964). *Understanding Media: The extensions of Man.* Canada: McGraw-Hill.

Oakes, M. P., & Ji, M. (Eds.). (2012). *Quantitative methods in corpus-based translation studies: A practical guide to descriptive translation research (Vol. 51).* Amsterdam/Philadelphia: John Benjamins Publishing.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation* (pp. 392-395). Lisbon: Association for Computational Linguistics.

Post, M. (2018). A call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186-191). Brussels: Association for Computational Linguistics.

Post, M., & Vilar, D. (2018, April 18). Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. *Proceedings of the 2018 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1314-1324). New Orleans: Association for Computational Linguistics. Retrieved from arXiv preprint arXiv:1804.06609: https://arxiv.org/pdf/1804.06609.pdf

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101-108). Online: Association for Computational Linguistics.

Rei, R., Farinha, A.C., Stewart, C., Coheur, L. & Lavie, A. (2021). {MT}-{T}elescope: {A}n interactive platform for contrastive evaluation of {MT} systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 73-80). online: Association for Computational Linguistics.

Roselló Verdeguer, J. (2017). El uso de tú y usted en el área metropolitana de Valencia. Un enfoque variacionista. *ELUA, 31,* 285-309.

Salvador, V. (1989). L'anàlisi del discurs, entre l'oralitat i l'escriptura. *Caplletra. Revista Internacional de Filologia (7)*, 9-31.

Sánchez Martínez, S. (2015). La escritura de los jóvenes en los chats en el siglo XXI. *Didáctica (lengua y literatura), vol. 27*, 183-196 .

Saussure, F. d. (1915). *Cours de Linguistique Générale.* Paris: Payot.

Schwenk, H., & Douze, M. (2017). Learning joint multilingual sentence representations with neural machine translation. *Proceedings of the 2nd Workshop on Representation Learning for {NLP}* (pp. 157-167). Vancouver: Association for Computational Linguistics.

Sennrich, R., Haddow, B. & Birch, A. (2015). Neural Machine Translation of Rare Words with Subword Units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1712-1725). Berlin: Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016). Controlling Politeness in Neural Machine Translation via side constraints. *Association for Computational Linguistrics* (pp. 35-40). San Diego: Proceedings of NAACL-HTL.

Shannon, C. E. (1948, July 1). A mathematical theory of communication. *The Bell System Technical Journal*, pp. 379-423.

Simpson, E. H. (1949, April 30). Measurement of diversity. *Nature*, pp. 688-688.

Smith, R. (2018, March 19). *World Economic Forum*. Retrieved from A million WhatsApp messages were sent in the time it's taken you to read this headline: https://www.weforum.org/agenda/2018/03/internet-minute-whatsapp-facebook-emails/

Świątek, A. (2012). *Pro-drop phenomenon across miscellaneous languages.* Retrieved from Academia.edu.

Tekwa, K. (2018). *Increasing Willingness and Opportunities to Communicate in a Foreign Language with Machine Translation and Instant Messaging.* Ottawa: Doctoral dissertation, Université d'Ottawa/University of Ottawa.

Templin, M. (1957). *Certain language skills in children.* Minneapolis: University of Minnesota Press.

Thompson, B., & Post, M. (2020). Automatic machine translation evaluation in many languages via zero-shot paraphrasing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 90-121). online: Association for Computational Linguistics.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (pp. 2214-2218). Istanbul: European Language Resources Association (ELRA).

Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? reassessing claims of human parity in neural machine translation. *Proceedings of the Third Conference on Machine Translation: Research Papers"* (pp. 113-123). Brussels: Association for Computational Linguistics.

Vanmassenhove, E., Shterionov, D., & Gwilliam, M. (2021). Machine translationese: Effect of algorithmic bias on linguistic complexity in machine translation. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2203-2213). Online: Association for Computational Linguistics.

