UNIVERSITAT OBERTA DE CATALUNYA

Machine Translationness: a Concept for Machine Translation Evaluation and Detection

by

Joaquim Moré López

A thesis submitted in partial fulfillment for the
degree of Doctor in Information and Knowledge Society
Programme: Information and Knowledge Society
Supervisor: Dr. Salvador Climent Roca

April 2015

Come, your answer in broken music, for thy voice is music and thy English broken

W. Shakespeare (Henry V. Act 5, Scene 2)

Always be a first-rate version of yourself, instead of a second-rate version of somebody else

Judy Garland

Abstract

Machine translationness (MTness) is the linguistic phenomena that make machine translations distinguishable from human translations. This thesis intends to present MTness as a research object and suggests an MT evaluation method based on determining whether the translation is machine-like instead of determining its human-likeness as in current evaluation approaches. Therefore we present an evaluation method that assesses machine translations according to what they are (translations produced by a machine) and not to what they resemble (human translations).

The method rates the MTness of a translation with a metric, the MTS (Machine Translationness Score). The MTS calculation is in accordance with the results of an experimental study on machine translation perception by common people. MTS proved to correlate well with human ratings on translation quality. Besides, our approach allows the performance of cheap evaluations since expensive resources (e.g. reference translations, training corpora) are not needed.

Machine translationness ratings can be applied for other uses beyond machine translation evaluation. The MTS metric can be an important indicator to prevent the consequences of the massive use of MT, such as plagiarism and other forms of cheating, or the detection of unsupervised MT documents published on the Web.

Acknowledgements

I write the acknowledgements when I put a final stop to my permanent thesis state, and now I think of my uncle writing the acknowledgements section of his PhD long time ago. Sharing this instant is the fulfillment of a task inspired by my admiration and love for him.

I want to thank my director, Salvador, for encouraging me to write this thesis. My gratitude reaches beyond his suggestions, his support at hard times, his concern in disseminating my work from the very beginning, and his contribution in the writing of this thesis, obsessed as he was in making it 'readable' and 'understandable'.

I would also like to thank friends at UOC, UB, UPC and other universities for their concern and good advice for this thesis. I know that they are happy for me, and I feel proud to share this moment with them.

I also acknowledge the important contribution of people who distracted me from this hard work. Some of them are not with me on this finishing line, but they deserve all my gratitude.

And last but not least I want to thank my family for their support and patience. And, obviously, I want to thank Laia for her care, love, cheerfulness, sensibility, and easy-going.

In sum, thanks to all the people who helped me to sow the silver seeds of machine translationness. Now I will enjoy with them the ripe and golden apples of the sun.

Agraïments

Escric aquest apartat quan poso punt i final al meu estat de tesi permanent. Un estat iniciat ja des de ben petit, quan vaig decidir ser doctor, com el meu tiet. Per fi he assolit una tasca inspirada per la meva admiració cap a ell.

Vull donar les gràcies al meu director, en Vadó, per animar-me a fer aquesta tesi. La meva gratitud va molt més enllà de les seves suggerències, el seu suport en temps durs, la seva implicació en donar a conèixer aquest treball des de bon començament, i la seva contribució en l'escriptura d'aquesta tesi, a més de la seva obsessió en fer-la 'llegible' i 'comprensible'.

M'agradaria també agrair els meus amics de la UOC, la UB, la UPC i d'altres universitats pel seu interés per la tesi i pels seus molt bons consells. Sé que estan molt contents per mi, i em sento orgullós de que estiguin aquí per compartir-ho.

També agraeixo la gran contribució de la gent que m'ha distret de la tesi. Alguns no estan amb mi a la línia d'arribada però mereixen tota la meva gratitud.

Per últim, però per això no menys important, vull agrair a la meva família el seu suport i paciéncia. I, evidentment, vull agrair-li a la Laia el seu amor, els seus ànims, el seu seny, i que hagi intentat sempre posar-me les coses ben fàcils, amb una pacient alegria. Hi ha molt d'ella en aquestes págines.

En resum, gràcies a tots els que m'han ajudat a plantar les argentades llavors de la machine translationness. Ara gaudiré amb ells de les daurades pomes madurant-se al sol.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI) machines are evaluated for their capacity to emulate human behaviour and they are said to exhibit intelligent behaviour if they pass the Turing test (Turing, 1950). The test is passed if the judge, confronted with output performed by a machine and by a human, cannot tell the machine from the human.

A machine translation system (MTSy) is an example of a machine that emulates intelligent human behaviour. Therefore, the interest in machine translation (MT) output has lain in determining whether this output is indistinguishable from a human translator output. This is what we will call the *human translationness criterion (HTC)* which is applied in machine translation evaluations. Our thesis, on the contrary, will deal with the linguistic features that make machine translations *distinguishable* from human translations. These features contribute to what we call *machine translationness (MTness)* which we define as *the quality of machine translations that makes them distinguishable from human translations.* MTness is the flavour of machine translations, and is perceived as something odd, queer and difficult to be attributed to a human being.

The ALPAC report already forecasted, in the 1960s, the impossibility for MT techonology to emulate fully human behaviour (Melby, 1995). In other words, the report warned about the inevitableness of machine translationness. This was so not only because of the state-of-the art technology of the time but also for AI limitations. AI should succeed in emulating human reasoning, inferences, communication strategies, sentiment monitoring, application of common sense, and other cognitive and communicative aspects which are widely agreed to be very difficult to attain. Since the ALPAC report MT specialists have been aware that passing the Turing test is utopian. However, features that differentiate MT systems from human translators have not deserved much attention as an object of research. Specialists consider these features as errors of the system but not all errors are actually distinguishable from humans. For instance errors according to grammar normative prescriptions are common among speakers of the language. Non-human MT features especially draw the attention of Luddites[1], columnists and commentators that present absurd and hilarious translations to mock at the technological attempt to emulate a human task (Budiansky, 1998) and also to warn about its perils (Porsiel, 2011). Absurdity and foolishness in MT is even the topic of a website[2]. Far from this intention, the objective of this thesis is to present errors distinguishable from human translations as an object of study and show its practical applications in ICT.

Currently, machine translation systems are extensively used and popular. Machine translation has become available for everyone. By clicking a button we can translate web pages, subtitle videos on the web in different languages, understand the menu in a foreign language captured by a mobile phone, or communicate with someone whose language we do not know. For this reason MTness is also widely extended and is currently an everyday phenomenon people live with. MTness is then a phenomenon worth to be studied. In section 1.1 we will present different aspects where MTness must be taken into account and in sections 1.2 and 1.3 we will concrete what this thesis will be about.

1.1 MTness and its importance

In this section we will demonstrate the importance of MTness in four aspects:

- Translation quality
- MT use
- Internet as a corpus of language use
- Information retrieval

Translation quality is the absolute requirement for both human and machine translations. Specially in business scenarios, where the clients expect high-quality translations. MT use is also a key factor, not only because of its universal and free access. A very important consequence of the widespread use of machine translation is the presence of MTness as a phenomenon one must live with. We will focus on the consequences of MTness when the Internet is both taken as a corpus of language use and a source for information retrieval. This is specially important because nowadays the Internet is the principal source of information.

[1] See (Hillas, 2009) about the negative attitude towards MT

[2] http://www.fortunecity.com/business/reception/19/index.html. [Accessed 8 April 2012]

1.1.1 MTness and translation quality

MTness may seem to affect only the quality of translations generated by a machine but, in fact, it affects the quality of translations by humans as well. Some professional translators use online MT tools that free them from tiring tasks (Vlasta, 2012) and saves them time. If MT reached acceptable quality levels, human translators would become correctors of MT systems. This is unlikely to happen in careful translations in distant language pairs (e.g. patents translated from Japanese into English), but it is true for closer language pairs (say, Spanish-Catalan). However, no matter how carefully the professional translator revises the translation, a route of entry is open to MTness.

There are real workflows where professionals only correct MT output. One well-known example is the daily publication of the Catalan version of the newspaper *El Periódico*, originally written in Spanish (Fité, 2006). The challenge of publishing separate daily editions in Spanish and in Catalan would not be possible if the translation workflow were not supported by MT, and professionals were not limited to correcting (postediting) MT output. However, the *El Periódico* experience is also an example of how MTness is disseminated, and scattered. MTness dissemination was especially important during the first days of publication, when the MTSy was not tuned enough and human posteditors were not yet trained to detect all the mistakes under time pressures. As MT has been used to provide other publications, websites, etc. with the Catalan and Spanish versions, readers have made public in the media their anecdotic reading of MT errors. Even the digital newspaper *El Punt* organised a contest. The reader who detected the most amusing MT error in the Catalan version of any publication was awarded with the Catalan version of a Tom Stoppard′s play, translated by a renowned Catalan writer [3].

When flaws in human translations are found, the blame is often put on the MTSy that hypothetically was used by the translator. So machine translation is a sort of goat scape that may hide the translator′s carelessness or any other circumstance. Evidences of MTness should be demonstrated (Multilizer, 2011). However research in MT has not dealt with characterizing these evidences. MT research has been more concerned in evaluating the resemblance of machine translations to human translations. That is, machine translations have been assessed on what they are not (human translations) rather than on what they really are (an output generated by a machine).

1.1.2 MTness and MT use

Unversal and free access to MT systems facilitate the non appropiate use of this technology. For instance, language teachers often feel they are correcting a machine translation

[3] http://www.vilaweb.cat/www/capde7mana/forums/ftop7?forum=2749335. [Accessed 8 April 2012]

instead of a composition written by a student (Sommers et al., 2006). *Machine Translator Detector*[4] is a tool that can be used to detect whether a text is a machine translation but MTness is not taken into account. In fact, the tool only determines how likely a text is a translation performed by one of these systems: Google Translate, Yahoo!, Babel Fish and Bing. On the other hand, it requires the text in the source language. Unfortunately, the source text is seldom available when testing if someone has not been fair. For this reason an MTness automatic detector, needless of the source text, would be as useful as plagiarism detectors to avoid inappropiate use of tools on the Web. The MTness detector would also be useful for detecting spam and phishing messages because most of them have been translated automatically with no postedition and their degree of MTness is very high.

As for appropiate use, we already mentioned that translators use MT systems. We also said that translators just postedit the translation when the source and the target languages are very close and the translation quality is quite acceptable. However, close language pairs and the fact that the target document keeps the layout of the source document may have disadvantages. Errors may go unnoticed in the first and even the second reading. Here are some examples of frequently overlooked errors in Catalan-Spanish translations

- The presence of the Catalan conjunction *i* (and) instead of the Spanish conjuction *y*
- *Corria* (was running) instead of *corría*
- No preposition *a* before a human direct object

Other non-corrected errors come from multilingual quotations. Translators often have to translate a quotation in a language they do not know. Then the translator uses an MTSy but is not able to postedit the translation as a speaker of the language would. When the translation is published readers that *do* know the language complain about fragments that lead to misinterpretations [5].

Translators, teachers and other professionals who are concerned about the linguistic correctness of their publications must spend time in a through and careful postedition. This task is rarely performed in a single reading. For this reason, a postedition tool that detected MTness instances would be very useful.

[4] http://www.translatordetector.com/ [Accessed 18 May 2012]

[5] http://blog.fluenthistorian.com/2011/07/05/why-you-should-not-rely-on-machine-translation/ [Accessed 16-04-2012]

1.1.3 Internet as a corpus of language use

For some natural language processing applications the Web is considered a large representative corpus of language use. In (Moré and Climent, 2004) and (Sjöbergh, 2006) the Web is the reference for grammar checkers, although the authors warn about the uneven quality of published contents, so the examples of use from the Web should pass idoneity conditions.

Nowadays, the Web is the environment where multilingual spontaneous communication is more lively. There has been an explosion of informal, inmediate and collaborative publications (blogs, wikies, tweets, comments in social networks, and so on). MT plugins in blogs and social network environments, with a wide range of language pairs, have helped participants in a blog or a social network to overcome linguistic barriers. The fact that multilingual followers of a blog machine translate and publish their comments in the author′s language is not spurious, even when the author′s language is minoritarian. The author also machine translates the reply in the native language of the commentator. Authors and participants think in worldwide terms. They are aware that followers and casual readers may live anywhere and belong to very different linguistic contexts. This is the reason why blog writers who are not fluent in English machine translate their posts into this language in order to reach a worldwide target. On the other hand, there is an affective bond between the author and the follower. Both of them show interest in the other′s culture and sometimes, as a token of respect, write in the addresee′s language by means of MT. I was reported an experience that exemplifies this. A Portuguese artist was a follower of a blog written in Catalan. He read the posts by using an online MTSy. Then he machine translated his comments from Portuguese into Catalan, and the author replied in Portuguese, by using a MTSy, as a courtesy for having written in his native language.

Twitter provides the most spontaneous, fastest and liveliest communication in Internet among individuals from all over the world. MT in tweets is regarded as an obvious feature, because MT provides the possibility to make large amounts of up-to-date information accessible for journalists, researchers and enterprises(Jehl, 2010). However, the instancy of communication and the awareness that a tweet′s life is very short do not favour people to correct the MT output. Besides, as in the case of posts in blogs, the authors do not have the knowledge of the language to *postedit* the messages. So machine translated tweets contribute to spreading MTness throughout the Web.

MTness is becoming more and more present on the Web and developers of linguistic services that take the Web as a corpus are aware of this. One example is the bilingual online dictionary service *Linguee* that shows contexts of use of a word or phrase in the source language and pairs each context to its translation in the target language. Both source and target contexts are retrieved from publications on the Web. The context pairs are filtered according to a machine-learning algorithm that distinguishes good and bad contexts. Translations with MTness are among the bad contexts. One thing the engine learned is that web sites with the word *Wordpress* are likely to have been translated by an MTSy so they are not good contexts[1]. Not only is this criterion too radical (all published post in Wordpress, machine translated or not, become not good) but also discriminates posts that have been correctly postedited.

Inmediate and non revised publications on the Web causes a paradox exemplified in the consequences of Google Translate′s success. Google′s MT system also uses the Web as a corpus. It learns to translate from the huge number of Web documents parallelised with their translations. When the learning corpus contains documents translated by the system itself, and these documents have MTness instances, the engine is also learning to translate as machines do. To overcome this paradox is a challenge for Microsoft researchers to enhance their technology by solving the deadlock of their competitor-Google. Therefore, (Rarrick et al., 2011) present an algorithm that identifies machine-translated content in Webscraped parallel corpora. The main goal is to clean polluted corpora for training statistical MT systems, and they suggest the application of the algorithm for improving search engines. Another motivation for filtering out machine translated sentences in Web parallel corpora is to obtain resources for training statistical MT systems when there are few available billingual parallel corpora (Antonova and Misyurev, 2011).

Although (Rarrick et al., 2011) say that for high-density languages (such as English, Japanese, German), only a small percentage of web pages are generated by MT systems, the percentage of non-revised machine translated sentences they found in parallelised web pages was considerable (15%). For low-density languages, the numbers increase dramatically. For instance, the same authors estimated that nearly 50% among all Web content (not only bilingual) in Lithuanian was machine translated. So MTness is a significant problem that affects not only multilingual communication on the Web but also threatens the presence of minority languages and languages with fewer resources than the ones with high presence in the Internet.

1.1.4 MTness and information retrieval

Nowadays there is strong concern about retrieving information from informal, collaborative and inmediate communication spaces (social networks, blogs, wikies). We highlight

[1] http://www.linguee.com/english-spanish/page/about.php?source=ES [Accessed 09-03-2014]

two reasons in particular. The first one is the fact that blogs and wikies are becoming references to explore specific subjects in detail. The second reason is the commercial interest in sentiment analysis, that is, in determining opinions about a product or service. Therefore there is interest in retrieving information from sources that are, as we have seen before, vulnerable to MTness.

Although participants in a collaborative project control the quality of a publication, MTness can momentarily misguide the readers. For instance, a Wikipedia entry which is the machine translation of the same entry in a different language. If the translation is non-revised, MTness may mislead readers until errors are fixed by a contributor. Nowadays, it is possible for a visitor to read *los Bocados* (the Bites) in the Spanish translation of a Catalan blog, wiky or tweet, referring to the *Mossos d′esquadra*, the police guard in Catalonia, just because *Mossos* is also the plural of *Mos* which means ′bite′. As long as this error is not fixed many people may believe that policemen in Catalonia are known in Spain as *los Bocados*.

MTness on the Web is also a challenge for search engines. The quality of the search results depends on the quality of the documents found. In a competition between search engines (Google, Yahoo, Bing) the filtering of odd documents is crucial. In fact, Google noticed that if they continued to permit the massive and indiscriminate use of their Google Translate API they were shooting themselves in the foot, because non-postedited contents would impact negatively on the results of their search engine. As Kirti Vashee says (Vashee, 2011) *Google is in reality polluting its own drinking water.*

In sentiment analysis, the algorithms that detect opinions are based on the recognition of words that connote appreciation, satisfaction or, on the contrary, rejection, or disappointment. Here MTness may mislead. Imagine someone reads the following headline in a Spanish newspaper: *el Máster de Software Libre se fue a hacer puñetas hace 1 mes* (the Master in Free Software went to hell one month ago) when the real meaning was that the Master in Free Software started one month ago, just because the Catalan verb ′engegar′ was mistranslated. The reader has the wrong idea that the Master was a failure, and the negative consequences of this interpretation for the institution that organized it are obvious.

An MTness detector, as a module of an evaluator of Web publications would be very interesting to improve search results, information retrieval and dissuade people, in the middle and long term, from the nocive and abusing use of MT.

1.2 MTness as a field of research and a novel proposal for MT evaluation

In this thesis we deal with MTness in translation quality. Our objectives are to characterize MTness from a linguistic point of view and to suggest a method of determining how machine-like a translation sounds. Scoring the MTness of a translation would help to prevent the negative consequences of MT use and abuse we explained in the previous section. The thesis puts MTness into the light of a research object and also presents a novel proposal for MT evaluation.

MT evaluation is currently based on the HTC. Therefore MTness is not the focus of research yet. Actually MT evaluation research is focused on overcoming the following drawbacks:

Expensive Methodologies Human evaluations are very costly in money and time. The conditions to get reliable results demand the designing of the evaluation procedures, hiring more than one evaluator, training them, selecting appropriate translations for the evaluation corpus, compiling and analysing data, and so on. Actually, automatic evaluations appeared as a cheap alternative to human evaluations. However, many human translations are needed in order to assess whether the MTSy′ behaviour is distant or close to a human translator′s behaviour. Compilation and revision of these human translations also require high costs.

Not fully reliable results Translation quality perception is very subjective. Therefore the results in human evaluations must be presented noting the subjectivity of human judgements. In automatic evaluations, it has been proved that metrics do not provide a global measure of quality (Giménez, 2008). They focus on partial aspects, such as lexical choice but are not suitable to capture the quality at the sentence level (Blatz et al., 2003) and (Turian et al., 2003). Even the assumption of the reliability of a measure may lead to wrong conclusions about the quality of a system (Callison-Burch et al., 2006). The rationale for calculating metrics seems not to correspond to a well defined explanation of human translationness nor translation quality perception by humans.

Non-reusable results In human evaluations, the judgement about the quality of a translation is represented by a numerical score. Scores by themselves say very little about problems that should be solved and improvements. On the other hand, the current guidelines for human evaluations suggest that the system should be evaluated in its context of use. Therefore evaluations in different contexts of use become different problems and the data obtained in one evaluation are difficult to be reused in another. In automatic evaluations, the metrics also say very little about things to be improved and the type of errors a system generates.

Our proposal is to face MT evaluation not by measuring qualities of something translations are not (human translations) but, on the contrary, by measuring features of what they really are (translations generated by a machine). MTness features are machine-idiosyncratic features so our approach lies on an assumption we call the machine translationness criterion (MTC) which, informally, can be expressed as *the less MTness in a machine translation, the better.* This is so because, as we said before, texts with MTness are odd and queer.

The main contributions of this thesis in this research line are:

machine translationness score (MTS) MTS indicates the degree of MTness in a translation. Contrary to HTC scores, MTS indicates how far translations are distant from human translators′ behaviour. Therefore, this score can be straightforwardly interpreted in terms of translation quality. The higher the score, the worse the translation is.

Multi-task evaluations MTC evaluations can be performed for tasks other than testing MT systems (detection of inappropiate use of MT, quality evaluation of Web contents, reliability evaluations of information retrieval)

Cheaper evaluations MTC evaluations can be performed automatically, with free and open-source resources, and need not collect nor generate human-made translations. Cost-saving is remarkable and evaluations can be performed on-the-fly.

Reliable evaluations MTC is intuitive and widely agreed, and MTS values are more informative about translation quality than values of HTC metrics.

Reusable results MTC evaluation results are obtained by a previous detection of MTness instances in the translation. The MTness instances detected are data that can be reused in an automatic postedition module and are straight evidences of the system′s errors that need to be solved in new versions

Another side of this thesis is the linguistic study on MTness features. This study has been necessary for the detection of MTness instances. Besides, the study provides a novel theoretical insight to MT, and opens up new paths in other fields of knowledge (psycholinguistics, artificial intelligence, cognitive linguistics) The main contributions of this study are:

Experimental work about MTness perception Although much has been said about queer, odd and foolish MT translations our study contributes in determining what linguistic features cause astonishment.

MTness Typology A linguistic typology of MTness instances has been established from real MTness perceptions. Apart from the interest in typifying the linguistic features that make translations sound machine-like, it serves to discriminate MT errors as those critical (MTness errors) and not-so-critical, which helps developers to focus in solving the errors with most impact.

1.3 Organization and overview

This thesis has three parts. The first part deals with MTness as a topic of research. The second part is about our proposal of evaluating translation quality by measuring MTness. We explain the method and the resources needed to calculate the MTS, and discuss the results obtained in an experiment carried out to evaluate the proposal. Reusability of the results for other tasks such as postedition or detectors of inappropiate use of MT will also be discussed. Finally, the third part is devoted to the conclusions and future work.

The thesis is organized as follows:

Part 1

Chapter 2: Contextualization of MTness in the MT Evaluation problem and advantages of our approach.

Chapter 3: Experimental study of MTness

Chapter 4: MTness typology

Chapter 5: Discussion about MTness perception

Part 2

Chapter 6: MTness automatic detection

Chapter 7: Calculation of the MTS score

Chapter 8: Experimental evaluation of our proposal

Part 3

Chapter 9: Conclusions and future work

Chapter 2

Evaluation of MT Quality and Machine Translationness

In this chapter we put the concept of *machine translationness (MTness)* in the light of machine translation evaluation (MTE), and remark the relevance of MTness rating in terms of a novel methodological approach.

Although the first MTE campaign for the ALPAC already disclosed the wide gap between human and machine translators, machine translations are still evaluated as if they were human translations. In fact, two of the most important MT quality items- target language fluency and fidelity to the original- are the outcomes of human capacities: language proficiency and understanding. Unfortunately human MT evaluation (HMTE) (HMTE), especially the criteria for scoring these items, are very subjective and vary among evaluators. The procedures to obtain as many reliable and objective data as possible have increased the costs of HMTE in time and money. As an alternative, automatic MT evaluation (AMTE) performs methods that lower costs and establish metrics whose values are not subjective perceptions. These methods have been based on the assumption which is stated in the following terms: a machine translation is as better as similar to a human translation of the same original. Other approaches have appeared, such as the ones based on the human-likeness assumption, that is a machine translation is as better as more human-like is. Costs have been reduced, development cycles of MT systems have been accelerated, but the reliability of these assumptions have not been critically analysed.

This chapter summarizes MTE and puts our proposal in context. Our proposal is to state the machine translationness criterion (MTC) which is the criterion by which a machine translation is evaluated according to qualities of machines. This criterion is distinguished from the human translationness criterion (HTC), which is the criterion on which MT evaluation methods have been based so far. HTC-based methods score qualities of human beings in machine translations.

The chapter is organised as follows. In section 2.1 we will overview the methods based on HTC. In section 2.2 the shortcomings of HTC-based methods will be discussed and in section 2.3 we will present the contributions of the MTC proposal.

2.1 HTC Methods

In this section we will overview the methods based on the HTC. First we will present the HTC in human evaluations and then in automatic evaluations. The presentation will be organized in a chronological order. This will help us to explain how quality rating in machine translation has changed and how evaluation methods have adapted to these changes.

2.1.1 HTC in human evaluations

2.1.1.1 ALPAC Evaluation

In 1966 the ALPAC report informed the sponsors of machine translation in the United States about the prospects of this technology. The report described how mature MT systems, especially the systems sponsored, were in reducing translation costs from Russian into English. The ALPAC report is well known because of its pessimistic prospects (Hutchins, 1996), which marked a turning point in MT evolution. But, apart from that, the ALPAC aimed to confectionate a report according to well-grounded and standard-intended evaluation procedures. The experiments performed, explained in (Carroll, 1966), were designed to know whether MT could satisfy the demands of human translation, especially those regading intelligibility and fidelity to the source. The ALPAC evaluation was based on the HTC because it implied the performance of a Turing test: both machine and human translations were presented to evaluators and, although they were told about the existence of machine translations, they did not know whether the translation to evaluate was human or not.

Evaluators scored intelligibility and fidelity with the numerical values of a scale. The scale for intelligibiliy had 9 values and the scale for fidelity had 10 values. In the case of intelligibility, each numerical value corresponded to a subjective appreciation when reading the translation. For instance, this is the description of the value corresponding to the *generally unintelligible* appreciation: *it tends to read like nonsense but, with* *a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.*

As for fidelity, the grades corresponded to the informativeness appreciated when monolingual and bilingual evaluators had to compare the translation to a reference. For bilingual evaluators, the reference was the source and for monolingual evaluators the reference was a human translation of the same source. Most informative translations were regarded as translations with the lowest fidelity, because their interpretation was different from the interpretation of the reference. Accordingly, the least informative translations were those with the highest fidelity, because their interpretation did not deviate from the source.

The use of scales was an important contribution, among other methodological proposals such as the association between subjective appreciations and objective measures. For instance, time reading was associated to intelligibility (unintelligible translations demand more reading time). Another important contribution was the participation of monolingual and bilingual evaluators in order to know how the knowledge of the source language influenced the intelligibility and fidelity appreciations. Lastly, some recommendations already indicated the costs of the methodology. For instance, three or four evaluators were recommended to be used.

2.1.1.2 The Van Slype Report

In 1979 the Marcel van Dijk Bureau wrote a report (Slype, 1979) for the European Commission with two objectives: firstly, to recommend the best methods to evaluate the SYSTRAN system, which was the system used by the Commission, and secondly, to assist in designing progress evaluations during the development of the MT system sponsored by the Commission (the EUROTRA Project).

The report puts an accent on the importance of establishing what the evaluation was for. So two types of evaluation were defined, depending on the finality: the **macroevaluation** and the **microevaluation**. The macroevaluation measures translation quality (inteligibility, fidelity and acceptability), which is consistent with the HTC, along with economic items (reading time and correction time), and is performed to validate a new system version or compare it with other systems. The microevaluation identifies translation errors and provides information to plan the production of an improved version of the system.

In constrast with (Carroll, 1966), the report suggested scales for intelligibility with just 4 grades, from *complete unintelligible* to *very intelligible* (*basely* and *fairly intelligible* in between). Fidelity was also measured with a 4-grade scale with descriptions ranging from *completely or almost completely unfaithful* to *completely or almost completely faithful*.

As for acceptability, this item indicates the shift from the Turing-test oriented evaluations to the context-of-use orientation the report points out. Acceptability is measured by asking the evaluator whether the translation quality is acceptable, taking into account the fact that the translator is a machine, and appreciating its capacity to perform fast translations and its suitability to the context of use.

2.1.1.3 The JEIDA contributions

In the 1980s, machine translation in Japan was emerging. The electronic industry was growing and exportation demanded the fast translation of huge amounts of documentation, user guides, and so on. In 1988 The Japan Electronic Industry Development Association (JEIDA) commissioned a report on the machine translation state of the art from 1966 onwards (Nagao, 1989). The report actually took the ALPAC report as a refence, which was explicitly stated in the title *A Japanese view of machine translation in the light of the Recommendations and Considerations Reported by ALPAC, U.S.A.*.

The conclusions of the JEIDA report were diametrically opposed to the ALPAC conclusions. Despite limitations in linguistic coverage, JEIDA reccomended to invest in machine translation because state-of-the-art systems were useful, and able to provide the gist of a topic so that the user could understand it without much effort. According to the report, systems were capable of satisfying a 800 billion ien market. Therefore utility displaced the human versus machine translator comparison as a basis for evaluations.

In the early nineties, the JEIDA worked on creating an evaluation method that answered the question *is it worth working on machine translation?*(JEIDA, 1992). This method had to consider all the necessary costs and confront them with profits. After modelling the user context, a number of questions were elaborated to be answered by evaluators. The questions were associated to the parameters of the user context model (translation needs, type of document, language pairs, time, installation conditions, etc.). The best system was the one that best fitted the user context model. Besides, the method also measured the relationship between the action of the system and the user′s satisfaction.

2.1.1.4 The DARPA Evaluation Campaigns

In 1991 the DARPA (Defense Advanced Research Projects Agency) wanted to establish a methodology of evaluation that was suitable for any system, regardless the language pair, translation environment, and engine (rule-based, statistical or human/not human assisted) ((White, 1994), (White, 1995)). The aim was to attain an overview of the translation quality of the systems they sponsored, which were quite heterogeneous.

Intelligibility and fidelity were the main translation qualities because their appreciation is independent from the language pair, the use context and the engine method. Fidelity was measured according to two criteria: adequacy and informativeness, whereas intelligibility was measured according to fluency. Adequacy was the degree in which a segment of the machine translation kept the meaning of the source segment. Informativeness was the degree in which the translation contained enough information to ask about. In order to measure this, evaluators had to answer some questions about the content of the text. Finally, fluency was the degree in which the translation was well formed according to the grammar rules and conventions of use in the target language. Both adequacy and fluency were graded in a 1-5 scale as is shown in table 2.1.

Score	Adequacy	Fluency
5	All information	Flawless English
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensible

Table 2.1: DARPA scale for Adequacy and Fluency

The DARPA evaluation campaign carried out in 1994 disclosed the problems of HTC evaluations based on subjective appreciations. There was little agreement in the subjective appreciations, so many data had to be collected and interpreted in order to get significant results. Besides, the costs were significant. For instance, each source was translated by a human translator because evaluators were not bilingual and had to appreciate adequacy by comparing the machine translation with the human translation. The size of the evaluation corpus augmented with the addition of human translations, which were used as control segments to assure that the appreciations were consistent regardless of the origin (human or machine). The process was slow and expensive. Many evaluators were involved, many translations had to be evaluated, and many human translators were enrolled to produce the human version for each machine translation. However, despite the efforts, the data obtained from human assessments were difficult to reuse. Yet a positive outcome was the huge corpora of machine translations paralleled with human translations.

