# Evaluation of Statistical Machine Translation Engines in the Context of International Organizations

Lubrina, Paula

# *Evaluation of Statistical Machine Translation Engines in the Context of International Organizations*

**Paula Lubrina**
Faculty of Translation and Interpretation
University of Geneva
August 2015

**Master in Translation Technologies**

Thesis Director: **Pierrette Bouillon**
Jury: **Johanna Gerlach**

J'affirme avoir pris connaissance des documents d'information et de prévention du plagiat émis par l'Université de Genève et la Faculté de traduction et d'interprétation (notamment la Directive en matière de plagiat des étudiant-e-s, le Règlement d'études de la Faculté de traduction et d'interprétation ainsi que l'Aide-mémoire à l'intention des étudiants préparant un mémoire de Ma en traduction). J'atteste que ce travail est le fruit d'un travail personnel et a été rédigé de manière autonome. Je déclare que toutes les sources d'information utilisées sont citées de manière complète et précise, y compris les sources sur Internet. Je suis conscient-e que le fait de ne pas citer une source ou de ne pas la citer correctement est constitutif de plagiat et que le plagiat est considéré comme une faute grave au sein de l'Université, passible de sanctions. Au vu de ce qui précède, je déclare sur l'honneur que le présent travail est original.

Lubrina Paula

Geneva, le 26 août, 2015

*No intentes convencerme de torpeza*
*Con los delirios de tu mente loca,*
*Mi razón es a la par luz y firmeza,*
*Firmeza y luz como el cristal de roca.*
*Los claros timbres de que estoy ufano*
*Han de salir de la calumnia ilesos:*
*Ha plumajes que cruzan el pantano y no se manchan;*
*¡Mi plumaje es de esos!*

**Salvador Díaz Mirón**

**Acknowledgements:**

I would like to thank those that made this thesis possible:

Professor Pierrette Bouillon for her support and guidance;

Dr. Johanna Gerlach for having accepted so willingly to participate as a jury;

Azéddine Sefrioui-Benzerrou, Bruno Pouliquen and Christophe Mazenc for taking time out of their busy schedule to answer all of my questions;

My dearest friends, Marcela and Josefina, who are always present despite the distance;

My parents, Olga and Roberto, my brothers, Esteban and Daniel, and my aunt and uncle, Marcela and Javier, for believing in me all along these years;

And Diego, for holding my hand every step of the way.

# Index

## I. Introduction

Machine Translation (MT) is defined as the mechanization of the translation process, that is to say, the use of computers to translate written and oral texts from a source natural language to one or more target natural languages. In its relatively short history, MT has gone through numerous stages and overcome significant obstacles, some of which, at the time, seemed unsurmountable (**Chapter II. Machine Translation History**…). We can distinguish different types of MT according to their approach to translation and to the degree of mechanization. With respect to the system's approach, MT can be classified into three groups: (1) Rule-Based Machine Translation (RBMT), (2) Statistical Machine Translation (SMT) and (3) Hybrid Machine Translation. According to the degree of mechanization, it can be classified into (1) Fully Automatic High Quality Translation (FAHQT), (2) Human Aided Machine Translation (HAMT) and (3) Machine-Aided Human Translation (MAHT). The first designates a type of MT tool capable of producing high quality translation without any human intervention, while the second and third encompass varying degrees of human intervention (**Chapter III. Machine Translation Architectures)**.

In the present, more and more people are leaving old prejudices against MT behind, and they are gradually realizing of the great array of new potential applications for MT. In this way, different types of MT engines are entering into the life of individual users, companies, and most importantly for the present study, international organizations. With greater expectations come greater exigencies on the part of users: MT engines are introduced into the daily workflow of companies or organizations in the hope that they will streamline the communication process: reducing translation and revision costs and time. Nevertheless, that is only possible if the system responds to the exigencies and particularities of institutional communication and translation (**Chapter IV. Machine Translation and Institutional Translation**). With this in mind, a number frameworks and guidelines have been designed to adapt the existing norms for software quality assessment to

the evaluation of MT quality (**Chapter V. Software Evaluation and Machine Translation System Evaluation**).

The work presented in this thesis has three objectives: First, presenting a practical framework to carry out comparative (context-oriented) evaluations of machine translation systems for their introduction into institutional settings. Second, drawing conclusions about the performance of generic versus customizable MT. Third, establishing a link between machine translation and institutional translation by showing how international organizations can benefit from the use of MT systems as an additional type of computer assisted translation (CAT) tool.

In order to achieve those objectives, the present study adopts a mixed qualitative-quantitative approach: data is collected from an automatic test (quantifiable score) and a series of manual tests.

Since the first objective is designing a context-oriented evaluation, the tests were developed to serve a specific purpose in a real-life scenario: an international organization based in Geneva, Switzerland searching for a suitable SMT system to incorporate into their web portal. The case study was carried out during a four months internship in the Geneva headquarters of the International Social Security Organization (ISSA), from July to October 2014. The purpose of the internship (which is specific to the case study, and therefore differs slightly from the thesis main objectives enumerated above) was to assist the Member Services and Promotion Department in the first stage of the project to incorporate a service of automatic translation into the Association's multilingual web portal (www.issa.int). This first stage consisted in performing an evaluation of different MT systems in order to determine which one would match better the ISSA's needs and providing this information in a well described report for the Head of the Department to approve the choice. The stages that would follow the choice of one of the evaluated systems—design and development of the system's web interface, incorporation into the ISSA web portal, and the follow-up evaluation— were not performed in the framework of the internship and will not be discussed in the present work (**Chapter VI. Evaluation Description**).

The second and third objectives are expected to be achieved by extrapolating the results obtained during the case study so as to make assumptions that might serve as hypothesis for future studies.

The present thesis is organized in three main parts that reflect the order in which the different stages of the project were completed. The first part, the theoretical framework for the thesis, extends over **Chapters II to V** (mentioned above). The second part describes the methodological framework of the evaluation, including the selection of relevant quality characteristics, sub-characteristics and metrics, and the criteria for building the corpora (**Chapter VI. Evaluation Description**). In addition, this second part describes the execution of the tests and discusses its partial results (**Chapter VII. Running the Evaluation**). The third and last part of the thesis presents the final conclusions and the result of the evaluation, as well as some final remarks on the relation between machine translation and institutional translation (**Chapter VIII. Final Conclusions**).

## II. Machine Translation: History and State of the Arts

The field of Machine Translation, previously called "mechanical translation", is relatively young: from the attempts to mechanize dictionaries in the early years of the 20th Century to the elaborate rule-based and corpus-based systems of the last decades. This evolution has been documented in abstracts and papers from researchers worldwide, as well as in the literature of different authors that will be quoted as they appear in this chapter. Through the years, different approaches to MT arose as the result of advances in related fields: computer technology, artificial intelligence (AI) and languages sciences, including Computational Linguistics and Natural Language Processing (NLP). Moreover, the history of MT is tied to the historical events that lead to globalisation, information democratisation and the growing importance of the international community. This opening chapter gives a brief historical overview of MT, which provides the basis for starting going deeper into the different types of MT architectures (**Chapter III**) and the relation between MT and institutional translation (**Chapter IV**).

### 2.1. Brief History of Machine Translation Systems

The history of MT dates back to the beginning of the 20th Century and it is closely related to the history of computers and digital technology, as well as to the history of international and supranational organizations (e.g. United Nations Organization, the European Community, etc.) In order to present the evolution of MT through the years of the Cold War and the era of the information society, this historical overview will mainly follow the work of W.J. Hutchins (1992; 2000), who classified the history of MT into different stages, from the 1940s to the 1980s. Moreover, for the latest advances in the field (from the 1980s to our days), it will rely on the studies of different authors that will be quoted as they appear.

According to Hutchins (2000), the idea of automatizing languages and translation was present among philosophers and researchers much before computers where invented, as far back as the seventeenth Century. Nevertheless, he highlights that the twentieth Century brought about the first feasible possibilities to materialize those ideas. The first attempts to mechanise language were carried out in the field of cryptology during the 1930s. However, it was not until the end of the

1940s that the great potential of newly created computers inspired a number of pioneers around the world to start looking into the potentialities of language mechanisation. These pioneers came from different fields of study, from engineering and physics to linguistics, and pursued different aims: some were interested in developing a system capable of producing useful translations regardless of errors in grammar or style, mainly with the idea of overcoming the linguistic barriers for international communication; others saw it as an academic endeavour and a way of decoding the works of the human mind (Hutchins 2000).

Warren Weaver, from the Rockefeller Foundation, was one of those researchers interested in addressing the problem of translation in the context of international communication. In 1947, Weaver wrote a letter to his friend Norbert Wiener asking him if he had ever considered the possibility of developing a system capable of translating. In this letter, Weaver acknowledges the problems of semantics and draws a comparison between machine translation and cryptography. It is interesting to notice that, even at this early stage, Weaver considered that machine translation would be worth it even limited to the translation of "scientific material" with "inelegant (but intelligible) result". (Weaver 1955, 15-23). Although most efforts were directed towards the aim of limiting MT to certain scientific and technical areas and producing intelligible results (i.e. translations that could render meaning correctly, despite grammar and stylistic deficiencies), many researchers pursued the idea of fully automatic high quality translation (FAHQT).

According to Rod Johnson, from the Centre for Computational Linguistics at UMIST, Weaver's analogy between the process of translation and the decoding of military and diplomatic messages and his claim that both could be equably mechanized encouraged others to embark on research in this area (Johnson 1979).

During the 1950s, there was a great deal of enthusiasm on the potentialities of machine translation: research groups started to appear around the world (notably in the United States and the Soviet Union), mostly funded by governmental and military sources; in 1952, the first MT Conference was held; and two years later, a first MT journal appeared: *Mechanical Translation,* funded at MIT by William N. Locke and Victor Ygnve. By 1959, the possibility of translating a wider range of languages was being considered (Hutchins 2000). However, at the beginning of the 1960s, researchers starting to be overwhelmed by the problems that semantics and

pragmatics posed to MT, as well as by the slow progress of endless projects that seemed to end in disappointing results.

In 1963, the US National Science Foundation (NSF) requested the formation of a committee to evaluate the current state of MT and its future prospects. One year later, the Automatic Language Processing Committee (ALPAC) was created and, in 1966, they published the notorious ALPAC Report, in which it is claimed that useful machine translation was not feasible at the time nor in the foreseeable future (ALPAC 1966). The impact of the ALPAC report, mainly felt in the United States and other English-speaking countries, put an end to the MT boom. Nevertheless, research did not stopped completely: some research groups in The Unites States and Europe manage to get funds to continue working. Among the systems that resulted from those efforts, it is worth mentioning SYSTRAN, which would later on be used by NASA and the European Commission; and TAUM-MÉTÉO, a system developed by University of Montreal, that was capable of producing high quality translations of weather reports from English to French. (L'Homme 2008).

It was not until the end of the 1970s that interest in MT re-emerged. This "renaissance" of MT can be explained by many factors: on the one hand, the increasing political and economic importance of translation; and, on the other hand, the new practical possibilities brought about by relevant developments in the field of artificial intelligence, information retrieval, and linguistics models. The improvement and refinement of linguistic models, notably the emergence of new insights into how to treat semantic and pragmatic information, was one of the positive results of the ALPAC report, as it recommended to move away research from the field of MT into computational linguistics and general linguistics (Johnson 1979). Moreover, the research groups that continued investigating in the field of MT worked to tackle some of the problems pointed out in the ALPAC report. From those efforts rose a second generation of MT, which took an 'indirect' linguistic approach and focused on syntax (see **Chapter III. Machine Translation Architectures**).

As mentioned above, multilingual communication was gaining more and more political and economic importance, especially because of the linguistic needs of international and supranational organisations, such as the European Economic Community (EEC), where multilingualism was established by its founding treaties. In 1978, the European Commission set up a committee of experts to work on the

design of a common European MT system; and, in 1983, the project Eurotra became official. (Johnson 1979). During the 1980s, MT was no longer constraint to the sphere of the academic and the theoretical research, some systems were already being used for practical assistance, as was the case of the previously mentioned SYSTRAN and TAUM-MÉTÉO (Hutchins 2000).

The 1990s was a decade of fast technological changes, especially in the field of digital technology: The World Wide Web[1] had been recently invented and was about to revolutionize information storage and retrieval. The Web offered the possibility of accessing to a large amount of machine-stored information, such as "reports, notes, data-bases, computer documentation and online systems help" (Berners-Lee & Cailliau 1990). This contributed to an important change in the field of MT: many researchers, inspired by the growing availability of texts and, particularly, of existing translations on the web, switched from rule-based to statistical models (**Chapter III**). Statistical models seemed to offer elegant solutions for some of the main problems of the traditional rule-based approach such as the difficulty of designing rules that could treat all possible sentences in a given language and the costs of having experts developing said rules. The underlying idea was to develop an algorithm for translation based on probabilistic distributions, making use of vast bilingual corpora. The *Candide* project, carried out by IBM at the beginning of the 1990s, is one of the first projects in the field of SMT (Berger et al. 1994). Later on, during the first decade of the 21th Century, a great range of web-based SMT systems appeared, such as Microsoft Translator, or Bing Translator, and Google Translator (see **Chapter III**).

## 2.2. Conclusion

This brief historical overview showed how MT grew from an academic endeavour to a highly profitable commercial project. For the purpose of this study, it is important to highlight that, from its beginnings, the history of MT has been linked not only to the development of computers and digital technology, but also to the history of international organizations. The first offered the practical means to materialize researcher's expectations, while the second provided a strong drive for

[1] Retrieved from the "Official Google Blog: On the 25th anniversary of the web, let's keep it free and open". Official Google Blog. (Consulted April 5, 2015).

continued funding and cooperation efforts. For this reason, it is only logical that, at the present, more and more international organizations search to integrate MT systems to their translation process. Next Chapter (**III. Machine Translation Architectures**) presents the main characteristics of the different types of MT engines.

## III. Machine Translation Architectures

Machine translation systems can be classified in three broad groups: Rule-Based Machine Translation (RBMT) (**3.1.**), Corpus-based Machine Translation (**3.2.**), and Hybrid Machine Translation (**3.3.**) This Chapter offers an in-depth overview of these main MT architectures, with a special focus on the statistical approach (**3.2.2.**). Figure 1 summarizes the different approaches that will be discussed:

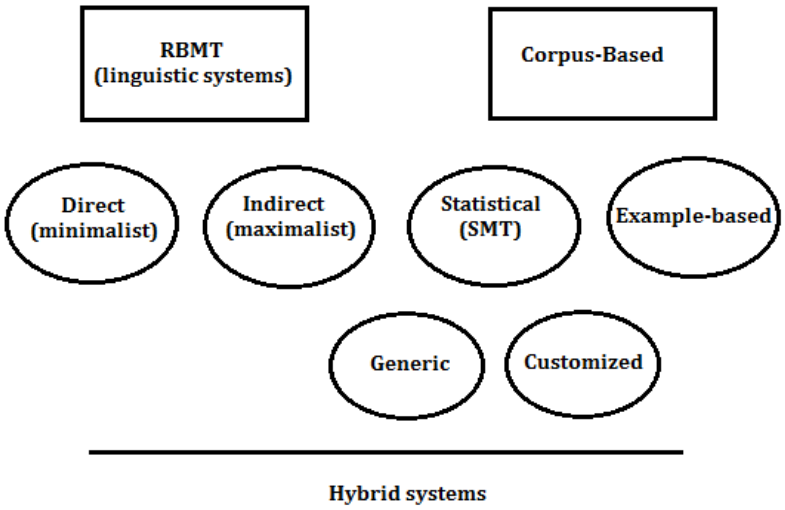



Fig. 1: MT Architectures

### 3.1. Rule-Based Machine Translation: First and Second Generation Systems

Among RBMT, there are two main approaches: the direct approach (first generation systems) and the indirect approach (second generation systems). RBMT is characterized by the use of linguistic rules, created manually by linguists and language experts. These experts base their work on the principles of formal linguistics, which aims at encoding linguistic phenomena with explicit unequivocal rules (L'Homme 2008). Early research on MT, focused on first generation systems and worked with direct translation models. These systems treated specific language pairs and consisted of well-developed bilingual dictionaries and rules for morphological analysis and pos-disambiguation. They also included reordering rules. Direct systems did not have an intermediate stage and they translated directly from a source language (SL) to a target language (TL). Figure 2 illustrates the direct process:

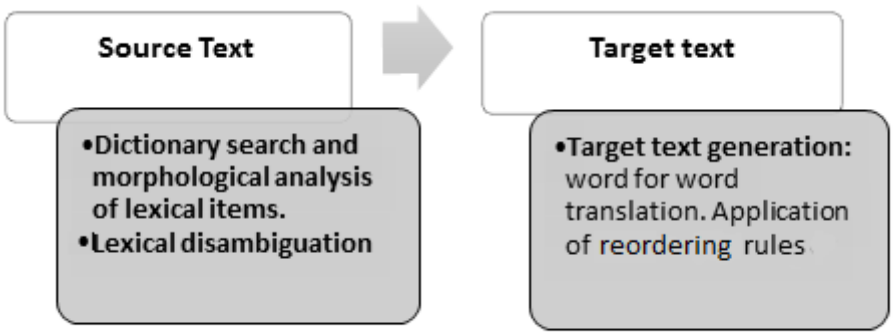



Fig. 2: Direct Systems

By the ends of the 1980s, the direct approach was already considered a "historical benchmark", and most researchers turned to indirect strategies (Tucker 1987). Nevertheless, in the present, some system (e.g. Reverso) continue to apply direct strategies.

After the general disappointment brought about by the notorious ALPAC report, research moved from a lexical to a syntactic focus. Second generation systems were named "indirect systems" because they generated an abstract representation of the source text on the basis of which the target text is produced. These systems are also known as linguistic knowledge systems, because they apply morphological, syntactic and/or semantic rules, called "grammars" (Quah 2006; L'Homme 2008).

Indirect systems can adopt a "transfer" or "Interlingua" approach. The former is language dependent and relies on the use of three dictionaries: two monolingual dictionaries (one for the SL and another for the TL) and a bilingual transfer dictionary containing rules that allow the system to go from one language to the other. The transfer model involves three phases: **analysis**, **transfer**, and **generation**. (Jurafsky and Martin 2009). First, the system analyses the SL, parsing each sentence according to the rules contained in the SL grammar; then, it converts the source representation into an intermediate representation or "machine-readable code" (Choudhury and McConnel 2013, 75), applying transfer rules (i.e. rules containing contrastive knowledge). This process is described in Jurafsky, and Martin (2009), as the transformation of a parse tree suitable for the SL into another parse tree, suitable for the TL. One example of this transfer process is the translation of a simple noun phrase (NP) from English to Spanish: if the engine is asked to

translate a NP like "retroactive measure" into Spanish, it will first parse the phrase into a tree identifying "retroactive" as an adjective (working as direct modifier) and "measures" as a noun (working as head of the NP). Second, it will transform this tree into one reflecting the Spanish form of the phrase, by performing all necessary transformations. Figure 3 illustrates the transfer phase described by Jurafsky and Martin (2009):

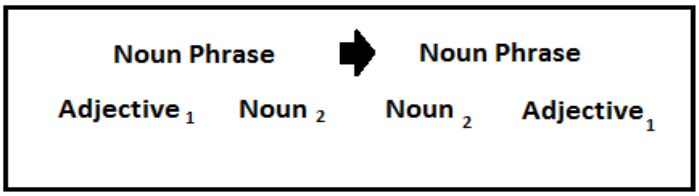



Fig. 3: Transfer Model: Syntactic Transformations (adapted from Jurafsky and Martin 2009)

Finally, in order to generate an output sentence in the TL, the engine performs a lexical transformation by searching SL lexical items in a dictionary and assigning TL equivalents. This process is not always simple, and may include a disambiguation process (e.g. in the case of polysemic words). In the present example, supposing that the dictionary lookup found only one translation for each item, the engine will not carry out a disambiguation process, but it will need to inflect the Spanish adjective "*retroactivo*" from masculine to feminine, to achieve adjective-noun agreement with the feminine noun "*medida*".

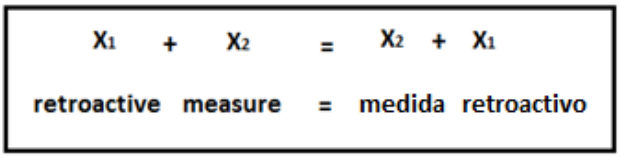



Fig. 4: Transfer Model: Lexical Transformation

In addition, some system (e.g. Lucy Software) rely on other mechanisms that can affect word ordering, lexical choices or even the whole structure of a sentence: *test* and *actions* (Thurmair, 1990). The former allow us define the conditions under which a given translation can be used. For example, certain English adjectives can have multiple translations into Spanish in accordance with the context in which they appear. In that case, the transfer dictionary contains a list of possible translations and we can associate different tests to each of them so that the system can select the

correct one. In this way, the system can disambiguate words and choose a translation adequate to its context. The latter modifies the source representation by means of SL annotations and TL annotations for certain words. In this way, we can influence the way in which a word is used in the target sentence: e.g. changing word ordering, syntactic structures (inverting the order of verbs, complements, etc.) (Ibid.). In total, there are three types of actions: those that modify the attributes of a single word, those that modify the context of a word, and those specific to multiword structures (Ibid.; Schneider 1991; *Introduction à Lex-Shop- Lexique de transfert - Test et Actions*. Traduction Automatique 2011[2])

In conclusion, this type of translation process requires an effective annotation of SL and TL linguistic knowledge and a careful design of transfer rules for the system to be able to map the attributes and values from the source to the target representation.

An alternative approach to RBMT is Interlingua, which treats translation as a process of "extracting" and "expressing" meaning by "performing a semantic analysis on the input from language X into the interlingual representation and generating from the interlingua to language Z." (Jurafsky and Martin 2009: 812). Since that meaning representation is language independent, interlingual architectures do not depend on contrastive knowledge and are suitable for multilingual translation. Nevertheless, this approach requires an extensive analysis and formalization of the semantics of different domains (Ibid). Moreover, remains the fundamental question of whether there is a universal meaning for all (or even most) languages.

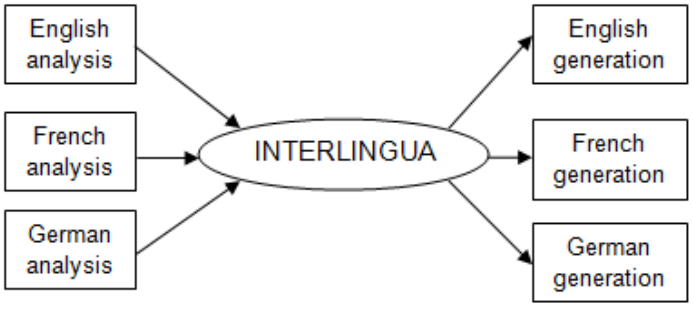



Fig. 5: Interlingua Model with Six Language Pairs (Hutchins and Somers 1992, 74)

[2] Master Class, Automatic Translation (2011); Master in Translation, Concentration in Translation Technologies. Faculty of Translation and Interpreting, University of Geneva. Retrieved 10/08/15 (09:52).

In brief, RBMT engines offer a series of advantages such as high quality results for translations in restricted domains and the possibility of correcting recurrent errors by manipulating rules. Nevertheless, they also entail some disadvantages concerning the need to handcraft linguistic rules and the capacity of a system to store such rules, as well as the generation of literal translations. During the lasts decades, with the increasingly amount of free available data, another approach to machine translation has become popular: Corpus-Based Machine Translation, and particularly, the statistical model.

### 3.2. Corpus-Based Machine Translation: Third Generation Systems

Third generation systems arose in the mid-1980s with the Candide Project at IBM (see Chapter **II. Brief History of Machine Translation Systems**). The underlying principle is that a phrase present in a SL corpus has a higher or lower probability of been an adequate translation to a phrase in a TL corpus (Gerlach 2009). During the last decade, and with the increasing amount of online available data, many researchers turned to this model in the hope of developing systems capable of exploiting the enormous amount of already available translations.

These systems rely on the analysis of correlations in stored data (real and example texts, comparable corpora, bi-text, etc.) from which they acquire the knowledge necessary for translation. It is important to point out, that these systems have no real knowledge or notion of grammar rules. Among corpus-based systems, we can distinguish between example-based and statistical-based models. Although both approaches will be discussed in this section, we will focus on SMT due to its particular importance for the present thesis.

#### 3.2.1. Example-based models

Example-based machine translation (EBMT) systems first appeared in the 1980s. EBMT engines "learn" how to translate from a language to another from the patterns of a sentence-aligned or parallel corpus (source phrases and their translations), which constitute the system's training data. The disadvantage of EBMT engines is that they require a very large corpus of perfectly aligned segments, which usually entails manual effort (Choudhury and McConnel 2013, 38).

### 3.2.2. Statistical Machine Translation (SMT) Models

Similarly to EBMT, statistical approaches exploit stored data, but they apply computer algorithms to generate translations. At this point, it is worth mentioning that, by 2003, most research in the field of SMT was carried out in private settings (e.g. IBM or Microsoft). This slowed down progress because every effort had to be duplicated and because it prevented effective comparisons among the systems (Koehn et al. 2006). With this in mind, a research group led by Philipp Koehn developed the open source toolkit Moses, which served as a starting point for a wide range of projects. Although the Moses engine will not be discussed in this study, it is important to mention it, since it was a significant step towards the democratization of SMT research and development.

In the present, most translation theorists agree that translation entails a compromise between *faithfulness* to the original and *fluency* in the TL. Drawing from this principle, Jurafsky and Martin (2009, 81) define the goal of translation as "the production of an output that maximizes some value function that represents the importance of both faithfulness and fluency". The translation problem is formalized following the Bayesian Model and applying the Noisy Channel Model, according to which the input is treated as a corrupted version of the output. The resulting formulae is that the best translation, *T*, equals the product of fluency, P(T), and the faithfulness, *P*(T|S):

$$\text{best-translation } \hat{T} = \text{argmax}_T \; P(T)\, P(S|T)$$

On the one hand, Fluency is quantified by means of the **language model**, P(T), a statistical description of a language, which analyses the frequency of **n-grams occurrences** in a monolingual corpus[3]. Language models, and the concept of n-grams, are described in detail further below. On the other hand, faithfulness or *fidelity*, *P*(S|T), is quantified by analysing to which extend TL words are plausible translations of SL words in a given sentence (Jurafsky and Martin 2009). This

[3] Retrieved from http://www.statmt.org/moses/glossary/SMT_glossary.html#n-grams (Last accessed July 5, 2015)

information is obtained from parallel bilingual corpora and incorporated into the SMT engine in a bilingual module called **translation model**.