Vanmassenhove, E., Shterionov, D., & Way, A. (2019, June 28). Lost in translation: Loss and decay of linguistic richness in machine translation. *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222-223). Dublin: European Association for Machine Translation. Retrieved from arXiv preprint arXiv:1906.12068.: https://arxiv.org/abs/1906.12068

Verdeguer, J. R. (2016, August 12). *Factores que intervienen en los usos de tú y usted en español peninsular. Algunos ejemplos prácticos para E/LE.* Retrieved from Foro de profesores E/LE.: https://ojs.uv.es/index.php/foroele/article/view/9186/8704

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016, September 26). *Google's neural machine translation system: Bridging the gap between human and machine translation.* Retrieved from arXiv: https://doi.org/10.48550/arXiv.1609.08144

Yang, C. Y., & Lin, H. Y. (2010). An instant messaging with automatic language translation. *2010 3rd IEEE International Conference on Ubi-Media Computin*, 312-316.

Yule, C. U. (2014). *The statistical study of literary vocabulary.* Cambridge: Cambrdige University Press.

APPENDICES

1. ANALYSIS OF RESULTING SUBSETS

Correct	Incorrect
SOURCE: Now **you** do not have one either? *TARGET*: Ahora **usted** tampoco lo tiene.	*SOURCE:* did **you** have to leave two little boys soaking in a pool of their mother's blood? *TARGET:* ¿**Tenías** que dejar a dos chicos empapados en un charco de la sangre de su madre?
SOURCE: Sorry about **your** brother. *TARGET*: Lamento lo de **su** hermano.	*SOURCE*: When **you** knew who your neighbour was. *TARGET*: Cuando **uno sabía** quién era su vecino.
SOURCE: **You** okay, mister? *TARGET*: ¿**Se encuentra** bien, señor?	*SOURCE*: **You** want me to say I knew John Latner wasn't her baby daddy, and **you** want my source so you can track down Casey's biological father. *TARGET*: **Quieres** que diga que sabía que John Latner no era el padre de su bebé, y **quieres** mi fuente para poder averiguar quién es padre biológico de Casey.

Table 25: Some examples for correctly and incorrectly classified sentences in the formal subset (regex approach)

Correct	Incorrect
SOURCE: I guess that's all **you** want of me. *TARGET*: Supongo que ya no me **necesita**.	*SOURCE*: I think **you** got it cheap. *TARGET*: Creo que es barato.
SOURCE: But **you** didn't do anything to me. *TARGET*: Pero no me **hicieron** nada.	*SOURCE*: would **you** believe Chavis would be the one in here still trying to get right, and Money be the one not at school today and not doing what he's supposed to? *TARGET*: ¿**hubieras** dicho que Chavis sería el único que intentaría hacer todo bien?
SOURCE: Hey, Jack, I thought **you** was a farmer. *TARGET*: Pensé que **era** granjero.	*SOURCE*: Figured **you** must be starving. *TARGET*: Imaginé que **tendrías** hambre.

Table 26: Some examples for correctly and incorrectly classified sentences in the formal subset (parsing approach)

Correct
SOURCE: If **you** are who you say you are, tell me *TARGET*: Si **tú** eres ser quien dices ser, dime.
SOURCE: I warned **you** before. *TARGET*: Ya **te** lo advertí.
SOURCE: I told **you** - -I own it. *TARGET*: Ya **te** dije, soy el dueño.

Table 27: Some examples for correctly and incorrectly classified sentences in the informal subset (regex approach)

Correct	Incorrect
SOURCE: - and looked at it. *TARGET*: - y lo **miraste**.	*SOURCE*: You want to follow him? Not until we figure out what we're dealing with. *TARGET*: Una chance de revolucionar el estudio de la genética.
SOURCE: Kyle, **will you** go talk to Wendy for me? *TARGET*: Kyle, ¿**puedes** hablar con Wendy de mi parte? ¿Por qué?	
SOURCE: - **You had** me worried. *TARGET*: - Me **has dado** un buen susto.	

Table 28: Some examples for correctly and incorrectly classified sentences in the informal subset (parsing approach).