2.1.1.5 The FEMTI

In 1993 the European Commission was concerned in standarising linguistic technology production in order to speed up the creation of new products and their transfer to other projects. The Expert Advisory Group on Language Engineering Standards, known as EAGLES, was created and an evaluation working group designed a common method to evaluate natural language processing systems and products. For MT technology, evaluations assess systems according to HTC quality standards (consistency, fidelity, wellformedness, etc.), and also guide potential consumers to decide which system to use. Assessment and guiding criteria focus on the suitability of systems to their specific purpose, following the trend of previous approaches such as JEIDA′ and Van Slype′s and especially the ISO 9126 standards (ISO, 1991). Even bad and crummy machine translations are considered to be acceptable if front-end users prefer to postedit them rather than translating from scratch (Church and Hovy, 1993) as Wagner already said in 1985 when considering postedition costs of Systran systems (Wagner, 1985). The evaluation framework for machine translation systems is the FEMTI[1], which guides evaluation according to the context of use (Hovy et al., 2002a); (Hovy et al., 2002b)). For instance, in (Bruckner and Plitt, 2001) the evaluation is set in an evironment where translation memories are used.

The FEMTI organizes quality features in taxonomies (Popescu-Belis et al., 2001). The features of one taxonomy defines the use of the system (e.g. user profile, task, input). The second taxonomy defines the characteristics of the machine translation system. The taxons of the second taxonomy are pointed from taxons of the first taxonomy, and have metrics and measures which indicate the suitability of the system to the context of use. The more positive values, the better the system. The evaluation leader can choose the most suitable method to obtain these values. The FEMTI provides a collection of methods for each taxon.

Although FEMTI provides well-defined guidelines to perform an evaluation, the taxonomies and the pointing from axons from the first taxonomy to the second reflect the complexity of a rigurous method of human evaluations. Standard procedures demand costly actions (e.g. methods that avoid subjectivity, impact calculation of lexical and semantic errors (Marrafa and Ribeiro, 2001)).

2.1.2 HTC in automatic evaluations

Automatic evaluations appeared as an alternative to human evaluations, which proved to be very expensive in time and money and the results were not as objective as they

[1] http://www.isi.edu/natural-language/mteval/

were expected (c.f. 2.1.1.4). Since the early 2000s, automatic evaluation methods have been used to obtain objective and reliable results in a very short time. In this section we will explain three assumptions that have supported automatic evaluation methods so far:

- The Reference Proximity Assumption (RPA)
- The Accuracy Assumption (ACA)
- The Human Likeness Assumption (HLA)

These assumptions are relevant because, as we will explain more fully in this section, they are the principles on which the main state-of-the-art methods of automatic evaluations are based.

Finally we will discuss how translation quality is assessed in automatic evaluations.

2.1.2.1 The Reference Proximity Assumption (RPA)

The reference proximity assumption (RPA) is formulated as follows: *the closer a machine translation is to a professional human translation, the better it is* (Papineni et al., 2002). The human translation of the original is the quality reference to evaluate the machine translation, in the same way as the human translation was also the reference for evaluating fidelity and adequacy in ALPAC and DARPA methods (c.f. 2.1.1.1 and 2.1.1.4) The quality degree is expressed with a metric obtained by an objective method: the distance between the machine translation, called **hypothesis**, and the human translation, called **reference**.

a) Translation reference and legitimate translation variations

If the distance were calculated with only one reference, good translations would be unfairly assessed as bad translations because legitimate translation variations (LTV) have not been taken into account. For instance, a hypothesis where words appear in a different order from the reference can be a legitimate variation. Let's see the example of the hypothesis in Catalan *li agrada la xocolata* ('he likes chocolate') and the reference *la xocalata li agrada* ('he likes chocolate'). Both translations are good but the different word order would penalise the hypothesis if LTV were not considered.

b) Metrics for measuring the distance between the hypothesis and the reference

The most common metrics are the ones called n-gram metrics, or methods based on the lexical similarity between a machine translation and one or more human references. We will summarise them according to the groups of measures explained in (Giménez, 2008).

b.1) Edit Distance Measures

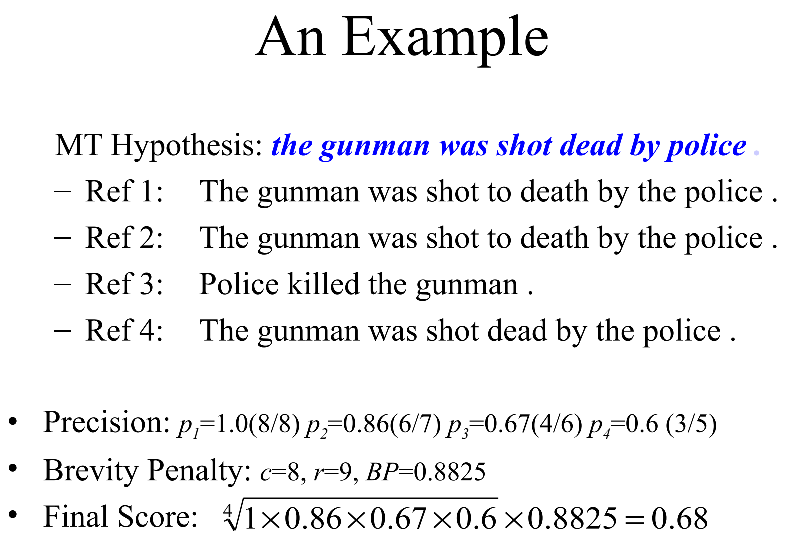
Edit Distance Measures are based on the number of changes to be applied to convert the machine translation into a reference. The most used metrics is the **WER** (Word Error Rate) (Nießen et al., 2000). WER is the Levenshtein distance (Levenshtein, 1966), which is the calculation of the minimum number of substitutions, deletions and insertions to be performed to convert the machine translation into a reference. WER, however, does not permit reordering of words, which is a drawback when facing LTV. So the **PER** (Position-Independent Word Error Rate) (Tillmann et al., 1997) is used when the edit distance is calculated regardless of word order.

Another solution for the LTV handicap is to calculate the distance of the hypothesis by collecting a set of references, as the **mWER** does. The mWER (**multiple** **W**ord **E**rror **R**ate) is calculated by obtaining the Levenhstein distance of the hypothesis with respect to each reference and the final result is the shortest distance. Another measure based on the Levenhstein distance with multiple references is **RED** (Ranker based on Edit Distances) (Akiba et al., 2001).

b.2) Precision-oriented Measures

These measures convey the proportion of n-grams in the hypothesis which are also present in a set of reference translations. The number of items in the n-gram generally ranges from 1 to 4 words. Therefore, the calculation of these measures is based on precision with respect to a representative sample of legitimate translations. **BLEU** (Bilingual Evaluation Understudy) (Papineni et al., 2002) has become the standard in automatic evaluation. In figure 2.1 an example of BLEU calculation is shown (Way, 2004). In this calculation, the precisions of unigrams (*p1*) bigrams (*p2*), trigrams (*p3*), and so on are combined. On the other hand, a *brevity penalty factor* (BP in the figure) is introduced for hypotheses that are shorter or as long as references, just to avoid examples such as *the the the* being regarded as good.

The National Institute for Standards and Technology (NIST) (Doddington, 2002) created the **NIST**, which is very similar to BLEU but weighs n-grams that are not very frequent because they afford high informativeness. Another criterion to weigh n-grams is statistical salience (Babych and Hartley, 2004), such as tf.idf for relevance and S-Score



An Example

MT Hypothesis: ***the gunman was shot dead by police .***

- Ref 1: The gunman was shot to death by the police .
- Ref 2: The gunman was shot to death by the police .
- Ref 3: Police killed the gunman .
- Ref 4: The gunman was shot dead by the police .

- Precision: p_1=1.0(8/8) p_2=0.86(6/7) p_3=0.67(4/6) p_4=0.6 (3/5)
- Brevity Penalty: c=8, r=9, BP=0.8825
- Final Score: $\sqrt[4]{1 \times 0.86 \times 0.67 \times 0.6} \times 0.8825 = 0.68$

FIGURE 2.1: Example of BLEU calculation (Way, 2004)

for the degree of pertainance of a n-gram to a specific document. As salient n-grams are expected to be common in all translation variations, the method allows to consider just one reference.

Since (Papineni et al., 2002) experiments have been carried out to prove that values correlate with HTC-based quality perceptions. This correlation is interpreted as *human acceptability* (Giménez, 2008).

b.3) Recall-oriented measures

Recall-oriented measures compute the proportion of n-grams in the references covered by the hypothesis. The aim is to capture translation quality, especially fluency, at one-sentence level, not at multi-sentence level as precision-oriented measures do.

ROUGE (**R**ecall-**O**riented **U**nderstudy for **G**isting **E**valuation) is a recall-oriented measure ((Lin, 2004), (Lin and Och, 2004a)) that was compared with BLEU, NIST, WER and PER for the ORANGE (**Or**acle **Ran**king for **G**isting **E**valuation) method, a method to automatically evaluate automatic metrics (Lin and Och, 2004b). ROUGE is, in fact, a package of measures (ROUGE-L, ROUGE-W and ROUGE-S) based on two notions: the *longest common sequence (LCS)* and the *skip-bigram co-occurrence*.

A *sequence* S is a subsequence from another sequence Z extracted by eliminating some of the elements of Z but not altering the relative position of the remaining elements. For instance,

[<¡B, C, D, B>]

is a subsequence of

[<A, C, B, D, E, G, C, E, D, B, G>]

As we see, elements of the subsequence are not necessarily contiguous with respect to the sequence Z. As for common subsequence, given two sequences X and Y, G is a *common subsequence* of X and Y if G is a subsequence of X and Y as well. Finally, the LCS will be the longest common subsequence of X and Y. The intuition is that the longer the LCS of two translations is, the more similar the two translations are. On the other hand, the more consecutive matched elements in the LCS the better. So, given a set of hypothesis to evaluate, the one with the longest LCS, with more consecutive matched elements (ROUGE-L), will be the best. ROUGE-W is a metric that favors consecutive LCSes.

As (Lin and Och, 2004b) point out a *skip-bigram is any pair of words in their sentence order, allowing for arbitrary gaps. Skip-bigram co-occurrence statistics measure the overlap of skip-bigrams between a candidate translation and a set of reference translations.* Therefore the more co-occurrent skip-bigrams the better (ROUGE-W).

ROUGE intends to capture differences in fluency due to grammatical structure. For instance, let′s see the following example:

S1 (Reference): police killed the gunman

S2 (Hypothesis): police kill the gunman

S3 (Hypothesis): the gunman kill the police

ROUGE-L scores better S2 than S3[2]. However, the BLEU value for bigrams would be the same.

b.4) Measures combining precision and recall

When the hypothesis is equal to a reference both precision and recall are also equal, so the highest matching degree between hypothesis and reference corresponds to a good translation. When there is no coincidence, the F-measure is taken into account because it indicates the degree of precision and recall. The F-measure is the key value for metrics such as **GTM** (**G**eneral **T**ext **M**atching) and **METEOR**. GTM (Melamed et al., 2003) relates precision and recall in unigrams. METEOR ((Lavie et al., 2004), (Banerjee and Lavie, 2005)) also accounts for word ordering and calculates matchings between hypothesis and references in lemmas and synonyms, based on WordNet (Fellbaum, 1998).

[2] ROUGE allows for word-stemming, as in *kill* and *killed*

Other measures are BLANC (**B**road **L**earning and **A**daptation for **N**umeric **C**riteria) (Vlad et al., 2005) and SIA (**S**tochastic **I**terative **A**lignment) (Liu and Gildea, 2006).

b.5) The syntactic approach

Translations should be evaluated as wrong because of their ungrammaticality. However, all the n-gram metrics have difficulties in capturing grammaticality at the sentence level. (Liu and Gildea, 2005b) present a method to calculate fluency based on a previous syntactic representation of the sentence, which is a dependency tree. The method is based on the notion of HWCM (Head Word Chain Matching), and calculates the fraction of matching head word chains of a given length between the hypothesis and the reference trees. A headword chain is defined as a sequence of words which corresponds to a path in the dependency tree.

HWCM has variants, such as the ones in the Asiya MT Evaluation Toolkit[4], which distinguish HWCM of words (HWCM-w), grammatical relations (HWCM-r) and grammatical categories (HWCM-c). So DP-HWCMw-4 retrieves the matching proportion of length-4 word chains and DP-HWCMc-3 computes average accumulated proportion of category chains up to length 2.

Other variants are provided by (Amigó et al., 2006), who present the following three metrics:

- TREE: overlapping between the words hanging from non-terminal nodes of a type tree, for instance a predicate of a clause
- GRAM: overlapping between the words with a gramatical category, for instance the word overlapping between adjectives or adverbs
- LEVEL: overlapping between the words hanging at a certain level of the tree, or deeper.

2.1.2.2 The Accuracy Assumption (ACA)

The accuracy assumption (ACA) can be explained in these terms: good translations are those that are accurate with respect to the source or the translation reference in the sentence level. Evaluating sentence accuracy is not new as we have seen in human evaluations. The novelty is the automatic calculation of semantic similarities between machine translations and references(Giménez and Márquez, 2008). The similarities are over named entities, semantic roles and other basic items of a semantic tree.

[4] http://asiya.lsi.upc.edu/

Semantic analysis still have several drawbacks. On the one hand, semantic parsers have been mainly developed for English and Chinese ((Palmer et al., 2005), (Fung et al., 2006), (Liu and Gildea, 2005a)). Therefore, this approach cannot be performed for other languages until reliable semantic parsers are produced. On the other hand, the approach is very complex, and, as (Lo and Wu, 2010a) consider, semantic labels are difficult to annotate automatically and deterministically. Therefore, Lo and Wu suggest the *utility approach*. This approach consists in transferring to humans the task of checking whether the most basic items in the semantic analysis of the machine translation are properly translated. By doing this, humans evaluate whether the translation is useful for a reader to *succesfully understand at least the basic event structure (who did what, to whom, when, where, and why)*. This reduces the costs of collecting references (luwo2011).

2.1.2.3 The Human Likeness Assumption (HLA)

A methodological alternative to evaluating with reference translations is based on the human likeness assumption (HLA). According to the HLA, a machine translation that resembles a human translation is good.

A method based on the HLA evaluates translation quality straightforwardly, and does not depend on an indirect procedure such as the comparison between the translation and a set of reference strings. On the other hand, (Reeder, 2001) says that there is no need to compile a huge amount of output to distinguish a system that performs like a human the same way as there is no need to read many texts to distinguish a native speaker from a learner. So the assumption suggests saving the costs of elaborating reference translations and confectioning huge evaluation corpora.

We will present two HLA strategies. The first one consists in regarding evaluation as a classification problem. A human/non human translation classifier is trained to perform this task. The second strategy, despite not being strictly based on RPA, takes advantage of RPA measures to identify the features that make machine translations resemble human translations.

a) Human/non-human translation classifier

The strategy turns evaluation into an automatic classification problem. Given a translation, how exact is our prediction about its human/non-human nature? The evaluator would emulate the human capacity of distinguishing a human translation. In fact, the proposal is the counterpart of a Turing Test: it is the machine, not a human being, the one who must recognise a human translation (Jones and Rusk, 2000)).

(Corston-Oliver et al., 2001) used decision trees to distinguish ′good′ translations (human-generated) from ′bad′ translations (machine generated). (Kulesza and Shieber, 2004), as in (Jones and Rusk, 2000), suggest that the problem can be solved via machine learning. They trained a support vector machine (SVM) that produced a separator that splitted a feature space in two. One half space corresponded to features of machine translations and the other half corresponded to features proper of human translations. The features were numerical and were data obtained from hypothesis and reference translations, inspired in RPA metrics (e.g. fraction of hypothesis n-grams appearing in any reference, word error rate between hypothesis and references). Given a translation example, the classifier computed the side the example fell to and also the distance to the boundary between both spaces. Worse translations were to be found in the machine translation side and far from the boundary. The results correlated better with human sentence-level assessments than RPA methods.

(Gamon et al., 2005) followed the Kulesza and Shieber′s approach but their proposal uses a language model instead of a set of human references. This proposal was motivated by the fact that a method for posteditors to detect wrong translations with no references was challenging.

b) Human Likeness and combination of RPA measures

Human Likeness is suggested by (Amigó et al., 2006) as a meta-evaluation criterion that captures syntactic improvements which are not captured by any single RPA measure. Their proposal is to combine the RPA measures that are good to distinguish machine translations and human translations in one metric. The rationale for regarding a metric as good is the following: all reference translations are good, so the best metrics are those that identify and use features which are common in references and distinguish them from machine translations. So given a translation reference T, the best metric is the one that makes T more alike to any other reference than a machine translation that would not be a reference.

2.1.2.4 Translation quality in automatic evaluations

The main human translation qualities assessed in automatic evaluations are fluency and adequacy. Both qualities are assumed to be assessed although the metrics are more fluency or adequacy oriented. For example, ngram based and syntax based metrics are regarded as fluency-oriented (Lo and Wu, 2010b). However, the calculation of the distance to translation references, which are assumed to keep the meaning of the source, theoretically implies also the adequacy assessment. For example, (Papineni et al., 2002) states, *a translation using the same words*[5] *(1-grams) as in the references tends to satisfy adequacy. The longer n-gram matches account for fluency.* When no references are used, a metric that assesses for both fluency and accuracy is the one we suggest in this thesis. Translations rated as machine-like are related to the perception of non-fluent translations and also to the suspicion of lack of fidelity to the source.

2.2 The shortcomings of the HTC

An important shortcoming of HTC methods is their costs. In human evaluations, to obtain objective results, despite subjective appreciations of translation quality, is very expensive (c.f. 2.1.1.1 and 2.1.1.4). Besides, context-of-use-oriented methods such as FEMTI (c.f. 2.1.1.5) turn each evaluation a new problem, which demands a complex and costly design from scratch. We already explained how automatic evaluations arose as a cheaper and faster alternative (c.f. 2.1.2). However, costs are also considerable in automatic evaluations, both in RPA, accuracy and HLA methods.

The representativeness of translation variations for RPA methods implies spending time and money to obtain references. According to (Callison-Burch et al., 2008), the expenses to obtain reference translations for the evaluation campaign in 2008′s ACL Workshop on Statistical Machine Translation amounted to 17,200 euro. Besides, other expenses should be taken into account: hiring professional translators or financing training courses for translators to adopt evaluation-oriented guidelines which are different from the ones they are used to (Cully and Riehemann, 2003). For instance, to translate as literally as possible, and keep the same syntactic structures as the source if the target grammar allows it. Translators must be aware that MT systems cannot be creative and have no stylistic taste.

An important expense, which is generally overlooked, is the revision of references. The revision of all the references should be budgeted because it is necessary to check whether there are some inadequate references that may unfairly affect the evaluation result. However, this revision is hardly carried out because of what (Jelinek, 2004) calls *the human translation myth*. By influence of the human translation myth, references are not revised because their appropiateness is taken for granted. Nevertheless, not always does reality match the truth. Human translators also use machine translation systems and may send a machine translation (revised or not) as if it was theirs. On the other hand, some human translations may be worse than machine translations because of

[5] We should add 'synonyms' as in METEOR. However, there are examples that contradict this assumption. For instance, in Spanish the meanings of *la guerra fría* (the Cold War) and *la fría guerra* (the merciless/bleak war) are not similar at all.

factors beyond translator professionality. Some decisions- sometimes too conservative or, on the contrary, too risky- may affect the reference idioneity. This is generally the case when translating idioms. If the idiom is in the bilingual dictionary of the system, the machine translation will be better than the translation by a professional who, for not having understood the context, translated it literally or too freely.

Another important expense which is seldom taken into account is the compilation of references from different subject domains. Such compilation would guarantee that the quality of the system is independent of the domain chosen. As references for each hypothesis are difficult to compile in different domains, it is usual to reuse as references the corpora that served for training the system. This is the case of stochastic machine translation systems that were evaluated with references from Europarl, which also provided the corpus to train the system(Koehn and Schroeder, 2007). However, no conclusion could be drawn about their performance in a different domain (Offersgaard et al., 2008).

Alternative methods to save the costs of references have not proved cheaper. Automatic evaluation of accuracy (c.f. 2.1.2.2) is so far limited to very few languages and the classification task of HLA methods (c.f. 2.1.2.3) is very expensive. The reason is the huge cost of creating a training corpus for the human/non-human classifier. For instance, in (Gamon et al., 2005) the training corpus amounted to 198,771 machine translations and 260,601 human translations produced by Microsoft. This bulk of translation is cheaply available if the developer works in an institution or a company like Microsoft with an overwhelming production of human and machine translations. On the other hand, the classifier learns from domain-specific corpora. If the sources were from other domains the results would probably be different.

The training data for SVM classifiers must be provided by mature nature language processing (NLP) tools (e.g. parsers and PoS taggers). Most NLP tools have been developed for few languages, so the method is not exportable to other languages than English, Spanish or Chinese. Yet even when their performance is quite acceptable, the consequences of the errors that must be assumed in any automatic processing are difficult to prevent. Manual revision of the training data would make HLC methods very expensive to be carried out.

Apart from high expenses another important shortcoming affects the reliability of automatic measures. These measures would be reliable if the references or the training corpus would contain all the legitimate variations of a translation. However, it is not possible to list aprioristically all the legitimate variations. Therefore, there always will be good translations scored as bad just because these translations were not in the references nor in the training corpus.

Finally, reusability of the results is another drawback. When explaining the DARPA method we already mentioned the non-reusability of results (c.f. 2.1.1.4). The same stands for automatic evaluations metrics. Measures such as BLEU (**Bi**Lingual **E**valuation **U**nderstudy), NIST and so on say little about the system′s limitations and errors and they are not informative enough about the aspects to be improved.

2.3 The machine translationness criterion and its contributions

We called the machine translationness criterion (MTC) the criterion by which a machine translation is evaluated according to qualities of machines. From this criterion we draw the machine translationness badness assumption (MTBA) which we will define as follows: *translations with machine translationness are bad.* Therefore, a translation with MTness will be worse than a translation that resembles a human translation or its degree of MTness is minor.

We hypothesize that MTC evaluations are more consistent than HTC evaluations because translations are evaluated according to *what they are* (sentences generated by a machine) and not to what they resemble (sentences written by a human being). Besides, the MTBA is an assumption which is actually generally agreed.

MTBA and HLA seem to be two sides of the same coin. To say that a translation resembling a machine translation is bad entails saying that a translation that resembles a human translation is good. In fact, we have presented approaches such as (Corston-Oliver et al., 2001), (Kulesza and Shieber, 2004) and (Gamon et al., 2005) that identify ′machine translations′ as ′bad translations′, which is the counterpart of the ′human translations - good translations′ identification. However, instead of seeking human resemblances, the MTBA approach seeks machine resemblances and this shift causes cost reductions in time and money. A MTBA-based evaluation need not references, just the detection of the linguistic phenomena that produce MTness. These phenomena are automatically detected thanks to an MTness typology and, once the resources necessary to perform the detection are ready, evaluations can be performed repeatedly with no costs. In chapter 7 we will prove that the MTness detection can be performed with freely available resources.

On the other hand, MTBA-evaluations save time considerably. Following the idea in (Reeder, 2001), fast MTBA-evaluations can be performed with short texts because in just a few words we may appreciate whether a system translates worse (in terms of MTness) than another. Only when the evidences of an MTBA-evaluation are not clear enough should a more exhaustive evaluation be performed. On-the-fly MTBA evaluations would then save time and money compared to traditional evaluations. This is in line with the quality estimation (QE) trend (Specia et al., 2010). QE is useful to predict costs and to know beforehand if it is worth the effort of fixing the translation (e.g. postedition) to get a human quality version. QE defines quality according to the task at hand, which is close to the FEMTI contexts of application. Besides, the quality standards are machine learned, so the prediction depends ultimately on the domain(s) of the learning corpora in QE. However, in the MTBA approach, quality is not defined according to the task and context of use. A translation with MTness is bad, regardless the domain and context of use. Moreover, MTBA evaluations have a wider scope, not only to be applied to save production costs (c.f 1.1), and MTness ratings do not depend on machine learning techniques[6]. Although QE and MTBA are not the same, MTBA-evaluations have a predictive power about translation quality which is important to be taken into account by QE.

Another important contribution of MTBA-evaluation is reusability. The list of detected MTness phenomena is very useful to confectionate automatic postedition modules, which substitute the MTness instance for the right translation solution. Besides, this list provides with very useful information to improve the system, so MTBA-evaluations are suitable for performing microevaluations. On the other hand, resources can be easily updated with new MTness instances not detected by the MTness detector.

Table 2.2 shows the advantages of MTC compared to HTC in automatic evaluations.

HTC	**MTC**
References and training corpus	No references and training corpora
Long compilation and revision of evaluation and training corpora	Fast evaluations for a small evaluation corpus
Hiring and formation costs for translators	No human translators required
Macro and microevaluations are not intertwined	Macro and microevaluations can be closely related
Results are not reused	Results can be reused for other tasks (e.g. postedition)
More costly resources	Free linguistic tools and resources
Subject-domain-dependent	Subject-domain-independent

Table 2.2: Advantages of MTC compared to HTC evaluations

[6] Some labels from which the QE systems learns and predicts are postedition-cost centered (Specia, 2013)

2.4 Conclusion

In this chapter we have presented the Human Translationness Criterion as the criterion on which machine translation evaluations, both human and automatic, have been based. Evaluations in terms of what machine translations resemble (human translations) involve huge costs in time and money when compiling the necessary data to rate objectively the *nearness* of the machine translation to a human production. Other shortcomings of the HTC are reliability and reusability.

In contrast, we propose the Machine Translationness Criterion. This criterion establishes evaluating machine translations according to what they are- translations generated by a machine. This approach saves much of the costs derived from testing the human resemblance and qualities. Apart from that, costs are also saved because our proposal depends on detecting MTness linguistic phenomena with resources that are freely available or demand very low cost. Detection of linguistic MTness phenomena is important because the list of MTness instances can be reused for other uses (e.g. automatic postedition module).

Chapter 3

MTness Experimental Study

As said in the previous chapter, MTC evaluations depend on the detection of MTness linguistic phenomena. However, MTness may seem an impression that cannot be embodied in a list of concrete and typified linguistic features. MT error typologies certainly exist but they have not been elaborated to explain why a translation sounds as it had been generated by a machine. Therefore, we had to answer two methodological questions, which were *can MTness be valued by detecting concrete linguistic features?* and *can MTness features be typified?*. We also had to answer two more methodological questions: *does the perception of these types only affect machine translations?* and *is MTness perception universal enough so that it is perceived regardless the informant′s background and the system′s methodology?*. These questions will be dealt with in chapter 4 and chapter 5. This chapter is devoted to describe the experimental study we performed to answer them. The study allowed us to analyse the perception of MTness by people with different learning backgrounds and different reading skills. So we put MTness perception as an experimental object of study.

The chapter is organised as follows: in section 3.1 we explain the goals of the study and in section 3.2 we explain the methodology and how it was carried out.

3.1 Goals of the Empirical Study

The goals of the empirical study were the following:

Goal 1 To assess there are MTness linguistic instances that can be typified

Goal 2 To assess that MTness perception only affects machine translations

Goal 3 To assess that MTness perception is not dependent on contingent factors

The fulfilment of these goals have the evidences shown in Table 3.1. Henceforth, these evidences will be referred to with a reference code.

Goals	**Evidences**	**Reference Code**
Goal 1	1. MTness typology	E11
Goal 2	2.1. Agreement among informants, regardless of their background	E21
	2.2. Types most agreed are particular of MT	E22
Goal 3	3.1. MTness perception is not language-pair dependent	E31
	3.2. MTness instances are not the errors of a particular MT methodology	E32

Table 3.1: Goals and evidences of the experimental study

3.2 Empirical Study Methodology

The empirical study consisted in performing a sort of Turing test. 100 people read a number of machine and human translations. For each translation they had to guess whether the translation had been produced by a machine or by a human. If the former, they had to underline the pieces that sounded machine-like for them.

Our main concern was to typify the linguistic phenomena responsible for MTness. The MTness phenomena had to be detected by at least two people, regardless of their reading skills and learning background. For this reason, each translation was read by three people with different cultural backgrounds, individually and in isolation. The typology was built by analysing the segments of the same translation that at least two people underlined. The experiment was performed without any computational support in order not to condition the experiment to technological skills.

Detailed information about the informants, the corpus, and the source and the target languages is provided in the following subsections.

3.2.1 The Informants

The informants were people living in Catalonia, literate and, in order to avoid bias in the results because of expertise, they were not language experts and were not familiar with computational linguistics. Each translation was evaluated by three people of different ages and levels of reading comprehension. These levels were established according to their reading habits for both professional and leisure interests, and also according to studies, as we assumed that the higher the level of studies, the higher their reading skills were to understand textual complexity and abstract contents.

Data relative to the informants are shown in the following tables.

Group Distribution According to Age			
From 16 to 30	From 31 to 45	From 46 to 60	61 and older
25	25	25	25

Table 3.2: Data about the age of informants

Group Distribution According to Educational Level			
No Studies	Primary and Secondary Studies	Secondary Studies (non compulsory)	University Studies
25	25	25	25

Table 3.3: Data about the educational level of informants

Group Distribution According to Sex	
Male	Female
25	25

Table 3.4: Data about the sex of informants

In order to avoid age bias, informants were balanced according to age groups. Each group spanned 15 years of age, and included individuals with reading skills acquired in similar educational systems. People of 61 and older were in the same group because their educational systems when they were young did not differ much as far as reading skills were concerned. The informants were older than 16 because we considered that people under this age are still developing their reading skills.

The groups were balanced in age and gender but it was not balanced in educational level, where people with university degrees amounted to 25%[1]. The reason was to avoid bias caused by a large number of participants with highly elaborated reading skills.

The number of informants was settled according to the data saturation technique (Guest et al., 2006), which consists in interviewing people as far the results do not vary much despite widening the sample. With a number of 100 informants we saw that MTness phenomena could be profiled.

[1] In order to obtain a balanced sample of university graduates, the number of people with university degrees were distributed according to whether their degree belonged to technology, natural sciences, social sciences, or humanities.

3.2.2 Corpus of the Experiment

The corpus consisted of 3750 translations. The elaboration of the corpus took into account the evidences of the goals we pursued (see Table 3.1), as we will show.

Criteria for getting evidence E11

In order to obtain data about machine translation perception for an MTness typology, we de-contextualized the translations. MTS translate sentence by sentence with no contextual information, so we wanted the informants not to fill in their comprehension gaps with the help of the context. Contextual interpretation distinguishes humans from machines so we did not want this human capacity to condition the results. Therefore, translations were single sentences with no contextual relation to the previous and the following sentence. Under these conditions, we assumed that machine-human differences would be more highlighted.

Criteria for getting evidence E21

In order to assess agreement in MTness perception across informants, each machine translation was replicated three times. We wanted three people, with different age and level of studies, to read each translation. Then we measured their agreement. Each informant read around 38 translations, which was a reasonable number. A larger number would have caused fatigue and lack of attention that would have undermined objectivity.

We wanted translations to be comprehensible for the informants, regardless of their learning and professional background. So we collected translations from news and tourism magazines. Moreover, in order to check how the knowledge of a domain may influence the detection of MTness, we mixed in sentences from articles about computers, economics, speeches from the Europarl corpus, and provisions and acts published in the official gazette of the Catalan government. These sentences were written to be understood by the general public.