In brief, we can summarize the basic requirements of a standard statistical model as follows: an aligned bilingual corpus (or bi-text), a language model, a translation model, and a decoder. This last part of the section will go deeper into these elements.

#### 3.3.1.1. The Translation Model

Translation Models are defined as bilingual module containing the data extracted from the bilingual aligned corpus. Said data is obtained by means of different algorithms, designed to calculate (1) the most likely translation for a given SL unit (**translation probability**); (2) the probability of a SL unit to be translated as a longer unit in the TL (**fertility probability**); and (3) the probability of a SL unit of changing its position in a TL sentence (**distortion probability**) (Nielsen 2009).

In the last paragraph, the expression "SL unit" is used in a general way to refer to the minimum syntactic unit considered when carrying out the translation. First approaches to SMT used word-based models, while later approaches used phrase-based models, which are the most popular nowadays.

##### 3.3.1.1.1. Word-Base Models and Phrase-Based Models

The first approach to SMT was the **word-based model**, which consisted in mapping individual SL words in a large bilingual (aligned) corpus to one or more elements in the TL corpus. Broadly speaking, this model implied the use of statistics regarding the count of times a given SL word was translated into a TL word. This resulted in a translation probability and a corresponding translation table or **T-table** (Koehn 2010). Different values were assigned to each translation, depending on the number of times they appeared in the corpus. For example, if we consider an English-Spanish corpus were the word "fiscal" appears 2,000 times: 1,000, translated into Spanish as "*presupuestario*"; 500 as "*fiscal*"; and 500 as "*tributario*", the translation probabilities for this word can be expressed in the following way:

- Given an English word (fiscal), the Spanish word "*presupuestario*" returns a probability of 0.5, and the other, a probability of 0, 25 of being the translation of "fiscal".

With the previous example, one of the major difficulties of this approach becomes clear: single words do not account for context that usually provides important information on translation choices, e.g. helping to explain why a single SL word can adopt different forms in the TL.

In order to solve this difficulty, researchers turned to **phrase-based models**, since phrases seemed more suitable as minimum translation units. In this context, a "phrase" refers to "any multiword unit" (Koehn 2010, 128). In phrase-based models, the input is broken up into multiword units, which might not correspond to a grammatically correct phrase (e.g. determiner and noun forming a noun phrase). Similarly to word-based models, these models estimate the probability of a TL phrase to be a suitable translation for a SL phrase. These probabilities are computed using the statistics obtained from a bilingual corpus. Several times now, we have mentioned that translation probabilities are computed from bilingual corpora. To be more specific, SMT engines extract data from **aligned corpora** or **bi-text**. L'Homme (2008) explains the concept of bi-texts as the mapping of segments (words, phrases, etc.) between a written source text and its translation.[4] Sentence alignment constitutes a complex subject in itself and will therefore not be described in detail.

#### 3.3.1.1.2. Factored Translation Models

Translations Models can also be classified in non-factored or factored models. The former analyse the phrase at the surface level, without accounting for part of speech (POS) or other syntactic features. The latter, factored models, optimize the mapping of words and improve translation quality by integrating additional linguistic features, called ***factors*** (Koehn and Hoang 2007, quoted in Koehn 2010). The additional information added in form of a factor can be, POS (useful for word reordering and agreement), semantic fields (useful to treat polysemy), and so forth.

[4] For more information on types of corpora and the different tools to exploit them, see Bowker and Pearson (2002) and Koehn (2010; Chapter 4.5)

The following figure (Fig. 6) illustrates linguistic annotations contained in a factored model:

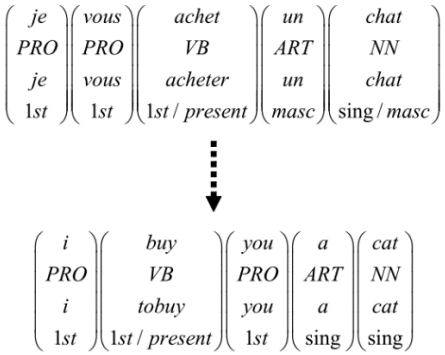



Fig. 6: Factored Translation Models (Koehn et al. 2006, 178)

Additional information can also be integrated into the source text, so as to improve the efficiency of the language model (prediction of right inflections, word order, etc.) (Avramidis and Koehn 2008, quoted in Koehn 2010).

### 3.2.2.2. The Language Model

Section 3.3 (**Statistical Machine Translation (SMT) Models**) explained that fluency was estimated by means of the language model. From this, we can infer that this model measures a sentence's probability of belonging to a given language. For example, if we translate a sentence from Spanish to English using Google Translate, we might find that the output looks very much alike a sentence originally produced in English:

Ginebra es especialmente linda durante el verano > Geneva is especially nice in summer

5

This is so, because Google's language model for English is highly efficient (in part due to the fact that this model is trained with an enormous English monolingual corpus). This is generally an advantage, since it enhances quality and pleases users; however, it can also be problematic for non-specialized users, or even for novel translators, as fluent results might hide meaning distortions.

[5] Google Translate: https://translate.google.com (July 05, 2015)

The most common models are n-gram language models, based on statistical descriptions of common characteristics of a language (e.g. the most probable word order for an English phrase) (Koehn 2010). An **n-gram** is a subsequence of *n* words embedded in a larger sequence[6]. If we consider the phrase "my sister goes to school" as the larger sequence, the subsequence "my sister" is a 2-gram occurrence. By analysing these occurrences, the system can compute how likely a sentence is to belong to the TL.

According to Koehn (2010), in order to calculate the probability of a string of words ($w_n$) of being a fluent sequence in a given language, $_P$(W), it is necessary to collect a large corpus of texts in said language and count how often the value W appears in it. This is estimated by computing how likely a word is to follow another given a certain ***history*** of preceding words. Above we considered the example of a 2-gram (bigram) sequence, but the size of n-gram sequences can vary from unigrams to 4-grams, depending on the system. The size of the sequence is usually decided on the basis of the size of the training data, since larger training corpora allows the computation of larger histories (Koehn 2010, 182-183).

#### 3.2.2.3. The Decoder

Decoding is one of the core elements of a SMT engine. However, explaining the full process of decoding is complex and requires the introduction of a number of new concepts. For this reason, this section gives a brief overview on the basics of the decoding process and the search algorithm. Once the language and translation models have been trained, the system applies a search algorithm to find the best scoring translation from a large amount of options. This process is called *decoding*. Since retrieving all possible translations would be computationally impractical and too expensive, SMT engines apply **heuristic search** methods, which offer no guarantees with respect to finding the best translation, and might lead to **search errors**. Nevertheless, these errors are not always caused by a search error, but to a model error, as the result of a bad scoring of the options (Koehn 2010).

[6] Retrieved from http://www.statmt.org/moses/glossary/SMT_glossary.html the 25/07/2015. Last accessed July 05, 2015 (12.34)

### 3.2.3. Customizable SMT Engines

SMT engines are also classified with regard to the type of training data. According this classification, SMT can be divided in two groups: generic and customizable SMT. Commercial systems such as Google Translate and Bing are called "generic" SMT, because they are trained with a general corpus, i.e. they rely on the large amount of stored data available on the internet. Differently, customizable SMT are those systems that are trained with specific (sometimes) private parallel or comparable corpora, with the objective of obtaining translations that would suit better the style and terminology of certain companies, institutions or organizations. Customized systems are expected to be able to reduce post-editing effort, not to render publishable quality translations. This is particularly important for the present study, as this is the kind of system that some international organisations (such as the United Nations and some of its specialized agencies) are integrating to the translation process.

## 3.3. Hybrid Machine Translation Engines

Among the advantages of SMT over RBMT presented in sections 3.1. And 3.2, we saw that developers do not need to create grammars, dictionaries or rules, because statistical systems exploit already translated texts. Although this makes SMT more economical and flexible, it means that results are difficult to correct. This difficulty is particularly annoying for language experts, who can recognize the type of rule that could correct a recurrent problem in the system's translation, but are not able to add it. Hybrid MT was developed in an attempt to solve this problem, combining the strengths of linguistic and the corpus-based systems. There are different approaches to do so: some systems make use of traditional linguistic rules to analyse and translate the source text and then apply corpus-based strategies to select the best solution. Another natural way to combine the advantages of both approaches is using the data of the linguistic system to train the statistical system: "we use the RBMT to create artificial training data for an SMT model." (Rayner, Estella, & Bouillon, 2012). These systems are still being tested and they may constitute the future of MT.[7]

[7] For more information on hybrid systems, please refer to Wolf, Bernandi, Federmann et al (2011).

## 3.4. Conclusion

This Chapter was devoted to the presentation of the main types of machine translation architectures, classified in accordance to their approach to translation. In particular, we focused on the characteristics of SMT because the candidate systems that will be tested during the evaluation are both statistical-base engines. This chapter did not pretend to give a valorisation of the different techniques regarding their quality, advantages or disadvantages, nor did it provide exhaustive descriptions on the characteristics of each architecture. To learn more about this topic, see Arnold et al (1994) and Hutchins and Somers (1992), for non-statistical methods; L'homme (2008), and Jurafsky and Martin (2009), for an overview of MT principles and types; and Koehn (2010), for statistical methods. Next Chapter (**IV. Machine Translation and Institutional Translation**) presents a brief introduction to institutional translation and emphasizes the importance of technology applied to translation and the place of machine translation in institutional settings.

## IV. Machine Translation and Institutional Translation

International organizations and machine translation share a history together. As long ago as the 1940s, Warren Weaver wondered on the possibility of implementing a system capable of producing understandable translations in view of the problem that linguistic barriers posed to UN agencies (see Chapter **II. Brief History of Machine Translation Systems**). Nevertheless, despite this long history and the numerous projects being launched at present, there seems to be a gap between translators and professionals in the field of NLP. MT, in particular, faces great resistance from professional translators and this hampers its way into the institutional linguistic workflow.

Chapter IV presents the concept of institutional translation, taking the Constitutive Model for organizational communication and the Four Flows Theory (McPhee and Zaug 2000) as a starting point (**4.1.**). Then, it discusses the connection between institutional translation and new technologies with the objective of providing some answers to the following question: What place does (or can) MT occupy in an organization's translation process? (**4.2.**)

### 4.1. Institutional Communication and Translation: the Case of Multilingual Organizations.

Institutional communication —also known as *organizational communication*— is a subfield of communications studies that focuses on the analysis of the role of communication in organizational contexts. This includes interpersonal communication (communication between individuals) or external communication (communication with clients, partners, members, etc.). The Constitutive Model is a common approach to organizational communication which arose during the 1980s. According to this model, the particular acts of communication within an organization must be seen as a constitutive and shaping factor, rather than as a product of institutional dynamics. Figure 7 illustrates the Four Flows Theory, described in McPhee and Zaug (2000), from the Arizona State University. In accordance to this theory, organizations require four different types of *message flows* in order to respond to distinct types of relations with different audiences:

> They [organizations] must enunciate and maintain relations to their members through *membership negotiation*, to themselves as formally controlled entities through *self-structuring*, to their internal subgroups and processes through *activity coordination*, and to their colleagues in a society of institutions through *institutional positioning*. (McPhee and Zaug 2000, 1).

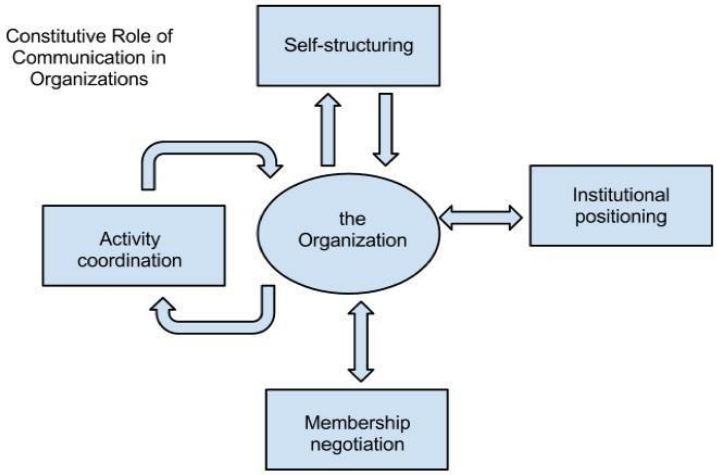



Fig. 7: Four Flows Theory (McPhee and Zaug 2000).8

If we consider this theory in the context of multilingual organizations, we can infer that: (1) institutional communication includes different types of interactions that require a variety of documents with different "status"; which, in turn, results in a (2) hierarchy of documentary sources; (3) since communication is a "shaping factor", the language used in a particular organization carries distinctive characteristics proper of its identity; (4) combined with this diversity of interactions, multilinguism in institutional settings entail an enormous workload for translators, and therefore, the need to maximize efficiency. Additionally, multilingual communication often entails (5) language asymmetry. According to Muñoz y Valdivieso (2002), some languages enjoy a higher status within organizations, notably English. It is widely known that English functions as a sort of *lingua franca* during negotiations, conferences and even internal correspondence. This last point is particularly important because one of the major oppositions from translators to MT is that it is said to encourage literal translation, benefitting source language (often English) structures and form over the target language.

[8] Image retrieved from http://creativecommons.org/licenses/by-sa/3.0), Wikimedia Commons, the 10/04/2015 (13:47)

The first and second point are closely related: because of the organization's limited time and resources, documents with a higher "value" will go through a more careful translation and editing process than documents with a lower value. Consequently, the later might have stylistic, terminological and grammar problems. Translators need to take this into account when consulting the organization's translation memories, as this tools do not filter texts by quality. There are some ways in which organizations can deal with this issue, such as creating filters by date (*texts added or modified after a certain date*) or name (*texts added or modified by certain users, e.g. revisers*.[9]), and then adding cleaned texts with the values defined on those filters. In this way, the tool can rely on the available filters to recognize the texts we want to use. The previous example describes a rather superficial solution; however, going deep into this subject will take us far from our main topic. If the organization uses, or plans to use, a SMT customizable engine (see section **3.2.2.**), using low quality corpora (e.g. one presenting high inconsistency in the translation of terms) for the training will most likely affect the performance of the engine.

The third point introduces the concept of *institutional identity*. Broadly speaking, this means that institutional messages express the voice of particular institutions, which can even be reflected in its own variety of a language. For example, the so-called *eurotalk,* refers to the variety of a language (English, Spanish, French, etc.) used in the European Union (EU), which does not respond to the variety of any particular region, but to the needs and peculiarities of the organization (Wager, Bech y Martinez 2002). In the case of the United Nations, some authors have stated that translation in the context of the UN constitutes a separate field of specialized translation, due to the particular characteristics and demands that derivate from the special nature of international diplomacy (Cao & Zhao 2008). It could be argued that the last observation is also true for other international organizations that might not strictly belong to the United Nations. From the point of view of translators, it means that they have to respect those peculiarities and follow certain conventions that affect their choices. Consequently, translating in

[9] Revisers do not normally add texts to the organization's text bases; their names can, however, appear in term bases or glossaries to mark "correct" or "preferred translations", especially when there are several available translations and the organization does not have time or resources to clean the bases. Additionally, the person in charge of CAT tools can create a mock user name like "edited text" to mark good quality texts produced by senior revisers.

institutional contexts often calls for a whole new training and adaptation process for many translators. The daily work of institutional translators involves the use of CAT tools, which help them make informed decisions on aspects of general language that might vary within the organization.

The fourth point is particularly important, as high volumes of work imply the need for effective administration and project management, as well as for the minimization of time and effort to take on a higher workload. For example, in a short presentation by the Directorate-General for translation (DG Translation)[10] of the European Commission, it was stated that, during 2013, over two million pages were translated by 1700 translators working for the European Commission. However, it was pointed out that the Commission would in fact need to translate almost 6.8 million documents a year for its webpage (Europa.eu) to be fully multilingual, this without taking user generated content into account.

The purpose of this section was not to provide a detailed explanation of organizational communication (which is a complex field in itself) or institutional translation, but rather to establish some grounds and list some of its main characteristics that will help us to understand the place of MT in the institutional context.

### 4.2. Institutional Translation and New Technologies: The Case of MT

For decades international organizations have turned to new technologies in order to reduce the time of the translation process, while keeping high quality. The debate over the introduction of new technologies into the translation process is no longer about whether it should or not be a part of the process, but rather over what type of tools should be exploited and how. The use of translation memories, corpus, terminological databases, and other CAT tools, is now commonplace in most, if not all, organizations. However, there has been some resistance to incorporate MT into the translation process. CAT tools can improve work efficiency, optimizing time and improving quality (terminological coherence, preserving style and institutional identity, correcting grammar mistakes and typos, and so on). Nevertheless, each

[10] Retrieved from http://ec.europa.eu/isa/documents/presentations/european-commission-machine-translation-for-public-administrations-in-the-eu-member-states_en.pdf. Las accessed August 7, 2015 (14.52)

organization needs to choose and customize its tools to suit their needs and available resources. Although organizations generally offer certain tools to their employees with the objective of achieving uniformity and improving efficiency, translators enjoy some freedom in how to make use of these tools. Considering the case of MT, translators tend to resist its introduction into their workflow, claiming different problems such as serious grammar mistakes, senseless sentences, poor style, etc. Svoboda (2013) discusses the paradigm shift brought about by MT and points out that some of MT engines have improved considerably, in some cases producing translations of acceptable quality for certain language combinations. Svoboda claims that the introduction of CAT tools did not produce a paradigm shift, because they assisted translators in their usual tasks; while MT did, since it produces translated text, turning translators into post-editors. However, it is interesting to make ourselves a question: Can MT be integrated into the translation process in the same way as other CAT tools? During the last few years, it has been suggested that MT (particularly customizable SMT) can work as a "translation accelerator", providing translators with context for different terms and expressions and helping organizations to handle large amounts of data (Pouliquen et al 2013). For example, the Interinstitutional Committee for Translation and Interpretation of the European Commission launched the MT@EC[11] project (machine translation for the EU) mainly with the purpose of developing a robust customizable SMT engine to improve efficiency in terms of internal communication and access to online content. When it comes to text that need to be produced and translated with publishable quality, the purpose of this engine is that of assisting translators, as a source of ideas, a tool to search or verify terminology, and so on. For example, the United Nations' headquarters in New York has been using SMT for some time, producing drafts for post-editing. This is mainly used by senior revisers (P5), since it has been observed that SMT output is sometimes highly fluent, but not very accurate (see section **3.2.2. Statistical Machine Translation (SMT) Models**) (Elizalde et al 2013). Despite the cognitive effort required to sort out post-editing problems, translators in the UN office in New York, are said to have welcomed the use of SMT as an additional tool

[11] Released in 2013, and upgraded to its 2.0 version in 2014, the MT@EC is a SMT customizable system, co-funded by EU Framework Programmes for research and innovation and the ISA Programme. (DGT-MT@ec.europa.eu)

to help speed-up the translation process for some document types. This office first started using generic SMT engines, especially Google Translate. The output was generally consistent because Google Translate had been trained with a large corpus of UN documents. Nevertheless, with the incorporation of new training data, it was perceived that the quality and consistency of translations had decrease. Therefore, they starting testing a customizable open source SMT system developed by a research team at the WIPO: TAPTA4UN. The project envisages the incorporation of this systems for all language combinations and all UN headquarters (Elizalde et al 2013).

### 4.3. Conclusion

This chapter started explaining that institutional translation is a complex multi-step process, and that, as any other process, it requires collaboration and openness at every stage of the process in order to accomplish its goal. Moreover, it showed that several international organizations have started offering SMT systems (particularly customizable statistic engines) to their translators in addition to other CAT tools. Systems such as MT@EC or TAPTA4UN share some common characteristics: they are open source and free, Moses-based engines, trained with the organization's corpora in order to support translators in their daily tasks. One fundamental lesson drawn from the experience of incorporating these two SMT systems to the institutional translation workflow of the EU and the UNO (New York headquarters, mainly) is that translators and revisers are much more likely to appreciate these engines if they are incorporated as machine translation *accelerators*, i.e. as an additional CAT tool.

## V. Software Evaluation and Machine Translation Systems Evaluation

Philipp Koehn opens the eighth chapter of his book *Statistical Machine Translation* (Koehn 2010, 217) with two compelling question: "How good are statistical machine translation systems today?", and "[...] how should we evaluate machine translation quality? Koehn focuses on the assessment of the system's output (i.e. translations generated by an MT engine) and presents a series of methods used to that purpose. Nevertheless, evaluating MT implies more than just assessing the quality of its output. According to the ISO/IEC norms (set of norms developed by the International Organization for Standardization and the International Electrotechnical Commission), evaluating software quality also requires the assessment of its internal characteristics.

Chapter V is divided in two parts: an analysis of different methods for MT quality evaluation (**5.1. Evaluating the Quality of Machine Translations**); and an overview of the ISO/IEC norms for software evaluation, as well as its importance in the assessment of MT systems' internal characteristics (**5.2. Evaluating MT Engines as Software Products**).

### 5.1. Evaluating the Quality of Machine Translations

Experts in the field of translation have given much thought to what it means for a sentence to be the translation of another, and most importantly, they have identified different *variables* that need to compromise in order to obtain an acceptable translation (Jurafsky and Martin 2009). This has been deeply studied by experts in the field of NLP, who decomposed translation quality in two main criteria: *adequacy* (or *faithfulness*) and *fluency*. The former refers to the quality of the output as a text belonging to the target language system, independently of the source text; while the latter refers to the correspondence of meaning between the source and the target texts. (Ibid; Koehn 2010). Nevertheless, and despite these criteria, evaluating translation quality is difficult, even for human judges, partly, due to the fact that a source text can have many adequate or valid translations. Generally, the output of a SMT system is evaluated in order to optimize the system's performance or to compare the quality of two or more systems. As underlined by Koehn (2010), MT is not an end in itself, it is always expected to support a task; and therefore, the system

needs to be evaluated in terms of its adequacy to the final task. SMT output evaluation can be carried out manually (by human judges) or automatically. This section presents both methods (**5.1.1. Manual Evaluation**; and **5.1.2. Automatic Evaluation**), as well as an overview of the main arguments for and against them (**5.1.3.**).

### 5.1.1. Manual Evaluation

The first method, manual evaluation, requires the participation of a group of monolingual or bilingual evaluators who generally analyse and grade the output of one or more SMT engines. Human evaluators usually grade the translation segment by segment, but a broader context is often provided to help them interpret the sentence. Generally, reference translations are also included, especially in the case of manual evaluations carried out by monolingual judges.

According to Koehn (2010), it is impractical to judge machine translations in absolute terms (e.g. correct or incorrect). It is more common to apply a scale measuring the output's degree of correctness in terms of *fluency* and *adequacy*. One of the disadvantages of evaluating translations in terms of these attributes, especially when offering reference translations, is that the human mind is capable of filling in missing information. By reading the reference, evaluators get a gist of the segment's meaning and might not notice that the output is confusing or that it does not transmit the whole meaning of the original.

Two common manual evaluation tests are (1) grading the translation segment by segment in terms of fluency and adequacy, and (2) ranking the output of the SMT engines one against the other, according to a given scale. In this study, the first method is used to assess fluency and adequacy (described in sections **7.1.1.2.1.** and **7.1.1.2.2.**). The second method is used to compare the output of both candidate systems in terms of grammatical correctness (described in section **7.1.1.2.3**).

Evaluating translations in this way gives a valuable insight into the engine's performance (e.g. revealing regular errors), especially when evaluators have experience not only in the field of languages, but also in NLP. However, manual evaluation poses a number of problems, such as the subjectivity of human evaluators and the lack of precision when it comes to the metrics applied. Sometimes it can be hard to distinguish whether the score reflects the system's

quality or the evaluator's leniency. In addition, it is expensive and time consuming: SMT researchers need to carry out tests very frequently, and evaluating machine translations manually can take weeks or even months. For this reason, researchers prefer to evaluate the output of SMT using automatic measures of different sorts (Koehn 2010).

### 5.1.2. Automatic Evaluation

Automatic evaluation offers the possibility to assess a translation quickly, precisely —at least in terms of metrics—, and at low cost. These systems analyse the output by comparing it with one or more reference translations. Although perfect matching is barely impossible —and therefore, not expected—, in theory, the best translation is the one that is closest to the reference translation. This "closeness" is calculated in different ways depending on the automatic metric. In general terms, the quality of a SMT system's output is reflected by the fall or rise of the automatic scores. Although the effectiveness and true value of these measures is constantly called into question, researchers continue relying on automatic methods and much progress has been made in this area. In order to get a clearer idea of how system can evaluate translation quality, we will analyse the following methods: Precision and Recall (**5.1.2.1.**), Levenshtein-Based Methods (**5.1.2.2.**), and N-grams-based Methods, including BLEU, METEOR and NIST (**5.1.2.3.**). The main sources consulted are Koehn (2010); Mauser et al. (2008); and Papineni et al. (2002).

#### 5.1.2.1. Precision and Recall.

Broadly speaking, these metrics are based on **word matches.** *Precision* measures the output of a MT engine against a reference translation in order to verify how many words they share. For example, if the input has six words, three correct words out of six represents a ratio of 50% (Koehn 2010). However, focusing on word matching alone suffers from many shortcomings (e.g. words can be out of order). The metric *recall* measures how many of the words the system should have generated are correct (Ibid.). Some systems apply a combination of recall and precision: a common way of combining these two methods is the **f-measure**, defined as the "harmonic mean of the two metrics" (Koehn 2010, 224):

$$\text{precision} = \frac{\text{correct}}{\text{output-length}} \qquad \text{recall} = \frac{\text{correct}}{\text{reference-length}} \qquad \text{f-measure} = \frac{\text{correct}}{(\text{output-length} + \text{reference-length})/2}$$

Fig. 8: Precision, Recall and the f-Measure. Adapted from Koehn (2010, 223-224.)

Similarly to recall, position-independent error rate (PER) uses reference length as a divisor, but it measures mismatches instead of matches. This measure also considers superfluous words that need to be deleted (Koehn 2010, 224; Mauser et al. 2008).