Correct	Incorrect
SOURCE: Word on the street is they had snipers on the roof last night. *TARGET*: Se dice en la calle que tenían francotiradores en el techo anoche.	*SOURCE*: He went away, **all right?** *TARGET*: Él se fue, ¿**entiendes**?
SOURCE: They will look great in court. *TARGET*: Se verán geniales en la corte.	*SOURCE*: In fact, ladies and gentlemen, for the continuity of the film, I'd like to **reintroduce** the radio. *TARGET*: De hecho, damas y caballeros, para seguir con la película, **les** vuelvo a presentar a la radio.
SOURCE: - Tell them? *TARGET*: - ¿Se lo digo?	SOURCE: And the other will be the offense, **how does that sound**? *TARGET*: Y el otro será la ofensiva, ¿qué **le** parece?

Table 29: Some examples for correctly and incorrectly classified sentences in the neutral subset (regex approach)

Correct	Incorrect
SOURCE: I'm real happy. *TARGET*: Estoy realmente feliz.	*SOURCE*: THAT'S A COP KILLER. *TARGET*: ¡**Quítense**!
SOURCE: Will no one? *TARGET*: ¡Nadie?	*SOURCE*: **Areyou** sure? *TARGET*: **Estas** seguro?
SOURCE: Damn felon. *TARGET*: ¡Maldito criminal!	*SOURCE*: **hear** that, Casey? *TARGET*: ¿**Oíste** eso, Casey?

Table 30: Some examples for correctly and incorrectly classified sentences in the neutral subset (parsing approach)

2. LING_TEST

2PERSON	
Phrases which contain the pronoun *you* in English	- You should go to the doctor if you are feeling sick. - What did you do yesterday? - We are avaiable via Whatsapp to solve any questions you may have during the purchase.
Phrases which should contain the pronoun *you* in the Spanish translation (without ellipsis)	- It was you who started the fight. - Who did it? Was it you? - Yesterday, we went out for a couple of drinks downtown. What about you guys? - Is it you, Tom? - You need to be the one that picks up the parcel.
Phrases containing possessives	- Can you check your agenda and let me know when you are free? - How was your experience with us? - Did you break your arm? - I believe that T-shirt was yours. - Let's take my car, not yours. - Please enter your address. - Where do you wish to receive your items? - Your purchase is almost done! - How was your experience with us?
Phrases containing imperatives	- Come with us, please! - Contact us at XXXXX. - Call me when you get home. - Click on the item you wish to purchase. - Look at this. - Please, do not hesistate to contact us and ask for a refund.*
Phrases containing the construction preposition + personal pronoun	- Can I come with you? - We have all these new items for you! - No, thank you.

	- I made all this for you. - We would love to go to the cinema with you tonight. - Did she come with you? - I was waiting for you guys forever!
NO_FORMS	
Phrases containing no verb	- Nonsense! - Why not? - How cool! - Seriously? - Postal code - Next item - Hey, there! - Welcome! - Where? There? - Customized delivery services
Phrases containing no forms of *you*	- We are delighted to be here today. - I am really happy to be here today. - They were suppose to come today.* - We enjoyed it so much! - Personally, I think that is not true. - He was such a nice person. - She moved to Madrid to attend University. - Offering customized delivery services since 1996. - They asked me whether I wanted a refund. - Let's go together.

*Table 31: Ling_test created for human evaluation and specific politeness evaluation. Segments with a * present some spelling mistakes in the source, which were intended to cover also IM content problems. However, these did not seem to present a problem in the translations generated by any of the engines.*

3. OPPOSITE_TEST

Clearly formal sentences	- Dear Mr. Smith, please contact us if you have any doubts. - We are really looking forward to receiving Miss Wright. - Everything is ready for you, Sir. - Everything is ready for you, Madam. - It was lovely having you over, Mrs. Taylor.
Clearly informal sentences	- Seriously, just piss off! - Damn! Your hair looks bloody amazing in that photo. - You look awful, mate. -It was sooooo nice to see you!!! -Would you like to come with us????