Criteria for getting evidence E22

In order to assess whether MTness instances with the most agreement were not found in human translations, we included 750 human translations, which corresponded to 25% of the amount of machine translations (3000).

Criteria for getting evidence E31

In order to assess whether MTness perception depends on the language pair we decided to collect the translations of a close language pair and also the translations of a distant language pair. The close language pair was Catalan-Spanish and the distant language pair was English-Spanish. The analysis of the language-pair influence on MTness perception did not take into account whether the informants knew both languages, since they only read the Spanish translation. However, the distribution of these informants in groups, as we explained before, was planned to obtain data from informants with an equivalent level of perceptiveness with regard to lexical or syntactical congruency.

The reasons why we chose Spanish as the target language were, on the one hand, the availability of MT systems with different methodologies for Catalan->Spanish and English->Spanish directions, which was important to get evidence E32. On the other hand, the informants lived in Barcelona and most of them felt more confident in judging translations in Spanish than in Catalan, let alone in English. Older participants′learning was basically in Spanish, as teaching and publishing in Catalan was prohibited when they were children.

Criteria for precondition E32

In order to assess whether MTness instances were errors of a particular MT methodology, we collected 250 sentences in Catalan and 250 sentences in English anb both the Catalan and the English sentences were translated into Spanish by a rule-based system (RBMT) and by a statistically-based system (SBMT). Thus, we collected 1000 machine translations (250 from the Catalan-Spanish RBMT system, 250 from the English-Spanish RBMT system, 250 from the Catalan-Spanish RBMT system and 250 from the English-Spanish SBMT system), which were replicated three times, amounting to 3000.

3.2.3 The survey

The survey was performed without any computational support in order not to restrict the informants to those with technological skills. Therefore, in order to process the data, all the translations, along with the segments underlined, had to be manually transferred onto a digital support.

On the other hand, since many informants were not capable of reflecting on linguistic aspects and explain the MTness phenomena they perceived, they were told to just underline the segments they thought as generated by a machine.

3.2.4 Collection of Data

Data were organized in contrast units. A contrast unit is a tuple that registers the perception of MTness in one translation by three informants. *Contrast* indicates that the translation segments underlined by the informants were contrasted. The tuple contains the following items:

Identity Tag A tag that describes the translation

Interview Ids The codes of the interviews where the translation appeared

MTness Segments Segments the informant underlined as not produced by a human translator.

The Identity Tag contains the following information:

- Id: A mumerical code that identifies the tuple.
- Translation: A code that indicates whether the translation was human or generated by a machine
- Methodology: Methodology of the MT system (RBMT or SBMT).
- MT System: The name of the MT system.
- Language Pair: Source and target languages.

This is an example of Identity Tag

<172,mt,rbmt,XMT,en-es>

which means that tuple 172 corresponds to a machine translation generated by the system XMT (invented name), which is a RBMT system for the English-Spanish language pair.

Figure 3.1 shows a schematic representation of a contrast unit. The segments that were underlined by at least two informants are in bold between the <MTN></MTN> tags. The segments underlined by at least two people were analyzed in order to discern the linguistic phenomena that caused their bewilderment and typify them.

Contrast units allowed us to see the different precision among informants when underlining the segments. Sometimes informant A and informant B underlined two different segments and informant C underlined a segment that included both. This raised the dilemma whether the number of linguistic phenomena detected were 1 or 2. We decided the following: if an informant underlined a segment which contained segments underlined separately by other informants, the number of coincidences of each isolated segment was increased in one.

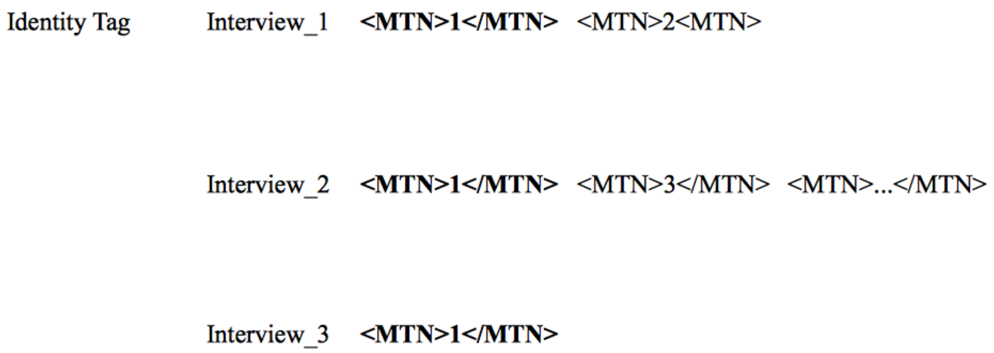


Figure 3.1: Schematic representation of a contrast unit

3.3 Summary

As a summary, table 3.5 presents the goals of empirical study, their evidences and the procedures .

Goals	Evidences	Procedures
Goal 1	1. MTness typology	Analysis of segments underlined by at least two informants from decontextualised translations. Segments underlined are the translation pieces they think not to have been generated by a human translator
Goal 2	2.1. Agreement among informants 2.2. Types most agreed are particular of MT	3 informants, with different ages and level of studies, test and underline each translation Comparison of segments underlined in human and machine translations
Goal 3	3.1. MTness perception is not language-pair dependent 3.2. MTness instances are not the errors of a particular MT methodology	Comparison of segments underlined in Catalan-Spanish translations and English-Spanish translations Comparison of segments underlined in translations by a RBMT system and translations by a SBMT system

Table 3.5: Goals, evidences and procedures of the experimental study

Chapter 4

MTness Typology

In this chapter we will present the MTness typology drawn from analysing the segments underlined by the informants in the experimental study (c.f. 3.2.3, 3.2.4). Each translation was read by three informants and the segments studied were those that were underlined by at least two of them. As we explained in the previous chapter, the typology evidences the fact that MTness can be valued by detecting concrete and typified linguistic features. However, the typology describes the MTness instances of the language of study- Spanish. Although we avoided to typify phenomena that are dependent of this particular language, the fact that the typology can be applied to all language pairs is not proved yet.

In section 4.1 we will explain the MTness types. Each type is followed by a code, which will be used in later references. The types are exemplified with translations into Spanish because they are taken from the corpus of the experimental study. In section 4.2 we discuss the pertinence of the typology.

4.1 MTness Types

The typology has 13 types classified in four groups:

- Lexical
- Syntactic
- Semantic
- Formatting

4.1.1 Lexical types

Lexical MTness instances belong to one single lexical type: words not pertaining to the target language.

4.1.1.1 Word not pertaining to the target language (NO-L2)

NO-L2[1] words are those which are not recognised as pertaining to the target language and are not loan words (see table 4.1).

LEXICAL		
MTness Type	**Example**	**Explanation**
NO-L2	Acceso de **Missatges** de Internet *(Access to Internet **Missatges**)*	*Missatges* is a Catalan word

Table 4.1: Examples of NO-L2 instances

4.1.2 Syntactic Types

Syntactic types are linguistic phenomena that affect the syntactic relations between words. The syntactic types are the following:

- Inadequate syntactic agreement (I-AGR)
- Inadequate part of speech (I-POS)
- Inadequate verbal form (I-VERBF)
- Inadequate constituent order (I-ORD)
- Word overgeneration (OVER-WRD)
- Word repetitions (WRD-REP)
- Syntactic gap (SYNT-GAP)

4.1.2.1 Inadequate syntactic agreement (I-AGR)

Morphological values that do not comply to the grammatical agreement restrictions between syntactic constituents. Table 4.2 shows some examples.

[1] L2 stands for target language, as L1 stands for source language

I-AGR		
Syntactic constituents	**Example**	**Explanation**
Subject and verb	Las ayudas estatales no **debe** seguir adelante (*State helps cannot go on*)	The verb *debe* in singular does not agree with the subject in plural (*Las ayudas estatales*)
Determiner-Noun modifier and noun	Los gobiernos son víctimas de sus propios **laberinto** (*Governments are victims of* ***their own*** *laberynth*)	The determiner (*sus*) and the adjective (*propios*) are in the plural whereas the noun (*laberinto*) is in the singular.
Determiner and noun - Anaphoric expression	En vista de **la doble desafío** de la ampliación, que debe permitir a Europa para llegar a términos **consigo mismo**... (*According to* ***the double challenge*** *of ampliation, which must allow Europe to reach terms with* ***himself****...*)	The determiner should be *el* (*el doble desafío*) because the noun is in the masculine. The reflexive should be in the feminine (*consigo misma*) because in Spanish the gender of *Europe* is femenine.

Table 4.2: Examples of I-AGR instances

4.1.2.2 Inadequate part of speech (I-POS)

The part of speech (PoS) of a word is inadequate according to the syntactic context in which it appears. Some examples are shown in table 4.3

I-POS		
PoS	**Example**	**Explanation**
Adjective	**He concreto** mencionado algunos de los factores (*I* ***have concrete*** *mentioned some of the factors*)	An adjective cannot appear (*concreto*) after the auxiliary verb (*haber*)
Noun	Se traen por ir a la playa, pero también por salir a **cena**. (*They were taken to the beach, but also to go to* ***supper***)	*Cena* is a noun. The verbal form *cenar* is the one expected

Table 4.3: Examples of I-POS instances

4.1.2.3 Inadequate verbal form (I-VERBF)

This type covers non-finite verbs that should have appeared in finite forms and viceversa. Inconsistencies in the verbal mood (indicative and subjunctive) are also covered in this type. Table 4.4. shows some examples.

I-VERBF		
Verbal form	**Example**	**Explanation**
Participle	Queremos, a petición de que se **diferida** por tercera vez por razones políticas (*We want, at the request that it **deferred** for the third time for political reasons*)	After the pronominal *se* a finite verb form is expected instead of a participle
Present tense	La Unión Europea ya , como se suele **hace** , se sitúa en un importante cambio de rumbo en su historia.(*the European Union already, as it is often **does**, is in an important turning point in its history*)	*hacer* should have appeared in the infinitive form, not in the present form
Subjunctive and participle	A fin de preservar el equilibrio entre las instituciones , creo que debemos actuar para que este defecto **es reparar** (*In order to preserve the balance between institutions, I think that we must act in order to this fault **is to repair***)	*es reparar* should be *sea reparado*, where the verb *ser* is in the subjunctive and the verb *reparar* is in the participle form.

Table 4.4: Examples of I-VERBF instances

4.1.2.4 Inadequate constituent order (I-ORD)

This is the type for queer orderings of syntactic constituents. Examples of constituents affected are shown in table 4.5

4.1.2.5 Word overgeneration (OVER-WRD)

A word, or a sequence of words, does not perform any syntactic or cohesive role in the sentence. By deleting them, the sentence makes more sense (table 4.6).

I-ORD		
Constituent order	**Example**	**Explanation**
Noun-Adjective	Víctimas de la **española represión** (*Victims of the* ***Spanish repression***)	The adjective should have appeared in a postnominal position (*represión española*)
Noun-Determiner	**la caída segundo mayor** en la satisfacción (**the fall second biggest in satisfaction**)	The ordinal determiner and the superlative adjective should be in a prenominal position (*la segunda mayor caída en la satisfacción*)
Noun-prepositional complement	He valorado mucho **del presidente Prodi declaraciones** (*I have valued much* ***of president Prodi declarations***)	The prepositional phrase should appear in a postnominal position (*declaraciones del Presidente Prodi*).
Collocations	**Temprano esta mañana** que embarcamos en un catamarán (***This morning early*** *we boarded on a catamaran*)	*Temprano esta mañana* should be *Esta mañana temprano* instead.
Adverbials	**el aproximadamente acuerdo de 150 página** (*the* ***about agreement*** *of 150 pages*)	*aproximadamente* seems to complementize a noun (*acuerdo*) instead of the numeral (*aproximadamente 150 páginas*).
Prepositional units displaced	su política **a los estados unidos subcontinente respecto** (*its politics* ***to the states united subcontinent respecting***)	The prepositional form *respecto a* is split in two units (*respecto* and *a*) and two noun phrases (*su política* and *subcontinente*) appear in between, which breaks the logical coherence

Table 4.5: Examples of I-ORD instances

OVER-WRD		
Constituent overgenerated	**Example**	**Explanation**
Unnecessary verb	Para mí , esto **es** también tiene una gran atención (*For me, this* ***is*** *also has a great attention*)	The verb *es* is overgenerated
Unnecessary pronoun	Os podéis dirigir**se** a Internet Content Rating Association (*You can address to Internet Content Rating Association*).	The pronoun *se* is overgenerated

Table 4.6: Examples of OVER-WRD instances

4.1.2.6 Word repetitions (WRD-REP)

Two identical word-forms in the same syntactic phrase or in two phrases which are close to each other, as is shown in table 4.7.

WRD-REP		
Constituent repeated	**Example**	**Explanation**
Determiner	Por consiguiente, negociamos en **un un** mínima base y tenemos una mínima carta, en particular respecto de los derechos sociales. (*Therefore we negotiate in* ***a a*** *minimum base and we have a minimal letter, in particular as regards social rights*)	The determiner **un un** is repeated.
Noun	El **paseo** resulta un duro **paseo** en la mesa de trabajo de Mac o dentro de Mi Ordenador. (*The* ***stroll*** *results in a hard* ***stroll*** *on the Mac desktop or in My Computer*)	The noun **paseo** is repeated

Table 4.7: Examples of WRD-REP instances

4.1.2.7 Syntactic gap (SYNT-GAP)

A missing constituent that should have appeared according to the argument structures of verbs and nouns and other syntactic constructions. Examples are shown in table 4.8.

SYNT-GAP		
Missing constituent	**Example**	**Explanation**
Direct object	El Senado veta los presupuestos y **retorna al Congreso** por primera vez en la democracia (*The Senate vetoes the budget and* ***gives back to the Congress*** *for the first time in democracy*)	The direct object of the verb *retornar* is missing
Verb	Uno de los mayores orfanatos **que en** el norte del país (*One of the largest orfanates which* ***in the north*** *of the country*)	A verb is missing.
Preposition	Actualmente, el consejo está **hablando incorporar** esos mecanismos en el artículo 7 (*Currently, the Council is* ***talking incorporate*** *these mechanisms in article 7*)	The preposition in the complement of *hablar* (talk) is missing (e.g *hablar de* (talking about)

Table 4.8: Examples of SYNT-GAP instances

4.1.3 Semantic Types

Semantic types are linguistic phenomena that affect the semantic relations between words. The semantic types are the following:

- Semantic gaps (SEM-GAP)
- Semantic incoherence (SEM-INCOH)
- Contextual incoherence (CON-INCOH)

4.1.3.1 Semantic gaps (SEM-GAP)

Semantic gaps are missing constituents that are necessary to understand the sentence. SEM-GAP is different from SYNT-GAP because the latter is not linked to the interpretation of the sentence. In fact, in SYNT-GAP cases, a correct interpretation of the sentence leads the reader to detect the missing syntactical constituent. Table 4.9 shows some examples of SEM-GAP.

SEM-GAP		
Missing constituent	**Example**	**Explanation**
Noun	Necesitamos una definición más precisa de **la relevantes** del mercado (*We need a definition more precise of **the relevant** in the market*)	The noun that the adjective (*relevantes*) is expected to modify is missing
Noun complement	Estas zonas sería destacado por la aplicación de criterios para definir este valor añadido en términos de creación rankings y la **exclusión** (*These areas would be remarked for the application of criteria to define this added value in terms of creation rankings and **exclusion***)	The noun complement with a reference to the excluded thing is missing.

Table 4.9: Examples of SEM-GAP instances

4.1.3.2 Semantic incoherence (SEM-INCOH)

This type covers absurd interpretations because arguments do not fit the semantic restrictions of the noun or the verb. Some examples are shown in table 4.10.

SEM-INCOH		
Unfit argument	**Example**	**Explanation**
Subject	Los **Bocados** detectan al Vallès una reavivada de asaltos nocturnos a viviendas (***The Bites*** *detect in the Vallès an arousal of night assaults to buildings*)	*Bites* is the mistranslation of the name for the Catalan police force (*Mossos*). The interpretation is absurd because the subject does not fit the selectional semantic restrictions of the verb *detectar* (*to detect*).
Noun complement	Los de la Generalitat Valenciana practican la política del valenciano **escondido** (*Those of the Valencian Generalitat practice the politics of the* ***hidden Valencian***)	*Escondido* (*hidden*) is not semantically consistent with *valenciano*, referring to the Valencian dialect.

Table 4.10: Examples of SEM-INCOH instances

4.1.3.3 Contextual incoherence (CON-INCOH)

This type covers arguments that do not violate semantic restrictions of the noun or verb but do not fit the context where they appear. This is shown in table 4.11.

CON-INCOH		
Unfit argument	**Example**	**Explanation**
Subject	Es el Estatuto que, a día de hoy, Catalunya necesita y los catalanes **volamos** (*It is the Estatute that, nowadays, Catalonia needs and the Catalans* ***fly***)	The translation of the Catalan verb *volem* (*we want* and *we fly*) into Spanish as *volamos* (*we fly*) makes the translation incongruent.
Noun complement	Vuelo en un balón (*Flight on a* ***ball***)	*Balloon* is translated as *balón* (*ball* in Spanish), instead of *globo*

Table 4.11: Examples of CON-INCOH instances

Other CON-INCOH are linguistic phenomena such as apocopation where context does not affect the meaning of words. The translation of the English *to be* into the Spanish *ser* or *estar*, or the Catalan preposition *a* into *a* (to) or *en* (in) in Spanish also cause CON-INCOH errors. Table 4.12 shows some examples

CON-INCOH		
Linguistic phenomenon	**Example**	**Explanation**
Apocopation	El **primero ministro** de Ucrania impugna las elecciones y reitera que no dimitirá (*Ucrania's* ***Prime Minister*** *impugnates the elections and again says he will not resign*)	*Primer*, as the apocopation of *Primero* is the right translation
Ser and *estar* and *a* and *en*	La versión 4.76 del Navegador en catalán ya **es al mercado** (*The 4.76 version of the Browser in Catalan is already* ***to the market***)	*está ya en el mercado* is the right translation

TABLE 4.12: Examples of apocopation and ser/estar CON-INCOH instances

4.1.4 Formatting Types

These types are TYPO-E, which covers type errors (e.g. inadequate use of upper case and lower case, missing or inadequate punctuation marks), and STR-CHAR, which are strange characters that appear because of an incorrect codification of the original text. In table 4.13 an example of STR-CHAR is shown.

Formatting		
MTness Type	**Example**	**Explanation**
STR-CHAR	Pa?s	Wrong decoding of accentuated vowels

TABLE 4.13: Example of a STR-CHAR instance

4.2 Discussion

In this section we discuss the pertinence of the typology according to its relevance, and its explicative and predictive capacity.

4.2.1 Relevance of the typology

From a theoretical point of view, the typology is relevant because it has been elaborated from the real perception of MTness by a large and varied number of people. From a practical point of view, its relevance could be questioned if it were not very different from already-existing MT error typologies. In order to check whether this is the case, we compared the MTness typology with the MT error typologies referenced in section 2.4 of Ariadna Font Llitjós' PhD thesis (Llitjós, 2007). This section is a state-of-the-art overview of MT error typologies. The approach of the review was interesting for us because the thesis suggests improving MT systems with the contribution of users who are neither MT nor linguistic specialists. So, its objective is also to typify MT errors appreciated by non specialists.

4.2.1.1 MTness typology and MT error typologies

Our typology captures the reader′s perception so MTness instances needn't be detected with the help of reference translations, as (Vilar et al., 2006) advises when detecting MT errors, nor with a tool for error analysis (Stymne, 2011). Our approach is not system-centered but reader-centered.

For instance, we have not seen WRD-REP in MT-error typologies[1], but readers tend to consider the repetition of a word in very close local contexts as machine-like. On the other hand, despite the similarities with other typologies, there is no straightforward relationship with the reader′ perception itself[2]

From Font Llitjós′ review we noticed that most of MT error typologies explain developers the cause of an error (Correa, 2003), give developers hints about how to improve the system's lexical and syntactical precision and recall, or are elaborated for postedition tasks ((Vilar et al., 2006), (Lingtech, 1996), (Lingtech, 1997) and (Elliot et al., 2004)). From our point of view, not all the linguistic phenomena in these typologies affect MTness.

[1] in (Vilar et al., 2006) word repetition in close local contexts are referred to as instances of the *style errors* type because the translation can be improved by a stylistic solution (use of a synonym or elisions) that is hardly for a system to perform

[2] In (Vilar et al., 2006) *Extra Words* seem to be similar to OVER-WRD, but as they say *This kind of error was introduced mainly when investigating the translation of speech input, as artifacts of spoken language may produce additional words in the generated sentence.*. *Word Order* is actually equivalent to I-ORD, but the detection relies on comparing orderings with reference translations instead of the reader′ perception itself. Finally, *Unknown words* seem to be equivalent to NO-L2 but the further classification (*truly unknown words/unseen forms of known stems*) refer to the possibilty for the system to copy the source word without processing it and errors due to bad processing by the system

Errors in the translation of terminology exemplify the dissociate character of MT errors and MTness. In (Loffler-Laurian, 1996) and Schffer′s typologies (Schäffer, 2001) there are categories for wrong terminological denominations. However, the fact that a doctor notices a mistranslated biochemistry term does not prove that the translation was generated by a machine, just because a non-specialist human translator could have been mistaken as well. For this reason, in our typology there is no special category for terminology not even for missing words in the dictionary (Llitjós, 2007). Terminology coverage could be further checked in case the MTness-based evaluation were not enough, but our typology is useful to detect words that are not of the language and phrases that the non specialist reader would consider flagrant errors such as *Ventanas* instead of *Windows* when referring to the operative system, or *fichero Excielo* (*Ex heaven file*) instead of *Excel file*.

MT errors are generally presented as faults of particular modules of the system′s architecture, especially the lexical and the syntactical components. The system-centered approach makes developers overlook, for example, the queer impression of readers when noticing a meaning that does not fit the context (SEM-INCOH), although it is well formed lexically and syntactically. Even Font-Llitjós′ typology is system-centered, despite the interest in error perception by non-specialists. More than this, it is methodology-centered, because the errors are presented in the light of improving the lexical and syntactical components of a transfer-based MT system.

On the other hand, MT typologies for developers and posteditors have language-dependent categories. For instance, in (B, 1997) there are error categories for German such as *case*. In (Flanagan, 1994) there is the category *accent* for French. In our typology, the types are not dependent on a particular language. Even I-AGR, which affects languages with syntactic agreement, cannot be said language-dependant since this type affects many languages. This is a consequence of the explicative status of the typology, where particular language errors are not focused but the linguistic phenomenona that cause the reader′*s* bewilderment.

Another example of the focus shift from particular errors to general linguistic phemonena, is the fact that the MT typology groups errors that in MT error typologies appear separated. For example, in (Flanagan, 1994) categories *pronoun*, *preposition* and *article* denote the missing or innecessary presence of a word with one of these morphsyntactic categories. In our typology these errors are under the general types SYNT-GAP and OVER-WRD.

In summary, despite the similarities between our typology and an MT-error typology, the latter’s objectives are different from ours and, for this reason, they do not fit our goal.

4.2.2 Explicative capacity of the typology

About 80% of the segments underlined in the experiment were categorized in the MTness typology. The rest were underlined by only one informant and we were not able to discern why. Although this proves that the explicative character of the typology is quite good, the fact that 16% of inexplicable segments were from human translations was of interest for us. In the following chapter we will deal with MTness perception in human translations.

On the other hand, the coincidence of MTness types with MT-error typologies among different languages is important. Although our typology is based on the experimental study for Spanish, the coincidence with error types in other languages suggests that the typology is cross-language. Besides, the coincidence with classes from Font-Llitjós′ typology, for non-specialist people, or Flanagan′s (see Table 4.14), proves that the MTness typology explains the people′s linguistic intuitions with or without specialized knowledge.

Font-Llitjós (2007)	**Flanagan (2004)**	**MTness Type**
Missing word	Elision	SYNT-GAP
Extra word		OVER-WRD
Wrong Word Order	Rearrangement	I-ORD
Incorrect Word/Selectional Restrictions		SEM-INCOH
Wrong Agreement	Agreement	W-AGR
	Capitalization	TYPO-E
	Verb Inflection	I-VERBF
	Category	I-POS

Table 4.14: Coincidences between MTness types and two MT-error types

4.2.3 Predictive capacity of the typology

The typology has predictive capacity for new translations. This is the consequence of the saturation technique when establishing the number of people to be interviewed (c.f. 3.2.1). As we added more people, the data did not change significantly. Therefore, it is expected that the typology can characterize the MTness of new translations.

4.3 Conclusion

In this chapter we have presented the MTness typology. Each type has been explained and we have discussed the relevance of the typology, compared to MT-error typologies, and seen its explicative and predictive capacity. The typology proves that typified linguistic phemonena cause MTness, which is the first objective of the experimental study. In the following chapter we will deal with the other two objectives, which we present in table 4.15 as a reminder.

Goals	**Evidences**
Assesment of MTness perception in machine translations only	2.1. Agreement among informants 2.2. Types most agreed are particular of MT
Assesment of MTness perception regardless contingent factors	3.1. MTness perception is not language-pair dependent 3.2. MTness instances are not the errors of a particular MT methodology

Table 4.15: Goals 2 and 3 and evidences of the experimental study

Chapter 5

MTness Perception

In this chapter we will deal with MTness perception and the evidences for the second and third objectives of the experimental study. These objectives are to assess that MTness is really a quality of machine translations and to evidence that this quality is perceived regardless the language pair and the methodology of the MT system.

We will see that MTness is mainly perceived in machine translations although a certain degree of subjectivity must be admitted. This subjectivity was evident in the informants′ underlining of MTness segments in some human translations. On the other hand we will see some examples of mismatches in the underlining of MTness errors by informants. The analysis of these mismatches provides useful information for automatic MTness detection and rating translation quality.

The chapter is organized as follows. Section 5.1 shows MTness types as phenomena that make machine translations be different from human translations. Section 5.2 is about the assessment of MTness as a quality that does not characterize a particular MT methodology nor the errors of a particular language pair. MTness affects both RBMT and SBMT systems alike in close and distant language pairs. In section 5.3 we will hypothesize the reasons why there were mismatches when informants underlined some MTness instances. We will introduce the ideas of MTness instance overlapping and MTness salience and will see whether there are MTness instances more salient than other types. In section 5.4 we will present MTness instance overlapping and MTness salience as two important notions to be taken into account in MTness detection and evaluation.

5.1 MTness and machine translations vs human translations

MTness instances must be perceived objectively, not on subjective and not clearly defined criteria which would blur the distinction between a human and a machine translation. Most typified MTness-instances which were more agreed among informants were found in machine translations. Only 1.85% of the segments underlined in human translations were agreed by at least two informants. Therefore, we concluded that MTness types really characterize machine translations.

By analyzing the underlined segments in human translations we hypothesize the two causes we list below

- Gaps in lexical knowledge
- Subjective linguistic and stylistic criteria

Let us discuss these hypotheses more fully.

Gaps in lexical knowledge

Many underlined segments that were not real MTness instances showed lack of knowledge about specific terminology and vocabulary. This influenced the perception of false NO-L2 instances in human translations. In table 5,1 some examples are found and the hypothesized disagreements are explained.

Gaps in lexical knowledge		
Lexical knowledge	**Example**	**Explanation**
Terminology	**Anonymity Proxy** para Windows es de dominio público ***Anonymity Proxy*** *for Windows is for the general public*	The informant probably did not know what *Anonymity Proxy* was
General vocabulary	Y aquí la Administración puede **aducir** pocas justificaciones. *And here the Administraton can* ***adduce*** *few justifications.*	The informant was not aware that *aducir* (adduce) is a right word in Spanish.

Table 5.1: Examples of lexical instances underlined in human translations

Terminology is not often translated compositionally, so proper denominations may look incoherent and queer for neophytes. As an example *fichero adjunto* (attached file) was underlined in a human translation. This proves that terms which are not of common use can be interpreted as incoherent.

Subjective linguistic and stylistic criteria

Different linguistic and stylistic different criteria among informants caused disagreement. These criteria are related to the use of a determiner, verbal mood, tense or aspect. Disagreement in lexical choices must be taken into account as well. See examples in table 5.2.

Subjective linguistic and stylistic criteria		
Linguistic use	**Example**	**Explanation**
Use of a determiner	EDICTO del Ayuntamiento de la Vall de Bianya, sobre **contratación de personal** *EDICT of the Town Hall of Vall de Bianya, on* ***staff contracting***	The informant considered that a noun phrase with no determiner was not adequate
Verbal aspect	Hace un mes, cuatro terroristas **viajaban** a Londres. *One month ago, four terrorists* ***were traveling*** *to London.*	The informant expected the perfective aspect (*viajaron*)
Lexical choice	La **pestilencia**, también, está llegando, dicen, en las alas de pájaros que emigran, en forma de gripe avian. *It is said that the pestilence, is also coming, on the wings of birds that migrate, as avian flu.*	It seems that the informant would have preferred *peste* (plague) instead of *pestilencia* (pestilence)

Table 5.2: Examples of underlined segments due to linguistic and stylistic criteria

Stylistic, rhetorical criteria may be very strict for a person who is not used to MT and is not aware of its limitations. Sometimes an expression which is different from what the informant would have said may seem an MTness instance. However, a person who is more used to MT may be less strict and underlines only the segments that affect intelligibility.

5.2 MTness in distant-close language pairs and MT methodology

It is generally expected for distant-language-pair systems to produce more errors than close-language-pair ones. Accordingly, we checked whether more MTness instances were perceived in the distant language pair. Figure 5.1 shows that more than twice underlined segments were found when English was the source language.

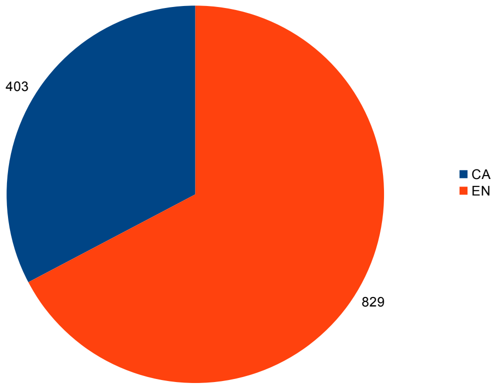


Figure 5.1: Underlined segments and source languages

The difference is considerable when the SBMT translates in the EN-ES language pair, as shown in figure 5.2. However, both SBMT-CA-ES and RBMT-CA-ES produced a similar number of results, which proves that the MT method and the language pair distance affect MTness when combined.