#### 5.1.2.2. Levenshtein-Based Methods

Word error rate (WER) was first used in the field of speech recognition. It applies the Levenshtein distance algorithm, defined as "the minimum number of editing steps —insertions, deletions and substitutions— needed to match two sequences." (Koehn 2010, 224; Mauser et al. 2008). This algorithm calculates the WER of a translation by dividing the number of editing steps with the reference length. In this way, it looks for the minimal cost solution: the lower the editing effort, the higher the translation quality. In order to determine the number of editing steps of a given translation, the system uses a word alignment matrix, displaying the output sentence on top of a grill and the reference on the left, as shown in Figure 9:

Output translation

Reference translation

| 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 2 | 3 | 4 | 5 |
| 2 | 1 | 0 | 1 | 2 | 3 | 4 |
| 3 | 2 | 1 | 1 | 2 | 3 | 4 |
| 4 | 3 | 2 | 2 | 2 | 3 | 4 |
| 5 | 4 | 3 | 3 | 3 | 3 | 4 |
| 6 | 5 | 4 | 4 | 4 | 3 | 4 |
| 7 | 6 | 5 | 5 | 5 | 4 | 4 |

Fig. 9: Levenshtein Distance Alignment Matrix. Adapted from Koehn (2010, 225)

Figure 9 illustrates the way in which the system assigns points to a translation and placed them in the matrix. Starting from the top left corner, it assigns points using word matches (cost 0) or editing cost (cost 1), in the following way (Fig. 10):

| | |
|---|---|
| Matches | Cost from the point to the left-top. |
| Substitutions, | Cost from the point to the left-top plus one. |
| Insertions | Cost from the point to the left plus one. |
| Deletions. | Cost from the point above plus one. |

Fig. 10: Levenshtein Distance, Points Assignment. Adapted from Koehn (2010, 225)

In order to understand this method better, let us go through the grill point by point. The points presented in Figure 9 correspond to the example studied in Koehn (2010, 225). Marked in grey, we see the "lowest-cost path" for translating a sentence.

| | |
|---|---|
| Output | ***Israeli officials responsibility of airport safety*** |
| Reference | ***Israeli officials are responsible for airport security*** |
| The first cell starts from "zero" | |
| **1.** Two word matches (two zeros starting from the top left corner). | |
| **2.** The third word is not a match: one point is added to the grill. | |
| **3.** Two editing steps: **substitution** (2 pts = the value diagonally to the left top plus one) and **addition** (3 pts = the value at the left plus one) | |
| **4.** Word match: zero pts (the points remain the same). | |
| **5.** Substitution: point to the left top (3 pts) plus one (4 pts) | |

Table 1: Levenshtein Distance, Points Assignment: detailed description based on Koehn (2010, 225)

The WER is calculated using the points obtained from the computation of editing steps. For instance, if a system A gets a WER of 60% and a system B, a WER of 75%, the difference in scores (in this case 15%) is a penalty given to the SMT system for errors in word ordering, deletions, and so on. If multiple reference translations are used, the reported error for a translation candidate is the minimum error over all the references (Mauser et al. 2008). This metric has a clear shortcoming, considering that a correct translation sometimes requires a different word ordering, or the

addition of new words. Another metric, position-independent word error rate (PER) is supposed to overcome this problem by comparing the words in the output and the reference without considering word-order (Popovi and Ney 2007). However, as word-order is usually important, both WER and PER should be calculated when measuring a system's output.

#### 5.1.2.3. N-Grams-Based Methods: BLEU, METEOR and NIST

The BLEU (Bilingual Evaluation Understudy) was first proposed by Papineni et al. (2002) and Papineni, Roukos, Ward, & Zhu (2001). BLEU is a precision-based metric that provides a more sophisticated solution to the **role of word-order**. It works similarly to PER, but takes into account matches of larger n-grams to the reference translation. In section **3.2. Corpus-Based Machine Translation: Third Generation Systems**, n-grams were defined as strings of tokens, where *n* is the number of items in the subsequence[12]. Two main elements of the BLEU metric are **n-gram precision** and the **brevity penalty**. The former is defined as "the ratio of correct n-grams of a certain order *n* in relation to the total number of generated n-grams of that order" (Koehn 2010, 226). The metric is called BLEU-4, because the maximum value set for n-grams is typically 4. It is important to point out that a precision of 0 for 4-grams is not common at sentence level, and therefore, BLEU is usually calculated over the entire corpus. The brevity penalty has a double purpose: compensate that BLEU does not calculate *recall* and address the problem of word drop. By applying this penalty, the metric assigns lower scores to outputs that are too short (Papineni et al. 2002; Koehn 2010). In addition to these two innovative elements, the BLEU score incorporates the use of **multiple reference translations**. The underlying idea is that, given the variability of translation options, evaluating the SMT output against multiple human translations helps the system to recognize all acceptable translations for ambiguous parts of the sentence. However, multiple reference translation make the computation of reference length more difficult, as the system has to choose among the references: if two reference lengths are equally close the shorter one is chosen (Koehn 2010, Papineni et al. 2002).

[12] Retrieved from http://www.statmt.org/moses/glossary/SMT_glossary.html#n-grams. Last accessed July 5, 2015 (13:14)

The METEOR metric is a variant of BLEU that puts more emphasis on recall. This measure overcomes some shortcomings of BLEU by incorporating the use of stemming and synonyms with the purpose of taking into account near matches. The main disadvantages of METEOR are that its matching process requires an expensive word-alignment and that it contains many parameters that need to be tuned (e.g. the weight of recall and its relation to the weight of precision) (Koehn 2010; Cancedda 2009). Another metric that uses BLEU as a basis is NIST, a precision metric developed by the National Institute for Standards and Technology in order to build on the BLEU measure. Similarly to BLEU, the NIST computes n-gram precision for different n-gram lengths and multiplies it by a brevity penalty. The main difference between the two is that NIST takes into account the *informativeness* of n-grams by assigning higher weight to less frequent ones (Cancedda 2009; Mauser et al. 2008).

### 5.1.3. Debate over Output Assessment: Manual vs. Automatic Evaluation

On the one hand, human evaluations of machine translation are extensive and thorough, and provide valuable information to researchers. Either by grading bilingual output segment by segment, or by assessing fluency by reading a monolingual test set, human judges can give a valuable insight into the system's performance, or make judgements on whether a system produces better results than another. Nevertheless, as mentioned before in section **5.1.1.,** human evaluations are expensive and time consuming; it can take weeks, or even months, to finish. Moreover, the work put into it cannot be reused. Developers need to carry out test frequently in order to assess how changes, even small ones, affect the SMT system's performance, and they benefit from automatic evaluation methods in that they are inexpensive, quick, and sometimes language-independent. In brief, whenever there is need for quick or frequent evaluations, automatic measurements are preferred (Papineni et al. 2002).

On the other hand, the use of automatic evaluation measures to assess translation quality is constantly under debate: Some common critiques are that automatic scores depend on too many factors (alignment, brevity penalty, n-gram precision, and so on) and that they do not truly reflect fluency and adequacy as a human evaluator would (Koehn 2010;Turian, Shen et al 2003). Other arguments against these measures are that their results, expressed in absolute numerical

scores, are difficult to read, and do not always offer clear and meaningful information on aspects like fluency, adequacy and grammatical coherence. What is more, n-gram based measures tend to privilege statistical MT output, due to the compatibility of their methods.

It is important to underline that, when choosing an evaluation metric, it is essential to keep in mind the adequacy between the purpose of the evaluation and the method used to achieve it. If the purpose of the evaluation is to check whether a small modification in the language model has contributed to improve the system's quality, the time required for the preparation of a manual test (searching and hiring suitable judges, preparing and distributing the test set, analysing the results, etc.) will hardly prove most adequate to the task than running an automatic test.

### 5.2. Evaluating MT Engines as Software Products

In order to decide which SMT system is the most appropriate for a particular user —in this case, the ISSA (see **Chapter I. Introduction**) — it is necessary to assess not only the quality of the output, but also the quality of the system as a software product (internal characteristics). This entails taking into account certain characteristics of its context of use, so as to design a quality model to measure its **quality in use**. This section includes a brief presentation of the main norms describing software quality (**5.2.1.**), the description of two frameworks for software quality evaluation: EAGLES (**5.2.2**) and FEMTI (**5.2.3.**). Next Chapter (**VI. Evaluation Description**) extends on these norms and frameworks, focusing on how they have been used to design the present evaluation.

#### 5.2.1. Norms and Frameworks for Software Evaluation

The most important standards for software evaluation were developed by the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC), which created a joint commission in order to satisfy the increasing need for a formal framework to measure software quality. According to the norms ISO/IEC, a system's quality is the "[d]egree to which a software product satisfies stated and implied needs when used under specified

conditions." (ISO/IEC 25000:2014)[13] Therefore, evaluating a system's quality provides information on the *value* this system has to its users. This information can be used for many purposes: to improve a software product, to assess the impact of changes made to a software, or, as in the case of the present evaluation, to establish a basis for decision-making, regarding the acquisition of a software product over others (Popescu-Belis, Estrella, King, & Underwood 2006).

#### 5.2.1.1. The ISO/IEC Norms and the EAGLES

According to the ISO/IEC Commission, in order to measure the quality of a software, evaluators need to identify the system's *necessary* and *desirable* characteristics according to the system's objectives and the tasks it is expected to perform (ISO/IEC 25000:2014)[14]. The first refers to the minimal requirements the systems needs to fulfil, while the second, to additional characteristics that will add value to the system. The analysis of those characteristics results in the definition of **quality models** containing a series of Quality Characteristics (QC), sub-characteristics and attributes, gathered in the norms ISO/IEC 9126-1:2001, and later revised by the ISO/IEC 25010:2011 and the ISO/IEC 25000:2014. It is important to point out, that this hierarchical organisation of QC implies that some attributes are considered more relevant than others depending on the objective of the evaluation, the purpose of the software and the need of the customer (this is further explained in section **6.6. Measurement Method: General Description**). In addition, ISO makes a distinction between different types of quality: *internal*, *external* and *in-context*. The first refers to features that evaluators can assess without running the software (i.e. languages supported, size of corpus or dictionary, etc. In the present study, we refer to these characteristics as "software product" (See **6.6.2.2. Operational Evaluation**). The second are those characteristics that can only be analysed by assessing the results produced by running the software (i.e. the quality of the resulting translation or output; see section **6.5.2.1. Declarative Evaluation**). Finally, the third assess the

[13] *Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE*. Retrieved the 03.05.2015 from https://www.iso.org/obp/ui/#iso:std:iso-iec:25000:ed-2:v1:en. Access to informative sections.

[14] *Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE*. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:25000:ed-2:v1:en the 03/05/2015. Access to informative sections.

software by placing it in a particular context of use (i.e. indirect and direct users, the translation's target audience, etc.)

In the 1990s, the ISO/IEC norms served as a starting point for a research project that was primarily concerned with adapting and extending these norms for applying them in the field of NLP. The EAGLES Evaluation Working Group (EWG) launched a project with the objective of identifying and defining "the components of a compendium of evaluation criteria and associated techniques, together with guidelines for their use, from which the individual evaluation user can select those techniques which are relevant to his purpose" (EAGLES 1996). EAGLES stands for Evaluation of Natural Language Processing Systems. In April 1999, The EAGLES EWG published *The EAGLES 7-step recipe*, which provides an overview of the maim steps to be followed in order to achieve a successful evaluation of language technology systems or components. The document consulted is a practical summary of the EAGLES Final Report, first published in 1996 and later on extended by a second report published in 1999[15]. This important research set the basis for other projects on the subject, among them, the FEMTI Framework, discussed in the next section.

#### 5.2.2.2. The FEMTI Framework: Evaluation Design

ISSCO, originally called "Dalle Molle Institute for Semantic and Cognitive Studies", was founded in 1972 in Lugano, Switzerland, by a private non-profit foundation, the Dalle Molle Foundation. Broadly speaking, the aim of this foundation was to conduct studies on the influence of new technologies on language and cognition, in order to provide answers on this matter in the midst of a rapidly changing world. Later on, ISSCO moved to Geneva and became attached to the former School of Translation and Interpreting (ETI) —nowadays, Faculty of Translation and Interpreting (FTI) — at the University of Geneva (UNIGE), which gave birth to The Multilingual Information Processing Department (TIM). During the last decades, TIM/ISSCO has participated in numerous projects at local, federal and European levels. Two of those projects were, precisely, the EAGLES I and II and the MTEval II (FEMTI).

[15] Ibid

The Framework for the Evaluation of Machine Translation in ISLE (FEMTI) is an online resource developed in the context of the research project Quality Models and Resources for the Evaluation of Machine Translation (SNSF Project n. ° 200021-103318; from October 2004 to September 2006; described in Popescu-Belis et al. 2006). This resource focuses on the evaluation of MT software in context, which entails that the *context of use* has a direct influence on the choice of the system's QC, sub-characteristics and attributes to be assessed during the evaluation (i.e. in the definition of a *contextual quality model*).

FEMTI provides a formalized framework for the design of particular evaluations: it consists of two screens: one in which the evaluators select the evaluation's characteristics and requirements (e.g. Evaluation Type and Context Characteristics); and another, which shows a list of system characteristics, with its sub-characteristic and attributes, as well as a set of proposed metrics (e.g. in Figure 11: 2.1 Functionality **>** 2.1.1 Accuracy **>** 2.1.1.1.1 Terminology). This is illustrated in Figure 11:

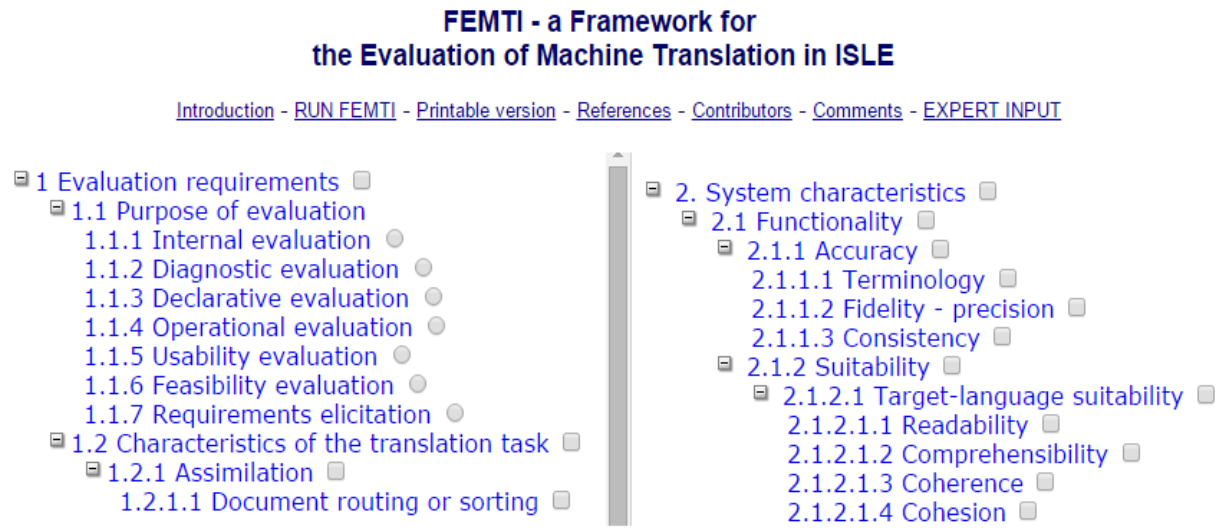



Fig. 11: FEMTI Design Environment[16]

These two screens are linked, so that FEMTI is able to suggest in the second one the characteristics that match the evaluation's requirements introduced in the first one. The results can be displayed in PDF, HTML or RTF format. Additionally, FEMTI

[16] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/st-home.html, the 03/08/2015.

provides: a glossary, defining key terms used in the field of software evaluation; a detailed description of each element presented in both screens, proposed metrics, references and comments. It also offers an *Expert Input* screen that is not been used during the present project, and therefore, it is not be described here. The following Chapter (**VI. Evaluation Description**) extends on FEMTI Framework, focusing on its application during the design of the present evaluation.

### 5.3. Conclusion

This Chapter described a number of existing resources, available for evaluators who wish to assess software quality. There are many reasons that can motivate a person, a company or an organization to carry out an evaluation of this sort: testing improvements on a system under development, pondering the plus points and short-comings of a number of systems to decide which one suits them better, writing a market report, or obtaining an international quality certification, to mention some. The ISO/IEC norms provide an extensive and detailed description of software quality (available for consultation or purchase in the ISO webpage[17]). This large description might not always result practical for individual users or novel evaluators. EAGLES and FEMTI provide the necessary information to prepare and carry out an evaluation, already "processed" and summarized into easy-to-follow guides and online resources. Even if the methods or definitions proposed by the EAGLES norms or the FEMTI Framework might not always exactly match the evaluator's needs, they provide valuable guidance and help to save time and effort. Next Chapter (**VI. Evaluation Description**) goes deeper into this subject, describing how the EAGLES and the FEMTI were used to build the present evaluation and discussing the obstacles that the researcher came across with while using them.

[17] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=45086

## VI. Evaluation Description

One of the major difficulties researchers face prior to undertaking a research project is choosing the right method of data elicitation to achieve the purpose of their study. In order to achieve the objectives defined in **Chapter I. Introduction**, this study adopts a mixed methods approach, i.e. a combination of qualitative and quantitative methods (Saldanha and O'Brien 2014), in which qualitative data is expected to complement the assumptions drawn from the scores obtained (quantitative data). This Chapter describes the overall structure of the present evaluation: evaluation context (**6.1.**), selection of candidate systems and control systems (**6.2.**), description of the corpus used for the evaluation (**6.3.**), application of the EAGLES Seven Steps for Software Evaluation (see **Chapter V**) (**6.4.**), application of FEMTI (see **Chapter V**) (**6.5.**), description of general metrics (**6.6.**), and definition of quality characteristics, sub-characteristics and attributes proposed by FEMTI (**6.8.**).

### 6.1. Evaluation Context

In **Chapter I**, we saw that the evaluation was designed to serve a specific purpose in a real-life scenario, which constituted our case study: The International Social Security Association (ISSA). The ISSA was created in 1927 after a series of meetings organised by the International Labour Organisation (ILO). It was founded in response to the poor working conditions of European industrial workers in the aftermath of the First World War. Although it still keeps close bounds with the ILO, the ISSA is now an independent organisation, funded by its members[18]. The organisation has its headquarters in Geneva, Switzerland, and counts with 268 affiliate members and 72 associate members in 160 countries. Having members coming from different parts of the world, the ISSA works with multiple languages, including English, French, Spanish, German, Arabic, Russian, Japanese, and Chinese[19]. Its web portal (www.issa.int) was created in an effort to provide a common platform to facilitate communication and interaction with and between members and staff, addressing them in all eight languages.

[18] Retrieved from https://www.issa.int/the-issa/history/timeline. Last accessed August 11, 2015 (16:02)
[19] Retrieved from https://www.issa.int/the-issa/mandate and https://www.issa.int/the-issa, Last accessed August 11, 2015 (16:30)

Moreover, during the last triennium, the ISSA has seen important structural transformations: the introduction of the Guidelines for Social Security Administration, designed with the purpose of helping members to achieve higher service quality and a more efficient governance; and the creation of the ISSA Centre for Excellence, in charge of promoting said Guidelines[20]. This structural change resulted in a significant rise of the amount of material published by the Association and the reduction of time for translation and editing. Due to this limitation in time, a great amount of publications produced by the ISSA on a regular basis (reports, updates, pieces of news, calls for upcoming events, to mention some), remain untranslated in the web portal. Source texts are mainly written in English or French, and translations from those languages to English, French, Spanish and German are generally given priority. For these reasons, the ISSA is currently in search of a machine translation system to integrate into its web portal.

### 6.2. Choosing Candidate Systems and Control Systems

Following the positive experience of other organizations with regards to the incorporation of customizable statistical engines (see **Chapters III. Machine Translation Architectures**, and **IV. Machine Translation and Institutional Translation**), the Head of the Association's Member Services Promotion Branch (MSP)[21] requested that the candidate systems should meet that criteria. After analysing the Association's needs and resources, two candidates were chosen: Microsoft Translator Hub (**6.2.1.**) and TAPTA (**6.2.2.** See TAPTA4UN in **Chapter IV**).

In this study, the term "candidate" is used to denote the systems that will be compared in order to see which of them is more suitable for the ISSA. In addition, three other SMT systems were used during the evaluation to test some assumptions over the performance of customizable in comparison to generic ones. Those system are referred to as "control systems" (**6.2.3.**). The tests and their results are described in **Chapter VII.**

[20] Retrieved from https://www.issa.int/the-issa/centre-for-excellence. Last accessed August 11, 2015 (16:37)

[21] To learn more about the Association's structure see https://www.issa.int/the-issa/organigramme

### 6.2.1. Microsoft Translator Hub (MTH)

Over the years, Microsoft has developed a wide offer of offline and online software services. In 1995, when Microsoft released its new operating system, Windows 95, it took a big step into the World Wide Web, launching its online service: the Microsoft Network, which after became MSN. Within the services offered by MSN, users could benefit from weather forecast and news, e-mail services, and chat rooms, among others. During the second part of the decade, new services were introduced, such as MSN Games, Outlook Express, refreshed versions of Internet Explorer, and so on. Soon, Microsoft started creating partnerships with other online providers in order to offer services such as online shopping (MSN Shopping), encyclopaedias (Encarta), and news broadcasts (NBCNews, ESPN.com, etc.)[22]. In 1999, it released MSN Search, which later on became the widely popular search engine Bing. By the late 1990s, another powerful search engine was growing at fast-pace: Google, developed by Stanford PhD students Larry Page and Sergey Brin. In 1998, it was already the investors' favourite.[23] At that time, Microsoft and Google were said to be discussing the possibility of merging their services, but no decisions were reached. Microsoft and Google soon became major competitors, developing overlapping services to counter each other's position in the market[24]. Among these services was: Microsoft Translator[25]. Both Microsoft Translator and its major competitor, Google Translate, are generic web-based SMT engines. The particular characteristics of Google Translate are not be discussed in the context of the present evaluation because it is not one of the candidate systems. Nevertheless, more information on this system, as well as its similarities and differences with Microsoft Translator, can be accessed following the links presented in the footnotes.

---

[22] Retrieved from http://www.microsoft.com/misc/features/features_flshbk.htm, https://www.microsoft.com/fr-ch. Last accessed May 3, 2015 (11.00).

[23] Retrieved from http://arstechnica.com/uncategorized/2003/10/31/3050-2/. Last accessed May 3, 2015 (11:43).

[24] Retrieved from http://www.pcmag.com/article2/0,2817,1682902,00.asp, Last accessed May 3, 2015 (16:00)

[25] Retrieved from http://intellogist.wordpress.com/2012/01/04/google-translate-vs-bing-translator-part-1. Last accessed May 3, 2015 (16:05).

The MTH, released in 2012, is one of the language services offered by Microsoft and it is based on the Microsoft Translator technology[26]. It differs from previous services in that the Hub allows users to customize their translations using their own corpus of parallel or bilingual texts in addition to Microsoft's "language knowledge"[27]. By creating a Microsoft account, users can build their own SMT engine and incorporate it to their websites or apps via the Microsoft Translator API. By means of this API (application programming interface), users can translate their websites, or parts of it, by adding a code, generated by Microsoft, in their website's matrix. Users with a Windows Live ID, can get a Bing App ID that will facilitate the creation and storing of those codes, so they can be used whenever needed.[28] Moreover, MTH is connected to Microsoft Translator's Collaborative Translation Framework (CTF), a sort of public forum that provides proposals and corrections with the purpose of improving translation quality.[29]

In conclusion, this system was chosen as a candidate due to the language combinations it supports, the possibility of training the system for it to respect the organization's style and terminology, and the option of incorporating it into the ISSA webpage.

### 6.2.2. Translation Assistant for Patent Titles and Abstracts (TAPTA)

TAPTA stands for Translation Assistant for Patent Titles and Abstracts. It is a customizable SMT engine built with the open-source system Moses. It was developed by the WIPO (World Intellectual Property Organization) to facilitate users' access to patent and abstract information written in languages that they do not know, as well as to help translators working for the organization (Pouliquen, Mazenc and Iorio 2011). Users were offered this system as a way to "accelerate" the translation of patents into English and French (Elizalde et al 2013). In the paper "Tapta: A User-Driven Translation System for Patent Documents based on Domain-Aware Statistical Machine Translation", Pouliquen, Mazenc and Iorio (2011)

[26] Retrieved from http://webtrends.about.com/od/profi3/p/Microsoft-bio.htm and http://www.zdnet.com/microsofts-latest-machine-learning-poster-child-microsoft-translator-hub-7000000804/. Last accessed May 5, 2015 (10:45)
[27] Retrieved from https://hub.microsofttranslator.com. Last accessed May 5, 2015 (12:00)
[28] Retrieved from http://www.microsoft.com/web/post/using-the-free-bing-translation-apis. Last accessed May 5, 2015 (11:00)
[29] Ibid.

highlight that TAPTA differs from other systems in that it is a domain-aware SMT thanks to the use of factors. This means that the system is able to recognize different translations for a term according to the domain of the source text, by means of domain-tags attached to each word. (Pouliquen et al. 2011; Pouliquen, Mazenc and Iorio 2014; Pouliquen and Mazenc 2011). For more information on factored translation models, see section **3.2.2.1.2. Factored Translation Models**, or see Koehn & Hoang (2007) and Koehn (2010).

In 2011, the representative of Documentation Division of the UNHQ attended a presentation on TAPTA, in the framework of the Association for Information Management (ASLIB) Conference, and became interested in incorporating the system to the UNO's linguistic resources. This resulted in a joint project: the UNHQ provided the TAPTA team with a corpus containing 11 years of UN documentation (English<>Spanish) in the form of html bi-text, which was then used to train TAPTA4UN (Elizalde et al 2013). Some of the advantages pointed out by the UN staff were the system's capacity to match existing terminology, the possibility of integrating it to SDL Trados Studio by means of a plug-in, and the savings in translation and editing time and effort. The promising results of this prototype, as well as the possibility of adding information in the form of factors, soon called the attention of some other United Nations agencies, including the ILO (International Labour Organisation) and the ISSA. Although the successful incorporation of TAPTA by the UN was one of the arguments to choose this system as a candidate, it is worth to underline that the version evaluated in this study is not TAPTA4UN, but the "basic" form of the system trained with the ISSA corpus. In brief, this engine was chosen as a candidate due to its compatibility with institutional settings, the language combinations it supports, the possibility of eventually defining translation domains, its trainability and flexibility.