Table 32: Opposite test containing segments with a clear distinction in register

4. WORD CLOUDS

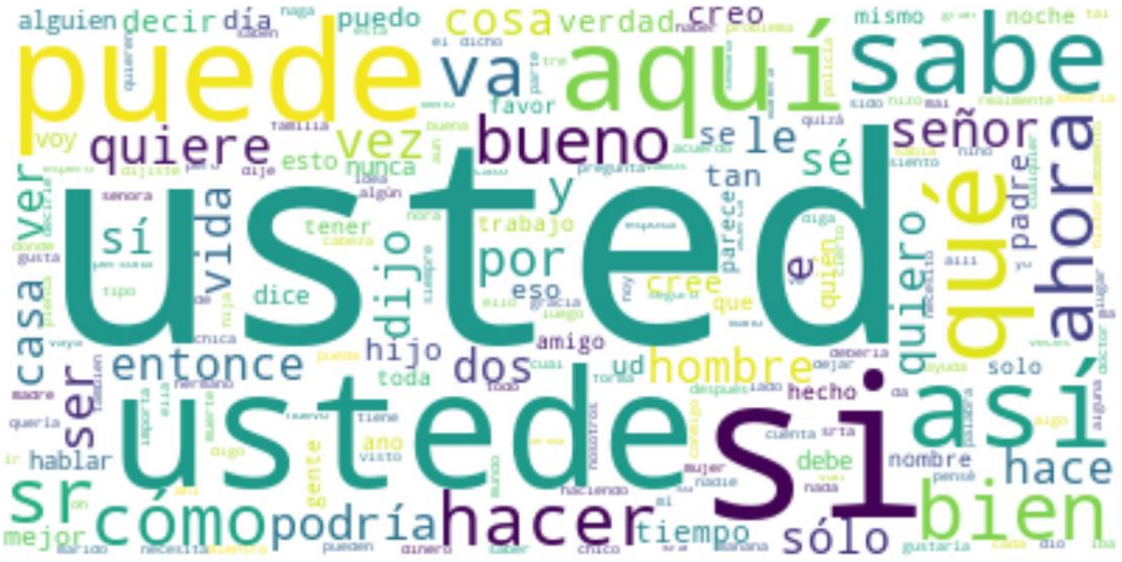

Figure 7: Word cloud from formal segments extracted from the OpenSubtitles corpus

Figure 8: Word cloud from informal segments extracted from the OpenSubtitles corpus

Figure 9: Word cloud from neutral segments extracted from the OpenSubtitles corpus

5. FINE-GRAINED EXPLORATION OF RESULTS BY TYPES OF SEGMENTS

	You_En	You_Es	Possessives	Imperatives	Pronouns	No_verb	No_you
FINE systems							
FINE_strict_baseline	4.17	4.47	**4.79**	4.4	4.1	4.37	**4.82**
FINE_strict_inf	4.06	4.6	4.31	4.53	4.45	3.62	3.97
FINE_strict_frm	4.28	4.37	4.1	4.23	4.24	4.42	3.75
FINE_loose_baseline	4.44	4.47	4.46	4.6	4.38	**4.43**	4.72
FINE_loose_inf	4.22	4.4	4.43	4.03	4.31	4.15	3.73
FINE_loose_frm	**4.83**	4.33	4.73	4.6	4.33	4.4	4.33
MULTI systems							
MULTI_Own_neutral	4.67	3.53	4.1	4.23	4.38	4.4	4.07
MULTI_Own_inf	4.56	4.87	4.58	**4.77**	**4.62**	3.58	4.15
MULTI_Own_frm	4.5	4.57	4.60	4.4	4.21	4.27	4.67
MULTI_Sen_neutral	**4.83**	4.73*	4.69*	4.0	4.38	4.28	4.15
MULTI_Sen_inf	4.22	4.7	4.5	3.97	4.31	4.25**	4.4
MULTI_Sen_frm	4.33	**4.9**	4.52	3.8	4.31	4.08	4.47

*Table 33: Breakdown of human scores per different type of segment. * marks that score is statistically significant with respect to the same register and technique, but different training strategy with alpha set to 0.05, while ** marks that score is statistically significant with alpha set to 0.1.*