5.3 Mismatches in underlined MTness instances

In section 5.1 we showed that some informants considered human translations as machine-like. Although the agreement among informants was low, a certain degree of subjectiveness in MTness perception must be taken into account. In this section we will see that gaps in lexical knowledge and subjective stylistic criteria also explain mismatches

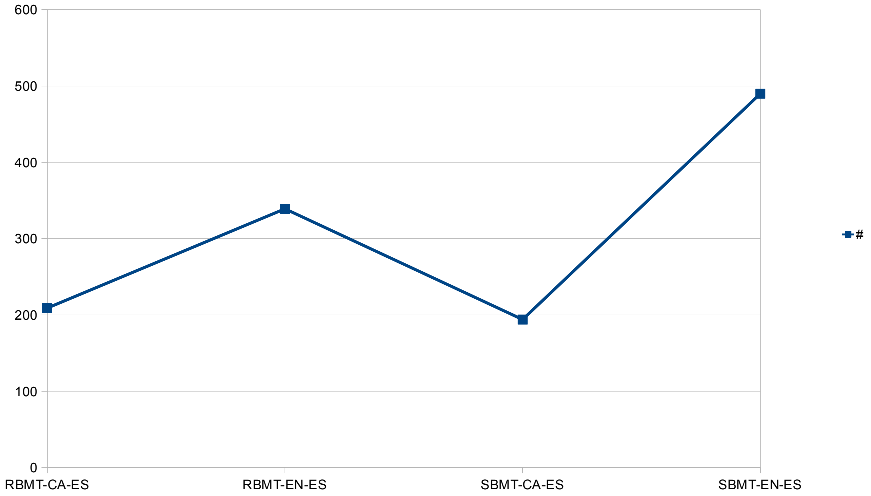


FIGURE 5.2: Underlined segments according to MT methodology and language-pairs

in machine translations. Besides, we will also present the notions of *MTness instance overlapping*, *non-salient MTness instances* and *source language evocation* as other important causes of the mismatched underlined segments.

5.3.1 Vocabulary knowledge

The mismatches related to vocabulary knowledge were closely related to source language evocation. A word in the translation evoked the source language equivalent in the informant′s mind. When the source word evoked was very similar to the translation, the informant sometimes thought that the translation was wrong, although not always was so. For example, the right Spanish word *convidamos* (we invite) was underlined probably because it was very similar to the Catalan verb *convidem* which means the same.

5.3.2 Subjective linguistic and stylistic criteria

In section 5.1 we explained how subjective linguistic and stylistic criteria affected the judgement of human translations. Orthotypographic errors (TYPO-E) were influenced by these criteria. Figure 5.3 shows the matching degree of underlined MTness types in the two language pairs. Notice that TYPO-E segments had little matching degree in machine translations in the two language pairs.

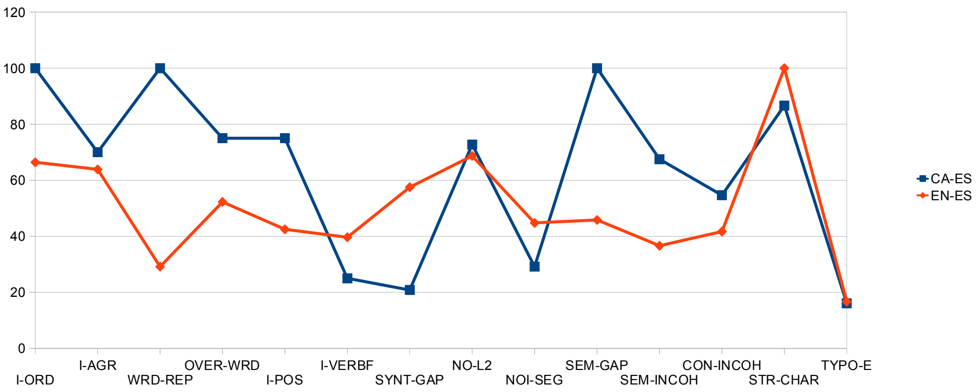


Figure 5.3: Perception agreement in CA-ES/EN-ES MTness types

On the other hand, matching of SYNT-GAP instances was low in CA-ES, no matter the MT methodology (see figure 5.4). In fact, some of the disagreements affected the non-use of a definite article, which is a matter of stylistic criteria in certain contexts in Spanish.

5.3.3 MTness instance overlapping

Sometimes, discerning distinct MTness instances was very difficult because they overlapped in a single translation. This overlapping explained underlining mismatches among informants. An important difficulty in measuring agreement for particular MTness types was to face a hen/egg problem when two MTness phenomena affected each other. For instance, we had to consider whether the reader perceived a case of I-ORD or I-AGR in a noun plus adjective combination, because the wrong position of the adjective also implied the wrong agreement with the noun. The lengths of the segments underlined by the informants were different and we had to guess whether the informant focused on the wrong agreement of one constituent or rather on the order of both constituents.

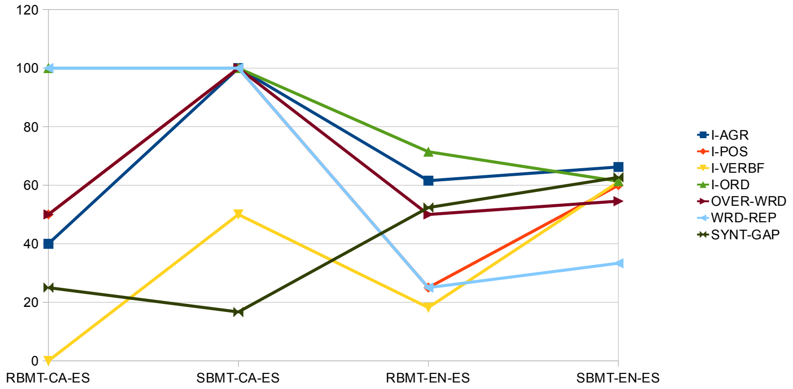


Figure 5.4: Perception agreement in syntactic MTness instances (language pair and MT methodology)

For example, table 5.3 shows a translation where an informant underlined the strange character (in bold), whereas another informant underlined a segment utterly incomprehensible, from the strange character until the end of the sentence. This segment contained other MTness instances that were not distinguished, such as the NO-L2 *pestanya* (*tab* in Catalan) and *Opcions* (*Options* in Catalan), which are a NO-L2 and a TYPO-E instance respectively.

Translation	**Underlined segments (informant A)**	**Underlined segments (informant B)**
Finalmente quin al ? última pestanya Opcions avanzadas *Finally who to last tab Advanced Options.*	Finalmente quin al **?** última pestanya Opcions avanzadas.	Finalmente quin al **? última pestanya Opcions avanzadas.**

Table 5.3: Example of disagreement in underlined segments

For some informants, the overlapping produced a saturation effect and the oddness of the translation segment led them not to worry about distinguishing MTness instances.

Other informants, on the contrary, provided more precise information and underlined each instance. Laziness and imprecision were also a consequence of the fatigue caused by MTness saturation when reading odd and incomprehensible segments. Odd and bizarre translation segments that produce MTness saturation will be referred as noisy segments (NOI-SEG).

The matching degree of underlined noisy segments is also displayed in figure 5.3. Distant language pairs are likely to produce more MTness overlappings. On the contrary, in close language pairs, MTness phenomena are more easily found out and delimitated. Notice in figure 5.3 that matched underlined noisy segments is higher in EN-ES. So MTness saturation, and consequently underlining mismatches, affect specially the distant language pair, which may explain the flatter line in the plot for EN-ES. Figure 5.4 illustrates the fact that matchings in most of the syntactic MTness types were generally lower in the distant language pair, both in the RBMT and SBMT systems[2]

Figure 5.5 shows the matching degree of semantic MTness types according to the language pair and the MT methodology. The matching of MTness types is contrasted with the matching of noisy segments, because of their bizarre and odd meanings and even their incomprehensibility. Notice that the language-pair distance and the MT methodology are the two factors that increase the agreement in NOI-SEG perception and hence the mismatches in the underlined MTness instances, as seen in figure 5.5. The matched noisy segments in the output of the CA-ES systems were fewer than in the output of the EN-ES systems, and the SBMT systems produced more matched noisy segments than the RBMT counterparts. This affected especially the underlining of SEM-INCOH and CON-INCOH instances. In SBMT-based output, informants generally agreed in expecting the missing constituents necessary to understand the sentence (SEM-GAP).

5.3.4 Non-salient MTness instances

MTness salience is a notion to be taken into account. There are MTness instances that are not as salient as other instances in the same translation. Weak MTness salience explains why some SYNT-GAP instances were not underlined by all the informants. When prepositions, clitics or articles are missing, the reader unconsciously fills in the blanks. It is the consequence of the Gestalt′s *law of closure* in reading, on which the *Cloze test* is based for reading comprehension (Soudek and Soudek, 1983). Actually, as

[2] I-POS and I-VERBF matchings in the distant language pair were higher in the SBMT system than in the RBMT system. This may be due to the inconsistencies with linguistic intuition and grammar knowledge of some SBMT translations. See 5.4.1 for the relationship between MTness salience and linguistic intuition. On the contrary, the percentage is lower for I-ORD in the SBMT; probably because rules in RBMT systems are sometimes too constrained to generate more natural word orderings over others.

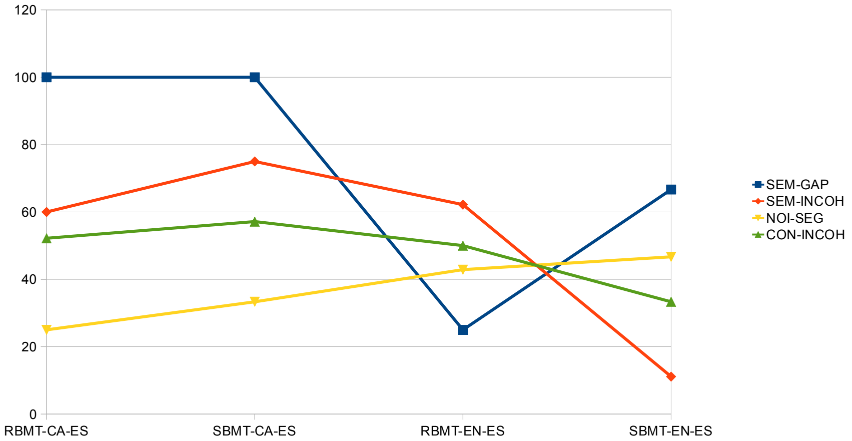


FIGURE 5.5: Perception agreement in semantic MTness instances

we pointed out in chapter 1, posteditors admit they do not notice the absence of function words in their first reading[2]

Non salient syntactic MTness instances affect especially the MTness perception in close language pairs. SYNT-GAP instances in language pairs with similar syntactic structures are not so salient as in syntactically dissimilar language pairs. We hypothesize that this explains the fewer SYNT-GAP matchings in CA-ES, apart from the reasons pointed out in 5.3.2. The number of matched SYNT-GAP instances in CA-ES was much lower than in EN-ES, no matter the MT methodology (see figures 5.3 and 5.4).

Source and target language similarity also favours non-salience in other MTness types, such as TYPO-E. For instance, a Catalan informant did not notice that the initial question mark was missing in some Spanish translations.

[2] Overgeneration of function words (OVER-WRD) was not perceived by all the informants as well. This was also due to the discreet salience of function words. Some informants did not underline prepositions, auxiliar verbs and articles that were not necessary to understand. It seems that missing or repeated function words may pass unnoticed if the words that guarantee a nominal or verbal argument structure are present

5.3.5 Source language evocation

When the reader knows the source language, some MTness instances give hints about how the source sentence was and call the right translation to the reader′s mind. When this happens, the MTness instance is more visible. In fact, the knowledge of the source language helps the reader to focus phenomena. This was evident in the CA-ES language pair with informants that knew the two languages, whereas MTness underlining in EN-ES was not as precise because informants did not know the source language and soon lost the thread because of MTness saturation .

Source language evocation can be considered a particular case of MTness salience. A source-language-evoked MTness instance is more visible if this instance triggers the right translation because it is overwhelmingly more used by the community. This is what we call the expected translation contrast. Let us see two examples. As a first example, imagine that *Windows operative system* is translated into Spanish as *Sistema operativo Windows.* The untranslated word *Windows* is not a NO-L2 instance. However, the translation *Sistema operativo Ventanas* is indeed an MTness instance, and so is appreciated by readers. As a second example, the translation *morir de siete* ('to die of seven'), instead of *morir de sed* ('die of thirst'), is an MTness instance because *siete* triggers the right translation of the original Catalan word *set.*

The more different the word was from the word expected the more agreement in perceiving it. Table 5.4 shows an I-POS example.

Translation	**Underlined segment**	**Expected translations**
Épocas de viajes y de descubiertas. *Times of travels and of discovered.*	**descubiertas**	A nominal form was expected: **descubrimientos** (discoveries)

Table 5.4: Example of I-POS agreement

On the other hand, some informants overlooked cases where the difference lied in one tilde (e.g. *donde* and *dónde*(where), or the expected forms were very similar.

5.4 MTness salience and MTness overlapping in MTness detection and evaluation

In this section we will discuss how MTness salience and overlapping must be taken into account in MTness and machine translation evaluations.

5.4.1 MTness salience

From the analysis of the results we first conclude that there are MTness errors more salient than others. Therefore it is important for posteditors and proof readers to use a tool that highlights the errors that can be overlooked. Our detection method aims to cope with non-salient instances but we are aware that non-salient instances are often caused by subtleties that require high precision natural language processing tools. Therefore, our method is focused on detecting at least the most salient instances if current natural language processing tools are mature enough to detect them.

From the analysis of the mismatched underlined segments and the ranking of the MTness types with more agreement, we drew some conclusions about what made MTness instances be more salient. Firstly we worked on confectioning a chart where MTness types were sorted according to the percentage of matched underlined segments (figure 5.6).

By comparing the ranking of MTness types, we drew the conclusion that salient MTness instances fulfill at least one of these conditions

- There is no evidence of use of the MTness instance in the target language
- The MTness instance forces a structure that is not recognised by a native speaker′s linguistic and intuitive knowledge.
- The MTness instance relates words in a way that is not consistent with the language use of native speakers
- The MTness instance triggers an expected translation contrast.

The first condition affects lexical units. Lexical units not identified as used in the target language caused the most impact on quality perception with high agreement among informants (over 90% for STR-CHAR and 70% for NO-L2[3]). Therefore translations with NO-L2 and STR-CHAR instances are to be rated the highest.

[3] Some disagreements were due to the fact that the reader was not able to recognize the source language word as a loan word in the target language

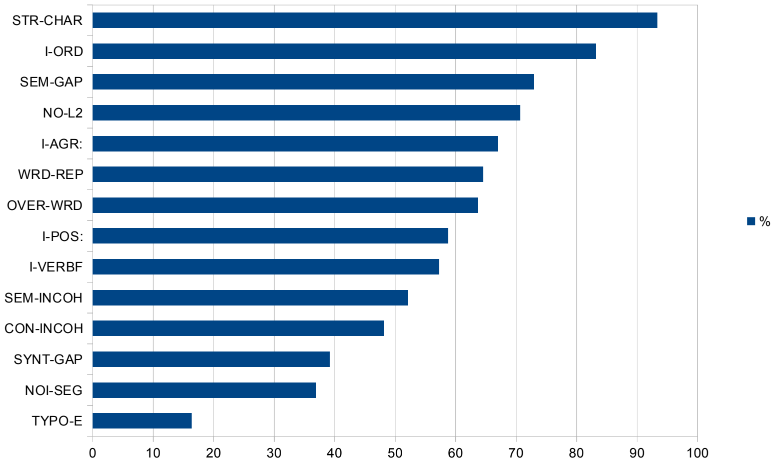


Figure 5.6: Underlining agreement in MTness types

The second type in the ranking is I-ORD (over 80%), a type which is salient because of either the second or fourth condition; that is, when the word order does not fit the intuitive linguistic knowledge or the I-ORD instance triggers an expected translation contrast (e.g: *unidad de cuidados intensivos* (intensive care unit) vs *unidad de intensivos cuidados* (unit of intensive cares). The third type, SEM-GAP, is salient because its instances shatter the semantic structure that would otherwise be recognisable in the target language model.

The second condition also explains the cases of I-VERBF caused by the mappings from the source verbal system into a very different one. For example, some transfers from the English non-finite verbal forms (participles, infinitives, gerunds) into Spanish were not grammatical. In I-POS instances, the transfers were established from the English morphological model, where a single form may correspond to more than one part of speech, into the Spanish model, where the part of speech of a word is recognised by its morphosyntactic features (e.g *supper* vs *cena* (noun) and *cenar* (verb)). Outstanding cases of SYNT-GAP, WRD-REP and OVER-WRD can also be explained by the second condition because the missing and added words form a syntactic structure which is not recognised by the intuitive linguistic knowledge.

The third condition affects the semantic and grammatical coherence of particular words which co-occur in a phrase, despite this phrase is recognised in purely syntactic terms. For example, in the Spanish sentence *Vuelo en un balón* (flight on a ball), the co-occurrence of *vuelo* and *balón* is semantically odd, although the combination *noun - preposition - noun* is right. Another example is the Spanish sentence *el acto de emergentes* (act of emerging) where *emerging* is translated as an adjective instead of a verb (*acto de emerger*), although the combination *noun - preposition - adjective* is right (e.g. *reunión de ricos* (a meeting of rich men)).

As for the fourth condition, MTness instances are very salient if they call up the right translation to the readerś mind. This can be explained by a musical analogy. Imagine a non-skilled piano player who plays the wrong key at the end of the famous initial tune of Beethoven′s *Fur Elise*. The wrong note would be appreciated by everyone because they were expecting the note they had heard many times before. In fact, following the musical analogy, the expected translation is like the missing key whose expectation outstands the discordant note.

Translations often come to mind when the reader knows the source and target languages. The reader guesses the source language form through the MTness instance and understands the reason for the mistranslation. We call these MTness instances source evokers. Some examples of source evokers are CON-INCOH instances like *sistema operativo Ventanas* (Windows operative system) or SEM-INCOH instances like *mueran de siete* (they die of seven), as we explained in 5.3.5.

5.4.2 MTness overlapping

An MTness instance can be metaphorically seen as a pebble dropped on a pond. The perception is more altered by the effect, the expanding ripples, rather than the cause, the tiny splash of the stone. Some instances produce the strongest effect by themselves. This is the case of NO-L2 and STR-CHAR cases.

In MTness overlapping, when more than one MTness instance is present, the perception is not affected by the instances in isolation. Following the metaphor, it seems that the reader′*s* perception is affected by the intertwining and expanding ripples produced by each MTness instance. In NOI-SEG cases, the ripples intertwined blur the pond completely and produce the saturation effect. As we will see in the following chapter, an MTness instance may impact across different linguistic levels and the more levels impacted the more outstanding is the MTness effect in translation quality appreciation. For example, SEM-GAP perception, which was widely agreed (around 70%), is the consequence of missing subcategorized complements that affect both syntax and semantics.

On the contrary, the impact on syntax and semantics by the overlooked SYNT-GAP instances shown in 5.3.4 is very low.

5.5 Conclusion

In this chapter we proved that MTness is an objective quality that characterizes machine translations and do not affect a particular MT methodology for a particular language pair. The overlapping of MTness instances produces a saturation that makes distinguishing single MTness instances difficult. MTness saturation often happens in the distant language pair.

On the other hand, there are MTness instances more salient than others. Among the salient instances there are words and syntactic structures that evoke the right translation, words and expressions which are not used in the target language, and finally phrases with words whose co-occurrence is not coherent at the syntactic or semantic level. The detection of these instances is the topic of chapter 6.

Finally, a new method of MT evaluation rating must be outlined. A method that captures the fact that one single word can spoil the translation and, on the other hand, a method where the rating reveals the cumulative effect of translation errors across different linguistic levels. This will be the topic of chapter 7.

Chapter 6

Automatic MTness Detection

In this chapter we will explain the automatic detection of MTness instances. In section 6.1 we will present our first approach to MTness detection. This approach was based on the use of an Internet search engine and took the Web as a representative corpus of use. The approach was lexically based, was intended to detect source evokers (c.f. 5.4.1), and was applied for the comparison of MT systems. Section 6.2 presents an approach applied to evaluate translations, rather than systems, based on detecting mismatches with the native speaker′s knowledge of the target language, since these mismatches make MTness instances more salient, as stated in chapter 5.

6.1 MTness detection based on Internet searches

In (Moré and Climent, 2006) and (Moré and Climent, 2007) we presented an MT evaluation method as a cheap alternative to RPA methods (c.f. 2.1.2.1). We focused on the detection of instances according to the first, third and fourth conditions of MTness salience (c.f 5.4); that is, words not used in the target language, inconsistent relations between words, and source evokers. We took the bulk of documents published on the Web as the representative corpus of use of the target language.

We were aware that evidences of use of whole sentences were unlikely to be found on the Web. Language is creative and our writing is not a compilation of sentences that have already been published on the Web. However, sentences can be split into phrases that are found in the Internet. For instance, if we Google the sentence *our writing is not a compilation of sentences that have already been published in the Internet* there are currently no results but the phrase *a compilation of sentences* has more than 5,000 results. So an Internet search engine and a parser that splits sentences into phrases were the resources of our evaluation method. Since the resources were available for everyone the cost of performing the evaluation was minimal[1].

The evaluation of a translation started by splitting the machine translation into syntactic chunks (MT chunks). The chunks corresponded to syntatic phrases and combination of phrases, as it is explained in (Moré and Climent, 2006) and (Moré and Climent, 2007). The translation was evaluated according to the number of MTness instances detected: the more MTness instances detected the worse the translation. An MT chunk was considered an MTness instance if it did not pass the use-evidence test; that is, the confirmation of its common use from the Internet.

The use-evidence test was inspired by Greffenstette′s method for a system to select the best solution among a set of possible MT translation alternatives (Grefenstette, 1999). For each word in an MT chunk, if the corresponding source word had other possible translations in the system′s bilingual dictionary, a new translation chunk (NMT chunk) was created by replacing the word with one of the other possible translations. So as many NMT chunks as alternative translations for each word were obtained. Then, each NMT chunk was turned into a query of the search engine. If the results of at least one NMT chunk overwhelmed the results of the MT chunk, then the MT chunk was considered an MTness instance.

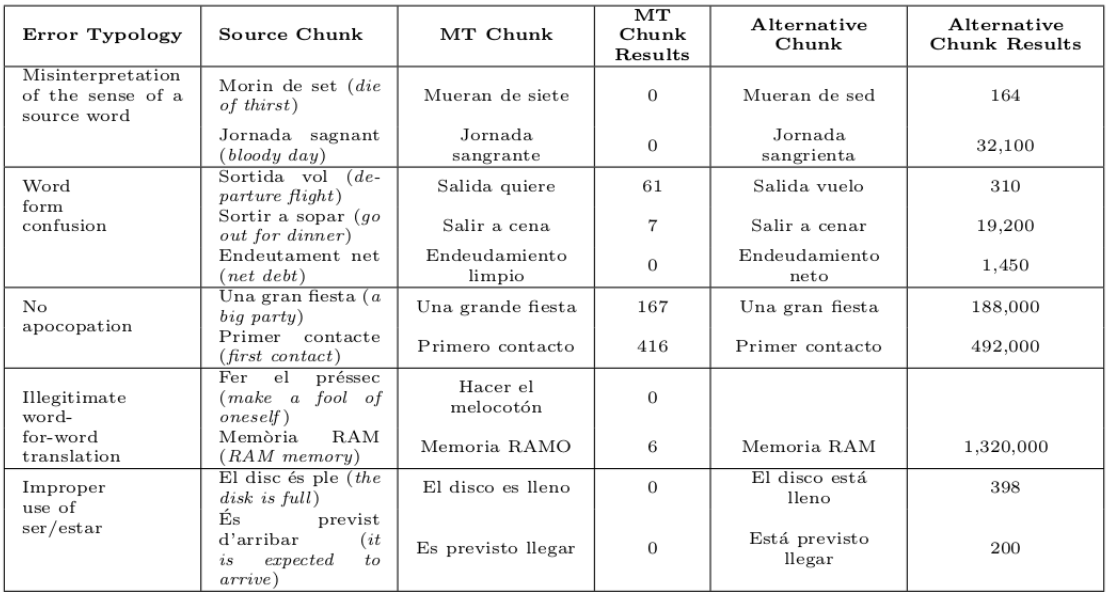
By performing this method cases of SEM-INCOH and CON-INCOH were detected. Figure 6.1 shows some examples in Spanish like *mueran de siete* (die of seven) or *memoria ramo* (bouquet memory)[2]. These translation solutions are odd because they are inexistent or spurious compared to other solutions found on the Web, the largest representative corpus of use currently available. The result comparison modeled the expected translation contrast, since the contrast was established between the actual MT solution and the NMT chunk with more results.

In order to detect NO-L2 instances, the source chunk was among the NMT chunks. Then if the results of the source chunk overwhelmed the number of results of the MT chunk, the latter was considered an MTness instance. For example, *Anonimidad Proxy*, with no results, is a MTness instance instead of *Anonymity Proxy*, with 63,900 results.

This method was a first approach to MTness detection, and had shortcomings. First, the method only detected lexical MTness instances. On the other hand, the method was not ready for evaluating MT systems whose dictionaries were under property licenses that would forbid their use for derived outcomes such as MT evaluation reports. Finally,

[1] The parsing was performed by *FreeLing*, a language analyzer under the GNU General Public License

[2] *siete* is the wrong translation of the Catalan word *set*, which also means thirst. *Ramo* is the wrong translation of RAM (random-access memory), which in Catalan also means *bouguet*



Error Typology	Source Chunk	MT Chunk	MT Chunk Results	Alternative Chunk	Alternative Chunk Results
Misinterpretation of the sense of a source word	Morin de set (*die of thirst*)	Mueran de siete	0	Mueran de sed	164
	Jornada sagnant (*bloody day*)	Jornada sangrante	0	Jornada sangrienta	32,100
Word form confusion	Sortida vol (*departure flight*)	Salida quiere	61	Salida vuelo	310
	Sortir a sopar (*go out for dinner*)	Salir a cena	7	Salir a cenar	19,200
	Endeutament net (*net debt*)	Endeudamiento limpio	0	Endeudamiento neto	1,450
No apocopation	Una gran fiesta (*a big party*)	Una grande fiesta	167	Una gran fiesta	188,000
	Primer contacte (*first contact*)	Primero contacto	416	Primer contacto	492,000
Illegitimate word-for-word translation	Fer el préssec (*make a fool of oneself*)	Hacer el melocotón	0		
	Memòria RAM (*RAM memory*)	Memoria RAMO	6	Memoria RAM	1,320,000
Improper use of ser/estar	El disc és ple (*the disk is full*)	El disco es lleno	0	El disco está lleno	398
	És previst d'arribar (*it is expected to arrive*)	Es previsto llegar	0	Está previsto llegar	200

Figure 6.1: Examples of MTness detection based on Internet searches (Moré and Climent, 2007)

the method was designed for comparative evaluations only: the worst system was the one with the more MTness instances detected.

The focus of this thesis is different from the focus of our first approximation. The thesis′ interest lies on translation quality rather than the quality of machine translation systems. This point of view widens the scope of our study. For example, in chapter 1 we presented MTness as a parameter to score the quality of documents. Besides, this thesis presents the detection of syntactic and semantic MTness instances from the perspective of the native speaker′s linguistic and intuitive knowledge

6.2 MTness detection based on the native speaker′*s* linguistic and intuitive knowledge

According to the conclusions from the experimental study explained in chapter 5, as far as MTness salience is concerned, MTness detection can be stated as: *detect the use of words and dependencies between words that do not match the native speaker′s knowledge of the target language*. The challenge is to detect as many mismatches as possible by using state-of-the-art natural language processing resources.

Parsers are basic NLP resources that model the processing of a sentence by native speakers according to their intuitive knowledge. Ideally, if parsers always behaved perfectly, the linguistic representations of human sentences generated by parsers would always be consistent with intuitive knowledge. Although state-of-the-art parsers do not behave perfectly, in this thesis we assume that parsing representations are as consistent with intuitive knowledge as possible, bearing in mind their limitations. So an MTness instance is regarded as the linguistic item that forces the parser to build a non-recognisable representation.

The MT sentence is parsed and the result is a parse tree. The MTness detection is executed by analyzing the lexical, syntactic, morphosyntactic and semantic information annotated in the tree.

In this section we will first explain the parse tree representations. Then we will explain the detection of MTness instances in the use of words. Finally we will explain the detection of instances in the relations between words.

6.2.1 The parse tree representations

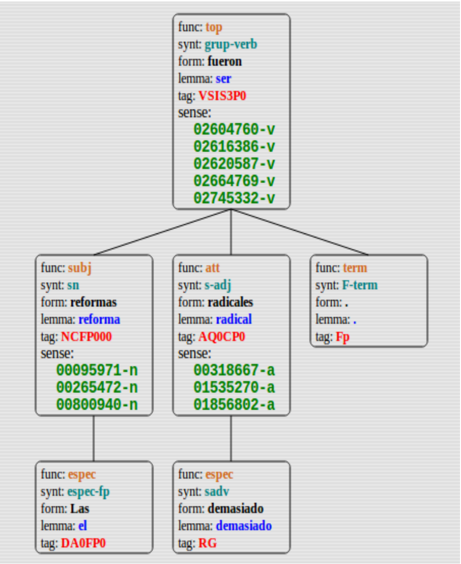
Relations between words are modeled as dependencies, as in the dependency grammar introduced by (Tesniere, 1959). The dependencies can be represented in a *dependency structure*, which, as defined by (Melĉuk, 1988), consists of a set of planar directed arcs among the words that form a tree. Each word (except the root word) has an arc out to exactly one other word, and no arc may pass over the root word. The root word is the governor, and the rest are the dependents.

Following (Melĉuk, 1988), three types of dependencies are established: syntactic, morphological and semantic. Syntactic dependency affects lexemes and the governor determines whether a dependent is optional or obligatory, and its syntactic function as well. Morphological dependency affects one lexeme and the values of its grammatical categories. The grammemes[3] of the dependents such as number or part of speech, depend on the restrictions of the governor. For example, in Spanish the head of the subject has a value for the category *number* that must agree with the number value of the verbal governor. Finally, the semantic dependency corresponds to the predicate-argument relation.

The three types of dependencies between the words in a sentence can be automatically represented in a tree by a dependency parser. When the tree is labelled with information about the three types of dependencies the structure is a typed dependency tree (Reichartz et al., 2010). This tree is displayed with indented lines where the indentation represents an arc.

[3] en.wikipedia.org/wiki/Grammeme

Let us explain the format of the typed dependency tree of the Txala parser, which is the parser we used for the experimental assessment of our evaluation method (see chapter 8). The Txala parser is a tool of the open source NLP Freeling library(Lloberes et al., 2010). Figure 6.2 shows an example.



```
1  grup-verb/top/(fueron ser VSIS3P0 02604760-v) [
2    sn/subj/(reformas reforma NCFP000 00095971-n) [
3      espec-fp/espec/(Las el DA0FP0 -)
4    ]
5    s-adj/att/(radicales radical AQ0CP0 00318667-a) [
6      sadv/espec/(demasiado demasiado RG -)
7    ]
8    F-term/term/(. . Fp -)
9  ]
```

FIGURE 6.2: Typed dependency tree of the Spanish sentence *Las reformas fueron demasiado radicales* (The reforms were too radical) with arcs (on the left) and indented lines (on the right)
.