### 6.2.3. Control Systems

The candidates are both customizable SMT systems, as requested by the Association. The underlying idea for this request is that the systems that can be trained with a certain corpus are more likely to produce faithful and consistent translations. What is more, the possibility of training SMT system with a carefully cleaned corpus

brought about the idea that these systems can be used not only as "translation accelerators", but also as terminological tools within a particular organization (see **Chapter IV: Machine Translation and Institutional Translation**). In an attempt to verify whether the use of the Association's corpus of aligned translations does impact the quality of translation significantly, the output of two general SMT: Google Translate and Bing Translator, indicated simply as "MT 3" will be included in the tests designed to evaluate the candidate systems' translations. In one of the tests (Terminological Test, described below in **Chapter VII. Running the Evaluation**), the output of TAPTA4UN (customizable open source systems, trained with the corpus of the UN, see **Chapter IV.**) will be used to test how much interference does a different corpus cause to the translation of terms. This is explained in detail when describing each test in Chapter VII.

### 6.3. Corpora Description

For the purpose of this evaluation, two corpora were built: one for training the systems and another for testing the SMT systems translations. Each corpus respond to different characteristics and will be described in two different sub-sections **6.3.1.** And **6.3.2.** These sub-sections will also extend on the reasons for using different corpora for training and testing the systems' performance: (1) protecting the privacy of unpublished publications and (2) testing the systems' capacity to treat a variety of texts similar, but not identical, to the ones used to train them.

#### 6.3.1. Corpus for Training

The process of building the corpus for training was simplified by the fact that the Association already had a base of aligned texts for different pairs of languages. None the less, there were some concerns about the total size of the ISSA corpus: on the one hand, the size varied greatly depending on the language pairs; on the other hand, even the largest corpus was rather small compared to the ones normally used to train SMT systems. Adding texts from other sources (not belonging to the ISSA) is not a suitable solution, since it would affect the SMT engine's translation choices. In the end, it was decided to conduct the assessment with the resources available, with the prospect that the Association would manage to increase the size of the corpus in the near future. The systems' capacity to improve their results by increasing the size

of its corpus, as well as their flexibility to be re-trained is discussed below in Chapter **VII**.

The texts were exported in XML format and placed into folders according to their language combinations. Due to certain limitations of the present study (see **Chapter VIII**), only the English-Spanish combination was tested. Table 2 presents the total extracted sentences and words for that language pair.

| Languages | Sentences | Words |
|---|---|---|
| English <> Spanish | 30674 | 696714 |

Table 2: Corpus for Training

### 6.3.2. Corpus for Testing

This corpus is formed by three publications extracted from the ISSA web portal, which, at the time (July 2014, see Chapter I.), were not aligned and stored in the Association's corpus. At this point, it is important to highlight that the web portal contains some publications for public access and other publications with restricted access (only for members and staff). The texts for public access are mainly informative, descriptive, and or persuasive; semi-specialized, and mostly targeting ISSA members or potential members. They include news about the Association's programmes and projects, reports on the state of social security in the world, and news about the member organizations. Table 3 shows the main characteristics of the publications used to build the corpus. Due to the limitation of time and resources of the present study (see **Chapter VIII**), random extracts of the publications were selected and incorporated into the tests. Since those samples were meant to be distributed to voluntary participants outside the Association (the participants profile is Chapter VII), this corpus only contains public texts. Table 3 presents a brief description of the corpus main characteristics.

| Title | Type of Text | Words | Source |
|---|---|---|---|
| Unemployment insurance systems and youth employment policies | Report: Informative & descriptive. Semi-specialized | 10.000 (Cleaned text: without index, references, etc.) | Corine MAEYA. From the National Employment Office (Belgium). Available at https://www.issa.int/all-conference-reports |
| How to Meet the Needs Arising from Sociological Changes in the Family? | Report: Informative & descriptive. Semi-specialized | 5.000 (Cleaned text: without index, references, etc.) | Technical Commission on Family Benefits. Summary of findings 2008-2010 Available at https://www.issa.int/all-conference-reports |
| Information note: Centre for Excellence. | Information note: Informative Semi-specialized | 4.000 (Cleaned text: without index, references, etc.) | 107th Bureau Meeting Geneva, 27 June 2013. ISSA Bureau on the Centre for Excellence. |

Table 3: Description of Corpus for Testing

### 6.4. The EAGLES Seven Steps for Software Evaluation

Following the first steps defined by the EAGLES, this section presents the evaluation's purpose and object, and the task model.

**Purpose of the Evaluation**:

Helping the ISSA to make an informed decision over which SMT engine will better suit their current needs and limitations.

**Object of the Evaluation**:

The object of software evaluation can be either the system treated as a whole (e.g. the evaluation of a text processor), a particular function of the system considered in isolation (e.g. the text editor's grammar correction engine) or as a part of a more complex process (e.g. the use of said editor in the editing of a translation, as part of the workflow of a freelance translator). In this case, the candidate systems (see sections **6.2.1** and **6.2.2**) are evaluated as a whole. The output of the control systems is only used to test some hypothesis on the quality of customizable vs. generic systems (see section **6.2.3**)

**User Description**:

Users are classified as direct and indirect users:

- **Direct Users**: ISSA staff in charge of training, incorporating and maintaining the chosen system: two IT assistants and the staff in charge of the management of the ISSA web portal. High computer literacy, no background knowledge in the fields of linguistics or NLP.
- **Indirect Users**: affiliate and associate members from around the world; ISSA general staff. Indirect users will benefit from the language service provided by the chosen SMT system, but they are not involved in the processes of training and maintaining it. Due to the great amount of indirect users and the impossibility of assessing their particular background IT or linguistic knowledge, they will be considered as having a minimum IT background and none or limited knowledge of the SL.

**Purpose of the SMT System**:

In a first stage, the aim is to offer to the ISSA members and staff a free SMT system, accessible through the ISSA web portal, for them to carry out quick short translations that help them to grasp the meanings of texts that are not translated on the website. Additionally, it is also expected that the ISSA staff will be able to use it for internal communication, as well as for other administrative tasks. By no means, it is expected to replace the language services provided by human translators and revisers. Nevertheless, the system will be offered to them as an additional tool for translation support.

**SMT System's Requirements**:

Table 4 summarizes the general requirements that need to be met by the candidate systems, as requested by the Association:

| | |
|---|---|
| **Resource utilisation** | SMT system: there is no specialized staff in the Association to develop and maintain a RBMT. |
| **Languages** | It must to support all the relevant languages. |
| **Terminological compliance** | It must be customizable, so it can be trained to respect the ISSA's terminology. |
| **Acceptable output** | The output is expected to properly transmit the gist of a text in the SL (*input*). See the purpose of the described in point **b.,** above. |
| **User friendly** | Easy to use, small learning curve, availability of technical assistance. |

| | |
|---|---|
| **Flexibility** | Capacity to be adapted to changes in the organization's needs. |
| **Cost** | The administration was interested in choosing the system with the best cost-benefit relation. |

Table 4: Summary of System Requirements (requested by the Association prior to the beginning of the evaluation)

In order to complete the evaluation design (selection of QC, sub-characteristics and attributes, description of methods and metrics, etc.), we relied on the FEMTI online tool described in section **5.2.1.**

### 6.5. Application of FEMTI

In Chapter V. (**Software Evaluation and Machine Translation System Evaluation**), it was explained that a software can be assessed on its external quality (the output) or its internal and in-context quality (software product). With the aim of considering all the requirements established by the ISSA, it was decided to test both aspects. Going back to Fig. 11: FEMTI Design Environment (section **5.5.1.**), we remember that the FEMTI interface is divided in two screens: one for choosing what we want to test in our software and a another one where FEMTI proposes the quality characteristics (together with a number of sub-characteristics, attributes and metrics does) that will help us obtain the information we are looking for. Section **6.5.1** focuses on the first screen, describing how the FEMTI online tool was set; section **6.5.2** focuses on the second screen and the final choice of quality characteristics, sub-characteristics and attributes.

#### 6.5.1. FEMTI Settings

**Main Value: 1. Evaluation Requirements**

**1.1. Purpose of the Evaluation**

From the seven options listed by FEMTI, two matched the type of evaluation we need to carry out:

- **Declarative Evaluation** ("external", concerned with linguistic phenomena)[30]

[30] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-105.html. Last accessed July 5, 2015 (11:59)

- **Operational Evaluation** ("internal", concerned with operational use)[31].

The system was run two times, one time for the declarative evaluation and a second time for the operation evaluation, the rest of the settings vary slightly:

### 1.2. Characteristics of the Translation Task

"Information flow intended for the output, from the point of view of the agent (human or otherwise) who receives the translation".[32] None of the three options listed (Assimilation, Dissemination and Communication) matched exactly our evaluation's purpose (see **6.4.**, point **b)**:

- **Communication**[33] > **Asynchronous Communication:** Although the communication task described in FEMTI focuses on oral communication, and the candidate systems are meant to assist in written communication, the relevant characteristics listed (intelligibility and comprehensibility) are adequate to evaluate if the systems meet the general requirements presented in Table 2.

This feature is relevant for both evaluations, and was set in the same way for both.

### 1.3. Input Characteristics (Author and Text)

This setting lists the characteristics of the source text (style of the writer, the genre of the text, the domain of specialisation, etc.)[34] Two options were selected:

- **Document Type** > **Genre** (form, style specific to a type of document) and **Domain** (field of specialization, e.g. scientific, medical, etc.).

[31] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-106.html. Last accessed July 5, 2015 (12:01)
[32] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-112.html. Last accessed July 5, 2015 (12:05)
[33] For a detailed definition of this option and the proposed relevant quality characteristic, see http://www.issco.unige.ch:8080/cocoon/femti/taxum-123.html
[34] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-310.html. Last accessed July 6, 2015 (09:23)

- **Author's characteristics** > **Proficiency in Source Language**: Advanced (since most of the ISSA articles are written by non-native speakers with varying degrees of proficiency, "advanced" was preferred over "superior"[35].

This feature is relevant for the declarative evaluation.

### 1.4. User Characteristics

None of three options listed (machine translation user, translation consumer and organisational user) match exactly the type of users of our evaluations (see **6.4.**).

For the declarative evaluation, this feature was set as follows:

- **Machine Translation User** > **Linguistic Education** >
  - **Proficiency in Source Language**: Novice.
  - **Proficiency in Target Language**: Advanced and Superior ("native" is not included among the options).
  - **Computer Literacy** (this setting does not offer different levels of literacy to select).

For the operational evaluation, this feature was set as follows:

- **Organisational User** > **Quantity of Translation**, **Number of Personnel** and **Time Allowed for Translation**.[36] (The option "organizational user" did not offer as many possible settings as "machine translation user", but the relevant quality characteristics proposed by FEMTI seemed to match the requirements for the evaluation of the SMT systems as software products.)

After analysing the data introduced, FEMTI suggested the following QC: functionality, usability, efficiency, maintainability and portability. The corresponding sub-characteristics and attributes will be presented further below in section **6.7.** To access the full report produced by FEMTI, please refer to Annex 1 and 2.

From the previous experience of designing a software evaluation using the web resource FEMTI, we can conclude that a tool like FEMTI is highly useful, as it helps

[35] For a detailed definition of this option and the proposed relevant quality characteristic, see http://www.issco.unige.ch:8080/cocoon/femti/taxum-138.html
[36] For a detailed definition of this option and the proposed relevant quality characteristic, see http://www.issco.unige.ch:8080/cocoon/femti/taxum-709.html

evaluators to organize a great amount of information related to the system's characteristics, the user's requirements, the context description, and so on, into a practical comprehensive framework. Moreover, the automatic suggestion of quality characteristics, with its sub-characteristics, metrics and evaluation methods, allows evaluators to save plenty of time, especially to those with few or none experience in the specific field of MT system evaluation. However, as usually happens with this type of frameworks, the categories provided by the system do not always match the evaluation's requirements. In addition, the final report does not keep the hierarchical order of the online tool, which makes it difficult to read.

### 6.5.2. Quality Characteristics, Sub-Characteristics and Attributes

After analysing the requirements introduced in the first screen, FEMTI propose a number of quality characteristics, sub-characteristics and attributes, as well as different tests to measure them. As we discussed in the last section, some of the options enumerated by FEMTI did not exactly match our evaluation's purposes, and consequently, some of the characteristics proposed were not relevant for the present study. Therefore, the options displayed in the second screen were analysed carefully, and a number of relevant characteristics were selected. Finally, two reports were produced by the online tool: one for each evaluation. This section presents the QCs that will be tested for the declarative evaluation (**6.5.2.1.**) and for the operational evaluation (**6.5.2.2.**). To access the full repots (which also contains the characteristics that were not selected), see Annexes 1 and 2.

#### 6.5.2.1. Declarative Evaluation (Output)

According to FEMTI, the purpose of the declarative evaluation is to measure the capacity of a MT engine to treat texts representative of a specific end-user.[37] Testing the system's performance while translating representative texts means that the corpus for testing needs to be built of extracts or samples of real texts (see section **6.3. Corpora Description**). This section presents the quality characteristic tested

[37] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-105.html. Last accessed August 5, 2015 (10:32)

to assess the quality of the output: functionality (**6.5.2.1.1.**), together with its sub-characteristics (suitability and accuracy).

#### 6.5.2.1.1. Functionality

ISO defines functionality as a "set of attributes that bear on the existence of a set of functions and their properties" (ISO 9126: 1991, 4.1)[38]. The sub-characteristics selected were the following:

**Accuracy**:

"The capability of the software product to provide the right or agreed results or effects with the needed degree of precision" (ISO 9126: 2001, 6.1.2, quoted in FEMTI)

- **Fidelity Precision** (correctness of the information transferred from ST to TT)[39], and **Terminology** (correct translation of terms)[40]

**Suitability**:

"The capability of the software product to provide an appropriate set of functions for specified tasks and user objectives" (ISO 9126: 2001, 6.1.1, quoted in FEMTI)

- **Readability**: Fluency (the text's naturalness in the target TL) and Adequacy (the extent to which the meaning of the SL text has been transferred into the TL text). Readability was defined following FEMTI[41], and Koehn (2010).

**Well-Formedness**:

Appropriate form in terms of grammar and syntax.

- **Grammatical Correctness**

Functionality is also important in terms of *interoperability* and *security*; however, these two sub-characteristics do not measure the output, but the SMT system as a software product, so they will be discussed in **section 6.5.2.2. Operability Evaluation**.

[38] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-601.html. Last accessed August 5, 2015 (15:54)
[39] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-179.html. Last accessed August 5, 2015 (18:00)
[40] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-175.html. Last accessed August 5, 2015 (18:05)
[41] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-172.html. Last accessed August 6, 2015 (10:30)

#### 6.5.2.2. Operability Evaluation

According to FEMTI, the purpose of the Operability Evaluation is to assess if a given MT system will actually serve its purpose in the context of use [42](see section **6.4.** The EAGLES Seven Steps for Software Evaluation). This section presents the quality characteristics, sub-characteristics and attributes tested to evaluate which of the candidate systems would serve best the purpose of the ISSA's context of use: Functionality (**6.5.2.2.1.**), Usability (**6.5.2.2.2.**), Efficiency (**6.5.2.2.3.**), Maintainability (**6.5.2.2.4.**), and Portability (**6.5.2.2.5.**)

##### 6.5.2.2.1. Functionality

Defined above in **6.5.2.1.1.**

**Security**:
"The capability of the software product to protect information and data so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them." (ISO 9126: 2001, 6.1.4., quoted in FEMTI)

**Interoperability**:
"The capability of the software product to interact with one or more specified systems." (ISO 9126: 2001, 6.1.3., quoted in FEMTI).

##### 6.5.2.2.2. Usability

According to ISO 9126 (2001, 6.3., quoted in FEMTI) usability is "[t]he capability of the software product to be understood, learned, used and attractive to the use, when used under specified conditions."[43] In the paper *Usability Evaluation Based on International Standards for Software Quality Evaluation,* the author emphasizes the importance of evaluating the system's capability in its context of use, which helps

[42] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-106.html. Last accessed August 5, 2015 (10:32).
[43] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-603.html. Last accessed August 5, 2015 (13:26)

determine the relevance of the different sub-attributes for a particular evaluation (Toshihiro 2008).

**Understandability**:

This measures the intuitiveness of the system's interface: to what degree can the user understand how to use it for particular tasks?

**Learnability**:

This measures the availability of reference material that will help the user understand how the system works (manuals, forums, video tutorials, etc.).

#### 6.5.2.2.3. Efficiency

ISO defines efficiency as "the capability of a software product to provide appropriate performance, relative to the amount of resources used, under stated conditions." (ISO 9126: 2001, 6.4., quoted in FEMTI)[44].

**Cost**:

ISO does not include cost as a sub-characteristic, since it is considered to belong to management decision-making. FEMTI defines it as an independent quality characteristic. However, it can be argued that, since it reveals a relationship between the software capacity to carry out a task and the amount of resources it consumes in order to do so, it is closely linked to resource utilisation. Consequently, in the case of the present evaluation, cost will be tested as a sub-characteristic within efficiency.

**Time behaviour**:

"The capability of the software product to provide appropriate response and processing time and throughput rates when performing its function under stated conditions."[45] This sub-characteristic will be analysed in terms of the time it takes to set up the system ready for use (installation and training), the time required for retraining the system (update time), and the time to carry out a translation (translation speed). The possibility of retraining the software regularly (approximately once every two months) is essential for the system to match the

[44] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-606.html Last accessed August 6, 2015 (15:56)
[45] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-206.html. Last accessed August 6, 2015 (16:00).

Association's needs, as it is in the middle of an important process of terminological updating. Consequently, it will be assigned a special weight. Given that the time that an SMT system takes to be trained or retrained highly depends on the size and characteristics of the corpus, as well as the characteristics of the computer, the assessment of this sub attribute will be carried out in the same computers and with the same corpus (see **6.3. Corpora Description**).

#### 6.5.2.2.4. Maintainability

ISO defines maintainability as "a set of attributes that bear on the effort needed to make specific modifications", and further clarifies that those modifications can include "corrections, improvements or adaptation of software to changes in environment, and in requirements and functional specifications." (ISO 9126: 1991, 4.5., quoted in FEMTI) This last clarification is important as it is to be noticed that the system's potential for change is not limited to updates made available by the distributor, but it also includes direct users' possibility to customize the software according to its changing needs. In this case, the evaluation will focus on the following sub-characteristics. The definitions were retrieved from FEMTI's web page.[46]

**Changeability**:

"The capability of the software product to enable a specified modification to be implemented"

**Stability**:

"The system's capability to avoid unexpected effects from modifications of the software".

#### 6.5.2.2.5. Portability

"The capability of the software product to be transferred from one environment to another." (ISO 9126: 2001, 6.6., quoted in FEMTI). In this case, the sub-characteristic

[46] Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-620.html. Last accessed August 6, 2015 (17:39).

*portability* will be tested under the QC portability because it only affects direct users (see **6.4.**)(ISO 9126: 2001, 6.6.2, quoted in FEMTI).

**Installability**:

"The system's capability to be installed in different environments"[47]. Although most computers in the Association run on Microsoft Windows; the ISSA also counts with computers with virtual boxes, where different operating systems (OP) can be installed. The Association technicians, for example, work on computers that have access to MS Windows and Linux.

### 6.5.3. Summary

Table 5 summarizes the quality characteristics, sub-characteristics and attributes to be evaluated, and introduces the metrics and Weights (see next section) that will be applied to measure each of them (described in detailed in forthcoming sections). Next Chapter (**VII. Running the Evaluation**), describes each test in detail, as well as their execution and results.

| QC | SC | Attribute | Metrics |
|---|---|---|---|
| **Declarative Evaluation** | | | |
| Functionality | Accuracy | Fidelity - Precision | BLEU Score |
| | | Terminology | Percentage of Terms correctly translated. |
| | Suitability | Readability | Subjective Rating of Fluency and Adequacy[48] |
| | Well-Formedness | Grammar - Syntax | Comparative Test: Subjective Rating of Correctness |
| **Operational Evaluation** | | | |
| Functionality | Interoperability | | Boolean Questionnaire |
| | Security | | Boolean Questionnaire |
| Usability | Learnability | | Boolean Multiple Choice |
| | Understandability | | Boolean Multiple Choice |
| Efficiency | Cost | | Boolean Questionnaire. |

[47]Retrieved from http://www.issco.unige.ch:8080/cocoon/femti/taxum-222.html. Last accessed August 4, 2015 (15:18)

[48] Although FEMTI proposed the attribute "comprehensibility", readability is used because evaluators analyse the translation segment by segment (Annex 1)

| | | | |
|---|---|---|---|
| | Time behaviour | Preparation Time | Boolean Questionnaire. |
| | | Translation Time | Record of time each system takes to translate (scale). |
| | | Update time | Record of time each system takes to be trained (scale). |
| Maintainability | Changeability | | Boolean questionnaire |
| | Stability | | Boolean questionnaire |
| Portability | Installability | | Boolean questionnaire |

Table 5: Summary of QC, Sub-Characteristics and Metrics

### 6.6. Measurement Method: General Description

One of the major difficulties of formal or scientific evaluations is to develop a measurement method able to capture the evaluator's judgement of the object of study, while reflecting the qualities of the object in an objective way. This means that, on the one hand, there is the object to be evaluated (in this case, two SMT systems); and on the other, the evaluator(s) in charge of assessing its quality, according to a series of requirements (context of use, including user's needs, etc.). Evaluators are inevitably, influenced by external factors such as deadlines or previous experiences. In this context, a well-defined measurement method should act as an intermediary between the two, reducing subjectivity and improving precision and validity.

According to the *EAGLES Seven Steps Recipe* (1999), after selecting the relevant quality characteristics, sub-characteristics and attributes (see Table 4), it is necessary to devise appropriate metrics to measure each of them. This includes: (a) defining a score scale for each attribute; (b) describing the way in which the values of the different attributes will combine to form the node value, (c) and the way of reflecting their relative importance (EAGLES 1999).

(a) In the present evaluation, two types of metrics are used: Boolean metrics, and scale metrics.

(b) To facilitate the computing of final scores, the questions of the Boolean tests were formulated in a way that "yes" is always the positive answer, giving the system one point, and "no", the negative answer, giving the system zero points. Multiple choice questions were designed with different scales, in most

cases starting from the lower value (0 pts), and moving up to the highest value (4 pts).

(c) The relative importance of attributes is reflected by the assignment of different Weights (W). Those W were defined in a scale from 1 to 4, according to the ISSA requirements (see section **6.4.**, Table 3). No attribute carries a W 0, since all of them are relevant to the evaluation. W 1 (Equal Importance) indicates that an attribute does not carry a special significance; W 2 (Slightly Higher Importance) and W 3 (Higher Importance) indicate that certain attributes were considered to be more significant than others, since the results of their quality would affect the decision making process in a greater measure. Finally, W 4 (Most important) indicates when an attribute stands out as fundamental for the ISSA.

Table 6 illustrates the elements listed above. Next Chapter (**VII. Running the Evaluation**) revisit these metrics and W, describing how they were used in each test.

| Metrics | | Weight |
|---|---|---|
| Binary | Yes= 1<br>No=0 | 4: Most important<br>3: Higher Importance<br>2: Slightly Higher Importance<br>1: Equal importance |
| Scale | Subjective assignment of points by human evaluators. | |

Table 6: Metrics General Definition

In brief, the quality of the system (Total Q) is equal to the sum of the total value of each of its quality characteristics (QC), which in turn, are equal to the sum of each attributes' values multiplied by their weight (W). Figure 5 illustrates the formulae:

$$QC1 = (A1.W) + (A2.W)$$

$$QC2 = (A1.W) + (A2.W)$$

$$\textbf{Total Q} = \textbf{QC1} + \textbf{QC2}$$

Fig. 12: General Metrics Formulae

# VII. Running the Evaluation

This Chapter is devoted to the description of the tests carried out during the Declarative (**7.1.**) and the Operational Evaluations (**7.2.**), as well as to the presentation of their partial results. These tests will be described in the same order in which they appear in **Chapter VI. Evaluation Description** (see Table 5). The quality characteristics and their attributes are presented in different sub-sections, each of them with summary of the purpose of the test, the method used to carry it out, the corpus used and the number of participants (description of corpus and participants only applies to the evaluation of the output, section **7.1. Declarative Evaluation**), the assigned W (see Table 6, section **6.6.**), as well as the test description and the results. The final results and conclusions are presented in **Chapter VIII**.

## 7.1. Declarative Evaluation

Both automatic and manual metrics are applied to carry out the Declarative Evaluation: automatic metrics are applied to test the attribute "Fidelity Precision" (**7.1.1.1.**), while manual metrics are used to Accuracy, Suitability and Well-Formedness (**7.1.1.2.)** Manual evaluations require the participation of a number of monolingual or bilingual judges: in this case, seven voluntary translators responding to a specific profile described in section **7.1.1.2.1**.

### 7.1.1. Functionality

This QC is tested in terms of three sub-characteristics: accuracy, suitability and well-formedness, with the intention of analysing the capacity of each candidate system to generate translations that are faithful to the original text (fidelity precision test, described in **7.1.1.1.**), correct and precise in their lexical choices (terminology test, described further on in **7.1.1.2.2.**) and that, to a certain extent, read naturally in the target language, reproducing the meaning of the original text (readability test: fluency and adequacy, **7.1.2.3.**).

#### 7.1.1.1. Automatic Evaluation: Fidelity Precision.

- **Purpose**: **Test 1**: Diagnose the system's quality after the first training. **Test 2**: corroborate the results of Test 1
- **Method**: BLEU Score (see section **5.1.2. Automatic Evaluation**)
- **Corpus**: **Test 1**: Corpus for Training; **Test 2**: Extract from Corpus for Testing.
- **Participants**: no voluntary participants.
- **Control Systems**: **Test 1**: no control system. **Test 2**: Google Translate, generic SMT.
- **W**: 1

**Test Description:**

**Test 1**:

Since the Translator Hub offers clients the possibility of using their own corpora to customize the system, the performance of MTH depends on the size and quality of each of the users' corpora. In order to show clients how well their personalized SMT engine will work, MTH carries out an automatic measurement of the system using the BLEU score and displays the results together with other information regarding the training.

In this case, the system, named **Trial_ISSA_AT [ENES],** obtains a BLEU Score of **37.39**. Since it is difficult to interpret whether that score is acceptable or not, MTH displays this information in green letters, preceded by a positive mark, in the way illustrated by Figure 13:

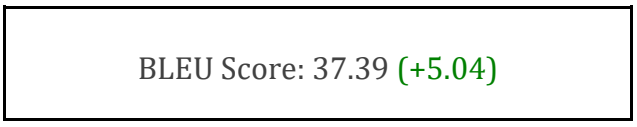



Fig. 13: MTH BLEU Score Display

If the user runs the cursor over the green numbers, the following explanatory message appears: "The score of this system with your test data is higher than Microsoft's general domain system." In this way, there seems to be no doubt as to the interpretation of the automatic measurement. This said, it is important to underline that this feature of MTH, as user-friendly as it seems, can result slightly misleading. First, because users who lack knowledge of BLEU are not able to judge whether the interpretation given to them by MTH is accurate or not; and second,

because comparing the performance against an unknown standard ("Microsoft's general domain system") does not offer much information.

TAPTA does not perform the evaluation together with the training in the same way as MTH does. In the case of TAPTA, now named **TAPTA4ISSA**, the results were only slightly higher than those obtained by MTH, **37.69**.

| | |
|---|---|
| MTH | 37.39 |
| TAPTA | 37.69 |

Table 7: BLEU Score Results

At that point of the evaluation, it was pointed out by TAPTA developers that the organization's corpus was rather small, and that, by adding a few sentences from the European Parliament Corpus (http://www.statmt.org/europarl/) —as an **additional test**— the results improved significantly: 38.62, instead of 36.67, for the language combination English>French. This additional test hinted at the system's improvement capacity. However, only the score from the first test are taken into account for the final score of this evaluation.

**Test 2**: It was executed using the online tool Asiya[49]. Figure 14 shows the results obtained by the candidates and the control system: (1) MTH, (2) TAPTA and (3) Google Translate (GT).


Fig. 14: BLEU Score Results. Second Test executed in Asiya

[49] To access Asiya, follow this link: http://asiya.lsi.upc.edu/demo/asiya_online.php

**Comments and Results**:

The result of this test is similar to the one of the first test: TAPTA's scores are just slightly higher than MTH's. The score for Google Translate is the highest of the three, but it does not differ greatly from the others.

A closer analysis of the sentences showed that, at segment level, the variation of BLEU scores did not reflected a true variation of quality. E.g. consider the first sentence and their scores: the three translations are very similar to the reference, and all of them are correct in Spanish, but the BLEU score for each of them varies significantly. All in all, it was decided to give **1 point to each systems**.

| |
|---|
| **Source:** *This report describes the main findings of the projects carried out by the Technical Commission during the triennium 2008-2010.* |
| **Ref**.: ***El presente informe*** *describe los principales* ***resultados*** *de los proyectos* ***llevados a cabo*** *por la Comisión Técnica durante el trienio 2008-2010* |
| **GT [BLEU: 0.91]**: ***Este informe*** *describe los principales* ***resultados*** *de los proyectos* ***llevados a cabo*** *por la Comisión Técnica durante el trienio 2008-2010* |
| **MTH [BLEU: 0.58]**: ***Este informe*** *describe los principales* ***hallazgos*** *de los proyectos* ***realizados*** *por la Comisión técnica durante el trienio 2008-2010.* |
| **TAPTA4ISSA [0.34]**: ***Este informe*** *describe los* ***hallazgos*** *principales de los proyectos* ***realizados*** *por la comisión técnico durante el trienio 2008-2010.* |

### 7.1.1.2. Manual Evaluation: Accuracy, Suitability, and Well-Formedness

Different manual evaluations were designed to test accuracy, suitability, and well-formedness. In total, seven voluntary evaluators participated in the tests (participants responding to the profile described below in section **7.1.1.2.1.**) Some of them participated in two tests, but most of them only in one. The participants' general information (years of experience, computer literacy, etc.) questionnaires are available in Annex 5. The terminological and the readability tests were designed using the online tool Google Forms[50], due to its user-friendly interface and its variety of facilities: automatic generation of answer sheet, possibility to restrict answers per row or column, variety of options to send the test to participants, to mention some.

[50] To access the home page of Google Forms, follow this link https://www.google.com/forms/about/

The well-formedness test was designed using MS Excel. All questionnaires are available in Annex 3.

#### 7.1.1.2.1. Manual Evaluators: Profile

Accuracy, Suitability and Well-Formedness are assessed manually by human evaluators matching the profile presented below:

*Professional translators and/or editors with at least four years of higher education in the field of translation and languages, and experience in the field of institutional translation at an international level. Their working languages must be English and their native language Spanish.*

To participate in the Terminological Test, the experience in institutional settings was not required, as participants were asked to compare the SMT system's outputs against a unique possible "correct" translation offered as a reference. It is worth to notice that no experience in the field of automatic translation or NLP was required. What is more, it was preferred that the participants did not belong to these fields, since it was interesting to elicit some data on the attitude of translators towards MT. In the end, the fact that none of the seven participants had experience in this area, resulted in some interesting findings about the MT and institutional translators (**Chapter VIII**) Nevertheless, the limitations of these findings due to the reduced number of participants is discussed in section **8.4.**

#### 7.1.1.2.2. Accuracy: Terminological Test

- **Purpose:** Evaluating the systems' accuracy when translating isolated terms.
- **Method**: Manual Test: Subjective Scale (see Fig.)
- **Corpus**: Corpus for Testing
- **Participants**: 3
- **Control Systems**: Bing (generic SMT).
- **W**: 3 (This is relevant to the present study because (1) one of the applications of MT in the institutional context is serving as a terminological tool; (2) one of the advantages of customized SMT engines is supposed to be its capability to

adjust to the organisations' terminology by reproducing the forms found in their corpora.)

**Test Description:**

Figure 15 presents the first page of the Terminology Test questionnaire: the different elements of the test are marked with colours and numbers: (1) in green, we can see the instructions for the test; (2) in blue, the English term; (3) in pink, the reference (Spanish translation provided by the ISSA Glossary); (4) in black, the scale: terms can be rated as "Incorrect" (0 points), "Acceptable" (1 point) or "Correct" (2 points); (5) in yellow, the systems' translations (note that participants do not know which systems are being tested).

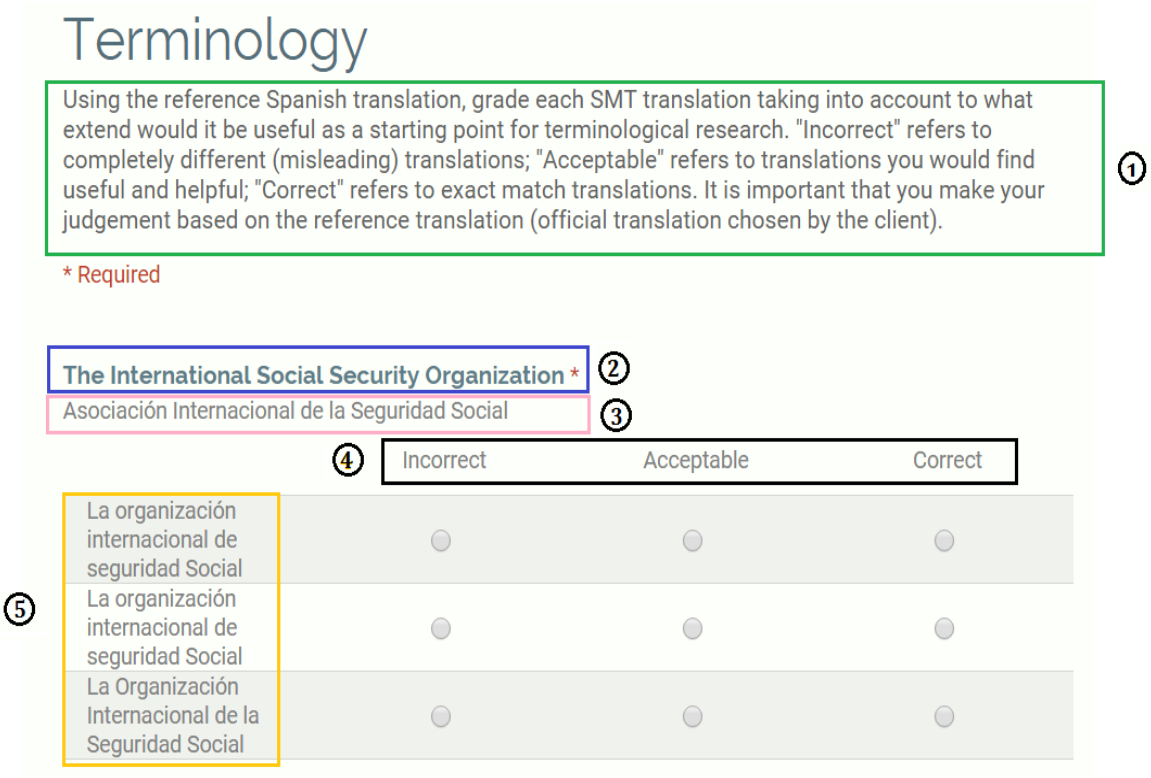



Fig.15: Terminological Test (Google Forms)51

**Comments**:

Participants' responses are recorded in a separate sheet, automatically generated by the tool, and then transferred into a new table, where each response is introduced in the form of the corresponding point, facilitating the computing of final scores (to access the Response Tables, see Annex 4). Figure 16 presents the Response Table

[51] To access the full test, follow this link: http://goo.gl/forms/xTsjfqIQ0U or see Annex 3.

that summarizes the points of one participant: (1) in light blue, we see the number of tested terms (sixteen in total); (2) in yellow, the candidate systems and the control system; and (3) in orange, the final scores of the participant.

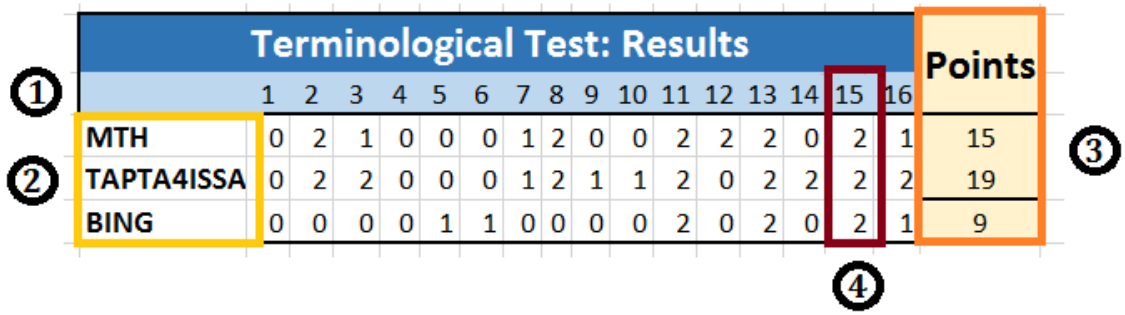



Fig.16: Terminological Test: Model of Response Table.

The last point in figure 16 (4) correspond to the points assigned by the participant to the term "occupational risks". As indicated in the instructions (Fig. 15, point 1) evaluators do not have to rate translations according to their own preferences (the way they would have translated the term), but only in accordance to the reference. In the case of term 15, the reference translation is "*riesgos ocupacionales*". The output of MTH and Bing "*riesgos laborales*" is correct and natural, but it is not the term used in the ISSA Glossary. Nevertheless, this evaluator graded those translations as "correct". This particular answer exemplifies one of the most common difficulties translators come across when using MT in institutional contexts: discriminating fluent output from strictly "correct" output, responding to the organization's terminology and style (see **Chapter IV. Machine Translation and Institutional Translation**).

Regarding the control system, its results were significantly lower in the case of the three participants.

**Results:**

Table 8 summarizes the points assigned by each evaluator and the total score (without **W**).

| SYSTEM | EVALUATOR 1 | EVALUATOR 2 | EVALUATOR 3 | TOTAL |
|---|---|---|---|---|
| **MTH** | **15** | **14** | **20** | **16.3** |
| **TAPTA4ISSA** | **19** | **25** | **22** | **22.0** |
| **BING** | **9** | **9** | **11** | **09.6** |

Table 8: Terminological Test: Summary of Points

#### 7.1.1.2.3. Suitability: Readability Test

- **Purpose:** Evaluate the quality of translations in terms of *fluency* and *adequacy*.
- **Method**: Manual Test: Subjective Scale. Ten phrases (see Fig. 16 below)
- **Corpus**: Corpus for Testing
- **Participants**: 3
- **Control Systems**: Google Translate (generic SMT, not trained) and TAPTA4UN (customized SMT, trained with a different corpus).
- **W**: 3

**Test Description**:

No reference translation was provided in this test, for the reasons explained in **5.5.1. Manual Evaluation:** reference translations can affect scores by giving judges a gist of the sentence's meaning before they have even read the SMT output. Figure 17 presents the scales used for the test, based on Koehn (2010, 219).

| Scale | | |
|---|---|---|
| Fluency | | Adequacy |
| Flawless Spanish | 3 | All Meaning |
| Good Spanish | 2 | Most Meaning |
| Understandable | 1 | Little Meaning |
| Indistinguishable | 0 | None |

Fig. 17: Readability Test: Adequacy and Fluency Scale (Based on Koehn 2010, 219)

Figure 18 presents the first page of the questionnaires: we can see (1) in green, the instructions provided to the evaluators; (2) in blue, the attribute and the number of phrase; (3) in pink, the English sentence; (4) in black, the scale as seen by the evaluators; and (5) the systems' translations (note that the evaluators do not know

which systems they are grading). In total, ten English phrases with its four translations (one for each system), were tested. The first two translations are those produced by the candidates, and the last two are those generated by the control systems. To access the full test, follow this link: http://goo.gl/forms/zhX4cedVPc or see Annex 3.

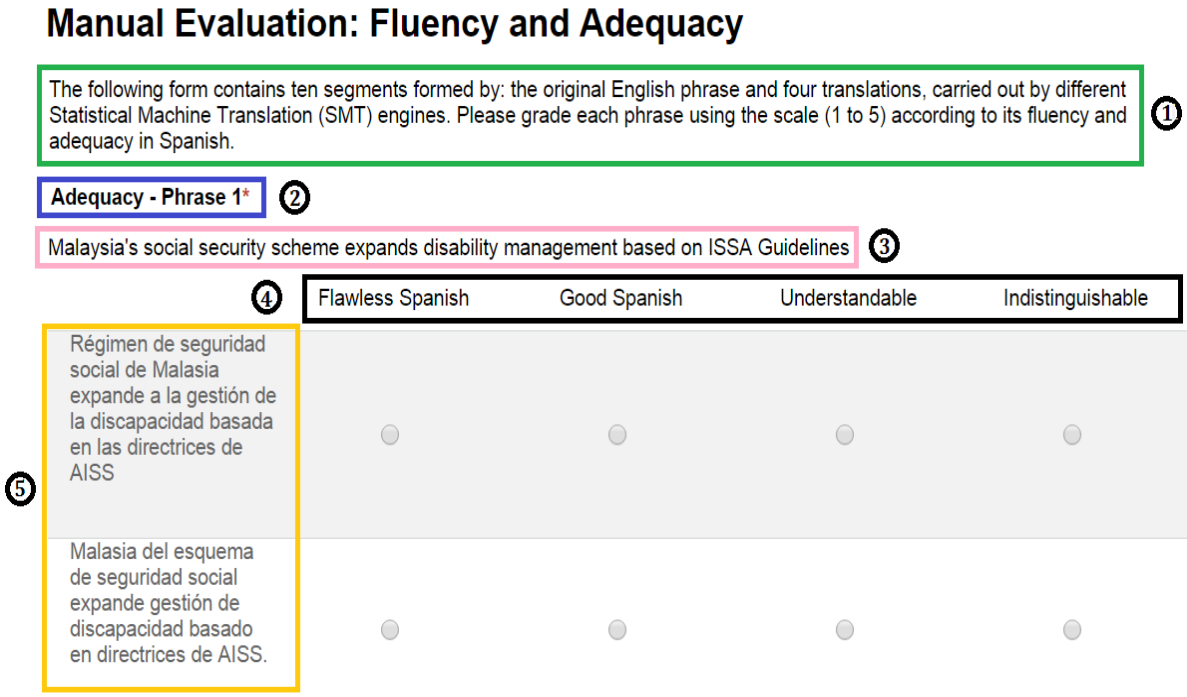



Fig. 18: Readability Test: Adequacy (Google Forms)

**Comments**:

The evaluator's responses are recorded in a sheet automatically generated by the tool, and, afterwards, transferred into a new table where each response is assigned its corresponding point. Figure 19 below shows the results of the test done by one of the participants. For the complete tables, see Annex 4:

| | Fluency and Adequacy | | | | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | | |
| | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | |
| MTH | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 2 | 3 | 2 | 2 | 29 |
| TAPTA4ISSA | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 3 | 3 | 2 | 2 | 0 | 0 | 2 | 2 | 2 | 2 | 35 |
| GT | 0 | 2 | 0 | 0 | 1 | 1 | 2 | 2 | 1 | 2 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 2 | 2 | 21 |
| TAPTA4UN | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 1 | 3 | 1 | 1 | 26 |

Fig. 19: Readability Test: Adequacy and Fluency Results

First, (1) in light blue, we see the number of phrases tested (ten in total), each of them graded in terms of adequacy (A) and fluency (F); (2) in yellow, the name of the systems; (3) and, in orange, the total points for the participant.

Let us consider to the results of phrase one, marked in dark red (4). Generally, it is expected that A and F values present a certain correlation: i.e. that evaluators are likely to assign similar or equal scores for both to the same sentence. However, if we pay attention to sentence 1 in Fig. 19, we can observe that the output of GT (Google Translate) is graded as "undistinguishable" (0 pts) in terms of adequacy, and as "Good Spanish" (2 pts), in terms of fluency. Giving a closer look to the original English sentence, we can see that virtually every word in the phrase carries an important piece of information, which is very common in English headlines and titles. GT managed to keep three out of five meaningful elements: the actor (marked in blue), the object (marked in green) and part of the complement (adverbial phrase indicating means or method: *How did the social security scheme of Malaysia expand disability management*? By following the ISSA Guidelines). Therefore, it is fair to judge that it maintained most of the sentence meaning. However, due to the omission of the initial article "*el*" and the translation of "expand" (transitive verb) as "*se expande*" (reflexive verb), the sentence does not read fluently in Spanish, and might result very confusing for the reader.

| |
|---|
| EN: **Malaysia's social security scheme** expands **disability management** based on **ISSA Guidelines** |
| GT: **Régimen de seguridad social de Malasia** se expande **gestión de la discapacidad** basado en las **Directrices de la AISS** |

Consequently, it can be concluded that, although both attributes are closely related, there is not always correspondence between them.

Before moving on to the final scores, we will discuss a few conclusions drawn from the performance of the two control systems: Google Translate and TAPTA4UN. The motivation for including the output of TAPTA4UN was to test if customized engine performs better than a generic one, when dealing with texts different from the ones used for training them. The final scores show that, in most cases, the

customized systems performed better than the control systems, but curiously, within the latter, Google Translate obtained higher scores than TAPTA4UN (Fig. 20).

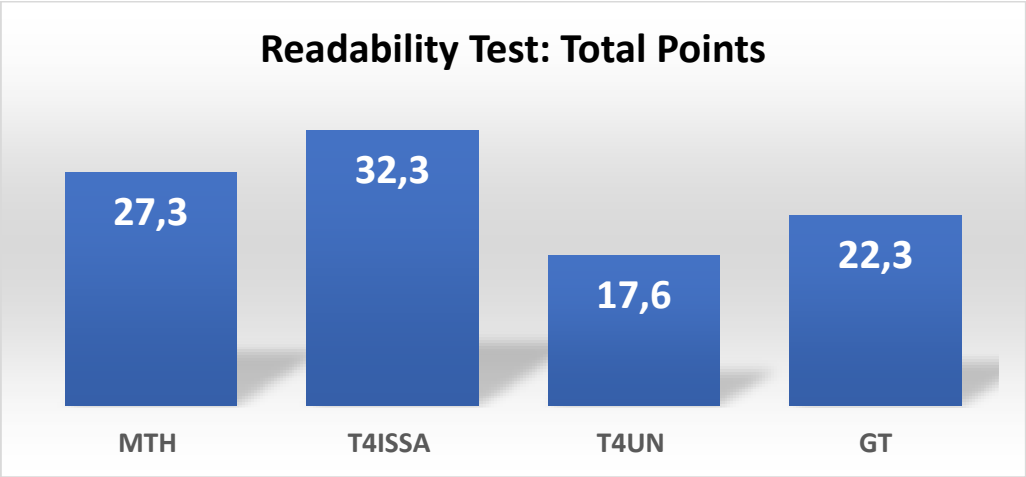



Fig. 20: Readability Test: Total Points

Lastly, if we consider fluency and adequacy scores separately, as we explained in the example presented in the last page (Fig. 19, point 4), it is clear that these scores are not always balanced (i.e. participants might consider that a sentence is difficult to read in Spanish, but still judge that it reflects the meaning of the source sentence, in the way illustrated by the chart below Fig. 19). If we consider the points assigned to the total of segments evaluated by all of the participants (ten segments, assessed separately in terms of adequacy and fluency, see **Fig. 19,** "Readability Test: adequacy and fluency Results", for an example of the result tables; and **Annex 4**, for the complete tables), we can observe that, in general, the four engines seem to performed better in terms of adequacy. It is important to underline that fluency problems might be correlated with grammar correctness problems (see section **7.1.1.2.4. Well-Formedness Test**). Nevertheless, this correlation was not analysed in depth in the framework of the present study. Next figure (Fig. 21) shows the percentage of segments that the translators considered to be useful output to work on (i.e. that they graded from one to three. See Fig. 19) for each system. For example, we see that 34% of segments translated by TAPTA4ISSA are considered to reflect adequately the meaning of the source sentences, while 31% are considered to read fluently. Consequently, we can estimate that there is a 3% of segments translated by TAPTA4ISSA in which adequacy and fluency values do not correlate.

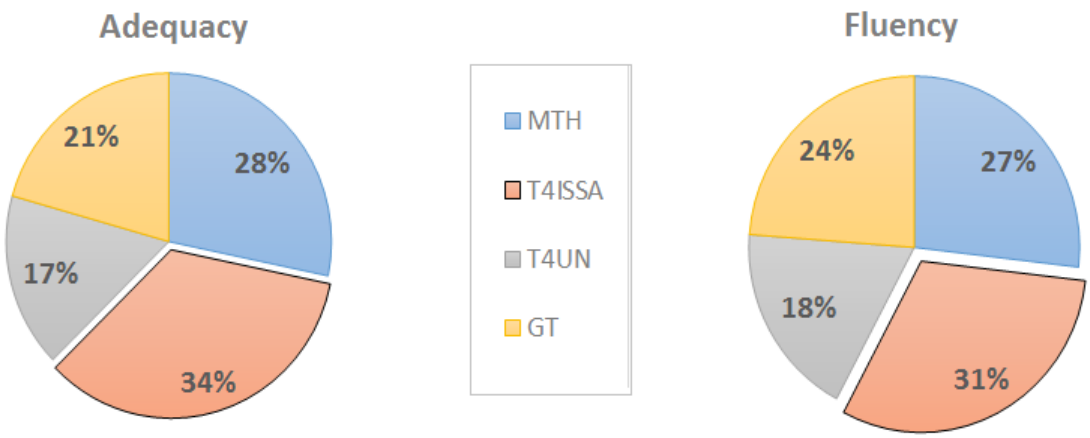



Fig. 21: Final Scores: Adequacy and Fluency

**Results:**

Table 9 shows the final points for the three participants (without W):

| SYSTEM | EVALUATOR 1 | EVALUATOR 2 | EVALUATOR 3 | TOTAL |
|---|---|---|---|---|
| **MTH** | 29 | 23 | 30 | 27.3 |
| **TAPTA4ISSA** | 35 | 20 | 42 | 32.3 |
| **TAPTA4UN** | 21 | 20 | 26 | 22,3.6 |
| **GT** | 26 | 4 | 23 | 17,6 |

Table 9: Final Scores: Adequacy and Fluency

### 7.1.1.2.4. Well-Formedness Test

- **Purpose**: Evaluating the quality of translations in terms of grammar correctness.
- **Method**: Comparative Test: one system against the other.
- **Corpus**: Corpus for Training
- **Participants**: 5
- **Control Systems**: Google Translate (generic SMT engine)
- **W**: 2

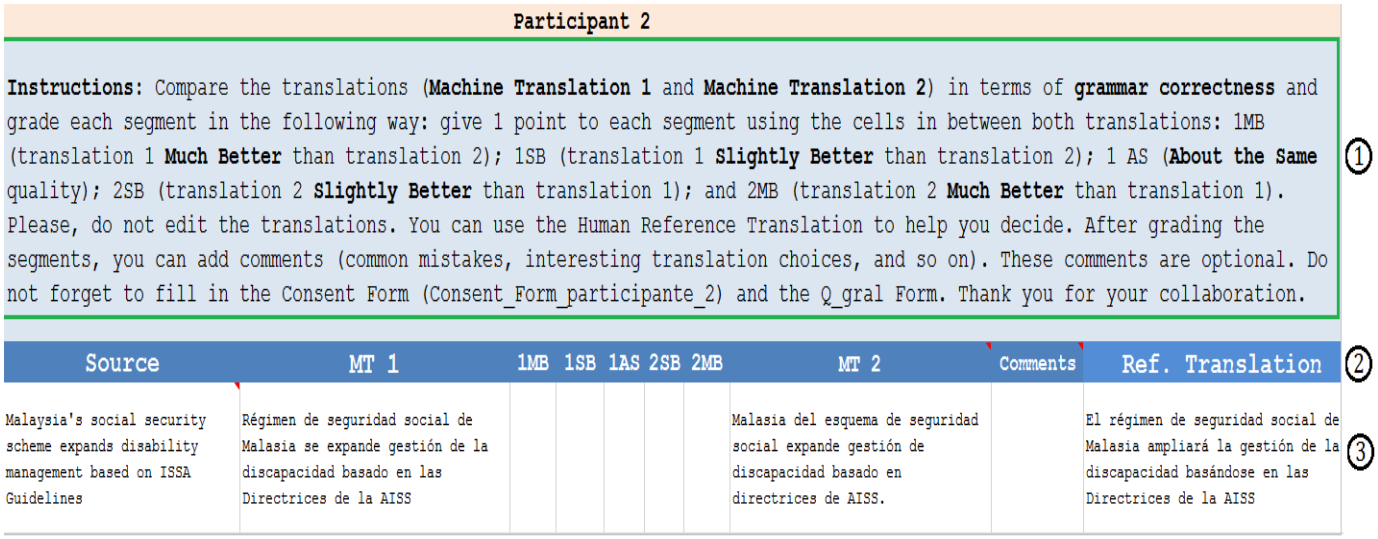
**Test Description:**

The test is a MS Excel table containing the original English sentence (29 segments, 1,214 words), the translations of the candidate systems of the translations of the

control system and a one of the candidates, and a reference translation. The scale is the following:

- 1**MB**: first translation **Much Better** than the second [W: 2]
- 1**SB**: first translation **Slightly Better** than the second [W: 1]
- **AS**: both translations **About the Same** quality [W: 0][52]
- 2**SB**: second translation **Slightly Better** than the first [W: 1]
- 2**MB**: second translation **Much Better** than the first [W: 2]

Figure 22 presents the test as viewed by participants: **(1)** marked in green, we see the instructions; **(2)** the titles indicate the order of the elements in the table: first the English sentence ("source"), followed by the systems translations (anonymized as "MT 1" and "MT 2"). Between the translations, there is the scale presented above: evaluators grade each segment by placing "1" in the corresponding cell. After, the W is added to the final count of those points (e.g. if MT 2 obtains a total of 3 segments translated much better than MT 1; the 3 points are multiplied by the W 2.). Lastly, there is a space for optional comments (common errors or surprising translation choices, etc.) and the reference translation. (3) The third line contains the first sentence and its translations. To access the full test template, see Annex 3.