Each line describes the syntactic, morphological and lexico-semantic information of a word in a tree node. The syntactic function is introduced first, and the word form, lemma, grammatical category values[4] and the Wordnet synset(s) are between parentheses[5]. The line indentation corresponds to the arc from the dependent to its governor, and the indentation length of a dependent node is two spaces longer than the indentation of the governor. So, in figure 6.2, the dependent nodes of the root are described in lines 2, 5 and 8. Notice that the dependent nodes of the governor are grouped between brackets and they share the same indentation length. These nodes are respectively the governors in smaller typed dependency trees whose dependent nodes are grouped between brackets as well. We will call these smaller typed dependency trees dependent subtrees. The grouped nodes are ordered according to the X-bar theory (Chomsky, 1970). The node for the head, which is generally the governor, appears first. Then it follows the node for

[4] Grammatical categories are annotated according to the EAGLES guidelines (http://www.ilc.cnr.it/EAGLES96/annotate/)

[5] The Txala parser allows the presentation of one synset after disambiguating the word

the word which functions as the specifier (e.g. subject of the verb, noun determiners), if any, then the complements and finally the adjuncts.

6.2.2 Detection of MTness instances in the use of words

MTness instances in the use of words will henceforth be referred as lexical MTness instances. Lexical MTness are words that do not match the intuitive knowledge of native speakers about words in the target language. Therefore, these words fulfil the first condition of salience, outlined in chapter 5 (c.f 5.4). The knowledge of native speakers about words is declared in the following sources of information:

- A monolingual dictionary of the target language
- A gazetteer of named entities
- A bilingual dictionary of the source and target languages

The monolingual dictionary contains the words used in the target language. The dictionary is simply a list of all the word forms of the target language lexemes (e.g. word forms in singular and plural). The monolingual dictionary is complemented with a gazetteer of named entities. The gazetteer is used in order to check whether a word form is the denomination of a named entity in the target language.

The bilingual dictionary represents the lexical knowledge of the source and the target languages. The bilingual dictionary consists of a list of pairs where the first element of the pair is the lemma of a source word and the second element is a list of translation equivalents, no matter the sense of the source word[6].

The use of the bilingual dictionary prevents loan words such as *golf e-Book* or *hip-hop* from being considered NO-L2 instances. Loan words are source words that also appear in the list of equivalents in the bilingual dictionary. As an example, the English word *Web* is considered a loan word in a translation because, in the bilingual dictionary, *Web* is a source lemma whose list of translation equivalents contains the same lemma.

The words that do not match lexical intuitive knowledge fulfill these conditions:

- The lemma of the word does not match any lemma in the monolingual dictionary

[6] For example, the English word *grave* has more than ten Spanish equivalents if all the senses are taken into account. The bilingual pair contains all these equivalents, regardless the sense of the source word, because nowadays word sense disambiguation techniques are not reliable enough to automatically pair source-target words according to their senses

- The lemma of the word does not match an equivalent in the bilingual dictionary

The detection procedure consists in traversing the typed dependency tree and for each node the word form is verified whether it matches a word form in the monolingual dictionary. If not, a pair is searched for in the bilingual dictionary where the lemma of the word matches the source and an equivalent. If no pair is found, then the word form is an MTness instance. This strategy allows the detection of both NO-L2 and STR-CHAR instances.

The detection algorithm is the following:

1. Dependency parse the translation to get its typed dependency tree
2. For each line in the tree, with information about a word form (word form_i) and its lemma (word lemma_i),
   (a) Check whether word form_i is in the monolingual dictionary
   (b) If word form_i is in the monolingual dictionary then do nothing
   (c) Else, if word form_i is in the gazetteer of named entities then do nothing
   (d) Else, if the bilingual dictionary contains the following pair P2 (word lemma_i, target equivalents) where target equivalents = {... lemma_a, ... , **word lemma_i**, ...} then do nothing
   (e) Else, mark the dependency tree as a representation of a lexical MTness instance

Table 6.1 shows how some words are checked about its MTness.

Word	**MTness?**	**Explanation**
mónadas	NO	The word form matches a word form in the monolingual dictionary
Nassau	NO	The word form matches a word form in the gazetteer
ostinato	NO	Bilingual pair: ostinato - ostinato
allegiance	YES	Bilingual pair: allegiance - fidelidad, lealtad
Caribbean	YES	Bilingual pair: Caribbean - Caribe

Table 6.1: Examples of lexical MTness assessment

6.2.3 Detection of MTness instances in the relations between words

In this thesis we focused on relations between words when the governor is a noun or a verb because the three types of dependencies are more clearly appreciated. The method consists in checking whether there are linguistic items that force the parser to build a dependency tree representation (DTR) which is not consistent with linguistic intuition. The DTR can represent the root tree or a dependent subtree.

We distinguish three types of DTRs, each corresponding to a type of dependency.

1. Syntactic DTR
2. Morphological DTR
3. Semantic DTR

A DTR from a machine translation is called hypothesis DTR because its consistency with linguistic knowledge must be assessed. The consistency is assessed by one of the following actions

1. Matching hypothesis DTRs with reference DTRs
2. Testing the real use of co-occurrent words

The first action is focused on detecting MTness instances that fulfill the second and fourth conditions of salience (c.f. 5.4). That is, linguistic items that force a structure that is not recognized by a native speaker of the target language, and linguistic items that trigger an expected translation contrast. The second action is focused on detecting MTness instances that fulfil the third condition of salience: words that co-occur in a way that is not recognised by a native speaker′s linguistic and intuitive knowledge.

Let us explain these actions more fully.

6.2.3.1 Hypothesis DTR matching with reference DTR

The linguistic consistency of a hypothesis DTR is assessed when the hypothesis DTR matches a reference DTR; that is, a DTR that represents a recognizable dependency structure in the target language. On the contrary, MTness instances are detected in non-matched hypothesis DTRs.

a) Creation of reference DTRs In order to obtain reference DTRs, the dependency parser parses corpora with texts that are representative of the speakers′ use of the target language. The resulting typed dependency trees displays all the lexical, syntactic, morphological and semantic information of the sentences in the representative texts. For each root tree and dependent subtree, the DTR creation puts linguistic annotations from the typed dependency tree onto a tuple with at most three elements. The central element is for the linguistic annotation of the head. The other two are for the annotations of the specifier or complements, if any.

The creation of reference DTRs requires a highly reliable parser. According to (Lloberes et al., 2010), the evaluation of the parser we used showed that around 80% of the dependency trees had a correct head and a correct head and dependency relations. Despite the high reliability of the parser, we are aware that bad reference DTRs may be present. Yet, after counting the number of times each DTR describes the representative sentences, a frequency threshold of reference DTRs was established for the detector to consider a reference DTR reliable.

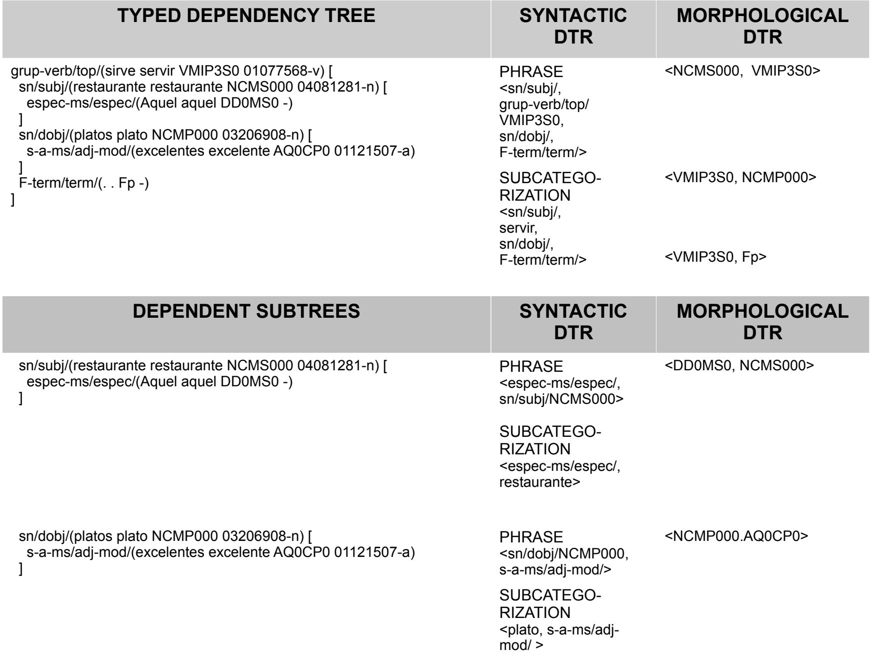
Although the percentage of possible ill reference DTRs is not significant enough to dismiss the use of a dependency parser, we noticed that detecting MTness instances by matching DTRs was more reliable for detecting syntactic and morphological MTness instances than for finding semantic MTness instances. Automatic semantic labeling depends on procedures such as word-sense-disambiguation whose results are still far to be reliable in any language. So we leave open for the future the possibility of detecting MTness instances by matching semantic DTRs, when semantic disambiguation and the calculations of semantic similarities between reference and hypothesis DTRs are more developed.

We will explain how syntactic and morphological reference DTRs are created, which are the more reliable representations for performing the MTness detection by matching hypothesis with reference DTRs.

a.1) Creation of reference syntactic DTRs There are two types of syntactic DTRs. The first one is the phrase DTR (DTR$_p$), which is the DTR that describes the dependency relations in terms of syntactic phrases and syntactic functions. The first position holds the phrase and function of the specifier, the central position holds the type and function of the phrase that dominates the head, and the third position holds the type and syntactic functions of the complements. The second type of syntactic DTR is the subcategorization DTR (DTR$_s$). The subcategorization DTR describes subcategorization relations when the governor has a specific lemma.

a.1) Creation of reference morphological DTRs Reference morphological DTRs are tuples with two positions. When the head has a specifier, the first position of the tuple is for the category values of the specifier, and the second position is for the category values of the head. Then, for each complement, a DTR is created whose first position is for the category values of the head, and the second position is for the category values of the complement. The reason is the assumption that the head restrictions are not applied at the phrase level but from head to dependent individually.

b) Creation of hypothesis DTRs Hypothesis DTRs are created from the typed dependency tree of a machine translation. Syntactic and morphological hypothesis DTRs are created the same way as reference DTRs. Figure 6.3 shows the hypothesis DTRs of the sentence *Aquel restaurante sirve platos excelentes* (That restaurant serves excellent dishes).



TYPED DEPENDENCY TREE	SYNTACTIC DTR	MORPHOLOGICAL DTR
grup-verb/top/(sirve servir VMIP3S0 01077568-v) [sn/subj/(restaurante restaurante NCMS000 04081281-n) [espec-ms/espec/(Aquel aquel DD0MS0 -)] sn/dobj/(platos plato NCMP000 03206908-n) [s-a-ms/adj-mod/(excelentes excelente AQ0CP0 01121507-a)] F-term/term/(. . Fp -)]	PHRASE <sn/subj/, grup-verb/top/ VMIP3S0, sn/dobj/, F-term/term/>	<NCMS000, VMIP3S0>
	SUBCATEGO- RIZATION <sn/subj/, servir, sn/dobj/, F-term/term/>	<VMIP3S0, NCMP000> <VMIP3S0, Fp>

DEPENDENT SUBTREES	SYNTACTIC DTR	MORPHOLOGICAL DTR
sn/subj/(restaurante restaurante NCMS000 04081281-n) [espec-ms/espec/(Aquel aquel DD0MS0 -)]	PHRASE <espec-ms/espec/, sn/subj/NCMS000>	<DD0MS0, NCMS000>
	SUBCATEGO- RIZATION <espec-ms/espec/, restaurante>	
sn/dobj/(platos plato NCMP000 03206908-n) [s-a-ms/adj-mod/(excelentes excelente AQ0CP0 01121507-a)]	PHRASE <sn/dobj/NCMP000, s-a-ms/adj-mod/>	<NCMP000.AQ0CP0>
	SUBCATEGO- RIZATION <plato, s-a-ms/adj- mod/ >	

Figure 6.3: Examples of hypothesis DTRs

c) Detection of MTness instances in syntactic DTRs An MTness instance is a linguistic item that is responsible for a hypothesis syntactic DTR not to be a reference DTR. The mismatched items can be:

1. the type of phrase of the dependents
2. the part of speech of the governor
3. the number of subcategorized dependents

If the types of phrase of the dependents do not match those of a reference phrase DTR, then the syntactic relation is against the linguistic intuition of native speakers and noisy segments (NOI-SEG) are produced. On the other hand, a mismatch in the part of speech of the governor explains, for example, I-POS and SYNT-GAP instances where a noun phrase dominates an adjective and the noun is missing. This is the reason why the category values of the head appear in phrase DTRs.

The mismatched number of subcategorized dependents are detected in subcategorization DTRs by calculating an expected syntactic DTR. The expected syntactic DTR is the reference subcategorization DTR whose edit distance to the hypothesis is the shortest. The MTness instance is detected by analyzing the necessary operations on the hypothesis DTR to get the expected syntactic DTR. The edit distance calculation models the evocation of the right translation since the MTness instance is the unexpected element in a familiar arrangement, like the wrong key at the end of the *Fur Elise* tune (c.f. 5.4.1).

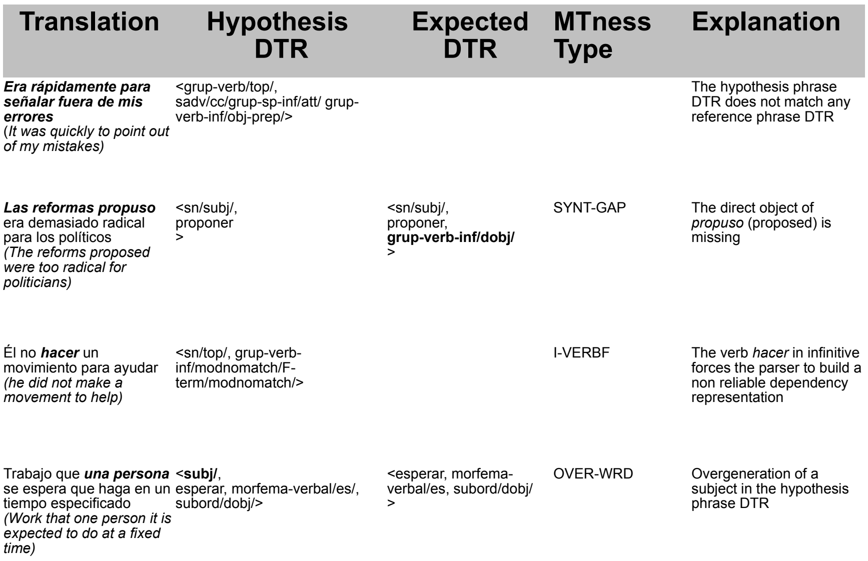
When a symbol must be inserted in order to get the expected syntactic DTR, and this symbol indicates a syntactic role (subject, direct object, indirect object or prepositional complement) then a complement with this syntactic role is expected. This is the case of SYNT-GAP and SEM-GAP instances. Finally, if a symbol must be deleted and this symbol indicates a syntactic role then a constituent with this role is not expected. This is the case of OVER-WRD cases, with an over generation of specifiers or complements.

The detection procedure is the following

1. Dependency parse the translation and obtain its typed dependency tree
2. Create all the syntactic hypothesis DTRs from the typed dependency tree.
3. For each hypothesis phrase DTR (DTR_p),
   (a) If the root of a verbal phrase is not a verb then DTR_p represents an MTness instance,
   (b) Else, if the root of a noun phrase is not a noun or pronoun then DTR_s represents an MTness instance
   (c) Else, if DTR_p does not match any reference phrase DTR then DTR_p represents an MTness instance.
4. For each hypothesis subcategorization DTR (DTR_s),
   (a) Obtain the expected syntactic DTR (DTR_e),
   (b) if a linguistic symbol must be inserted to get DTR_e and the symbol indicates either the subject, direct object, indirect object or prepositional complement functions then DTR_s represents an MTness instance.

(c) Else if a linguistic symbol must be deleted to get DTR_e and the symbol indicates either the subject, direct object, indirect object or prepositional complement functions then DTR_s represents an MTness instance.

Figure 6.4 shows some examples of MTness instances detected by the procedure explained above.



Translation	Hypothesis DTR	Expected DTR	MTness Type	Explanation
Era rápidamente para señalar fuera de mis errores (*It was quickly to point out of my mistakes)*	<grup-verb/top/, sadv/cc/grup-sp-inf/att/ grup-verb-inf/obj-prep/>			The hypothesis phrase DTR does not match any reference phrase DTR
Las reformas propuso era demasiado radical para los políticos *(The reforms proposed were too radical for politicians)*	<sn/subj/, proponer >	<sn/subj/, proponer, **grup-verb-inf/dobj/** >	SYNT-GAP	The direct object of *propuso* (proposed) is missing
Él no ***hacer*** un movimiento para ayudar *(he did not make a movement to help)*	<sn/top/, grup-verb-inf/modnomatch/F-term/modnomatch/>		I-VERBF	The verb *hacer* in infinitive forces the parser to build a non reliable dependency representation
Trabajo que ***una persona*** se espera que haga en un tiempo especificado *(Work that one person it is expected to do at a fixed time)*	<**subj/**, esperar, morfema-verbal/es/, subord/dobj/>	<esperar, morfema-verbal/es, subord/dobj/ >	OVER-WRD	Overgeneration of a subject in the hypothesis phrase DTR

Figure 6.4: Examples of MTness detection in syntactic DTRs

The SYNT-GAP instance exemplifies how the MTness phenomenon forces the parser to build an odd structure. The parser identifies *las reformas* as a subject, against the linguistic intuition of some people who told me that the subject was the missing element. All in all, the wrong subject forces the parser to build a structure with a missing direct object.

d) Detection of MTness instances in morphological DTRs Each morphological DTRs is assessed whether it represents an MTness instance. A morphological DTR represents an MTness instance if the DTR is not a reference DTR because of:

1. the grammatical category of the specifier, expressed in the morphological form
2. the grammatical category of a complement, expressed in the morphological form

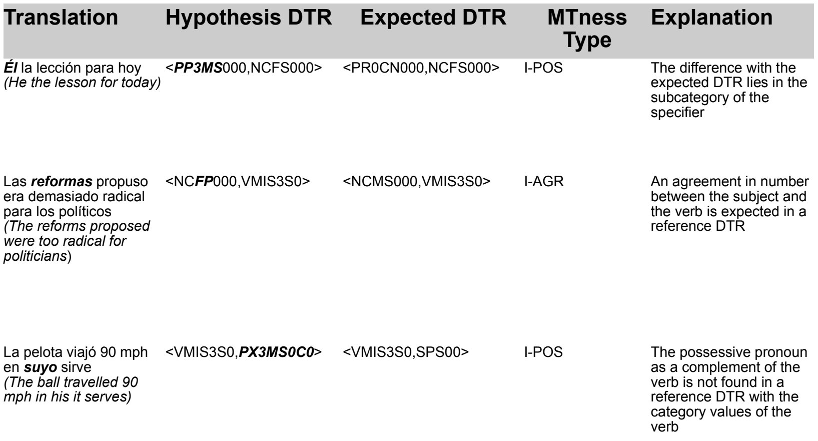
The first cause explains MTness instances such as I-AGR. In I-AGR instances, there is no reference DTR where the head, with its morphological form and grammatical values, governs a specifier holding the grammatical values of number, person or gender in the hypothesis DTR. The second cause explains MTness instances such as I-POS or I-VERBF. In these cases, there is no reference DTR where the head has a part of speech value or verbal form with the values of the complement as in the hypothesis DTR.

We assume that these MTness instances are salient if they trigger the right grammatical category values. The evocation of the right values is modeled by obtaining the expected morphological DTR, the same way the expected syntactic DTR was obtained. If the expected DTR is the result of replacing the value of part of speech, number, gender, or verb form/mood with a different value, then the DTR represents an MTness instance.

This is the detection algorithm:

1. Dependency parse the translation and obtain its typed dependency tree
2. Create all the morphological hypothesis DTRs from the typed dependency tree.
3. For each morphological hypothesis DTR_m if DTR_m does not match any reference morphological DTR
   (a) Retrieve the e_DTR (expected morphological DTR)
   (b) If the part of speech, number, gender or verb form/mood values of the governor or a dependent must be changed to get e_DTR then DTR_m represents an MTness instance.

Figure 6.5 shows some examples of MTness instances detected by the detection algorithm.



Translation	Hypothesis DTR	Expected DTR	MTness Type	Explanation
Él la lección para hoy *(He the lesson for today)*	<***PP3MS***000,NCFS000>	<PR0CN000,NCFS000>	I-POS	The difference with the expected DTR lies in the subcategory of the specifier
Las ***reformas*** propuso era demasiado radical para los políticos *(The reforms proposed were too radical for politicians*)	<NC***FP***000,VMIS3S0>	<NCMS000,VMIS3S0>	I-AGR	An agreement in number between the subject and the verb is expected in a reference DTR
La pelota viajó 90 mph en ***suyo*** sirve *(The ball travelled 90 mph in his it serves)*	<VMIS3S0,***PX3MS0C0***>	<VMIS3S0,SPS00>	I-POS	The possessive pronoun as a complement of the verb is not found in a reference DTR with the category values of the verb

FIGURE 6.5: Examples of MTness detection in morphological DTRs

6.2.3.2 Testing the real use of co-occurrent words

MTness instances can also be detected with the following assumption: in a dependency tree, if the governor does not occur with the specifier or one of the complements in the largest representative corpus available- the Web- then the dependency tree represents an MTness instance. Search engines are used to find evidences of two co-appearing word forms[7] and hence their grammatical or semantic coherence.

The function of the search engine is to retrieve contexts from the Web search engine where the governor coappears with the specifier or a complement in context, either together or with a distance between them. The requests are performed with queries interpretable by the search engine, and the key words are the word forms of the head and the word form of the specifier or the complement, because the search engine matches word forms, not lemmas. The queries have the following patterns:

- {specifier word form+near:d_1+ head word form}[8]

[7] The matching of co-occurring words on the Web is also the base for the *Normalized Google Distance*, which measures the semantic relatedness between words (Cilibrasi2007)

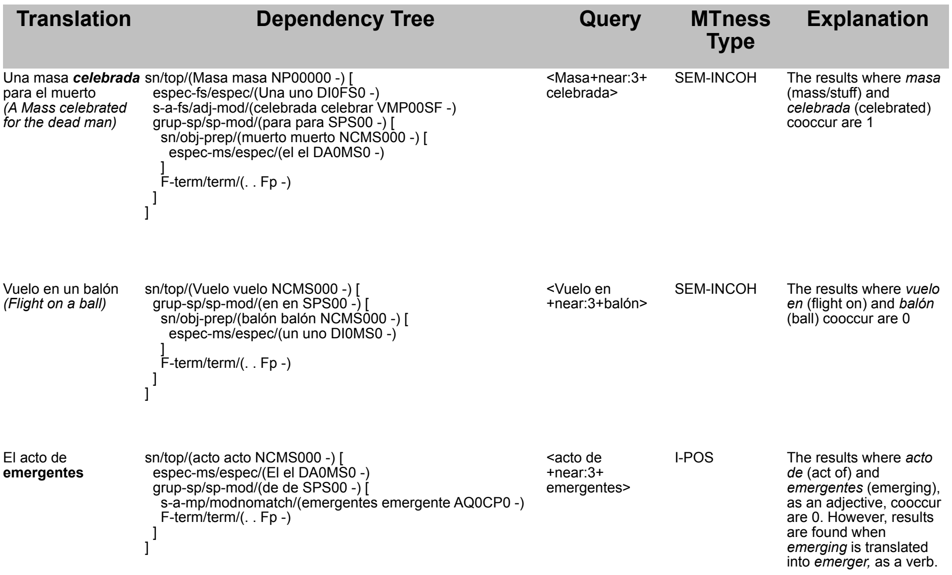
[8] d_1 indicates the maximum number of words that may be found between the specifier and the head. As an example *teacher+near:3+sings* is the query for the search engine to find documents where sings appears with teacher, which is its specifier (subject), separated by 3 words at most.

- {head word form+near:d_2+complement word form}[9]

The order of the key words indicates the order in which these words should appear in the matching contexts. The order specifier-head-complement is suitable for retrieving contexts in languages such as English, Spanish or Catalan. For other languages with a different argumental order, the queries should be adapted

The MTness instance is detected if the search engine retrieves no results or the number of results is below 3. We establish a threshold of 3 because it is possible that the co-occurring words match contexts of non revised machine translated documents. We assume that the number of these misleading contexts is generally below 3. If the number of results is over 3, then the number of matching snippets are counted. Matching snippets are those where the words co-occur with no punctuation marks that initiate another clause in between (e.g. period, semicolon, parenthesis). If the number of matching snippets is below 3 then the semantic DTR contains an instance of MTness.

The MTness types detected are mainly SEM-INCOH although other types are also detected, as shown in figure 6.6.



Translation	Dependency Tree	Query	MTness Type	Explanation
Una masa ***celebrada*** para el muerto *(A Mass celebrated for the dead man)*	sn/top/(Masa masa NP00000 -) [espec-fs/espec/(Una uno DI0FS0 -) s-a-fs/adj-mod/(celebrada celebrar VMP00SF -) grup-sp/sp-mod/(para para SPS00 -) [sn/obj-prep/(muerto muerto NCMS000 -) [espec-ms/espec/(el el DA0MS0 -)] F-term/term/(. . Fp -)]]	<Masa+near:3+ celebrada>	SEM-INCOH	The results where *masa* (mass/stuff) and *celebrada* (celebrated) cooccur are 1
Vuelo en un balón *(Flight on a ball)*	sn/top/(Vuelo vuelo NCMS000 -) [grup-sp/sp-mod/(en en SPS00 -) [sn/obj-prep/(balón balón NCMS000 -) [espec-ms/espec/(un uno DI0MS0 -)] F-term/term/(. . Fp -)]]	<Vuelo en +near:3+balón>	SEM-INCOH	The results where *vuelo en* (flight on) and *balón* (ball) cooccur are 0
El acto de **emergentes**	sn/top/(acto acto NCMS000 -) [espec-ms/espec/(El el DA0MS0 -) grup-sp/sp-mod/(de de SPS00 -) [s-a-mp/modnomatch/(emergentes emergente AQ0CP0 -) F-term/term/(. . Fp -)]]	<acto de +near:3+ emergentes>	I-POS	The results where *acto de* (act of) and *emergentes* (emerging), as an adjective, cooccur are 0. However, results are found when *emerging* is translated into *emerger,* as a verb.

Figure 6.6: Examples of MTness co-ocurrent words

This is the detection algorithm:

[9] d_2 indicates the maximum number of words that may be found between the head and one of its complements. For instance, *killed+near:3+Kennedy* is the query for the search engine to find contexts where the complement *Kennedy* apears with *killed*, separated by 3 words at most.

1. Dependency parse the translation and obtain its typed dependency tree
2. When the specifier of the main tree is not empty, generate the following query: ⟨specifier form+near:3, head form⟩
3. For each complement of the main tree
   (a) If the complement is an adjective, generate the query: ⟨head form+near:2+adjective form⟩ else the query is ⟨head form+near:3+complement form⟩
   (b) If the head of a complement is a preposition, attach the preposition to the head of the main tree and generate the following query: ⟨head form with preposition attached +near:3+prepositional complement form⟩
4. For each dependent subtree create queries as stated in steps 2 and 3
5. For each query
   (a) If the search engine retrieves no results or the number of results is below 3 then the tree out of which the query is generated describes an MTness instance.
   (b) Else, in the results page, count the number of snippets where the word forms in the query appear with no punctuation marks (initiating another clause) in between. If the number is below 3, then the tree out of which the query is generated describes an MTness instance.

6.2.3.3 MTness types with no specific algorithms

The algorithms we presented cannot take word positions into account, because DTR are hierarchically- not linearly- ordered. Therefore instances of I-ORD (inadequate order) and WRD-REP (the two same words very close together) are not detected by retrieving an expected syntactic DTR. Anyway, I-ORD and WRD-REP cause ill dependency structures that are assessed as having an MTness instance. Very recent annotations formalisms have appeared where this information is provided, such as *Kyoto Annotation Framework* (KAF) (Bosma et al., 2009), so we leave open the detection with position information.

6.3 Summary

In this chapter we have explained the MTness detection methods, which are summarised in table 6.2.

Condition of MTness salience	**Action**	**Resources**	**MTness types**
No evidence of use of the MTness instance in the target language	Checking word lemmas in the typed dependency tree	Monolingual and bilingual dictionary	NO-L2 STR-CHAR
Structure that is not recognised by a native speaker′s linguistic and intuitive knowledge. Expected translation contrast.	Hypothesis DTRs matching with reference DTRs	Reference DTR from a large sample of texts in the target language	I-POS SYNT-GAP SEM-GAP OVER-WRD I-AGR I-VERBF
Words related in a way that is not consistent with the language use of native speakers	Testing the real use of co-occurrent words	Web Search engine	SEM-INCOH I-POS
Expected translation contrast	-	-	I-ORD WRD-REP

Table 6.2: MTness detection methods

The detection is applied at the syntactic, lexical and semantic levels. This is a step further than our early proposal, which was lexically oriented. The current detection is performed by using information from a dependency parser. Several detection algorithms are run to find instances at each linguistic level. The overlapping of instances in more than one linguistic level produces the MTness saturation or at least the reader′s astonishment. This is reflected in the MTness score, as we will explain in the following chapter.

Chapter 7

Calculation of the MTS Metric

In this chapter we will explain the MTS metric and how it is calculated. The MTS rates the quality of machine translations according to the MTness instances detected. MTness-based evaluation is not a quantitative- the more instances the worse- but a qualitative evaluation; it depends on how salient the MTness instance is and the overlapping degree of MTness instances, as we stated in chapter 5.

7.1 The MTS score

MTS (MTness Score) is a metric that rates the machine translationness of a piece of text (translation unit). MTS values range from 0 to 1. 0 means that no traces of MTness were detected and 1 signifies that the piece of text was unquestionably produced by a machine. Values between 0 and 1 indicate how close the translation unit is to a piece of text where all the words are affected by machine translationness.

As we said in chapter 5, the score must be consistent with the fact that one single word can spoil the translation and, on the other hand, the score must capture the MTness saturation effect, caused by the overlapping of MTness instances at different linguistic levels. The single presence of a word not used in the target language spoils the translation. Therefore translations with STR-CHAR and NO-L2 MTness instances have the highest MTS value. The saturation effect of MTness instances at different linguistic levels are modeled by combining the values of the following three metrics:

- syntactic mts: Metric that rates MTness when only ill syntactic dependencies are detected

- morphological mts: Metric that rates MTness when only morphologically inconsistent dependencies are found.
- co-occurrent mts: Metric that rates MTness when only inconsistent co-occurrent words are found.

These metrics are considered partial MTS scores. For this reason, they are labelled as MTS but in lower case. The partial score illustrates how MTness in either the syntactic, morphological or word co-occurrence levels contribute in obtaining the final MTS value. We will explain how the values of the syntactic, morphological and semantic mts are calculated, and then we will explain how these values are combined in order to obtain the MTS.