**Participant 2**

**Instructions**: Compare the translations (**Machine Translation 1** and **Machine Translation 2**) in terms of **grammar correctness** and grade each segment in the following way: give 1 point to each segment using the cells in between both translations: 1MB (translation 1 **Much Better** than translation 2); 1SB (translation 1 **Slightly Better** than translation 2); 1 AS (**About the Same** quality); 2SB (translation 2 **Slightly Better** than translation 1); and 2MB (translation 2 **Much Better** than translation 1). Please, do not edit the translations. You can use the Human Reference Translation to help you decide. After grading the segments, you can add comments (common mistakes, interesting translation choices, and so on). These comments are optional. Do not forget to fill in the Consent Form (Consent_Form_participante_2) and the Q_gral Form. Thank you for your collaboration. ①

| Source | MT 1 | 1MB | 1SB | 1AS | 2SB | 2MB | MT 2 | Comments | Ref. Translation ② |
|---|---|---|---|---|---|---|---|---|---|
| Malaysia's social security scheme expands disability management based on ISSA Guidelines | Régimen de seguridad social de Malasia se expande gestión de la discapacidad basado en las Directrices de la AISS | | | | | | Malasia del esquema de seguridad social expande gestión de discapacidad basado en directrices de AISS. | | El régimen de seguridad social de Malasia ampliará la gestión de la discapacidad basándose en las Directrices de la AISS ③ |

Fig. 22: Well-Formedness Test Template

[52] Segments graded as "AS" are given 0 points because this is a comparative evaluation; and therefore, whenever the systems perform equally, the points are neutralized from the final count.

**Comments**:

Given that the results of the Readability Test (**7.1.1.2.3.**) show higher scores for TAPTA4ISSA, both in terms of adequacy and fluency, it is surprising to see that the results of the Well-Formedness Test reveal higher scores for MTH. The following figure (Fig. 23) shows the percentage of segments graded as MB, SB or AS for each candidate system.

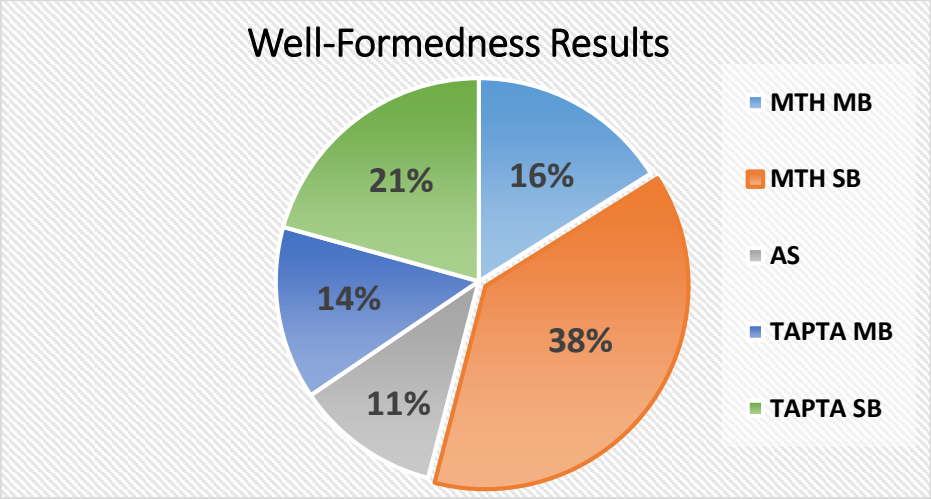



Fig. 23: Well Formedness Test: Results

From the total of sentences, the evaluators considered that MTH translated a 16% of segments **much better,** and 38%, **slightly better** than TAPTA4ISSA. While the difference between the segments considered to be MB translated is small (2%), the difference between those judged as SB is quite large (17%). Some comments from the evaluators reveal that the omission of articles and the mistakes regarding gender agreement are among the principal reasons explaining TAPTA's drop in scores.

It was also surprising that both candidates presented lower scores than the control system, as we can see in Figure 24.

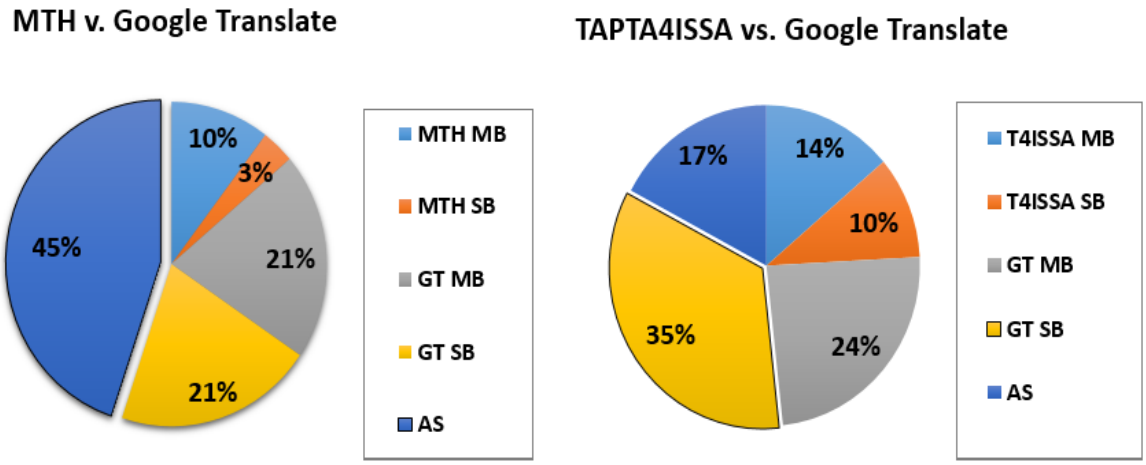



Fig. 24: Well Formedness Test: Results (2)

**Results**:

The points obtained in the tests that compare the candidates with the control system are not taken into account in the final score, since they do not indicate a relation between the candidates. In total, **MTH obtained 22.3 pts**, and **TAPTA4ISSA, 14 pts**.

### 7.2. Operational Evaluation

The operational evaluation is carried out by means of a series of manual Boolean and multiple choice tests. No external participation (e.g. voluntary evaluators) was required to accomplish this task. Section **6.6** above, shows an overview of the general metrics applied for this evaluation: 1 point for each positive answer and 0 point for each negative answer (Boolean). Further specifications are provided in the description of each individual test. The operational evaluation was carried out by a unique evaluator (the researcher). This evaluation was designed using the online tool Google Forms. To Access the full evaluation form, follow this link: http://goo.gl/forms/i4TzwegV5r or see Annex 3.

#### 7.2.1. Functionality

Functionality (defined in section **6.5.2.1.1.**) is evaluated in terms of interoperability (**7.2.1.1.**) and security (**7.2.1.2.**).

#### 7.2.1.1. Interoperability Test

- **Purpose**: Evaluating the formats compatible for training the systems (i.e. the formats that can be used when uploading the training data). The formats tested correspond to the organization's resources and preferences.
- **Method**: Boolean Questionnaire: five questions.
- **W**: 1 (The ISSA enjoys great flexibility in terms of computer resources)

**Test Description**:

Figure 24 shows the different parts of this test: in the image we can see: (1) in green a brief explanation of the aspect under evaluation; (2) in pink, the first question; (3) in blue, one of the formats tested and (4), in yellow, the options to grade.

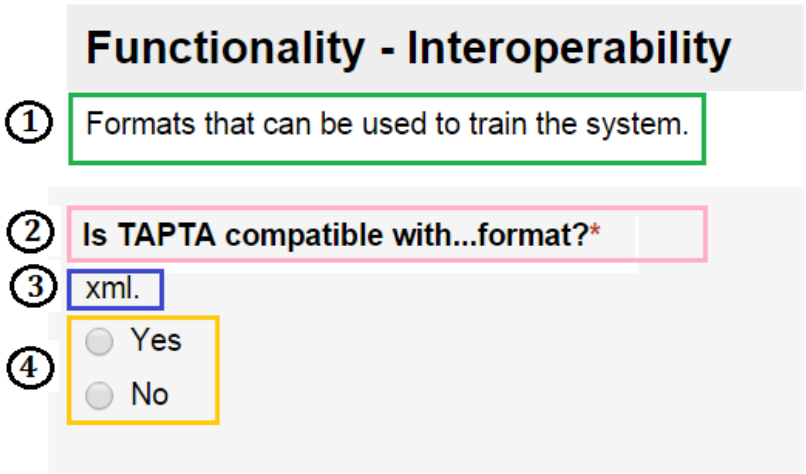



Fig. 25: Interoperability Test (Google Forms)

**Comments and Results**:

Table 10 summarizes the responses:

| FORMAT | TAPTA4ISSA | MTH |
|---|---|---|
| **XML** | 1 | 1 |
| **HTML** | 1 | 1 |
| **DOC** | 0 | 1 |
| **PDF** | 0 | 1 |
| **XLIFF** | 1 | 1 |

| FORMAT | TAPTA4ISSA | MTH |
|---|---|---|
| TOTAL | 3 | 5 |

Table 10: Interoperability Test (Summary)

#### 7.2.1.2. Security Test

- **Purpose**: Evaluating the system's capacity to protect information against unauthorized users.
- **Method**: Boolean Questionnaire: four questions
- **W**: 4 (This attribute is very important for the Association, since some of its documents are private)

**Test Description**:

Figure 26 shows the test template, as visualized through Google Forms. As we can see, the test starts with a brief explanation of the attribute (1), followed by the question (2) and a brief note with additional information (3), and finally, (3) the options to grade.

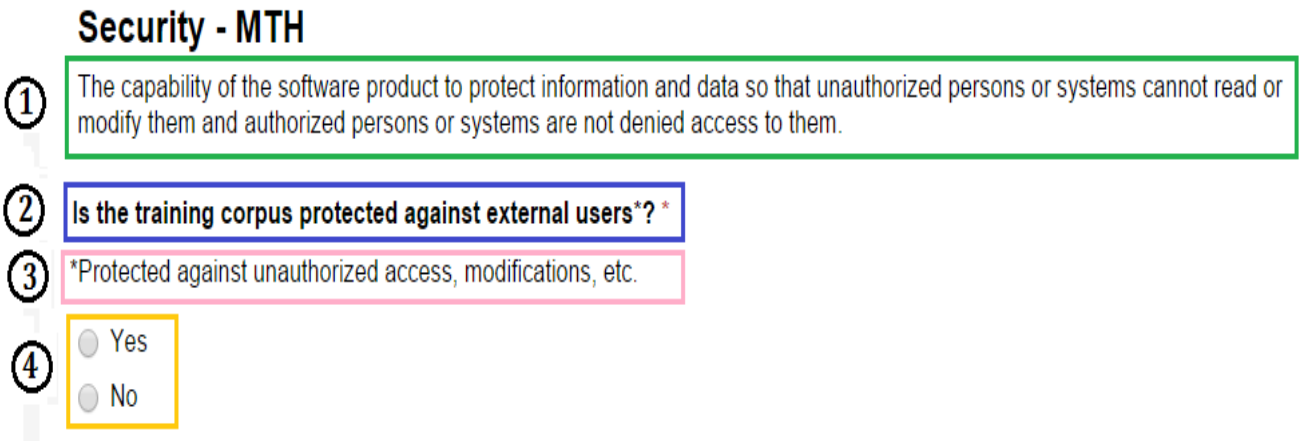



Fig. 26: Security Test (Google Forms)

**Comments**:

The resulting score is not completely surprising (without including W calculation): **TAPTA (4); MTH (3)**. Being a commercial system, MTH is linked to other Microsoft services: to train the SMT engine, it is necessary to create a MS Account and upload the whole corpus to MS server. Although it is not disseminated, uploading

information to the cloud or to external servers always makes information vulnerable, as we do not hold complete control over it anymore.

**Results**: The resulting scores (without including W) are: **TAPTA (4); MTH (3)**

### 7.2.2. Usability

Usability (defined in section **6.5.2.2.2.**) is evaluated in terms of the sub-characteristics learnability (**7.2.2.1.**) and understandability (**7.2.2.2.**).

#### 7.2.2.1. Learnability Test

- **Purpose**: Evaluating the resources available for learning how to use the system.
- **Method**: Multiple Choice Questionnaire: one question with five options.
- **W**: 1 (direct users are expected to have a high computer literacy)

**Test Description**:

Figure 27 shows the test template, as visualized through Google Forms. We can see that under the title of the test, marked in green, there is a brief definition of the sub-characteristic under evaluation (1), followed by the name of the system marked in blue (2), the question, marked in pink, (3) and the multiple choice options, in yellow (4).

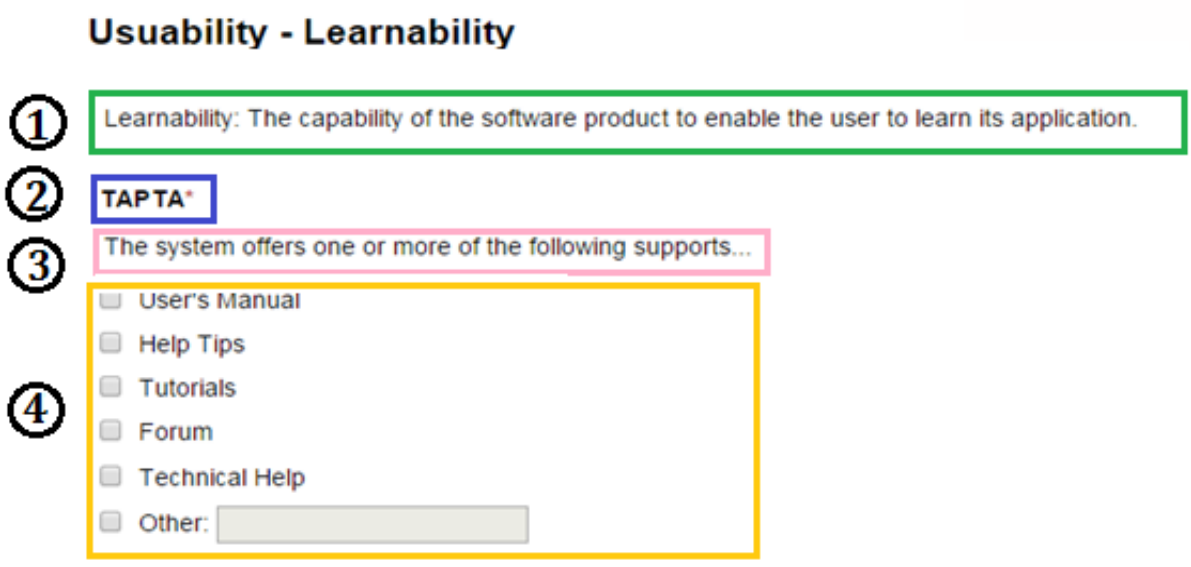



Fig. 27: Learnability Test (Google Forms)

**Comments**:

While TAPTA offers User's Manual, Tutorials, Technical Help (onsite technical help and personalized guides provided by developers); MTH offers User's Manual, Help Tips, Forum, and (online) Technical Help. It seems inadequate to give the same score to the systems' the technical help, since TAPTA provided a thorough onsite guidance during the training stage. In order to account for the extra quality support, an extra point is added, leaving both systems on equal terms.

**Results:**

The resulting scores (without including W) are: **TAPTA (4); MTH (4)**.

### 7.2.2.2. Understandability Test

- **Purpose**: The purpose of this test is evaluating how intuitive (or user friendly) are the candidate systems, considering the direct and indirect users described in section **6.4**.
- **Method**: multiple choice questionnaire, made up of two questions with 5 options.
- **W**: 1

**Test Description**:

Figure 28 shows the test template, as visualized through Google Forms. Similarly to the last test, under the title of the test, marked in green, there is a definition of the sub-characteristic evaluated (1), the name of the system (2), the question, (3) and the multiple choice options (4).

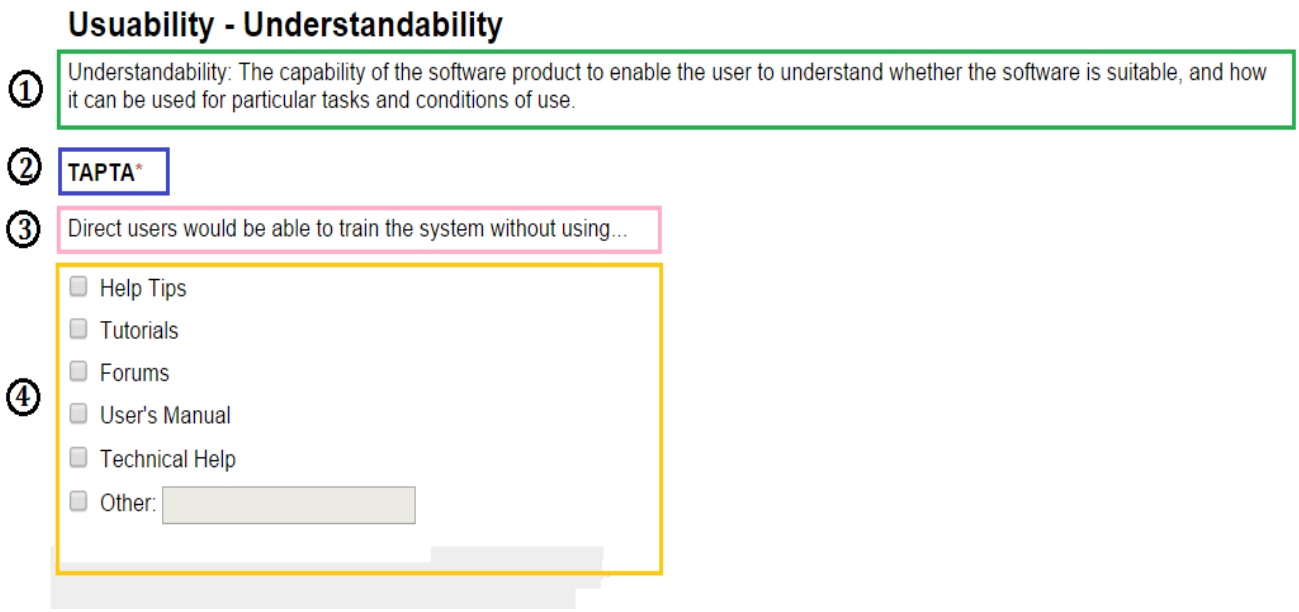



Fig. 28: Understandability Test (Google Forms)

**Comments**:

As a general rule, commercial systems are more user friendly than open source systems. For example, in this case, MTH counts with a much more intuitive training interface. It has a simple training interface, and therefore, it requires less support than TAPTA to learn how to customize the engine.

**Results:**

The resulting scores (without W) are: **TAPTA (5); MTH (7)**.

### 7.2.3. Efficiency

This QC (defined in section **6.5.2.2.3.**) is tested in terms of cost (**7.2.3.1.**) and time-behaviour (**7.2.3.2.**)

#### 7.2.3.1. Cost Test

- **Purpose**: Evaluating if the candidate systems match the Association's budget.
- **Method**: Multiple Choice Questionnaire: five questions (Three mandatory, and two derive from question one)
- **W**: 4 (At institutional settings, resource efficiency, in this case, lower cost, is highly preferable, even if the price is affordable.)

**Test Description**:

Figure 29 shows the test template, designed in a similar way to the previous tests: title and name of system (1); definition of the sub-characteristic evaluated (2), the question, followed by a brief clarification (3), and the multiple choice options (4).

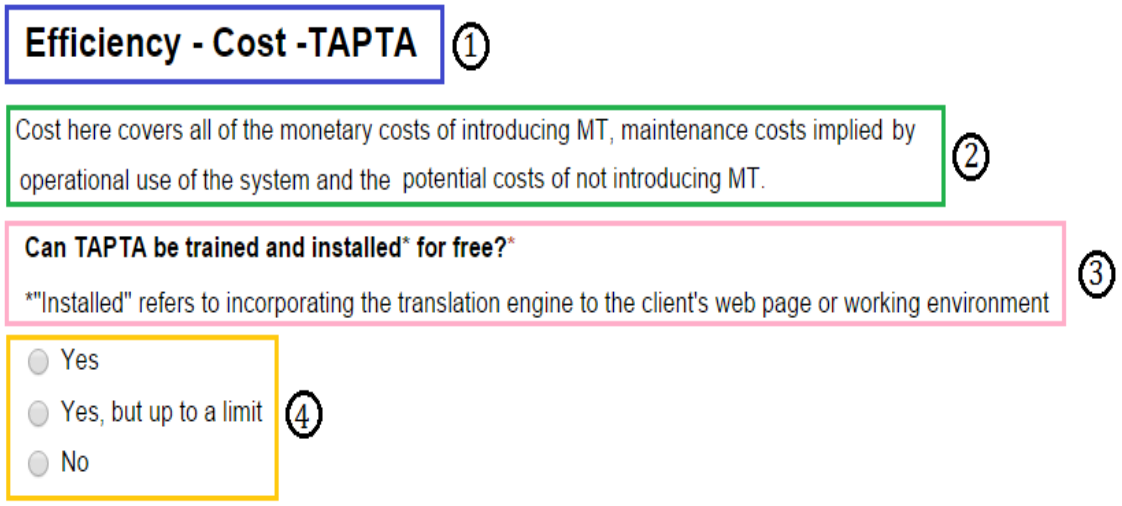



Fig. 29: Cost Test (Google Forms)

This test differs from the previous ones in that, depending on the answer, the evaluator will move on to different questions. Figure 30 illustrates the process:

| | | |
|---|---|---|
| Question 1 | Yes > 1<br>Yes, but... > 0 >> Question 2 Yes > 1<br>No> 0 >> Question 3 Yes > 1<br>No > 0<br>No > 0 >> Question 3 Yes > 1<br>No > 0 | Question 4 and 5 |

Fig, 30: Question Sequence (Cost Test)

For example: the first question "Can the system be trained and installed for free?" is mandatory; if the answer is "yes" (1 point), the evaluators moves to the fourth question: "Client can add more direct users (administrators, developers) for free."; and the fifth: "Are updates free?". If the answer to the first question is "no" (0 point), the evaluator moves to the third question: "is the price superior to the client's budget?" ("yes"=0; "no"=1). And then continue in the same way. Finally, if the

answer is “yes, but up to a limit”, the evaluator moves on to the second question: “If there is a limit of data usage*, is it superior to what the client would need?” (“Yes” and moves to third question; “no”, moves on to the fourth).

**Results:** The resulting score (without including W) is: **TAPTA (4); MTH (2)**.

#### 7.2.3.2. Time Behaviour Test

- **Purpose**: evaluating which of the two candidate systems is more efficient in terms of time.
- **Method**: Three Boolean questionnaire: (1) setting the system ready for use; (2) translation time; (3) update time
- **Corpus**: extract from corpus for testing (see Annex 6)
- **W**: **3**

**Test Description**:

The first aspect, setting the system ready for use, is tested with three Boolean questions (Fig. 31). In Figure 31, we see that the structure of the test is similar to the previous ones: definition (1); title (2); questions (3); and systems under evaluation (4).

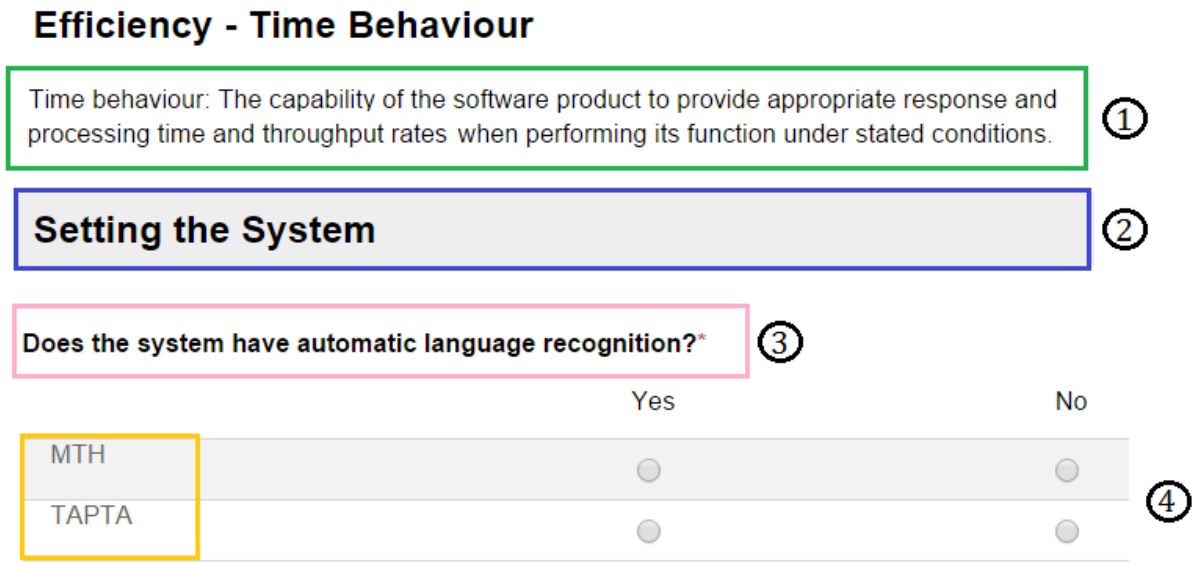



Fig. 31: Time Behaviour Test 1 (Google Forms)

The second and third questionnaires (Fig. 32) present a slight variation: instead of the binary options "yes" or "no", we can see that there is a scale from 0 to 4 points (marked in black, point 3). Translation time is measured on extracts (sentence, paragraph and text) of the corpus for testing (**section 6.3.1.**).

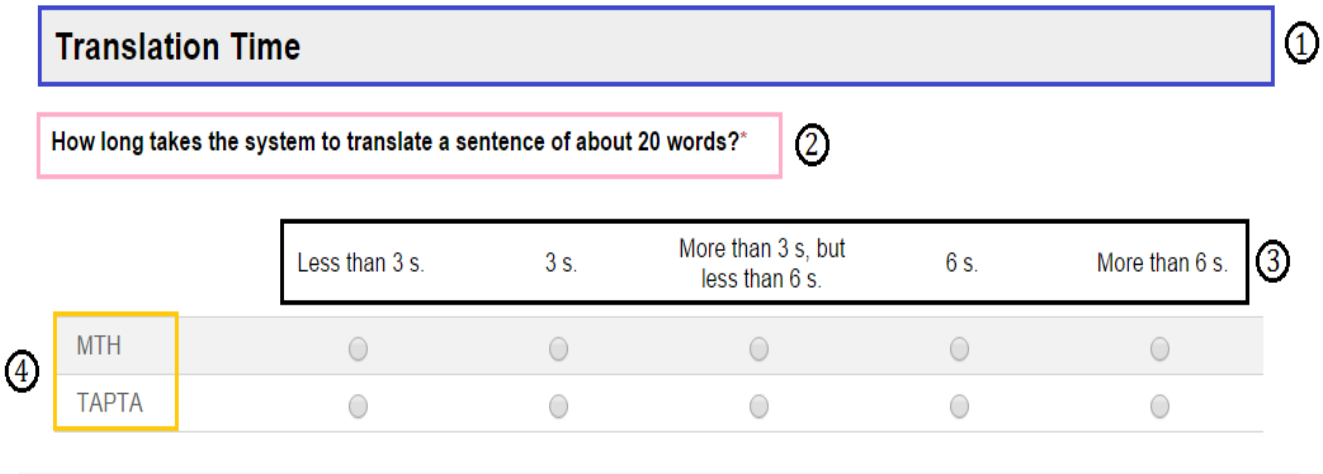



Fig. 32: Time Behaviour Test 2 and 3 (Google Forms)

**Comments:**

Both systems did well in terms of time behaviour. The resulting scores (without Weight) are: **TAPTA (5); MTH (4).**

### 7.2.4. Maintainability

Maintainability (defined in **section 6.5.2.2.4.**) is tested in terms of two sub-characteristics: changeability (**7.2.4.1.**) and stability (**7.2.4.2.**).

#### 7.2.4.1. Changeability Test

- **Purpose**: Evaluating to what extend the systems can be customized.
- **Method**: Boolean Questionnaire: five questions with two options each (Fig. 29)
- **W**: 4

**Test Description:**

In Figure 33, we see that the structure of the test is similar to the previous ones: definition (1); clarification (2); question (3); and systems under evaluation (4).

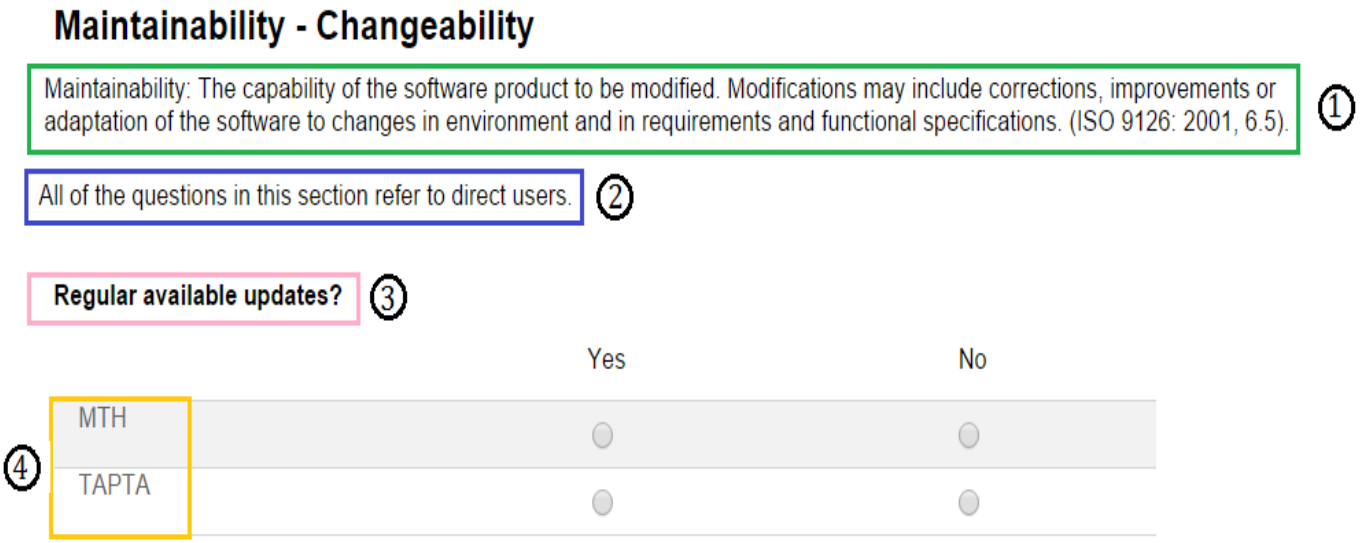



Fig. 33: Maintainability: Changeability Test (Google Forms)

**Comments**:

The results show a gap between both systems: while TAPTA obtained 5 pts (without W), MTH obtained only 1 (without W). This is easily explained by the fact that the latter is a commercial system, and therefore, direct users have limited freedom to introduce changes into the system.

**Results:**

The resulting scores (without Weight) are: **TAPTA (5); MTH (1).**

### 7.2.4.2. Stability Test

- **Purpose**: Evaluating the system's capacity to avoid undesired changes.
- **Method**: Boolean Questionnaire: Three questions with two options each (Fig. 29)
- **W**: 4 (Fundamental to the system's long-term quality)

**Test Description**:

Figure 34 shows the test template as visualized in Google Forms. The structure of the test is similar to the previous ones: definition (1); questions and clarification (2); and systems under evaluation (3).

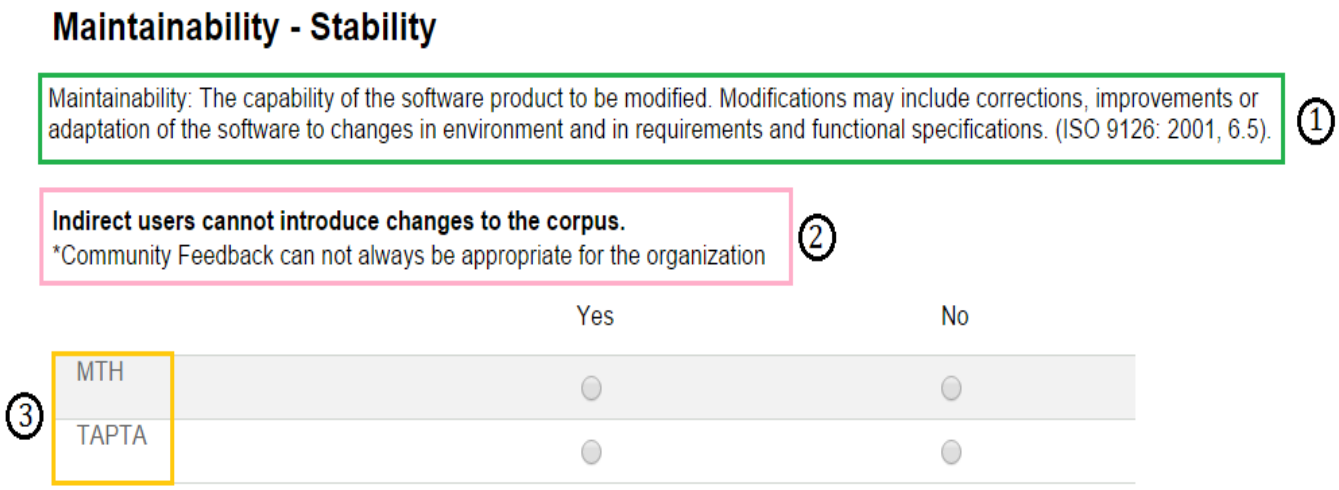



Fig. 34: Maintainability: Stability Test (Google Forms)

**Comments**:

The results were the same for both systems: MTH (3); TAPTA (3), as they both protect the engines functioning against external users. Although open source systems are usually more vulnerable to external users' modifications, TAPTA algorithm, corpora, models, etc. Can only be modified by those who are given access to the software (e.g.: TAPTA developers offered all of the necessary files to UN developers in New York, for them to introduce the modifications they considered appropriate.)

**Results:**

The resulting scores (without Weight) are: **TAPTA (3); MTH (3).**

### 7.2.4.3. Comments on Additional Characteristics

Both systems offer an online editing interface. TAPTA offers a series of possible translations drawn from the corpus. Users can select the translation that they judge most adequate and, additionally, incorporate more changes by hand. MTH offers the possibility of adding changes by hand, but does not propose different translations

from the corpus. Authorised users are able to modify translations and "send" the changes, which afterwards affects the system's translation choices. The Extent to which these isolated proposals affect the output (positively or negatively) are difficult to test in the framework of this short evaluation project. Consequently, this characteristic is presented with information purposes, but it is not included among those tested.

### 7.2.5. Portability

Portability (defined in section **6.5.2.2.5.**) is evaluated only in terms of the sub-characteristic installability (**7.2.5.1.**)

#### 7.2.5.1. Installability Test

- **Purpose**: Evaluating the compatibility of the candidate systems with the Association's operating systems.
- **Method**: Boolean Questionnaire: five questions with two options each.
- **W**: 1 (The ISSA enjoys of great flexibility in terms of computer resources)

**Test Description**:

Figure 35 shows that the structure of the test is similar to the previous ones: definition (1); questions and clarification (2); and systems under evaluation (3).

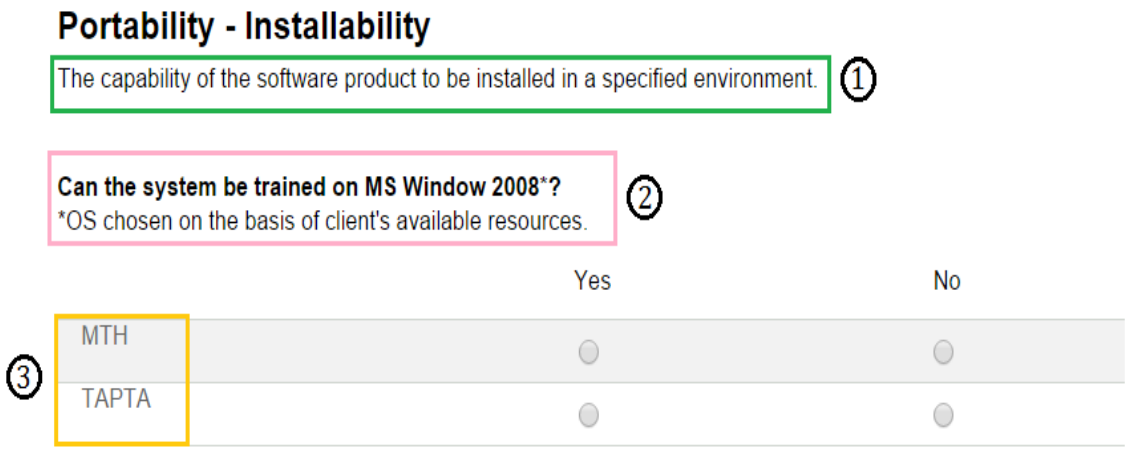



Fig. 35: Portability Test: Installability (Google Forms)

**Comments**:

The results were 2 pts for MTH and 1 for TAPTA: MTH can be trained in all Microsoft Windows Environments, while TAPTA, only in Linux environments. This test is only carried out with respect to training because the translation interface (that will be designed after one of the candidates is selected) will be accessible online through the Association's web portal.

**Results:**

The resulting scores (without Weight) are: **TAPTA (1); MTH (2).**

### 7.3. Partial Conclusion

This chapter presented the evaluation execution in detail, examining the results of each test. By going through these partial results, it is clear that both candidate systems present interesting characteristics that match the Association's needs. Some tests seem to turn the tide in favour of one of the candidates, while other tests reveal surprising results that call for further analysis. The following chapter (**VIII. Final Conclusions**) analyses the final results for the complete evaluation, and summarizes the scores in a table. In addition, it discusses the limitations of the present study, as well as possibilities for future research projects on the area. Next Chapter (**VIII. Final Conclusions**), contains a figure (**Fig. 37. Summary of Final Scores**) that summarizes the results of each test, with and without Weight.

# VIII. Final Conclusions

In the introduction, we mentioned three objectives for this thesis: (1) presenting a framework for comparative context-oriented evaluations of MT systems; (2) comparing generic engines and customizable ones; and (3) making assumptions on the link between MT and institutional translation. These objectives are wider than the specific purpose of the case study: assisting the International Social Security Association (ISSA) to select a suitable SMT engine for their multilingual web portal (see **Chapter I**). This chapter summarizes the results for each of these objectives.

## 8.1. Methodological Framework for Evaluating MT Engines in the Context of International Settings

The methodological framework for the present study was designed following EAGLES (Evaluation of Natural Language Processing Systems), particularly its summary report *The EAGLES 7-step recipe* (1999), and FEMTI (Framework for the Evaluation of Machine Translation in ISLE) (see **Chapters V.** and **VI.**). The underlying idea is designing a comprehensive, but practical, comparative evaluation, with a clear structure that researchers can reuse for future studies. The steps described in Chapter VI can be regrouped in the following way:

a. Definition of Evaluation Purpose
    - *What do we want* (*or expect*) *to find out with the evaluation*?
b. Definition of Application Context (Users' Needs and Resources)
    - *Who is are the users*? *What do they want*? *What do they need*? *What do they have*?
c. Definition of System's Requirements
    - *Minimum requirements that need to be met by all candidate systems*.
    - *QC, SC and Attributes to be tested*
d. Selection of Candidate Systems
e. Estimating Time and Cost for the Evaluation

- *How long will it take to carry out the evaluation and present the results?*
- *Resources needed: human resources, funding, equipment, etc.*

Moreover, Figure 36 summarizes the proposed quality model for the undertaking evaluations of MT engines for institutional settings:

| QC | SC | Attribute | Metrics | Weight |
|---|---|---|---|---|
| **Declarative Evaluation** | | | | |
| Functionality | Accuracy | Fidelity - precision | BLEU Score | |
| | | Terminology | Percentage of terms correctly translated. | |
| | Suitability | Readability | Subjective rating of fluency and adequacy | |
| | Well-formedness | Grammar - Syntax | Comparative test: subjective rating of correctness | |
| **Operational Evaluation** | | | | |
| Functionality | Interoperability | | Boolean questionnaire | |
| | Security | | Boolean questionnaire | |
| Usability | Learnability | | Boolean multiple choice | |
| | Understandability | | Boolean multiple choice | |
| Efficiency | Cost | | Boolean questionnaire. | |
| | Time behaviour | Preparation Time | Boolean Questionnaire. | |
| | | Translation Time | Record of time each system takes to translate (scale). | |
| | | Update time | Record of time each system takes to be trained (scale). | |
| Maintainability | Changeability | | Boolean questionnaire | |
| | Stability | | Boolean questionnaire | |
| Portability | Installability | | Boolean questionnaire | |

Fig. 36: Evaluation Structure

The first, second and third column remain practically unchanged with regards to the evaluation undertaken for the ISSA (see Table 5). The attributes marked in lilac in third and fourth columns are proposed as alternative values, i.e., if necessary, the organization can reduce the evaluation time and effort by testing one or the other. In addition, if the researcher does not count with the participation of human judges, he or she can rely on the score provided by one or more automatic metrics. As an

alternative to the BLEU metric, other automatic metrics such as NIST, METEOR can also be tested in Asiya, following this link http://asiya.cs.upc.edu/demo/.

Weights need to be assigned in a case by case fashion, since they change to reflect the particular needs and resources of an organization. For example, for organizations with limited resources, and particularly those intending to introduce rule-based or hybrid MT engines (see sections **3.1.** And **3.3.**), the quality characteristics efficiency and portability and the sub-characteristic resource utilisation (memory usage, program size, etc.)[53], are essential, and therefore carry a high W.

## 8.2. Evaluation Results

In total, fifteen tests were done during the evaluation: four larger tests to assess the quality of the output; and eleven shorter tests to assess the quality of the systems as software products. Figure 37 summarizes the final scores (including Weight):

| DECLARATIVE EVALUATION | | | | | | | |
|---|---|---|---|---|---|---|---|
| TEST | | | | MTH | | TAPTA | |
| | | | W | Without W | With W | Without W | With W |
| 1.Functionality | 1.1. Accuracy | 1.1.1. Fidelity P. | 1 | 1,00 | 1,00 | 1,00 | 1,00 |
| | | 1.1.2. Terminology | 3 | 16,30 | 48,90 | 22,00 | 66,00 |
| | 1.2. Suitability | 1.2.1. Readability | 3 | 27,30 | 81,90 | 32,30 | 96,90 |
| | 1.3. Well-formedness | 1.3.1. Grammar | 2 | 22,30 | 44,60 | 14,00 | 28,00 |
| Points | | | | 66,90 | 176,40 | 69,30 | 191,90 |
| OPERATIONAL EVALUATION | | | | | | | |
| TEST | | | | MTH | | TAPTA | |
| 2.Functionality | 2.1. Interoperability | - | 1 | 5,00 | 5,00 | 3,00 | 3,00 |
| | 2.2. Security | - | 4 | 3,00 | 12,00 | 4,00 | 16,00 |
| 3. Usability | 3.1. Learnability | - | 1 | 4,00 | 4,00 | 4,00 | 4,00 |
| | 3.2. Understandability | - | 1 | 7,00 | 7,00 | 5,00 | 5,00 |
| 4. Efficiency | 4.1. Cost | - | 4 | 2,00 | 8,00 | 4,00 | 16,00 |
| | 4.2. Time Behaviour | 4.2.1. Setting the System | 3 | 4,00 | 12,00 | 5,00 | 15,00 |
| | | 4.2.2. Translation Speed | 3 | 1,00 | 3,00 | 5,00 | 15,00 |
| | | 4.2.3. Update Time | 3 | 3,00 | 9,00 | 3,00 | 9,00 |
| | 5.1. Changeability | - | 4 | 1,00 | 4,00 | 5,00 | 20,00 |
| 5. Maintainability | 5.2. Stability | - | 4 | 3,00 | 12,00 | 3,00 | 12,00 |
| 6. Portability | 6.1. Installability | - | 1 | 2,00 | 2,00 | 1,00 | 1,00 |
| Points | | | | 35,00 | 78,00 | 42,00 | 116,00 |
| Total points | | | | 101,90 | 254,40 | 111,30 | 307,90 |

Fig. 37: Summary of Final Scores

[53] For more information on this attribute, refer to the FEMTI web page http://www.issco.unige.ch:8080/cocoon/femti/taxum-210.html

Figure 37 shows that TAPTA4ISSA performed better than MTH, with a difference of 53.5 points. Observing the results in detail, we observe that, in some cases, MTH obtained better scores than TAPTA4ISSA. One clear example is the Well-Formedness Test, resulting in 40.6 pts for MTH and 28 pts for TAPTA. In section **7.1.1. Functionality**, sub-section Well-Formedness Test (**7.1.1.2.4.**), we mentioned that evaluators assigned lower scores to TAPTA due to the drop of articles and problems with gender agreements. These problems might be a consequence of the corpus-size issue (discussed in section **7.1.1.1.**) TAPTA is designed to deal with large corpora, and during the training stage and the execution of the automatic tests, it was pointed out by TAPTA developers that the Association's corpus was rather small. Nevertheless, TAPTA4ISSA performed better in most tests, including many important ones such as the terminological, readability and security tests, to mention some. All in all, TAPTA5ISSA was chosen as the most suitable SMT engine for the Association, with the perspective that the corpus the organization will use for training the final version of TAPTA4ISSA would be much larger than the one used to train the prototype. The possibility of extending the corpus by merging it with the corpora of an associate organization is under discussion.

Concerning the quality of customized systems in comparison to generic systems, there is a widespread idea that the former perform better than the latter when dealing with certain types of texts (the type used to train them). From the results of the Declarative Test, we observe that, while MTH and TAPTA4ISSA obtained higher final scores in most tests, in some tests **(1)** the translations generated by the control systems rated higher than the ones produced by the candidates (e.g. see the Well-Formedness Test, **7.1.1.2.4.**); and **(2)** the final scores obtained by the candidates did not differ significantly from those of the control systems (e.g. see the Readability Test, **7.1.1.2.3.**)

However, when it comes to terminological accuracy, the superiority of the customized systems is undeniable. Figure 38 presents the comparative results between both candidates and the control system (Bing). "P1", "P2" and "P3", stand for participant 1, 2 and 3. The results were broken down in this way to reveal certain divergence between the evaluators.

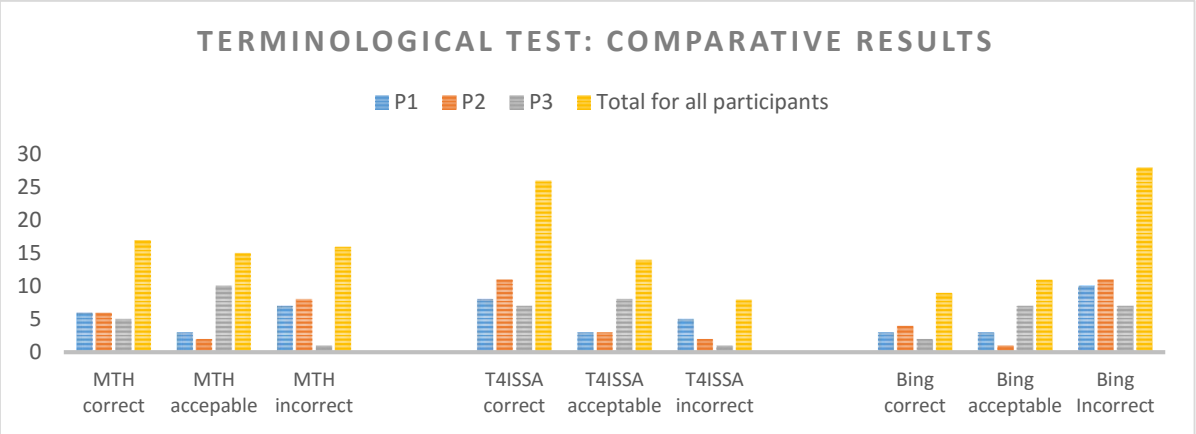



Fig. 38: Terminological Test: Comparative Results

Figure 36 shows that evaluators differ mostly in their judgment of what constituted an "acceptable" or an "incorrect" translation: e.g. while participants one (blue) and two (orange) considered that MTH have translated between a 19% and a 13% of terms acceptably, and, between a 44% and a 50%, incorrectly, participant three (grey) judged that 63% of MTH output was acceptable and only 5%, incorrect. The combined scores for the three participants, result in: 31% of acceptable output and 33% of incorrect output for MTH, and becomes clear that the judgment of participant three favoured MTH considerably. This divergence between participants results in *data dispersion*, a common impediment for research that hampers the elaboration of conclusions on trends or tendencies (Saldanha and O'Brien 2014, 197)[54]. In the case of the present study, it was decided to take the results into account, since despite some differences between the evaluators, there seemed to be a general agreement among them in terms of correct and incorrect output: TAPTA4ISSA (correct: 50%; incorrect: 15%), MTH (correct: 33%; incorrect 31% ), and Bing (correct: 17%; incorrect 54%). The limitations of these results are discussed in section **8.4**.

## 8.3. The Relation between MT and Institutional Translation: The Place of MT in Institutional Settings

From the linguistic results we have been examining up to now, we can make the assumption that, at least for the case of TAPTA4ISSA, MT output provides interesting translation choices, especially in terms of fluency and terminological accuracy.

[54] To read more about statistical tests and standard deviation, refer to Saldanha and O'Brien (2014) and Matthews and Ross (2010).

However, it is difficult to make a general statement on the advantages of using this output, since factors such as the experience of translators or editors working in the organization enter into play. If we consider the problem of discriminating fluent and adequate output (see section **7.1.1.2.3. Readability Test**), it is clear that MT output proves much more useful for senior revisers than to novel translators, who can be misled by highly fluent, but incorrect, output. In addition, the grammar problems found in the output of both candidates confirm that automatic translations cannot be used as final products for publication. The post-editing effort was not measured in this study; consequently, we cannot make assumptions in terms of how much time translators would actually save or lose by using those translations.

Regarding the position of institutional translators towards the use of MT tools, the general questionnaires distributed to participants included three points for them to provide their opinion and describe their experience with regards to the use of MT tools in their daily practice. Table 11 summarizes the answers:

| **Use MT** | | | **Do not use MT** | **No Answer** | **Interested** |
|---|---|---|---|---|---|
| Terminological Tool | Base for PE | Other | | | |
| 0 | 3[55] | 1 | 2 | 1 | 1[56] |

Table 11: Evaluator Comments on MT

Although the initial idea was obtaining answers from all participants in order to study their position towards MT, only five of them answered, and only three expanded on their opinions. Moreover, even if all seven participants had answered the question, the sample would have been too small as to reveal a general tendency (see **8.4.**). However, based on the previous findings, it seems fair to say that most participants showed certain resistance to the use of MT output.

[55] From this four: three said that they use MT output (integrated into their CAT tools) as a bases for PE and one did not specified in which way. One of them explained that, in her area of work, MT output meant more a burden than a help, since it took too long to post-edit. The second one, explained that, although sometimes it helped to accelerate the translation process, it prevented translators from offering their own translations. The third one expressed clear objection to the use of MT, explaining that it slowed down the translation process

[56] Only one participant showed interest in applying MT engines as translation accelerators.

## 8.4. Limitations of the Present Study and Future Research

The main limitations of this study were clear from the beginning: a small size sample (few participants, short samples for testing) and limited time and funding. On the one hand, it was difficult to find voluntary participants responding to the profile. While reducing the requirements for participations might have improved the odds of finding willing participants (recruiting students, for example) this would have affected the results negatively: students might have graded the translations in a way teachers grade theirs, not from the point of view of experience as an institutional translator. On the other hand, since participants were volunteers, asking them to grade larger samples would have probably resulted in a high dropout rate. The size of the Association's corpus is not considered a limitation of the research because it was a characteristic of the case study scenario. However, despite these limitations, the findings of this study let us generate two hypotheses that can be tested in future studies: (1) in terms of translation quality, customizable systems are more appropriate for institutional translation (see sections **7.1.1. Functionality** and **8.2.1 Evaluation Results** […]); (2) in terms of operability, open source systems are more compatible with institutional settings than commercial systems (see sections **7.2. Operational Evaluation** and **8.2.1**).