7.1.1 The mts calculation

The reasoning behind the mts calculation is to rate how close the translation is to the worst of the situations, that is a translation with the highest MTness saturation effect in a linguistic level. The highest syntactic mts indicates that all the nodes of the typed dependency tree appear in syntactic DTRs with MTness instances. The highest morphological mts indicates that all the nodes of the typed dependency tree appear in morphological DTRs with MTness instances. Finally, the highest co-occurrent mts indicates that all the dependent words of the typed dependency tree cannot co-occur with their governors.

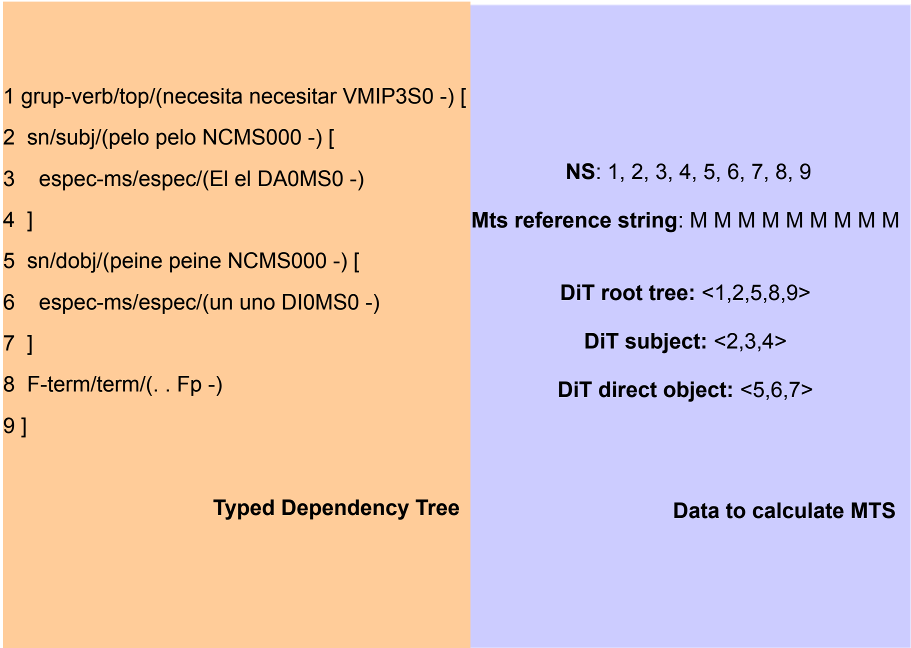
All the lines of the typed dependency tree representation, which represent the nodes of the tree, are indexed. We decided to index them with their line number. Then the ordered indexes are concatenated in a string. This is the node string (NS). The worst of the situations is modeled by a string where all the indexes of the NS are replaced by the symbol 'M' ('M' stands for 'machine translationness'). This is the mts reference string, which means that all the nodes are affected by MTness.

The mts indicates how close the mts hypothesis string, which models the actual MTness status of the translation, is to the mts reference string. The mts hypothesis string is generated by replacing the indexes of the NS affected by MTness with the symbol 'M'.

The distance of the hypothesis to the reference is rated by using a metric that takes precision and recall of a hypothesis string with respect to a reference. We chose ROUGE-L because this metric takes into account the consecutive positions of the indexes with MTness in the hypothesis. We assume that the more consecutive the positions are, the more impact in the perception of MTness.

The mts hypothesis string is generated with the dependency index tuple (DiT). For the root tree and for each dependent subtree a DiT is created with the indexes of the governor, the indexes of the dependents and the index of the end of the dependency tree. When the root tree or dependent subtree represents an MTness instance, the DiT indexes of the dependents are replaced with the symbol 'M' in the node string[1]. The mts hypothesis string is the result of these replacements for the root tree and all the dependent subtrees.

Figure 7.1 shows the NS, the reference string and the DiTs of the Spanish translation *El pelo necesita un peine* (the hair needs a comb).



```
1 grup-verb/top/(necesita necesitar VMIP3S0 -) [
2  sn/subj/(pelo pelo NCMS000 -) [
3    espec-ms/espec/(El el DA0MS0 -)
4  ]
5  sn/dobj/(peine peine NCMS000 -) [
6    espec-ms/espec/(un uno DI0MS0 -)
7  ]
8  F-term/term/(. . Fp -)
9 ]
```

NS: 1, 2, 3, 4, 5, 6, 7, 8, 9

Mts reference string: M M M M M M M M M

DiT root tree: <1,2,5,8,9>

DiT subject: <2,3,4>

DiT direct object: <5,6,7>

FIGURE 7.1: NS, reference string and DiTs of *El pelo necesita un peine*

7.1.1.1 The syntactic mts calculation

To calculate the syntactic mts the mts hypothesis string must be created. From the DTRs representing syntactic MTness instances, the DiTs of either the root tree or the dependent subtrees affected are retrieved. The hypothesis string is created by substituting the indexes of the dependents in the DiTs with the symbol 'M' in the node string. Then the ROUGE-L is calculated with the mts reference and the mts hypothesis string.

[1] The DiT index of the governor is not replaced by the symbol 'M' because the governor is not an MTness instance by itself but the dependents that are wrongly related to the governor

7.1.1.2 The morphological mts calculation

From the morphological DTRs representing MTness instances, the DiTs of either the root tree or the dependent subtrees affected are retrieved. The hypothesis string is created by substituting the indexes of the DiTs with the symbol 'M' in the node string. Then the ROUGE-L is calculated with the mts reference and the mts hypothesis string.

7.1.1.3 The co-occurrent mts calculation

From the DiTs of either the root tree or the dependent subtrees the indexes of ill co-occurring words are retrieved. The hypothesis string is created by substituting the indexes of the DiTs with the symbol 'M' in the node string. Then the ROUGE-L is calculated with the mts reference and the mts hypothesis string.

7.1.2 The MTS calculation algorithm

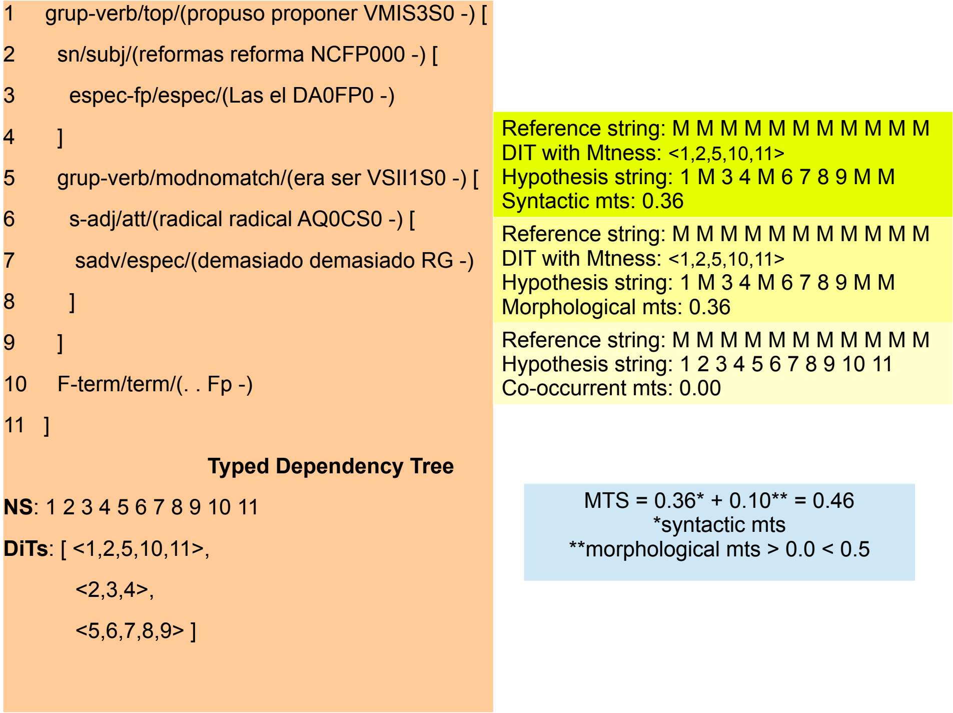
When the MTness instance is NO-L2 or STR-CHAR the MTS value is 1. Otherwise, the value is calculated according to the cumulative effect of MTness instances across different linguistic levels, which is modeled in the following algorithm:

- Score the partial MTS with the highest mts from the syntactic, morphological and semantic mts
- For each remaining mts over 0.5, increase the partial MTS by two tenths
- For each remaining mts over 0 and below 0.5, increase the partial MTS by one tenth
- Equal the definitive MTS to the current partial MTS value

Sometimes the mts in one type is so high that when added to other mts, the sum is over 1. In that case, the value is normalized to 1.

7.2 An example of MTS calculation

Figure 7.2 illustrates the MTS calculation of the translation *las reformas propuso era demasiado radical* (the reforms he proposed was too radical).



```
1    grup-verb/top/(propuso proponer VMIS3S0 -) [
2      sn/subj/(reformas reforma NCFP000 -) [
3        espec-fp/espec/(Las el DA0FP0 -)
4      ]
5      grup-verb/modnomatch/(era ser VSII1S0 -) [
6        s-adj/att/(radical radical AQ0CS0 -) [
7          sadv/espec/(demasiado demasiado RG -)
8        ]
9      ]
10     F-term/term/(. . Fp -)
11   ]
```

Typed Dependency Tree

NS: 1 2 3 4 5 6 7 8 9 10 11

DiTs: [<1,2,5,10,11>,
<2,3,4>,
<5,6,7,8,9>]

Reference string: M M M M M M M M M M M
DIT with Mtness: <1,2,5,10,11>
Hypothesis string: 1 M 3 4 M 6 7 8 9 M M
Syntactic mts: 0.36

Reference string: M M M M M M M M M M M
DIT with Mtness: <1,2,5,10,11>
Hypothesis string: 1 M 3 4 M 6 7 8 9 M M
Morphological mts: 0.36

Reference string: M M M M M M M M M M M
Hypothesis string: 1 2 3 4 5 6 7 8 9 10 11
Co-occurrent mts: 0.00

MTS = 0.36* + 0.10** = 0.46
*syntactic mts
**morphological mts > 0.0 < 0.5

Figure 7.2: MTS calculation of *las reformas propuso era demasiado radical*

The figure displays the DiTs of the root tree and the dependent subtrees of the subject noun phrase, and the verbal phrase whose head is the verb 'era'. The expected syntactic DTR of the root tree (table 7.1) indicates that the direct object of the verb *propuso* (proposed) is missing. So the hypothesis DTR describes a SYNT-GAP instance. In the DiT of the root tree, the dependents have the indexes 2, 5, 10 and the limit index of the tree is 11. Therefore, these indexes in the node string are replaced by the symbol 'M' to generate the hypothesis for calculating the syntactic mts.

Hypothesis DTR_s	**Expected DTR_s**
<sn/subj/,proponer/grup-verb/modnomatch/ >	<sn/subj/,proponer/grup-verb-inf/**dobj** >

Table 7.1: SYNT-GAP instance

One may wonder why all the indexes of the dependents are replaced by M when the MTness instance is a missing object while the relation between the head and the dependents can be correct. We explain this by using a simile. Imagine a chair with three legs in perfect condition but one leg is shorter than the others. The shorter leg makes the chair as a whole a useless object, regardless the quality and state of the other legs. The same happens in a dependent subtree with a single MTness instance. This instance affects the structure as a whole, and this consequence is modelled by substituting the indexes of the dependents with the symbol 'M'.

The fact that *las reformas* is wrongly typed as the subject of the main verb *propuso* has consequences at the morphological level. The expected morphological DTR of the root tree (table 7.2) indicates that the expected number feature of the subject should agree with the verb. So the hypothesis DTR describes an I-AGR instance.

Hypothesis DTR	**Expected DTR**
<NCF**P**000, VMIS3S0 >	<NCM**S**000, VMIS3S0 >

Table 7.2: I-AGR instance

So the indexes 2, 5, 10 and the limit index of the tree in the node string are replaced with the symbol 'M' to generate the hypothesis for calculating the morphological mts.

The highest mts value corresponds to the value of the syntactic mts and the morphological mts. If we take the syntactic mts value as the partial MTS, the definite MTS value is obtained by adding one tenth because the morphological value is over 0 and below 0.5.

7.2.1 Calculation of the MTS in multi-sentence translations

When the translation has more than one sentence, the MTS score is the result of merging the MTS of each sentence. The calculation is similar to the calculation in single translations but the factors are MTS scores instead of mts. First we score the partial MTS, which is the highest MTS, and then for each remaining MTS over 0.5, the MTS is raised by two tenths. On the other hand, for each remaining MTS over 0 and below 0.5 the MTS is raised by one tenth.

Let us see the MTS calculation for the translation *El acto social de montaje para un propsito comn. Su reunin con los vendedores fue el punto culminante de su da* (The social act of assembly for a common purpose. His meeting with salesmen was the peak of his day)

MTS sentence 1: 0.40

MTS sentence 2: 0.23

Then the MTS score is:

MTS = 0.40 + 0.10 (MTS sentence 2 >0 and <0.5) = 0.50.

7.3 Summary

In this chapter we explained how the MTS score is calculated, either for a one-sentence translation or a multi-sentence translation. The score displays the spoiling effect of one single word and the MTness saturation effect caused by the overlapping of MTness instances at different linguistic levels. In the next chapter, we will explain how this multidimensional score correlates to the human perception of translation quality.

Chapter 8

Experimental Evaluation of the Proposal

In this chapter we will explain the evaluation experiment of MTS as a metric for evaluating machine translations. In order to perform this experiment we developed an evaluation tool- MTness Eval. This tool will be described in the first section. Subsequently we will discuss the results of the experiment and the contribution of the MTS metric in relation to state-of-the-art metrics. The comparison will be presented from these points of view: evaluation costs, the rating of fluency and accuracy, and the distinction between human and machine likeness.

We will see that MTS scores significantly correlate with human judgements and the correlations are better than state-of-the-art evaluation metrics. Since the good correlation results were obtained with lower costs than the RPA (c.f. 2.1.2.1) and the classification approaches (c.f. 2.1.2.3), the applicability of the proposed method was positively assessed.

8.1 MTness-Eval: an MTness evaluation tool

MTness-Eval is a Perl-based application that assigns a MTS value to a machine translation. MTness-Eval detects MTness instances with the algorithms explained in chapter 6 and rates the translation accordingly.

MTness-Eval scores translations through three stages:

1. Parsing and hypothesis DTR generation and queries

2. MTness detection

3. Scoring

Let us explain each stage more fully, following the flow shown in figure 8.1.

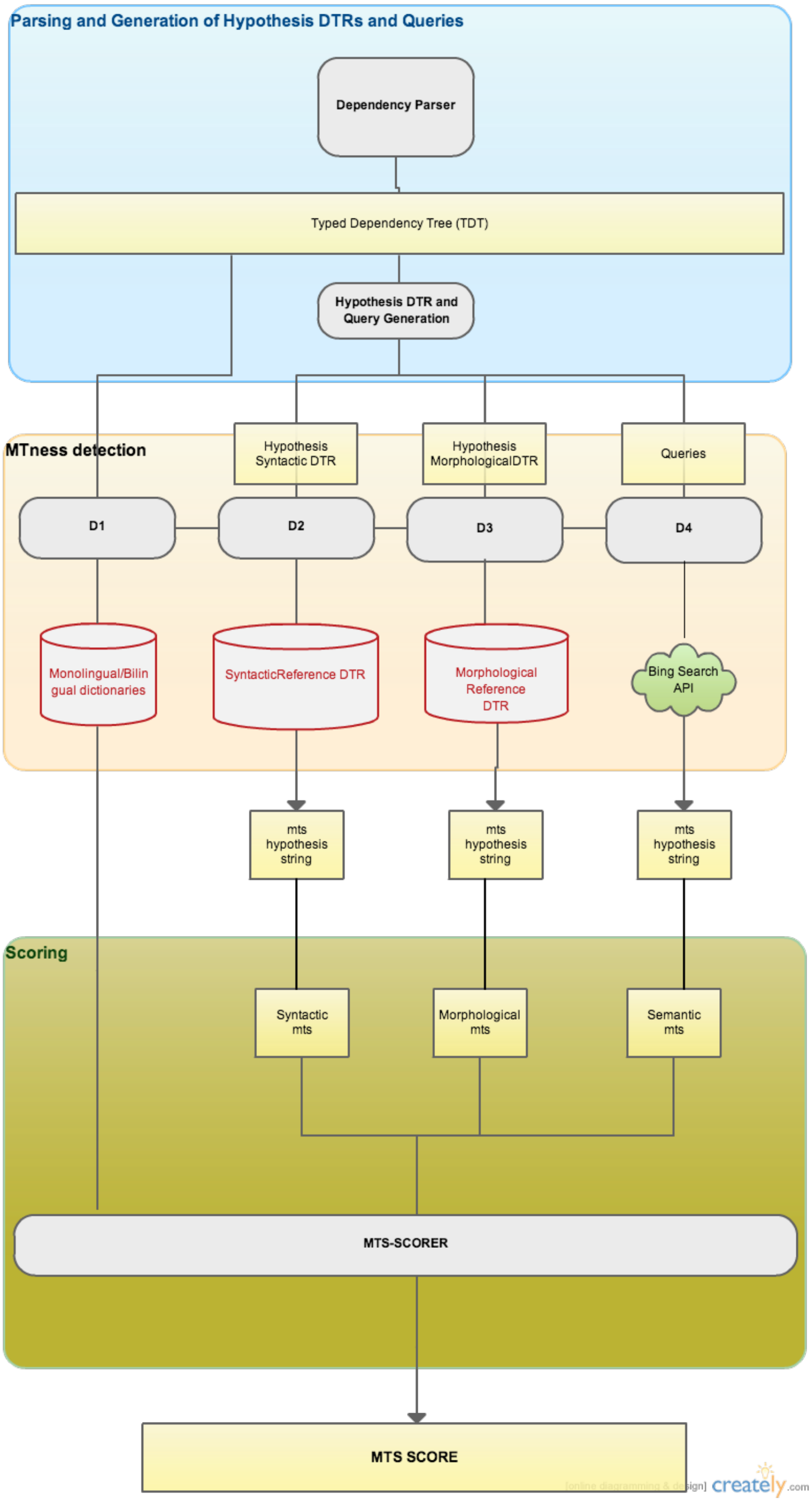


FIGURE 8.1: MTness-Eval flow

8.1.1 Parsing and generation of hypothesis DTR and queries

The translation is dependency-parsed. Once the whole type dependency tree is obtained, then the hypothesis syntactic and morphological DTRs are generated, as we explained in 6.2.3.1. The queries for the search engine to find contexts of co-occurring words are also prepared.

8.1.2 MTness detection

MTness detection is performed by four detectors:

D1: Detector of lexical MTness instances in the dependency tree

D2: Detector of MTness instances in the syntactic DTR

D3: Detector of MTness instances in the morphological DTR

D4: Detector of MTness instances in co-occurring words

The detection algorithms are the ones explained in chapter 6. The outputs of detectors D2, D3 and D4 are the corresponding mts hypothesis strings, explained in 7.1.1. The resources used by D1 are the monolingual and the bilingual dictionaries (c.f 6.2.2). The D2 detector calculates the edit distance of syntactic hypothesis DTRs with reference DTR stored in the Syntactic-DTR database. The D3 detector calculates the edit distance of morphological hypothesis DTRs with reference DTRs stored in the Morphological-DTR database. Finally D4 uses the API of the search engine in order to find evidences of word co-occurrence on the Web.

8.1.3 Scoring

If there is no lexical MTness instance, the scoring is performed by four components, each specialized in calculating a mts score with the output of each detector. The calculation is performed as explained in 7.1. Then, with all the mts scores, the MTS scorer draws the definitive value, the MTS score.

8.2 Evaluation costs

We will discuss evaluation costs from three points of view. The first point of view will be the costs of performing the experiment. The second point of view will be the expenses in obtaining the MTness-Eval resources and the third will be the reusability of the data obtained.

8.2.1 Expenses of the experiment

Test sets are generally very large and their compilation is expensive. Moreover in evaluation campaigns evaluators have to read many sentences thoroughly and attentively. Fatigue and cognitive saturation may affect their judgements and their motivation to be as fair as possible.

We adopted the reasoning of (Reeder, 2001). The reasoning is the following: readers are able to differentiate native from non-native language compositions by reading a short sample of texts. We assume that the same happens to machine-like translations. So there is no need to analyze a large number of translations for distinguishing systems that produce machine-like (bad) output. Our method then saves the effort of compiling a large test set.

The evaluation corpus had 196 machine translations into Spanish of Wordnet glosses originally written in English[1]. Our evaluators considered the corpus size large enough to perform the task with interest and motivation and, as we will see later, the correlation coefficient with human judgements was significant.

The translations compiled for the experiment were performed by a rule-based system and a statistically-based system, and the number of translations performed by either of them was balanced

8.2.2 Expenses in the MTness-Eval resources

The resources were, on the one hand, the monolingual and bilingual dictionaries and, on the other hand, the textual corpora used to generate the reference DTRs. The tool also uses NLP technology such as the dependency parser and the search engine. All these resources can be obtained and exploited with no costs. Let us present them, and we will discuss the contribution of our experiment to current evaluations as far as costs are concerned.

[1] To obtain machine translations of Wordnet glosses was a task of the KNOW-2 project. The project was funded by the former Spanish Science and Innovation Ministry (Ministerio de Ciencia e Innovación) TIN2009-14715-C04, and consisted in updating the Spanish and Catalan Wordnets to version 3.0 (Oliver and Climent, 2012). The use of machine translation was a fast and inexpensive way of updating the Wordnets in these languages.

8.2.2.1 Dictionaries

The monolingual dictionary used for the experiment was generated from the Spanish lexical database used by the parser. The dictionary also contains a gazetteer of named entities used by the parser as well. The bilingual dictionary was automatically generated from the Spanish-English and English-Spanish translations in the Wiktionary Spanish index[2]. These monolingual dictionary and the gazetter are licensed under the General Public License (GPL), because they are components of the open-source language suite FreeLing.. The data obtained from Wiktionary are licensed under the Creative Commons Attribution-ShareAlike 3.0 Unported License, as well as the GNU Free Documentation License.

8.2.2.2 The reference syntactic and morphological DTR

The reference syntactic and morphological DTRs were obtained by parsing a representative corpus taken from freely available sources described in table 8.1.

Source	Word definitions in Spanish Wiktionary (eswikitionary20120718) News and articles from newspapers (El País, ABC, El Mundo)
Conditions	Wiktionary: Creative Commons News and articles: Free Download
Total size	2.273.915 words (130.237 sentences)

TABLE 8.1: Sources of the syntactic refererence patterns

8.2.2.3 The dependency parser

The dependency parser- the Txala parser- belongs to the FreeLing suite[3] (Atserias et al., 2006). FreeLing is an open source language analyzer tool suite, released under the GNU General Public License (GPL). It is developed and maintained by the TALP Research Center at the Universitat Politècnica de Catalunya (UPC), with contributions from the community around the suite. The supported languages are Spanish, Catalan, Galician, Italian, English, Russian, Portuguese, Welsh and Asturian.

[2] http://en.wiktionary.org/wiki/Index:Spanish

[3] http://nlp.lsi.edu/freeling/index.php

8.2.2.4 The search engine

Currently, the MTness-Eval uses the Bing Application Programming Interface (API)[4], which allows developers to submit queries to and retrieve results from the Bing Search Engine. At the time of the experiment, the use of the API was free and there was no restriction in the number of queries. We are aware that these conditions may change. We expect that in the future consistency in co-occurring words will be evaluated by using resources whose conditions of use cannot change overnight.

8.2.2.5 Resource costs in comparison to state-of-the-art evaluations

The free availability of the resources is very important for the real application of our method. In fact reference-based and classification-based evaluations are very expensive, and only large institutions and companies that produce lots and lots of human and machine translations can afford the expenses of these methods (c.f. 2.2). (Callison-Burch, 2007) says that most of the evaluations are performed with only one reference because the cost of creating more references is prohibitively high and the available multi reference test suites are limited to a small number of languages. In section 2.2 we also noted the expenses in the revision of references. Nowadays, the growing use of machine translation by human translators is, by itself, an important reason to consider the revision.

Our method does not need training data and the DTRs are created from corpora which do not belong to specific domains. So the costs of adapting the resources whenever the domain changes are also saved. This contrasts with evaluations where the test set and the corpora used to train statistical MT systems share the same domain (evaluations for the annual Workshops on Statistical Machine Translation) or the MT evaluator of (Gamon et al., 2005).

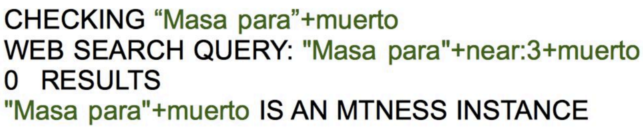
8.2.3 Reusability of the data

MTness Eval scored each translation and registered the MTness instances in a log file, with an explanation of the detection. The log file was conceived as a source of information for MT developers who need to locate the critical errors of their systems. For instance, the log file registers the words that were not recognized as target language words. The bilingual dictionary can be extended with the translation of these words. The detection of bad hypothesis DTRs is also useful for developers to know the syntactic and semantic drawbacks of the system in a fast and efficient way.

[4] http://www.bing.com/toolbox/bingsearchapi

The log file was also conceived as a repository of MTness instances that can be fixed in an automatic postedition module by performing easy regular expression operations. For instance, a systematically mistranslated word can be replaced with the right solution by means of a postedition tool. Besides, errors that are often overlooked can also be highlighted in the translation to catch the posteditor′s attention. The identification of systematic errors eases the work of human translators when dealing with large amounts of text, as (Gamon et al., 2005) and (Richardson, 2004) pointed out.

Figure 8.2 shows the explanation of an MTness instance detected, as it is registered in the log file.



```
CHECKING “Masa para”+muerto
WEB SEARCH QUERY: "Masa para"+near:3+muerto
0  RESULTS
"Masa para"+muerto IS AN MTNESS INSTANCE
```

Figure 8.2: Log excerpt explaining why *Una masa celebrada para el muerto* (A mass celebrated for the dead man) is an MTness instance

MTness Eval allows the user to enrich reference DTR with good hypothesis DTR that were not evaluated as such according to the log file. For example, we added 11 good syntactic hypothesis DTRs that did not match any reference DTR.

8.3 The rating of fluency and accuracy

The MTS calculation procedure permits to know if the translation is wrong because of fluency or adequacy. When the MTS is drawn from the syntactic or morphological mts then the translation is wrong because of fluency, whereas if the MTS comes out from ill word co-occurrence then the translation is wrong because of accuracy. To detect odd translations means to detect flaws in accuracy. At least the reader suspects about the fidelity to the original.

Our method detects very disfluent sentences and the inaccurate translations detected are those whose co-occurrent words lead to absurd, odd, and unintelligible sentences. We are aware that this method does not capture subtle grades of fluency and accuracy, as human evaluations and state-of-the-art metrics intend to capture. However, our more coarse-grained evaluation metric proved to be good enough to evaluate the translation quality of the machine output. The results correlated better with quality perception than state-of-the-art metrics.

This section deals with the experiment that evaluated MTS as a metric that captures fluency and accuracy. We will first present the evaluators. Then we will explain the translation quality scale the evaluators used to rate the translations. We will compare the correlation results of the MTS values with the correlation of state-of-the art metrics and, finally, a discussion will follow about the differences in the approaches to fluency and accuracy between our method and state-of-the-art methods. These differences explain the correlation differences between them.

8.3.1 Evaluators

Four people participated in the experiment. They had different educational and professional backgrounds and their ages spanned from 32 to 60 years of age (32, 35, 52, and 60 years-old respectively). They evaluated the translation in isolation and were not time-pressed. The evaluators did not have any experience in testing MT systems, and the professional backgrounds of three of them were not related to proofreading, postediting nor any other activity where language quality testing was involved. One of the evaluators had some experience as an editor and proofreader and her scores were used in order to assess to what extent judgements were influenced by linguistic expertise.

We planned to increase the number of participants in case the data collected were not enough to draw a relevant conclusion. However, the data obtained proved to be relevant about the pertinence of our conclusions.

8.3.2 Translation quality scale

Machine translation evaluation metrics are assessed by calculating the correlations with the human judgements on fluency and then calculating the correlations with the judgements on accuracy. So the metric can be analysed in terms of which of these items the metric correlates better. Generally evaluators only read the translation when rating fluency and read both the original and the translation when they read accuracy.

We were interested in the linguistic intuition of monolingual ordinary readers in detecting flagrant disfluent and inaccurate translations. So we did not need bilingual evaluators to judge disfluent translations and judge the inaccuracy of odd and absurd translations. We realized that a standard DARPA scale was suitable for our experiment. This scale has five points: 1- Incomprehensible, 2- Disfluent, 3- Non-native, 4- Good, 5- Flawless English (Spanish for our experiment). Although the scale is for fluency we considered that translations with MTness instances affecting accuracy could be rated as incomprehensible.

8.3.3 Correlation analysis

We calculated the Pearson correlation coefficient between the MTS scores and the evaluators′ judgements. The correlation coefficient indicated to what extent MTS scores matched the human judgements. Then we wanted to assess whether the MTS correlated better than state-of-the art metrics.

8.3.3.1 Correlation variables in MTS

The first variable is the H score and the second variable is the MTS value. The H score is the mean of the ratings of the three evaluators using the DARPA scale. The H score is an objective indication of the human quality appreciation despite discrepancies.

The agreement between the non-expert evaluators was low. The Fleiss Kappa[5] was 0.172. In the group of evaluators with an expert, the value was higher: 0.484. Curiously, the expert agreed more with the intuitions of the non-experts. That meant that linguistic expertise did not influence much in the perception of MTness. Therefore we decided to take the scores of the non-expert group to obtain the MTS correlation result and compare it with the correlations of state-of-the-art metrics.

Although the human agreement rate is normally low ((Koehn, 2012), (Koehn, 2010), and (Callison-Burch et al., 2007)) we wondered whether low agreements were caused by the descriptions of the points of the scale. The evaluators especially asked about the differences between 4 (good) and 5 (flawless Spanish) and the difference between 2 (disfluent) and 3 (non-native). If the DARPA scale was turned into a rougher scale, where the differences between good and bad translations were more clearly cut (see table 8.2[6]), the Fleiss Kappa for the non-expert evaluators was higher (0.301).

Quality Value	Mapped Value
1 (Incomprehensible)	1
2 (Disfluent)	1
3 (Non-native)	2
4 (Good)	3
5 (Flawless Spanish)	3

Table 8.2: Fluency measures mapped onto a 3-point scale

[5] Unlike Cohen′s Kappa coefficient, which works for two evaluators, Fleiss′ Kappa works for any number of raters. For the interpretation of Fleiss Kappa, see (Landis and Koch, 1977)

[6] The quality values *Incomprehensible* and Disfluent share the lowest value because they represent the worst appreciation for accuracy and fluency. *Non-native* value is taken as an intermediate value. See 9.3.3 about the differences between MTness and disfluent performance by native speakers

The MTS variable should decrease as the H score increases (the better the translation the lower the MTS). Therefore, the coefficient should be a negative fraction, away from 0 (the two variables do not vary together at all) and tending to -1, which is the value of the perfect negative or inverse correlation.

Correlation can be represented by means of a scatterplot that shows the relationship between the two variables. The correlation is a measure of the degree to which points (pair of numbers) in the scatterplot cluster together around a straight line. When two variables correlate, the line shows clearly the increasing or decreasing trend in value variations and the most predictable points are scattered around this line.

8.3.3.2 MTS correlation and comparison to state-of-the-art metrics

The MTS correlation coefficient was -0.71 ($t = -14.2268$, $df = 194$, $p\text{-value} < 2.2e\text{-}16$). In terms of Cohen′s *effect size*(Cohen, 1988), a Pearson′*s* correlation value above 0.5 has a large effect. Let us compare this correlation coefficient with the coefficients of state-of-the-art metrics. Contrary to MTS, the variable of the state-of-the-art metric should increase as the fluency variable increases (the higher fluency the higher the value), so the correlation should be positive.

a) Correlation in n-gram metrics In order to obtain the scores of the n-gram metrics (c.f 2.1.2.1.b) we prepared 4 translation references. One of the references was the published translation of a Wordnet gloss and the other three were translations performed by non-professional translators, with a good command of both the source and the target language. The reason why we chose non-professional translators was to prevent unnecessary deviations from the hypothesis because of the professional′s dislike towards literal translations (Cully and Riehemann, 2003).

Table 8.3 shows the scatterplot where, for each translation, an H score is paired with an MTS score. This scatterplot is followed by the scatterplots where the H scores are paired with the scores of n-gram metrics (BLEU, NIST, METEOR, GTM and ROUGE-L) with four references [7]. The values of the x axis correspond to the H scores of the non-expert evaluators. The values of the y axis correspond to the scores of the n-gram metrics .

Comparing the scatterplots, we can see that there are more dots in lexical metrics corresponding to false positives and negatives; that is, translations scored as very good that evaluators rated the lowest, and vice versa.

Table 8.4 shows the correlations of n-gram metrics with one reference and four references.

[7] The values were calculated by using the Asiya toolkit (http://asiya.lsi.upc.edu/)

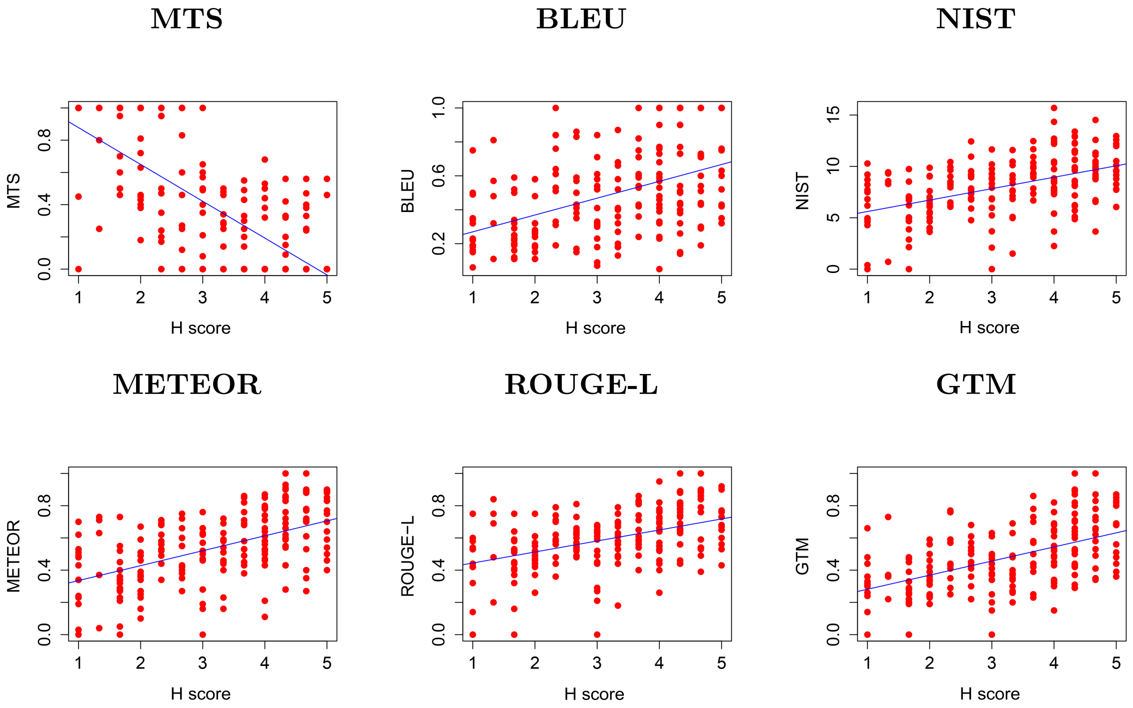


TABLE 8.3: Comparison of scatterplots for MTS and n-gram metrics

Metrics	Correlation index (1 reference)	Correlation index (4 references)
BLEU	0.31	0.45
NIST	0.34	0.46
METEOR	0.36	0.51
ROUGE-L	0.39	0.45
GTM	0.35	0.52

TABLE 8.4: Correlation of BLEU, METEOR, ROUGE-L and GTM with one and four references

The results with four references improved, some of them were slightly above 0.50. However, despite the effort of collecting references from four different translators, the results were significantly below MTS. In fact, the regression line of the MTS scatterplot shown in table 8.3 indicates that MTS correlates better. Therefore, the MTness approach proved to draw much better results than n-gram metrics, even when the expensive cost of obtaining more than one reference was paid.

b) Correlation in syntactic RPA metrics Since our method is based on a syntactic representation, specifically a dependency tree, we were also interested in comparing the correlation of MTS with the correlation of metrics drawn from a dependency tree. We took the metrics presented in 2.1.2.1b.5 and we calculated them[8].

Table 8.5 shows the syntactic metrics whose coefficients were over 0.30 with one reference. Notice that the highest correlation was 0.40, which was considerably below the MTS index.

Metrics	**Correlation index**
MTS	**-0.71**
Dpm-HWCM_w-2	0.40
Dpm-HWCM_i_w-2	0.39
Dpm-OL_*	0.38
Dpm-OL_1	0.38
Dpm-HWCM_w-1	0.37
Dpm-HWCM_w-3	0.37
Dpm-HWCM_w-4	0.35
Dpm-HWCM_i_w-2	0.39
Dpm-OL_2	0.35
Dpm-OL_3	0.32

Table 8.5: Syntactically-oriented results for one reference (correlation index over 0.30

The correlation increased with four references (Table 8.6) but the highest value (0.50) was also considerably below MTS. Metrics above 0.40, based on syntactic relations and syntactic categories, came up[9]. So the syntactic metrics needed the cost of obtaining four references to get significant results beyond the word level.

c) Correlation in classification-based metrics The correlation coefficients in (Gamon et al., 2005) and (Mutton et al., 2007) are considerable below the MTS index. This difference might be due in part to the different evaluation corpus sizes. It would have been interesting to compare the correlations with our evaluation corpus size. Actually, the 0.4 coefficient achieved by the classification approach in (Mutton et al., 2007) was not obtained with less than 300 sentences. Such a large test set makes human judgement a tiresome task. Besides, this task is costly in comparison to the correlation achieved. On the contrary, our correlation coefficient was significant with fewer translations (196).

[8] The values were also calculated by using the Asiya toolkit

[9] These metrics are distinguished with the symbols r and c respectively (c.f. 2.1.2.1.b.5)

Metrics	**Correlation index**
MTS	**-0.71**
Dpm-HWCM_w-2	0.50
Dpm-OL_*	0.50
Dpm-HWCM_i_w-2	0.49
Dpm-HWCM_w-3	0.48
Dpm-HWCM_w-4	0.46
Dpm-HWCM_i_r-2	0.45
Dpm-HWCM_r-2	0.45
Dpm-HWCM_w-1	0.45
Dpm-OL_1	0.45
Dpm-HWCM_r-3	0.44
Dpm-HWCM_c-3	0.43
Dpm-OL_2	0.43
Dpm-OL_3	0.43
Dpm-HWCMi_w-3	0.43
Dpm-HWCM_c-4	0.42
Dpm-HWCM_r-4	0.42
Dpm-HWCM_c-2	0.41

Table 8.6: Comparison between MTS and syntactically-oriented metrics for four references (correlation index over 0.40)

8.3.4 Differences in MTS and state-of-the art approaches

In this section we will compare our approach to fluency and accuracy with the approaches of RPA and classification-based metrics. We will see that MTS is more consistent to human judgements than state-of-the-art metrics.

8.3.4.1 Differences between MTness and n-gram approaches

The main difference between the MTS and n-gram metrics is the fact that MTS scoring is based on an experimental work on translation quality perception, whereas n-gram metrics evaluate a translation through a computational operation: string matching.

Apart from the shortcomings of BLEU explained in (Koehn, 2010), we note that n-gram metrics do not penalize very bad translations when they are quite close to the reference. On the contrary, our metric takes into account the real perception of machine translationness, no matter if the translation is near to a reference or not.

The different results due to the different methodologies are evident when rating a translation with a NO-L2 instance. NO-L2 instances affect fluency and accuracy. Table

8.7 shows the MTS value for *El centro alrededor del cual algo rotates* (the center around which something rotates) with the values of BLEU, GTM, NIST, METEOR and ROUGE-L. The values of these metrics were obtained with the following reference *Centro alrededor del cual algo rota* (center around which something rotates). Notice that the MTS value is 1 because, according to our experimental study, readers agree to consider a translation with a NO-L2 instance like *rotates* as very bad. In fact, all the evaluators scored the translation with the lowest value. As a contrast, notice that the BLEU and NIST values show a very moderate penalty, and the highest value is around 0.8 because there is only one mismatched element to the reference.

H SCORE	**MTS**	**BLEU**	**GTM**	**NIST**	**METEOR**	**ROUGE-L**
1	1	0.41	0.61	1.47	0.35	0.76

Table 8.7: H score, MTS and n-gram metrics for *Centro alrededor del cual algo rotates*

MTS is consistent to accuracy thanks to the fact that the values do not depend on string matching. Table 8.8 shows the H score, the MTS and the BLEU, METEOR, GTM and ROUGE-L for *Vuelo en un balón* (Flight on a ball). This time the values of the n-gram metrics were obtained with four references.

H SCORE	**MTS**	**BLEU**	**METEOR**	**GTM**	**ROUGE-L**
1.3	0.40	0.15	0.10	0.29	0.36

Table 8.8: H score, MTS and ngram metrics for *Vuelo en un balón* (flight on a ball)

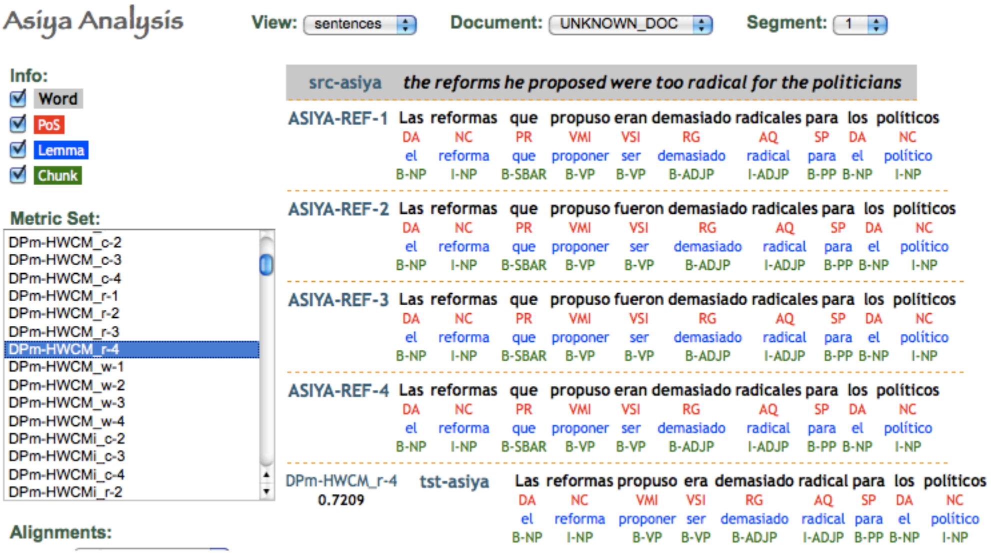
This translation has a considerable MTS score (0.40) because of the inaccurate meaning. The RPA scores were also low. Three of the reference translations were *volar en globo* (flying on a balloon), which is quite distant from the hypothesis. In the fourth reference *balloon* was translated as *globo aerostático.* However, if one of the references had been *vuelo en un globo* (flight on a balloon), which is acceptable, the score would have been very different (see table 8.9). With only one mismatched token (*balón*), the hypothesis would have been wrongly rated as good. As a contrast, the value of MTS is more consistent because of the detection of the semantic oddness when reading that a ball is a means of transport.

H SCORE	**MTS**	**BLEU**	**METEOR**	**GTM**	**ROUGE-L**
1.3	0.40	0.43	0.60	0.61	0.83

Table 8.9: H score, MTS and n-gram metrics of *Vuelo en un balón* with *Vuelo en un globo* as a reference

8.3.4.2 Comparison between MTS and syntactic RPA metrics

Syntactic RPA metrics indicate how close the syntactic representation of the hypothesis is to the representation of the references. So a translation that only differs from the reference in a missing element, such as a relative pronoun, is not as bad as in our method. As an example, figure 8.3 shows the Asiya HWCMr4 of *Las reformas propuso era demasiado radical para los políticos* (The reforms proposed was too radical for the politicians), with four references. Notice that the HWCMr4 score is quite high (0.7209) because the mismatch just lies in the missing relative pronoun. However, the MTS score (0.33) is more consistent with the grammatical ill formedness of the translation.



View: sentences

Document: UNKNOWN_DOC

Segment: 1

☑ Word
☑ PoS
☑ Lemma
☑ Chunk

DPm-HWCM_c-2
DPm-HWCM_c-3
DPm-HWCM_c-4
DPm-HWCM_r-1
DPm-HWCM_r-2
DPm-HWCM_r-3
DPm-HWCM_r-4
DPm-HWCM_w-1
DPm-HWCM_w-2
DPm-HWCM_w-3
DPm-HWCM_w-4
DPm-HWCMi_c-2
DPm-HWCMi_c-3
DPm-HWCMi_c-4
DPm-HWCMi_r-2

src-asiya *the reforms he proposed were too radical for the politicians*

ASIYA-REF-1	Las	reformas	que	propuso	eran	demasiado	radicales	para	los	políticos
	DA	NC	PR	VMI	VSI	RG	AQ	SP	DA	NC
	el	reforma	que	proponer	ser	demasiado	radical	para	el	político
	B-NP	I-NP	B-SBAR	B-VP	B-VP	B-ADJP	I-ADJP	B-PP	B-NP	I-NP

ASIYA-REF-2	Las	reformas	que	propuso	fueron	demasiado	radicales	para	los	políticos
	DA	NC	PR	VMI	VSI	RG	AQ	SP	DA	NC
	el	reforma	que	proponer	ser	demasiado	radical	para	el	político
	B-NP	I-NP	B-SBAR	B-VP	B-VP	B-ADJP	I-ADJP	B-PP	B-NP	I-NP

ASIYA-REF-3	Las	reformas	que	propuso	fueron	demasiado	radicales	para	los	políticos
	DA	NC	PR	VMI	VSI	RG	AQ	SP	DA	NC
	el	reforma	que	proponer	ser	demasiado	radical	para	el	político
	B-NP	I-NP	B-SBAR	B-VP	B-VP	B-ADJP	I-ADJP	B-PP	B-NP	I-NP

ASIYA-REF-4	Las	reformas	que	propuso	eran	demasiado	radicales	para	los	políticos
	DA	NC	PR	VMI	VSI	RG	AQ	SP	DA	NC
	el	reforma	que	proponer	ser	demasiado	radical	para	el	político
	B-NP	I-NP	B-SBAR	B-VP	B-VP	B-ADJP	I-ADJP	B-PP	B-NP	I-NP

DPm-HWCM_r-4 0.7209	tst-asiya	Las	reformas	propuso	era	demasiado	radical	para	los	políticos
		DA	NC	VMI	VSI	RG	AQ	SP	DA	NC
		el	reforma	proponer	ser	demasiado	radical	para	el	político
		B-NP	I-NP	B-VP	B-VP	B-ADJP	I-ADJP	B-PP	B-NP	I-NP

FIGURE 8.3: Asiya Toolkit screen showing the references and the HWCMr4 score for the translation *Las reformas propuso era demasiado para los políticos (The reforms proposed was too radical for the politicians)*

Besides, we noticed that subcategory codification was not as complete as in our method. Morphological labels describe just the category and subcategory, but they do not describe morphosyntactic features, which produce wrong verbal forms (e.g. mood and person). Therefore, bad translations with considerable high MTS scores are, on the contrary, evaluated as good in syntactic RPA metrics. The translation in figure 8.3 is also an example, since the wrong agreement in person and number between the verb *propuso* and the subject does not affect the syntactic RPA score.

8.3.4.3 MTS and the Legitimate Translation Variation

One might wonder whether we can guarantee that the reference corpus, from which reference DTRs are drawn, covers all the hypothesis space of good DTRs. Actually, the same question is relevant for RPA metrics, because we may also wonder if translation references cover all the legitimate translation variations.

On the other hand, how can we guarantee that something that is not in the reference corpus, but is correct in the target language, will not be considered as an error? The same prevention must be applied as in translation references. If the references do not cover all the hypothesis space, a good translation can be scored as bad.

8.3.4.4 Comparison between MTS and classification metrics

We wanted to compare our results with the works of (Gamon et al., 2005) and (Mutton et al., 2007) as representative of two classification methods. Unfortunately, we could not carry out this comparison. Their evaluation engines were not available and were trained for other languages than Spanish (French in (Gamon et al., 2005) and English in (Mutton et al., 2007)). So the training data for our comparison was to be created, which is time consuming and expensive. On the other hand, according to (Gamon et al., 2005), the training corpus′ domain was the same as the evaluation corpus. So in case we wanted to apply the method for other domains, we had to repeat the process. Therefore we verified that our method is much easier to implement.

The only reference of comparison was the correlation with human judgements presented by the authors. As we said in 1.3.3.2c the correlation values were lower.

8.4 MTS and machine-like distinction

Figure 8.4 shows the scatterplot of MTS and H scores of non-experts with continuous dots jittered and their sizes proportional to MTS values. The H scores are from the rougher scale, where the differences between good and bad translations were more clearly cut (c.f. table 8.2), and the evaluators agreed the most.

As can been seen, the densest region is located in the intersection of the lowest H score range (from 1 up to 1.5) and the highest MTS range (from 0.8 up to 1). This is consistent, since the worse the translation the higher the MTS value. A less dense region, but still crowded, corresponds to H score values of 1.8 approximately, whereas MTS values also range from 0.8 to 1.

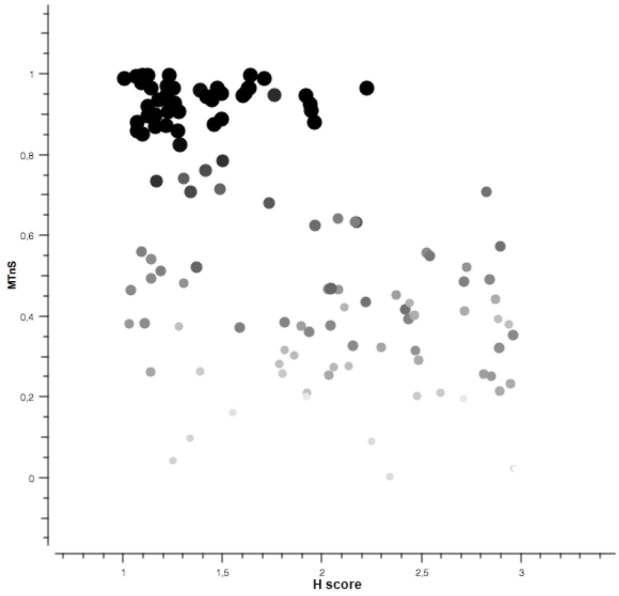


Figure 8.4: H score-MTS (MTS perspective)

Figure 8.5 shows the same relationship but the dots are jittered and sized from the H score perspective. In this case, the densest region is located in the intersection of the highest H score range (from 2.5 up to 3) and the lowest MTS range (from 0 to 0.175). This is consistent as well, since the better the translation the less the MTS value. Besides, notice that MTS values are not paired with H scores above 2.

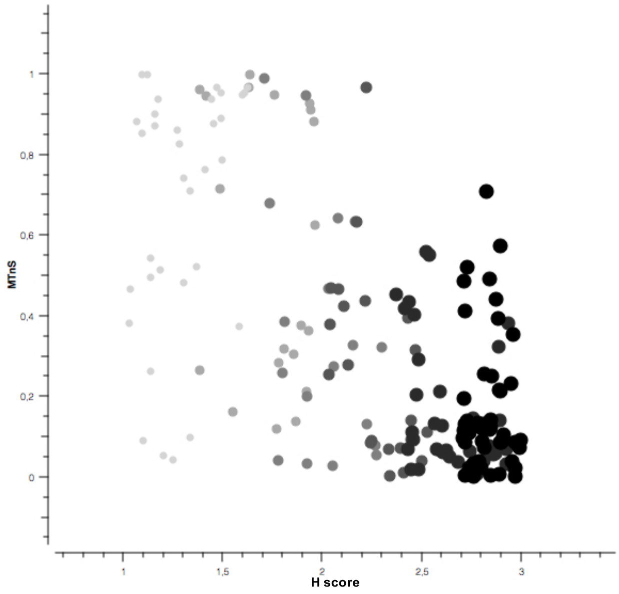
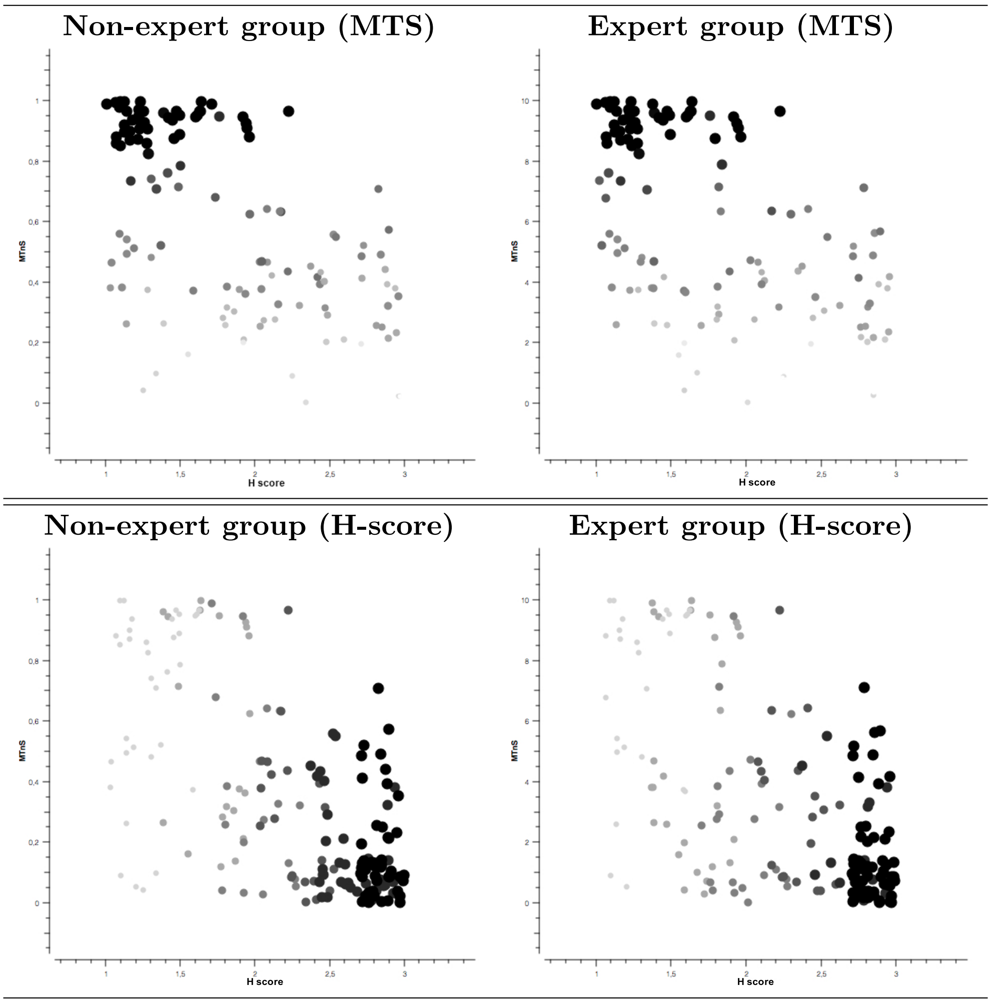


Figure 8.5: H score-MTS (H score perspective)

The differences between the judgements of the expert and the non-expert groups did not vary the results significantly, as is shown in table 8.10.


Non-expert group (MTS) **Expert group (MTS)**

Non-expert group (H-score) **Expert group (H-score)**

Table 8.10: Comparison of scatterplots between the non-expert and expert group appreciations and MTS

In sum, it is possible to establish a relationship between machine-like translations and low quality translations as perceived by humans.

8.5 Consistency of our method

In order to assess that our method is consistent and valid, we used a *normal Q-Q plot*. A normal Q-Q plot allows the comparison of two distributions. The assumption is that the distribution of the data obtained with our method is similar to the distribution in a linear model (lm) where the values are related in a linear trend. Since the distributions are assumed to be normal distributions, they are divided by means of *quantiles*. A *quantile* is the point below which a percentage of points in the distribution lies. This percentage is the same in the rest of the quantiles. As an example, the median (or 2-quantile) is the numerical value that allows the partition of the linearly distributed values in two, that is, 50% of the values below the median and 50% above the median. Another quantile is the *quartile*, which is any of the three points that divide the ordered distribution into four parts, each containing a quarter of the population[10].

The quantile values of the standard normal distribution (which is considered 'theoretical') are plotted on the x-axis. The quantile values of the distribution of the actual sample data are plotted on the y-axis. The points in the plot show the quantile values of the actual data set against the quantile values of the theoretical normal distribution. If the points are reasonably well approximated by a straight line, then both distributions are very similar and the data obtained are consistent. Q-Q plots also allowed us to locate deviations from linearity (s-shaped curves over the straight line).

The Q-Q plot in figure 8.6 shows that the points are certainly approximated by a straight line, so our method proves to be consistent.

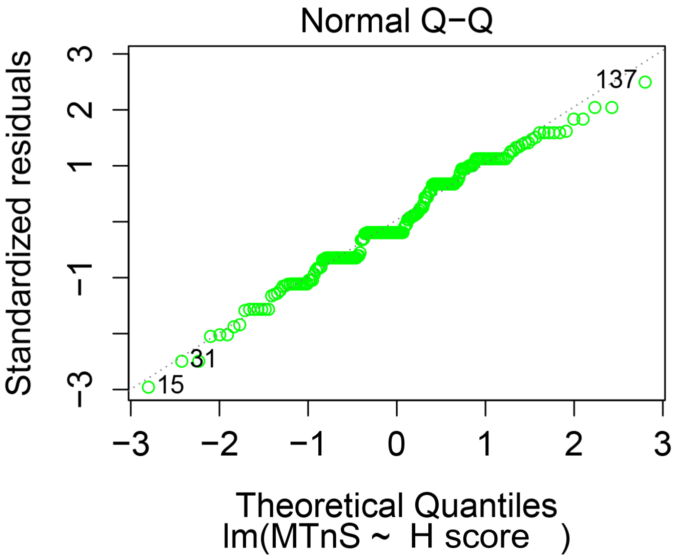


Figure 8.6: Q-Q plot for the MTness method

8.6 Conclusion

In this chapter we have explained the assessment of MT evaluation based on the notion of machine translationness. In order to produce the assessment we developed an application that performs the detection and scoring algorithms explained in chapter 6. Besides, we also had to prepare the necessary resources to perform MTness detection. Having the

[10] Definition of *quartile* from Wiktionary http://en.wiktionary.org/wiki/quartile. For information about *normal distribution* and *quantile* see http://en.wikipedia.org/wiki/Normal_distribution and http://en.wikipedia.org/wiki/Quantile

goal of performing cheap evaluations in mind, we collected resources that were freely available.

The results of the assessment indicated that an automatic evaluation based on the analysis of MTness perception was closer to human judgements than state-of-the art evaluation methods. Our method is a cheaper alternative because the experiment proved a stronger correlation between MTS values and human ratings with freely available resources. Apart from that, the method is not trained for specific genres nor domains, and can be adapted to any language, provided there are free linguistic resources available.

The method integrates linguistic analysis at different levels (lexical, syntactic and semantic). This contrasts with state-of-the-art metrics, which are generally specialized in one of these levels. Another distinctive feature is the fact that the MTness instances are also registered and explained in a log file. This file can be useful for developers and provides information to create automatic postedition modules. Besides, the user can check false positives and other failures, and tune the detectors by adding the information in the detector′s databases necessary to perform better. Our method does not rate translations according to what they look like (as state-of-the-art metrics do), but according to the evidences found of MTness instances. The participation of users in declaring non-misleading information in the databases will improve the MTness detection and the already good correlation with human ratings.

The method proved to be consistent and, despite some discrepancies between the automatic ratings and the human ratings, the correlation and consistency of the method has not been affected by the expertise of one of the evaluators. Some reasons for the mismatches lie on the expected limitations of any automatic task, for instance, the performance of the dependency parser.

Chapter 9

Conclusions and Future Work

9.1 Introduction

This chapter presents the main contributions of this thesis and outlines future research on the notion of *machine translationness.* Section 9.1 is a summary of the contributions and section 9.2 presents future work and research directions.

9.2 Contributions

The contributions of this thesis impact on the following topics

- MT evaluation
- Machine translation detection
- MT post edition
- Use and abuse of MT technologies

9.2.1 MT Evaluation

Our main contribution is the proposal of a novel direction, which is just the opposite direction of traditional methods. Instead of measuring how good a translation is, in terms of human likeness, our approach measures how bad a translation is, in terms of machine-likeness. This thesis contributes in introducing:

- A new notion, *machine translationness*

- A typology of machine translationness phenomena
- A new metric (*MTS*) for measuring how much a translation is machine-like.

We have shown that our evaluation method rates translation quality with stronger correlation to human judgements than state-of-the art metrics. Apart from this, our method has the following contributions:

A cheap method : the method does not demand the expensive procedures and resources of state-of-the art methods. Neither translation references nor large training corpora are necessary, and the evaluation can be performed with resources that are freely available.

An explicative method : State-of-the art metrics say little about translation quality perception. Besides, they are obtained by computational procedures (ngram matching, machine learning) that are applied for heterogeneous tasks. On the contrary, our method is based on an experimental study on machine translation perception and provides the reasoning behind the detection of MTness instances. Moreover, our proposal puts machine translation evaluation in a more consistent perspective, since machine translated texts are evaluated according to what they really are (texts generated by a machine) not according to what they are not and might resemble (texts generated by a human being).

Inexpensive multi-layered approach : Our score reflects the effect of MTness at different linguistic levels (lexical, syntactic, semantic), taking into account MTness salience. The score calculation is not expensive because it can be obtained by processing the output of a dependency parser, where information concerning several linguistic levels is declared.

Cross-linguistic method : The method can be adapted to any language, provided there are resources available for that language (dependency parser, multilingual dictionaries, large number of texts published on the Web).

Non-domain centered method : The results are domain independent. We proved this by creating reference DTRs from texts whose domain was not the domain of the test set, which was rather neutral in itself (Wordnet glosses).

Cross NLP evaluations : Non-detected MTness instances, because of wrong parse trees, provide data for developers to evaluate the drawbacks of their parsers. On the other hand, MTness detection provides dependency structures that parsers should not attempt to recognize. Contrary to current parsers, which do not evaluate their input and always provide a solution, new parsers would be capable of identifying typical MTness phenomena. Then they would judge the input and even warn the user about its quality. This would impact on grammar checkers, and even detectors of MT use for cheating purposes.

On the fly evaluations : Future MT evaluation campaigns would benefit from the results presented in this thesis. They would be cheaper, faster, and with fewer people involved. On the other hand, the automatic detection of MTness favours fast microevaluations that assess the MT system's capabilities and shortcomings. Our method is intended to be applied for on-the-fly evaluations, in situations where a rapid glance at the output is enough to judge its quality, so time and money invested in in-depth evaluations that lead to similar conclusions can be saved. The interest in on-the-fly evaluations is justified for language pairs whose poor quality is taken for granted and the goal is to know which system is more helpful for users.

9.2.2 Machine translation detection

The objective of our method goes beyond creating good parallel corpora from the Web (c.f. 1.1.3), which is the current application of machine translation detection(Kotani et al., 2008). Detection methods for parallelising source and target texts deal with information that can be both superficial (ngram combinations(Antonova and Misyurev, 2011)) and extralinguistic (URL and full HTML of the target pages(Rarrick et al., 2011)). The idea of applying MTness detection to assess the quality of Web pages was already suggested in (Moré and Climent, 2006) but our method is centered in analysing the fluency and semantic consistency of texts. This is the key feature that makes our method general-purpose.

On the other hand, the log file of our evaluation tool leaves open the possibility for the user to check the detection performance. So in case the tool detected a non real MTness instance, the evaluator can add its DTR in the reference DTR database of the system, and reevaluate the evaluation corpus. Another possibility could be to build a database of MTness DTRs, so when the detector failed to recognise an MTness instance, the user could add the DTR in this database. So the user takes an active role in improving the system and provides important information to be processed for further studies on MTness and its applications. This is in line with other proposals where the user participates in enhancing MT proposals (e.g. *LetśMT*(Vasiļjevs et al., 2012)).

9.2.3 MT Postedition

In (Moré and Climent, 2006) we already pointed out the link between machine translation detection and automatic postedition. Recently, (Aikawa and Rarrick, 2011) also suggest postedition as a potential application of machine translation detection.

Throughout this thesis we have warned about non-salient MTness errors, especially when published texts have not been revised by trained posteditors. Our method detects non salient MTness instances, so it can help specialists and non-specialists not to overlook them, especially when they are time pressed.

The log file of our evaluation tool can be a source of information for posteditors to tailor their own correction memories. By correction memories we mean collections of recurrent errors paired with the correct version. Posteditors- even the translators who postedit their machine translated drafts- could use the correction memories in translation-assistance environments, as if they were translation memory databases.

9.2.4 Use and abuse of MT technologies

Our method can be useful for detecting abusive use of MT engines. For instance, to detect cheating practices by foreign language learners. MTness has been uncontrollably spread on the Web so MTness detectors, like plagiarism detectors, can also help authors, publishers and academia to keep high quality standards. This would contribute in restoring the trust in the Web-as-a-corpus approach for developing linguistic applications (grammar checkers, as in (Moré and Climent, 2004)).

We are aware that the precision of our method, which triggers Web-queries, might be affected by the very phenomena we intend to avoid. It is clear that the more publications with MTness on the Web the more likely these publications will affect our query results. All in all, at least we provide methods that detect linguistically deficient publications on the Web and also prevent misunderstandings from Web pages that are not warned about their unreliability. So the utopic objective of our work is to wipe out MTness until our tool becomes meaningless.

9.3 Future Work

The future work can be developed in three directions. Firstly, the improvement of the resources of the detection method. Secondly, the application of our method in research beyond MT evaluation. Finally, the study of *machine translationness* as a linguistic phenomenon per se.

9.3.1 Improving resources

The future work can be oriented to improve the dependency parser and consequently improve the MTS results. We think that research that enhances a bidirectional relationship between MTness detection and dependency parsing should be carry out. MTness detection analysis may provide information for improving dependency parsers and better dependency parsers will improve MTness detection. This is an interesting path to pursue. However, the research should not be constrained to fix drawbacks of particular MTness detection tools and parsers. The research should have a wider scope, such as the possibility for these NLP tools to evaluate and warn users about the reliability of their own output.

Another interesting path to pursue is to draw semantic patterns from the dependency tree representations by combining the Wordnet semantic labels of the terminal nodes. Semantic inconsistencies could be detected by the same procedures applied to detect inconsistencies in syntactic and morphological DTRs. We have not tackled this approach because we thought that the detection of semantic inconsistencies by processing Wordnet labels was a field of research of its own. We hope that this research may succeed and be carried out in the detection of semantic MTness instances as a domain of application.

At the lexical level, we expect that further research in MTness triggers the elaboration of more complete bilingual dictionaries, freely available, that are suitable to cope with MTness instances that evoke expected translations at the lexical level.

It is true that the costs of our method are reduced if reliable parsers and huge corpora are available for the target language, Not all the languages have this kind of resources yet. So the large-scale application of the method for as many languages as possible is a challenge for MT and by extension for NLP.

9.3.2 Other applications of MTness detection

The main challenge is to put the potential applications of MTness into action. First of all, a promising path is to apply MTness detection in reranking raw output produced by other applications. For instance, our method could be optimized and integrated in an Internet search engine that would rank results according to the degree of MTness of the documents retrived. The higher the value, the lower the rank position. The reranker idea can be also applied for a machine translation system that evaluates its own output and dismisses raw translations with MTness patterns to favour other translation solutions. This is in line with (Aikawa and Rarrick, 2011)′ suggestions, but instead of working with lists of word ngrams, our method would manage combinations of abstract syntactic constituents. Finally, a promising path would be to create a sort of MTness community where crowd-sourcing resources could be shared by developers, translators and posteditors, in order to lessen the impact of MTness in everyday communication and optimize their translating and postediting work[2].

9.3.3 The study of MTness

We regard this thesis as a first approach to the notion of MTness and we are convinced that the study of machine translationness is a promising field of research that can be approached from different disciplines. Some aspects we believe are worth getting more insight are:

i) MTness salience

ii) A more clear-cut theoretical distinction between MTness instances and non-fluent use of the language (e.g by a foreign person)

iii) Exploration of the possibilities of MTness detection in learning foreign languages

iv) MTness typology across languages

From the perspective of cognitive linguistics, studies about MTness salience, why some linguistic features are more salient than others and also the subjectivity in MTness perception are promising. These studies can be complemented with experimental research such as relating MTness perception and eye-tracking, for instance, in line with studies on human evaluators (Bremin et al., 2010). From the perspective of applied linguistics, it would interesting to work on distinguishing more clearly MTness phenomena from disfluent performance of non-native speakers, or even bad performance by native speakers in a stressing and hasty situation, as interpreters do. This distinction would make sense if MT evaluations were performed according to (Way, 2012) and (Loehr, 1998)′ suggestion. For these authors, evaluations based on comparison would be fairer if the comparison was between MT and interpreters, who make mistakes, just like MT engines do, and not between the target language and the language of the MT systems, which is not representative of the language spoken or written anywhere (Bellos, 2011). So the distinction between human/intepreter translation, regarded as ’tolerable’, and machine translationness, regarded as ’not tolerable’ should be more clear-cut

[2] See in (Tatsumi et al., 2012) a crowd-sourcing initiative for optimizing postedition

In the context of language learning, methodological research might explore the possibilities of machine translationness as data for foreign language awareness and comparative analysis of the source and the target languages.

Finally, it would be interesting to provide a more complete MTness typology from the study of machine translationness phenomena across other languages than the one studied in this thesis. We hope that this thesis will contribute in starting studies on MTness perception in different languages. So the goal of presenting an MTness typology across languages can be fully fulfilled. Our typology, based on MTness perception for Spanish, is a first step.

Glossary

accuracy assumption (ACA) good translations are those that are accurate with respect to the source or the translation reference in the sentence level. The accuracy rating is based on an automatic calculation of semantic similarities between machine translations and references.

adequacy The degree in which a segment of the machine translation keeps the meaning of the source segment.

automatic MT evaluation (AMTE) Evaluation of machine translations by automatic means.

co-occurrent mts Metric that rates MTness when only inconsistent co-occurrent words are found.

CON-INCOH MTness type: CON-INCOH stands for **CON**textual **INCOH**erence.

contrast units tuple that registers the perception of MTness in one translation by three informants. These units served to organize the data of the experimental study on MTness perception.

dependency index tuple (DiT) A tuple representing the indexes of the root tree or a dependent subtree.

dependency tree representation (DTR) Tree representation of the dependencies between words drawn by a parser.

dependent subtrees Typed dependency trees whose root node is a dependent node in a larger typed dependency tree.

expected morphological DTR Reference morphological DTR whose edit distance to a hypothesis DTR is the shortest.

expected syntactic DTR Reference syntactic DTR whose edit distance to a hypothesis DTR is the shortest.

expected translation contrast Cognitive process triggered by a word or phrase in a translation which consists in contrasting the actual word or phrase with a more expected translation..

fluency The degree in which the translation is well formed according to the grammar rules and conventions of use in the target language.

H score Mean of the scores provided by human evaluators. The scores belong to a DARPA scale used for the human evaluation of machine translations.

human likeness assumption (HLA) Assumption in automatic evaluations according to which a machine translation that resembles a human translation is good.

human MT evaluation (HMTE) Evaluation of machine translations by humans.

human translationness criterion (HTC) The machine translation evaluation criterion based on determining whether machine translations are indistinguishable from human translations..

hypothesis DTR A tuple that represents a structure where words are related according to the dependency tree of a machine translation.

I-AGR MTness type: I-AGR stands for **I**nadequate **AGR**eement.

I-ORD MTness type: I-ORD stands for **I**nadequate **ORD**er of syntactic constituents.

I-POS MTness type: I-POS stands for **I**nadequate **P**art of Speech.

I-VERBF MTness type: I-VERBF stands for **I**nadequate **VERB**al Form.

machine translationness (MTness) The quality of machine translations that makes them distinguishable from human translations.

machine translationness badness assumption (MTBA) Assumption according to which translations with machine translationness are bad.

machine translationness criterion (MTC) The machine translation evaluation criterion by which a machine translation is evaluated according to qualities of machines..

machine translationness score (MTS) Score that rates the machine translationness of a translation.

morphological mts Metric that rates MTness when only morphologically inconsistent dependencies are found.

MTness instance overlapping More than one MTness instance confluate in a translation segment. One instance type may be the consequence of the presence of another type. The effect in quality perception is MTness saturation.

MTness salience The condition by which linguistic phenomena are more likely regarded as machine-like.

MTness saturation The effect produced by MTness overlapping. The perception of machine translationness increases although MTness instances are difficult to be distinctively discerned.

mts hypothesis string A string where the indexes of a node string affected by MTness are replaced by the symbol 'B' ('B' stands for 'bad'). It models the actual MTness status of a translation..

mts reference string A string where all the indexes of a node string are replaced by the symbol 'B' ('B' stands for 'bad'). It models a translation with the highest MTness score.

n-gram metrics metrics that score the similarity of a machine translation (hypothesis) to a set of references. The unit of comparison is the n-gram. Some of these metrics are BLEU, NIST and ROUGE.

NO-L2 MTness type: words which are not recognised as pertaining to the target language and are not loan words.

node string (NS) String where all the indexes in a typed dependency tree are concatenated.

noisy segments Translation segments that cause MTness saturation.

OVER-WRD MTness type: OVER-WRD stands for *word overgeneration*.

phrase DTR (DTR$_p$) Syntactic DTR that describes the dependency relations in terms of syntactic phrases and syntactic functions.

quality estimation (QE) A quality indicator addressed, by using machine learning techniques, to predict quality scores of translations without using translation references.

reference DTR A tuple that represents a structure where words are related according to the linguistic intuition of native speakers..

reference proximity assumption (RPA) Assumption in automatic evaluations according to which the closer a machine translation is to a professional human translation (reference), the better it is.

SEM-GAP MTness type: SEM-GAP stands for **SEM**antic **GAP**.

SEM-INCOH MTness type: SEM-INCOH stands for **SEM**antic **INCOH**erence.

source evokers Mistranslated words or phrases through which a source-language word or phrase is guessed and the cause of the mistranslation is understood.

STR-CHAR MTness type: STR-CHAR stands for **STR**ange **CHAR**aracter.

subcategorization DTR (DTR$_s$) Syntactic DTR that describes the subcategorization restrictions when the governor has a specific lemma.

SYNT-GAP MTness type: SYNT-GAP stands for **SYNT**actic **GAP**.

syntactic mts Metric that rates MTness when only ill syntactic dependencies are detected.

typed dependency tree Dependency parse tree labelled with information about syntactic, morphological and semantic dependencies between words.

TYPO-E MTness type: TYPO-E stands for **TYPO E**rrors.

WRD-REP MTness type: WRD-REP stands for **W**o**RD REP**etition.

Bibliography

Aikawa, T., and Rarrick, S. 2011. Are numbers good enough for you? A linguistically meaningful MT evaluation method. Pages 332–337 of: *MT Summit XIII: the Thirteenth Machine Translation Summit.*

Akiba, Y., Imamura, K., and Sumita, E. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. Pages 15–20 of: *Proceedings of Machine Translation Summit VIII.*

Amigó, E., Giménez, J., Gonzalo, J., and Márquez, L. 2006. MT Evaluation: Human-Like vs. Human Acceptable. Pages 17–24 of: *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL).*

Antonova, A., and Misyurev, A. 2011. Building a Web-based parallel corpus and filtering out machine-translated text. Pages 136–144 of: *Proceedings of the 4th Workshop on Building and Using Comparable Corpora.* 49th Annual Meeting of the Association for Computational Linguistics.

Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., and Padró, M. 2006. FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006).*

B, Crysmann. 1997. *Fehlerannotation.* Tech. rept. DFKI GmbH.

Babych, B., and Hartley, T. 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL).*

Banerjee, S., and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. Pages 65–72 of: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization.*

Bellos, D. 2011. *Is That a Fish in Your Ear: Translation and the Meaning of Everything.* Particular Books. Penguin Group.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. 2003. *Confidence estimation for machine translation. Final Report of John Hopkins 2003 Summer Workshop on Speech and Language Engineering.* Tech. rept. John Hopkins University.

Bosma, W.E., Vossen, P., Soroa, A., Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., and Aliprandi, C. 2009. Kaf: a generic semantic annotation format. In: *Proceedings of the GL2009 Workshop on Semantic Annotation.*

Bremin, S., Hu, H., Karlsson, J., Lillkull, A. Prytz, Wester, M., Danielsson, H., and Stymne, S. 2010. Methods for human evaluation of machine translation. Pages 47–48 of: *SLTC 2010. The Third Swedish Language Technology Conference (SLTC 2010).*

Bruckner, C., and Plitt, M. 2001. Evaluating the Operational Benefit of Using Machine Translation Output as Translation Memory Input in the Translation Process of Software Documentation. In: *MTEval Workshop ISSCO, Geneva.*

Budiansky, S. 1998. Lost in Translation. *Atlantic Monthly*, **282**, 80–84.

Callison-Burch, C. 2007. *Paraphrasing and Translation.* Ph.D. thesis, University of Edinburgh.

Callison-Burch, C., Osborne, M., and Koehn, P. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In: *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL).*

Callison-Burch, C., Fordyce, C. Shaw, Koehn, P., Monz, C., and Schroeder, J. 2007. (Meta-)evaluation of machine translation. Pages 136–158 of: *Proceedings of the Second Workshop on Statistical Machine Translation.*

Callison-Burch, C., Koehn, P., Fordyce, C., Monz, C., and Schroeder, J. 2008. Further Meta-Evaluation of Machine Translation. Pages 70–106 of: *Proceedings of the Third Workshop on Statistical Machine Translation, ACL.*

Carroll, J.B. 1966. *An experiment in evaluating the quality of translations.* Tech. rept. Washington DC National Academy of Sciences.

Chomsky, N. 1970. Reading in English Transformational Grammar. Waltham: Ginn.

Church, K. W., and Hovy, E. H. 1993. Good Applications for Crummy Machine Translation. Pages 239–258 of: *Machine Translation*, vol. 8.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates.

Correa, N. 2003. A Fine-grained Evaluation Framework for Machine Translation System Development. In: *MT Summit IX.*

Corston-Oliver, S., Gamon, M., and Brockett, C. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. Pages 140–147 of: *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL).*

Cully, C., and Riehemann, S. 2003. The limits of ngram translation evaluation metrics. In: *Machine Translation Summit IX.*

Doddington, G. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Pages 138–145 of: *Proceedings of the 2nd International Conference on Human Language Technology.*

Elliot, D., Hartley, A., and Altwell, E. 2004. *A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. AMTA: Machine Translation: From Real Users to Research.* Berlin/Heidelberg: Springer. Chap. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation, pages 64–73.

Fellbaum, C. 1998. *WordNet. An Electronic Lexical Database.* The MIT Press.

Fité, R. 2006. El Periódico, una experiència en traducció automàtica. *Revista Tradumàtica. Traducció i Tecnologies de la Informació i la Comunicació.*

Flanagan, M. 1994. Error Classification for MT Evaluation. Pages 65–72 of: *Proceedings of the Association of Machine Translation of the Americas (AMTA-94).*

Fung, P., Wu, Z., Yang, Y., and Wu, D. 2006. Automatic Learning of Chinese English Semantic Structure Mapping. Pages 230–233 of: *IEEE Spoken Language Technology Workshop.*

Gamon, M., Aue, A., and Smets, M. 2005. Sentence-Level MT evaluation without reference translations: beyond language modeling. Pages 103–111 of: *Proceedings of EAMT.*

Giménez, J. 2008. *Empirical Machine Translation and its Evaluation.* Ph.D. thesis, Universitat Oberta de Catalunya.

Giménez, J., and Márquez, L. 2008. A smorgasbord of features for automatic MT evaluation. Pages 195–198 of: *Proceedings of the 3rd Workshop on Statistical Machine Translation.* Association for Computational Linguistics.

Grefenstette, G. 1999. The WWW as a Resource for Example-Based MT Task. In: *Proc. Of Aslib Conference on Translating and the Computer.*

Guest, G., Bunce, A., and Johnson, L. 2006. How many interviews are enough? An experiment with data saturation and variability. *Field Methods*, **18(1)**, 59–82.

Hillas, B. 2009. *Why Human Translators Still Have a Job.* http://translation-blog.trustedtranslations.com/why-human-translators-still-have-a-job-2009-10-07.html [last checked: 17-05-2012]. Trusted Translations.

Hovy, E., King, M., and Popescu-belis, A. 2002a. Computer-Aided Specification of Quality Models for Machine Translation Evaluation. In: *in 'Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)', Las Palmas, Canary Islands.*

Hovy, E., King, M., and Popescu-Belis, A. 2002b. Principles of context-based machine translation evaluation. *Machine Translation*, **16**, 1–33.

Hutchins, J. 1996. ALPAC: the (in)famous report. In: *MT News International.*

ISO. 1991. *ISO/IEC IS 9126: Software Product Evaluation- Quality Characteristics and Guidelines for their Use.* Tech. rept. International Organization for Standardization.

Jehl, L. 2010. *Machine Translation for Twitter.* Tech. rept. University of Edinburgh.

JEIDA. 1992. *JEIDA Methodology and Criteria on Machine Translation Evaluation.* Tech. rept. Japan Electronic Industry Development Association.

Jelinek, R. 2004. Modern MT Systems and the Myth of Human Translation: Real World Status Quo. In: *Proceedings of the International Conference Translating and the Computer.*

Jones, D. A., and Rusk, G. M. 2000. Toward a Scoring Function for Quality-Driven Machine Translation. In: *Proceedings at the Conference for Computational Linguistics.*

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

Koehn, P. 2012. Simulating Human Judgement in Machine Translation Evaluation Campaigns. In: *International Workshop on Spoken Language Translation (IWSLT).*

Koehn, P., and Schroeder, J. 2007. Experiments in domain adaptation for statistical machine translation. In: *Proceedings of the ACL-2007 Workshop on Statistcal Machine Translation.*

Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I., and Isahara, H. 2008. A Method of Automatically Evaluating Machine Translations Using a Word-alignment-based Classifier. Pages 11–18 of: *Proceedings of the Workshop "Mixing Approaches to Machine Translation"(MATMT).*

Kulesza, A., and Shieber, S. M. 2004. A learning approach to improving sentence-level MT evaluation. Pages 75–84 of: *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI).*

Landis, J. R., and Koch, G. G. 1977. "The measurement of observer agreement for categorical data". *Biometrics*, **33**, 159–174.

Lavie, A., Sagae, K., and Jayaraman, S. 2004. The Significance of Recall in Automatic Metrics for MT Evaluation. Pages 134–143 of: *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004).*

Levenshtein, V. I. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady.*

Lin, C. Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In: *Proceedings of the Workshop on Text Summarization Branches Out.*

Lin, C. Y., and Och, F. J. 2004a. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: *ACL '04 Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics.*

Lin, C. Y., and Och, F. J. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In: *Proceedings of the 20th International Conference on Computational Linguistics.*

Lingtech. 1996. *Post-editor Survey 1: Analyse og sammenfatning af sprogrevisorsvar.* Tech. rept. Lingtechs spørgeskemaundersøgelse.

Lingtech. 1997. *Post-editor Survey 2: Evaluering af oversættelser af tekstkorpusser K; edb og kontormaskiner. Baseret på input fra Lingtechs sprogrevisorer og datalingvister.* Tech. rept. Lingtech.

Liu, D., and Gildea, D. 2005a. Semantic Features for Evaluation of Machine Translation. Pages 75–84 of: *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.*

Liu, D., and Gildea, D. 2005b. Syntactic Features for Evaluation of Machine Translation. Pages 25–32 of: *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.*

Liu, D., and Gildea, D. 2006. Stochastic Iterative Alignment for Machine Translation Evaluation. Pages 539–546 of: *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL).*

Llitjós, A. Font. 2007. *Automatic Improvement of Machine Translation Systems.* Ph.D. thesis, Carnegie Mellon University.

Lloberes, M., Castellón, I., and Padró, Ll. 2010. Spanish FreeLing dependency grammar. Pages 693–699 of: *LREC-2010.*

Lo, C., and Wu, D. 2010a. Evaluating maching translation utility via semantic role labels. In: *Proceedings of LREC 2010.*

Lo, C., and Wu, D. 2010b. Semantic vs. Syntactic vs. N-gram Structure for Machine Translation Evaluation. Pages 52–60 of: *Proceedings of SSST-4, Fourth Workshop on Syntax and Structure in Statistical Translation. COLING 2010.*

Loehr, D. 1998. Can simultaneous interpretation help machine translation? In: *Machine Translation and the information soup: third conference of the association for machine translation in the Americas AMTA′98.* Springer.

Loffler-Laurian, A.M. 1996. *La traduccion automatique.* Paris: Presses Universitaries du Septentrion.

Marrafa, P., and Ribeiro, A. 2001. Quantitative Evaluation of Machine Translation Systems: Sentence Level. In: *Proceedings of the MT Evaluation Workshop.*

Melamed, I. Dan, Green, R., and Turian, J. P. 2003. Precision and recall of machine translation. In: *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology companion volume of the Proceedings of HLTNAACL 2003–short papers Volume 2.*

Melby, A. K. 1995. *Why Can't a Computer Translate more Like a Person?* http://www.ttt.org/theory/barker.html [Last checked: 13-05-2012].

Melĉuk, I. 1988. *Dependency Syntax: Theory and Practice.* Albany, N.Y.: The SUNY Press.

Moré, J., and Climent, S. 2004. A Grammar and Style Checker Based on Internet Searches. Pages 1931–1934 of: *Proceedings of LREC.*

Moré, J., and Climent, S. 2006. A cheap MT evaluation method based on Internet searches. Pages 19–26 of: *EAMT-2006: 11th Annual Conference of the European Association for Machine Translation.*

Moré, J., and Climent, S. 2007. A Cheap MT Evaluation Method Based on the Notion of Machine Translationness. Pages 83–90 of: *METIS-II Workshop: New Approaches to Machine Translation.*

Multilizer. 2011. *Can We Always Blame the Machine Translation?* http://translation-blog.multilizer.com/can-we-always-blame-the-machine-translator/ [Last checked: 16-05-2012]. Multilizer Translation Blog.

Mutton, A., Dras, M., Wan, S., and Dale, R. 2007. GLEU: Automatic evaluation of sentence-level fluency. In: *Proceedings of ACL 2007.*

Nagao, M. 1989. *A Japanese View on Machine Translation in Light of the Considerations and Recommendations reported by ALPAC, USA.* Tech. rept. Japanese Electronic Industry Development Association.

Nießen, S., Och, F. J., Leusch, G., and Ney, H. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC).*

Offersgaard, L., Povlsen, C., Almsten, L., and Maegaard, B. 2008. Domain specific MT in use. In: *Proceedings of the 12th EAMT conference.*

Oliver, A., and Climent, S. 2012. Building WordNets by machine translations of sense tagged corpora. Pages 232–239 of: *Proceedings of the Global WordNet Conference 2012.*

Palmer, M., Gildea, D., and Kingsbury, P. 2005. The Proposition Bank: an Annotated Corpus of Semantic Roles. *Computational Linguistics 31*, 71–106.

Papineni, K., Roukos, S., Ward, T., and Zhu, WJ. 2002. BLEU: a method for automatic evaluation of machine translation. Pages 311–318 of: for Computational Linguistics, Association (ed), *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.*

Popescu-Belis, A., Manzi, S., and King, M. 2001. Towards a Two-stage Taxonomy for Machine Translation Evaluation. In: *Workshop on Machine Translation Evaluation at MT Summit VIII,.*

Porsiel, J. 2011. Machine Translation and Data Security. *ITI Bulletin.*

Rarrick, S., Quirk, C., and Lewis, W. 2011. MT detection in web-scraped parallel corpora. Pages 422–429 of: *MT Summit XIII: the Thirteenth Machine Translation Summit.*

Reeder, F. 2001. In One Hundred Words or Less. In: *MT Evaluation Workshop MT Summit VIII.*

Reichartz, F., Kortes, H., and Paass, G. 2010. Semantic relation extraction with kernels over typed dependency trees. 773–782.

Richardson, S. 2004. Machine Translation of Online Product Support Articles Using a Data-Driven MT System. Pages 246–251 of: *Proceedings of AMTA.*

Schäffer, F. 2001. *MT post-editing: How to shed light on the "unknown task". Experiences made at SAP.* Tech. rept. SAP.

Sjöbergh, J. 2006. *The Internet as a normative corpus. Grammar checking with a search engine.* Tech. rept. School of Computer Science and Communication, the Royal Institute of Technology, Stockholm, Sweden.

Slype, G. Van. 1979. *Critical study of methods for evaluating the quality of MT.* Tech. rept. BR 19142. European Commission/Directorate for General Scientific and Technical Information Management (DG XIII).

Sommers, H., Gaspari, F., and Niño, A. 2006. Detecting Inappropiate Use of Free Online Machine Translation by Language Students- a Special Case of Plagiarism Detection. Pages 41–48 of: *Proceedings of the Eleventh Conference of the European Association for Machine Translation.*

Soudek, M., and Soudek, L. 1983. Cloze after thirty years: New uses in language teaching. *ELT Journal*, **37**(4), 335–340.

Specia, L. 2013. *Machine Translation Quality Estimation.* Slides for the course "Machine Translation Quality Estimation" at the University of Sheffield.

Specia, L., Raj, D., and Tuchi, M. 2010. Machine Translation Evaluation versus Quality Estimation. *Machine Translation*, May, 39–50.

Stymne, S. 2011. Blast: A tool for error analysis of machine translation output. Pages 56–61 of: *Proceedings of the ACL-HLT 2011 System Demonstrations.*

Tatsumi, M., Aikawa, T., Yamamoto, K., and Isahara, H. 2012. How Good Is Crowd Post-Editing? Its Potential and Limitations. Pages 69–77 of: *In AMTA 2012 Workshop on Post-Editing Technology and Practice (WPTP 2012).*

Tesniere, L. 1959. *Éleménts de syntaxe structurale.* Klincksieck.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. 1997. Accelerated DP based Search for Statistical Translation. In: *Proceedings of European Conference on Speech Communication and Technology.*

Turian, J. P., Shen, L., and Melamed, I. Dan. 2003. Evaluation of Machine Translation and its Evaluation. In: *Proceedings of MT SUMMIT IX*.

Turing, A. 1950. Computing Machinery and Intelligence. *Mind*, October, 433–460.

Vashee, K. 2011. *Analysis of the Shutdown Announcements of the Google Translate API*. Tech. rept. eMpTy Pages, http://kv-emptypages.blogspot.com.es/2011/06/analysis-of-shutdown-announcements-of.html [Accessed 19-05-2012].

Vasiļjevs, A., Skadiņš, R., and Tiedemann, J. 2012. Let′s MT!: a cloud-based platform for do-it-yourself machine translation. Pages 43–48 of: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL2012)*.

Vilar, D., Xu, J., D'Haro, L. Fernando, and Ney, H. 2006. Error Analysis of Statistical Machine Translation Output. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)*.

Vlad, L., Rogati, M., and Lavie, A. 2005. BLANC: Learning Evaluation Metrics for MT. Pages 740–747 of: *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*.

Vlasta, S. 2012. *More Reflections of a Human Translator on Machine Translation in The Field of Patent Translation*. http://www.translationdirectory.com/articles/article2349.php [Last checked: 04-04-2012].

Wagner, E. 1985. Rapid Post Editing of Systran. In: Lawson, V. (ed), *Tools for the Trade, Translating and the Computer*. Oxford.

Way, A. 2004. *MT Evaluation*. Slides for the course "Machine Translation" at Upsala University.

Way, A. 2012. Is That a Fish in Your Ear: Translation and the Meaning of Everything – David Bellos, Book Review. *Machine Translation*, **26**(3), 255–269.

White, J. S. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. Pages 193–205 of: *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)*.

White, J. S. 1995. Approaches to Black Box MT Evaluation. Page 10 of: *Proceedings of Machine Translation Summit V*.