It is worth stressing that, since both candidates were SMT engines, no assumptions can be made of whether these type of engines are more suitable for international organisations than other types of engines, such as RBMT or hybrid systems.

In conclusion, it could be interesting to test; on the one hand, the replicability of the results by carrying out the same study (evaluation of MT systems in institutional settings) with a larger scope (more participants, larger samples, more candidate systems, etc.); and on the other hand, the transferability of the findings by undertaking a similar project in different international organizations, multilingual national institutions, or even multinational companies.

## XIX. Bibliography

ALPAC. (1966). *Language and Machines. Computers in Translation and Linguistics.* Washington, D. C.

Berger A., Brown P., Della Pietra S., V Della Pietra, Lafferty, J. Printz, H. and Ures L. (1994). The Candide system for machine translation. In *Proceedings of the ARPA Conference on Human Language Technology*.

Berners-Lee, Tim; Cailliau, Robert (12 November 1990)."World Wide Web: Proposal for a hypertexts Project". Retrieved 15 March 2015 from http://www.w3.org/Proposal.html

Cao, D., & Zhao, X. (2008). Translation at the United Nations as Specialized Translation. *The Journal of Specialised Translation*, (9), 39–54.

Cancedda, N., Dymetman, M., Foster, G, Goutte, C. (2009). A Statistical Machine Translation Primer in *Learning Machine Translation* (pp. 2-37). The MIT Press, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142.

Choudhury Rahzeb and McConnel Brian, (reviewers) Van der Meer Jaap and Lockwood Rose, TAUS, *Translation Technology Landscape Report*, April 2013. Funded by LT-Innovate

COMISIÓN EUROPEA. DGT. (2009) *Traducir para una comunidad multilingüe*. Luxemburgo: Oficina de Publicaciones Oficiales de las Comunidades Europeas.

COMISIÓN EUROPEA. DGT. (2009) *Translation Tools and Workflow*. Luxemburgo: Oficina de Publicaciones Oficiales de las Comunidades Europeas.

EAGLES: Evaluation of Natural Language Processing Systems. FINAL REPORT. Version of October 1996. Retrieved from http://www.issco.unige.ch/en/research/projects/ewg96/ewg96.html. Last accessed May 8, 2015.

Elizalde, C., Pouliquen, B., Mazenc C., and García-Verdugo, J. TAPTA4UN: Collaboration on machine translation between the World Intellectual Property Organization and the United Nations. In: *Languages and Translation. Machine Translation*. N. 6, February 2013, pp. 22-23

Estrella, P. S., Popescu-Belis, A., & Underwood, N. (2005). Finding the system that suits you best: Towards the normalization of MT evaluation. *27th international conference on translating and the computer (ASLIB)* (pp. 23-34) Retrieved fromhttp://archive-ouverte.unige.ch/unige:2289

FEMTI - a Framework for the Evaluation of Machine Translation in ISLE [online]. http://www.issco.unige.ch:8080/cocoon/femti/st-home.html (Last accessed: August 2015)

Gerlach, J. (2009). *Les Interlangues en TA: l'example de MedSLT*. Université de Geneve.

Hutchins, W. J. (2000). *Early Years in Machine Translation. Memoirs and biographies of pioneers*. *Amsterdam studies in the theory and history of linguistic science. Series 3, Studies in the history of the language sciences, v. 97*. http://doi.org/10.1162/089120102762671990

Hutchins, W. and Somers, H. An Introduction to Machine Translation, London: Academic Press, 1991

Hutchins, J. and Somers H.. 1992. *An Introduction to Machine Translation*. London; San Diego [etc.]: Academic Press.

Introduction à LexShop – Lexique de Transfert - Tests et Actions. Master Class, Automatic Translation (2011); Master in Translation, Concentration in Translation Technologies. Faculty of Translation and Interpreting, University of Geneva. Retrieved 10/08/15 (09:52).

ISSCO. MTEval II (FEMTI on-line application) - *Quality requirements and metrics for the evaluation of MT: analysis and integration of expertise* [online]. http://www.issco.unige.ch/en/research/projects/ (Last accessed: August 2015)

[ISO 01] Iso/Iec, ISO/IEC 0126-1: Software Engineering-Product Quality-*Part 1: Quality Mode*/ International Organization for Standardization/International Electrotechnical Commission (2001).Retrieved from http://www.iso.org/iso/catalogue_detail.htm?csnumber=35733. Last accessed August 6, 2015 (09:16 )

Johnson, R. (1979). Contemporary Perspectives in Machine Translation by Rod Johnson , Centre for Computational Linguistics ,. In V. Hanon, Suzanne and Hjørnager Pedersen (Ed.), *Human translation, machine translation. Papers from 10th annual conference on computational Linguistics,*. Odense, Denmark.

Jurafsky, D. and Martin, J. H. (2000*). Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River N.J.: Prentice Hall.

Koehn, P. (2010). *Statistical Machine Translation*. United Kingdom, Cambridge: University Press.

Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., … Moran, C. (2006). Open Source Toolkit for Statistical Machine Translation, (June), 177–180.

L'Homme, Marie-Claude. *Initiation à la traductique*. Montréal : Lingatech, 2008. 2e édition. 317 p. : ill.

McPhee, Robert D., Zaug, Pamela. The Communicative Constitution of Organizations A Framework for Explanation. In: *The Electronic Journal of Communication* [online], Volume 10, Numbers 1 and 2, 2000. Retrieved from http://www.cios.org/www/ejc/v10n1200.htm (10/07/2015)

Mauser, A., Mauser, A., Hasan, S., Hasan, S., Ney, H., & Ney, H. (2008). Automatic Evaluation Measures for Statistical Machine Translation System Optimization. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, *1*(1), 3089–3092. Retrieved from http://www.lrec-conf.org/proceedings/lrec2008/

MT@EC: European Commission Machine Translation for Public Administrations in the EU Member States. Updated: 26.09.2014. Retrieved April 1, 2015 (17:45) from http://ec.europa.eu/isa/documents/presentations/european-commission-machine-translation-for-public-administrations-in-the-eu-member-states_en.pdf

MUÑOZ MARTÍN, F. Javier y VALDIVIESO BLANCO, María (2002) «Traductores y especialistas en la Unión Europea. Hacia el binomio integrador», 410-427 en P. Hernúñez y L. González (coords.): Actas del I Congreso Internacional "El español, lengua de traducción", Comisión Europea y Agencia EFE, Almagro

Nielsen, M. (2009). Introduction to Statistical Machine Translation. (Online Blog). (URL http://michaelnielsen.org/blog/introduction-to-statistical-machine-translation/). (Last accessed August 05, 2011).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *… of the 40Th Annual Meeting on …*, (July), 311–318. http://doi.org/10.3115/1073083.1073135

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02* (pp. 311–318). http://doi.org/10.3115/1073083.1073135

Popescu-Belis, A., Estrella, P. S., King, M., & Underwood, N. L. (2006). A model for context-based evaluation of language processing systems and its application to machine translation evaluation. *Proceedings of the fifth international conference on language resources and evaluation (LREC)* (pp. 691-696) Retrieved from http://archive-ouverte.unige.ch/unige:3444

Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *TAL (traitement automatique de la langue), vol. 48, n. 1* (pp. 67-91) Retrieved from http://archive-ouverte.unige.ch/unige:3486

Popovi, M., & Ney, H. (2007). Word Error Rates : Decomposition over POS Classes and Applications for Error Analysis. *Computational Linguistics*, (June), 48–55.

Pouliquen, B., & Mazenc, C. (2011). Coppa, CLIR and TAPTA: three tools to assist in overcoming the patent language barrier at WIPO, 24–30.

Pouliquen, B., Mazenc, C., & Iorio, A. (2011). Tapta: A user-driven translation system for patent documents based on domain-aware Statistical Machine Translation. *Proceedings of Th 15th International Conference of the European Association for Machine Translation (EAMT)*, (May), 5–12.

Pouliquen, B., Mazenc, C., & Iorio, A. (2014). TAPTA : Translation Assistant for Patent Titles and Abstracts A . What is it ?, 1–6.

Pouliquen, B., Elizalde C., Junczys-Dowmunt, M., Mazenc C., & Garcia Verdugo, J. (2013), *Large-scale multiple language translation accelerator at the United Nations*. [MT Summit 2013], Proceedings of the XIV Machine Translation Summit, Nice September 2-6, pp. 345-352

Quah, Chiew Kin. 2006. Translation and Technology. Palgrave Textbooks in: *Translating and Interpreting*. Basingstoke: Palgrave.

Rayner, M., Estella, P., & Bouillon, P. (2012). *A Bootstrapped Interlingua-Based SMT Architecture*. Geneva.

Saldanha, G. & O'Brien, S. *Research Methodologies in Translation Studies*. 1st Ed. New York: Routledge, 2014. 277 p.

SVOBODA, Tomáš (2013) *Moving beyond CAT tools- The MT paradigm Shift from the Translators' perspective*. OPTIMALE, Rennes, June 6. Retrieved (n.d.) from http://www.ressources.univ-rennes2.fr/service-relations-internationales/optimale/conference/65-optimale-symposium-programme

*Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE*. Retrieved from https://www.iso.org/obp/ui/#iso:std:iso-iec:25000:ed-2:v1:en. Last accessed May 3, 2015.

Tucker, A., "Current Strategies in Machine Translation Research and Development," in Nirenburg, S. (ed), Machine Translation: Theoretical and Methodological Issues, Cambridge University Press (1987), Chapter 2. Pp 22-41.

Toshihiro, K. (2008). Usability evaluation based on international standards for software quality evaluation. *Nec Technical Journal*, *3*(2), 27–32. Retrieved from <Go to ISI>://WOS:000256954500004

Thurmair, Gr. Complex Lexical transfer in Metal, in: *Proceedings of TMI-90*, Austin (1990)

Wagner, E., Bench, S., and Martinez, J. M. (2002) *Translating for the European Union Institutions*. Manchester: St. Jerome.

Weaver, W. (1955): Translation. In: *Locke and Booth* (1955), 15-23.

Wolf, P., Bernardi, U., Federmann, C. and Hunsicker, S. (2011). From Statistical Term Extraction to Hybrid Machine Translation. In Proceedings of the 15th Annual Conference of the European Association for Machine Translation. Leuven, Belgium.
Xiong,

## Annex 1-FEMTI Report: Declarative Evaluation

EVALUATION TYPE

---

- Declarative evaluation: The purpose of declarative evaluation is to measure the ability of an MT system to handle texts representative of an actual end-user. It is concerned with coverage of linguistic phenomena and handling of samples of real text. Declarative evaluations generally test for the functionality attributes of intelligibility, (how fluent or understandable it appears to be) and fidelity (the accurateness and completeness of the information conveyed).

CONTEXT CHARACTERISTICS

---

Machine translation user: This refers to the person who interacts with the machine translation system and with the output produced by it.

Author characteristics: This set of characteristics covers writer attributes that are relevant to the writing task, which influence the unproofed text that is produced.

Genre: Genre refers to the characteristic or definitive form and style peculiar to a type of document. Examples of genre are: newspaper articles; scientific and technical articles; recipes and instructions; correspondence; business/commercial reports; marketing texts and advertisements; legal texts; literature: novels, poetry, etc.; and many others.

Document type: The type of the input document can greatly affect the output of an MT system. For example, inputs to the METEO system are specific and very restricted, mainly weather forecast texts, using a limited lexicon and particular syntactic constructions. As a result the system produces accurate output, comparable to human translation. In contrast, MT of arbitrary text invariably produces output of much lesser quality. Both the genre and the application domain determine the quality.

Evaluation requirements:

Domain or field of application: Domain refers to topic, the field of interest for which the document is relevant, and the potential sublanguage effects germane to MT, for example technical/scientific (specific field being biology, chemistry, automotive mechanics, etc.), social, etc.

Novice: " The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. [...] [May have] sufficient control of the writing system to interpret written language in areas of practical need. Where vocabulary has been learned, can read for instructional and directional purposes, standardized messages, phrases, or expressions, such as some items on menus, schedules, timetables, maps, and signs. " (ACTFL 1983 guidelines for reading proficiency).

Superior: " Able to read with almost complete comprehension and at normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on knowledge of the target culture. [...] Occasional misunderstandings may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms. [...] Material at this level will include a variety of literary texts, editorials, correspondence, general reports, and technical material in professional fields. Rereading is rarely necessary, and misreading is rare. " (ACTFL 1983 guidelines for reading proficiency).

Computer literacy: This refers to the degree to which the user is at ease in computer use and manipulation.

Proficiency in target language: This refers to proficiency in the source language as attested by some recognised measurement. The level of proficiency may be measured, for example by local education tests, internationally recognised examination schemes or organisation internal testing. Two of the best known language proficiency

scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines. Depending on the operations performed on the translation, it is either the reading or the writing proficiency which are more specifically relevant. We propose to use the ACTFL eading proficiency scale (1985) -- note that only the guidelines for writing/speaking have been recently updated.

Proficiency in source language: This refers to proficiency in the source language as attested by some recognised measurement. The level of proficiency may be measured, for example by local education tests, internationally recognised examination schemes or organisation internal testing. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines for reading (1985) -- note that only the guidelines for writing/speaking have been recently updated.

Asynchronous communication: In the case of asynchronous or delayed communication the interaction between participants occurs with interruption, for example by email.

User characteristics: This covers the characteristics of users in three senses: the end user who will interact with the machine translation system; the end user of the final product of the translation process which may include for example, post-editing; the organisation deploying the machine translation system. Note however that in the case when machine translation is combined with substantial post-editing, the resulting "system" might no longer fall under the scope of FEMTI, hence the end users are no longer users of a machine translation system.

Proficiency in source language: This refers to proficiency in the source language as attested by some recognised measurement, international or regional. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines.

Input characteristics (author and text): Input characteristics refer to the stylistic form or format of the source document, the topic domain, and both the competency and performance qualities of the author.

Advanced: " Advanced-level writers are characterized by the ability to: write routine informal and some formal correspondence, narratives, descriptions, and summaries of a factual nature; narrate and describe in major time frames, using paraphrase and elaboration to provide clarity, in connected discourse of paragraph length; express meaning that is comprehensible to those unaccustomed to the writing of non-natives, primarily through generic vocabulary, with good control of the most frequently used structures. " ACTFL 2001.

## QUALITY CHARACTERISTICS SUGGESTED BY FEMTI

---

- Fidelity - precision: Subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation (Van Slype). Measurement of the correctness of the information transferred from the source language to the target language (Halliday in Van Slype's Critical Report).

Normalized weight: 0.4

Metrics:

- *Rating of sentences*

Method: Rating of sentences read out of context on a 9-point scale.

- *BLEU*

Method: Bleu evaluation tool kit Automatic n-gram comparison of translated sentences with one or more human reference translations.

Terminology: Correct translation of technical (domain-specific) terms.

Normalized weight: 0.1

Metrics:

*Percentage of domain terms correctly translated.* Method:

Coverage of corpus-specific phenomena: Coverage refers to the ability of the system to deal satisfactorily with linguistic phenomena, both generally addressing known cross-language phenomena and specifically addressing phenomena in a corpus of interest. Coverage of corpus-based problematic phenomena concerns the ability of the system to deal with the particular challenges presented by a corpus of interest. Normalized weight: 0.1

Metrics:

Method: Subjective human scoring on a 10-point scale.

Corpora : The kinds and number of monolingual, comparable or parallel corpora available. The category of corpus will depend on the style of language modeling and statistical techniques used in the system.

Normalized weight: 0.1

Metrics:

*Types of corpora incorporated into the system.* Method: Report by the developer.

Functionality: The capability of the software product to provide functions which meet stated and implied needs when the software is used under specified conditions. Normalized weight: 0.1

Metrics:

No selected metrics for this quality characteristic

Languages: "The range of languages which the product supports is a vital selection criterion. In machine translation systems, the languages are classified according to source and target language pairs, due to the need for full linguistic processing capability. In translator workbench products, the languages are not necessarily classified by strict language pairs as these products are interactive and therefore require only partial linguistic information. Terminology products have little or no linguistic ability and therefore the information only relates to the character sets which the product supports." (OVUM report) Normalized weight: 0.1

Metrics:

- *Ability to add new languages*

Method: Study the documentation to discover whether it is possible to add new languages or language pairs and whether this can be achieved by a user or is task for the developer/vendor of the tool

- *Languages supported*

Method: For each component tool of the product (MT, terminology management, translation memory etc) run the tool on texts, or other relevant resources in a variety of languages, and record whether it was possible to treat that particular language.

ADDITIONAL QUALITY CHARACTERISTICS (NOT SUGGESTED BY FEMTI)

Maintainability: The capability of the software product to be modified. Modifications may include corrections, improvements or adaptation of the software to changes in environment and in requirements and functional specifications. (ISO 9126: 2001, 6.5).

Normalized weight: 0.0

Metrics:

No selected metrics for this quality characteristic

Linguistic resources and utilities: This characteristic is concerned with linguistic resources such as bilingual dictionaries (lexicons), vocabulary lists, terminology, grammars and corpora along with the utilities to enable the user to use or modify the resources as well as to add new resources. This #internal# characteristic considers the existence and availability of the resources and utilities. Questions of their usefulness, efficiency and ease of use are considered under the so-called external characteristics which are properties of the running system. "In order to provide users with a working system adapted to their environments, many translation technology products provide add-on dictionaries in certain subject areas and languages. Linguistic resources may also include the ability to create other bilingual, multi-lingual or reversible dictionaries to provide terminology quickly in other language pairs. The ability to enter additional information to the dictionaries or terminology database is also reviewed." In order to ensure that terminology is consistent between multiple translators working in the same target language, it is essential for the product to offer facilities whereby the terminology can be shared and re-distributed as required. The way in which the product provides multi-user access to terminology is documented, together with any utilities for generating printouts and reports of the dictionary or terminology database contents.(OVUM report).

Normalized weight: 0.0

Metrics:

No selected metrics for this quality characteristic

Grammar - Syntax: Degree to which the output respects the reference grammatical rules of the target language.

Normalized weight: 0.0

Metrics:

- *ALPAC measure*

Method: 5-point scale of syntactic correctness.

## Annex 2-FEMTI Report: Operative Evaluation

EVALUATION TYPE

---

- Operational evaluation: Operational evaluations generally address the question of whether an MT system will actually serve its purpose in the context of its operational use. The primary factors include the cost-benefit of bringing the system into the overall process (costs).

CONTEXT CHARACTERISTICS

---

Organisational user: An organisational user of MT may be a corporate user, a translation service, a translation agency or other provider of translation.

Asynchronous communication: In the case of asynchronous or delayed communication the interaction between participants occurs with interruption, for example by email.

Novice: " The reader can identify an increasing number of highly contextualized words and/or phrases including cognates and borrowed words, where appropriate. [...] [May have] sufficient control of the writing system to interpret written language in areas of practical need. Where vocabulary has been learned, can read for instructional and directional purposes, standardized messages, phrases, or expressions, such as some items on menus, schedules, timetables, maps, and signs. " (ACTFL 1983 guidelines for reading proficiency).

Quantity of translation: This concerns the volume of translation typically dealt with by the organisation.

Computer literacy: This refers to the degree to which the user is at ease in computer use and manipulation.

Proficiency in target language: This refers to proficiency in the source language as attested by some recognised measurement. The level of proficiency may be measured, for example by local education tests, internationally recognised examination schemes or organisation internal testing. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines. Depending on the operations performed on the translation, it is either the reading or the writing proficiency which are more specifically relevant. We propose to use the ACTFL eading proficiency scale (1985) -- note that only the guidelines for writing/speaking have been recently updated.

Proficiency in source language: This refers to proficiency in the source language as attested by some recognised measurement. The level of proficiency may be measured, for example by local education tests, internationally recognised examination schemes or organisation internal testing. Two of the best known language proficiency scales are the ACTFL guidelines (first proposed in 1983 by the American Council for the Teaching of Foreign Languages) and the the ILR (FSI) proficiency scale, a five-level scale originally developed by the Foreign Service Institute (FSI) of the United States government, and later adopted by other services under the name of Interagency Language Roundtable (ILR) scale. The scale proposed for use in FEMTI is based on the ACTFL guidelines for reading (1985) -- note that only the guidelines for writing/speaking have been recently updated.

User characteristics: This covers the characteristics of users in three senses: the end user who will interact with the machine translation system; the end user of the final product of the translation process which may include for example, post-editing; the organisation deploying the machine translation system. Note however that in the case when machine translation is combined with substantial post-editing, the resulting "system" might no longer fall under the scope of FEMTI, hence the end users are no longer users of a machine translation system.

Time allowed for translation.: This concerns the deadlines for translation production typical within the organisation.

Evaluation requirements:

Superior: " Able to read with almost complete comprehension and at normal speed expository prose on unfamiliar subjects and a variety of literary texts. Reading ability is not dependent on subject matter knowledge, although the reader is not expected to comprehend thoroughly texts which are highly dependent on knowledge of the target culture. [...] Occasional misunderstandings may still occur; for example, the reader may experience some difficulty with unusually complex structures and low-frequency idioms. [...] Material at this level will include a variety of literary texts, editorials, correspondence, general reports, and technical material in professional fields. Rereading is rarely necessary, and misreading is rare. " (ACTFL 1983 guidelines for reading proficiency).

Number of personnel: This concerns the number of personnel within the organisation who will be directly concerned with the use of the MT system.

## QUALITY CHARACTERISTICS SUGGESTED BY FEMTI

---

Input to Output Translation Speed: This characteristic concerns the amount of time it typically takes the system to carry out the whole translation process including any pre-processing which the system might perform automatically.

Normalized weight: 1.6
Metrics:
No selected metrics for this quality characteristic
Overall Production Time: This characteristic concerns the time between the request for a translation and reception of the final translation.
Normalized weight: 0.6
Metrics:
No selected metrics for this quality characteristic
Maintainability: The capability of the software product to be modified. Modifications may include corrections, improvements or adaptation of the software to changes in environment and in requirements and functional specifications. (ISO 9126: 2001, 6.5).
Normalized weight: 0.2
Metrics:
No selected metrics for this quality characteristic
Introduction cost: TBD
Normalized weight: 0.1
Metrics:
No selected metrics for this quality characteristic
Cost: Cost here covers all of the monetary costs of introducing MT, maintenance costs implied by operational use of the system and the potential costs of not introducing MT. Normalized weight: 0.1
Metrics:
No selected metrics for this quality characteristic
Portability: The capability of the software product to be transferred from one environment to another.
Normalized weight: 0.1
Metrics:
No selected metrics for this quality characteristic
Usability: The capability of the software product to be understood, learned, used and attractive to the use, when used under specified conditions.
Normalized weight: 0.1
Metrics:
No selected metrics for this quality characteristic
Other costs: TBD
Normalized weight: 0.1
Metrics:
No selected metrics for this quality characteristic

Languages: "The range of languages which the product supports is a vital selection criterion. In machine translation systems, the languages are classified according to source and target language pairs, due to the need for full linguistic processing capability. In translator workbench products, the languages are not necessarily classified by strict language pairs as these products are interactive and therefore require only partial linguistic information. Terminology products have little or no linguistic ability and therefore the information only relates to the character sets which the product supports." (OVUM report) Normalized weight: 0.1

Metrics:

- *Ability to add new languages*

Method: Study the documentation to discover whether it is possible to add new languages or language pairs and whether this can be achieved by a user or is task for the developer/vendor of the tool

- *Languages supported*

Method: For each component tool of the product (MT, terminology management, translation memory etc) run the tool on texts, or other relevant resources in a variety of languages, and record whether it was possible to treat that particular language.

## ADDITIONAL QUALITY CHARACTERISTICS (NOT SUGGESTED BY FEMTI)

---

Understandability: The capability of the software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use.
Normalized weight: 0.0
Metrics:
No selected metrics for this quality characteristic
Installability: The capability of the software product to be installed in a specified environment. Normalized weight: 0.0
Metrics:
No selected metrics for this quality characteristic

Adaptability: The capability of the software product to be adapted for different specified environments without applying actions or means other than those provided for this purpose for the software considered.

Normalized weight: 0.0

Metrics:

No selected metrics for this quality characteristic

Linguistic resources and utilities: This characteristic is concerned with linguistic resources such as bilingual dictionaries (lexicons), vocabulary lists, terminology, grammars and corpora along with the utilities to enable the user to use or modify the resources as well as to add new resources. This #internal# characteristic considers the existence and availability of the resources and utilities. Questions of their usefulness, efficiency and ease of use are considered under the so-called external characteristics which are properties of the running system. "In order to provide users with a working system adapted to their environments, many translation technology products provide add-on dictionaries in certain subject areas and languages. Linguistic resources may also include the ability to create other bilingual, multi-lingual or reversible dictionaries to provide terminology quickly in other language pairs. The ability to enter additional information to the dictionaries or terminology database is also reviewed." In order to ensure that terminology is consistent between multiple translators working in the same target language, it is essential for the product to offer facilities whereby the terminology can be shared and re-distributed as required. The way in which the product provides multi-user access to terminology is documented, together with any utilities for generating printouts and reports of the dictionary or terminology database contents.(OVUM report).

Normalized weight: 0.0

Metrics:
No selected metrics for this quality characteristic
Security: The capability of the software product to protect information and data so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them.
Normalized weight: 0.0
Metrics:
No selected metrics for this quality characteristic
Time behaviour: The capability of the software product to provide appropriate response and processing time and throughput rates when performing its function under stated conditions.
Normalized weight: 0.0
Metrics:
No selected metrics for this quality characteristic
Corpora: The kinds and number of monolingual, comparable or parallel corpora available. The category of corpus will depend on the style of language modeling and statistical techniques used in the system.
Normalized weight: 0.0
Metrics:
No selected metrics for this quality characteristic
Learnability: The capability of the software product to enable the user to learn its application.
Normalized weight: 0.0
Metrics:
No selected metrics for this quality characteristic

## Annex 3: Questionnaires

### Operational Evaluation

This evaluation has been designed using the FEMTI Framework

* Required

Functionality - Interoperability

---

Formats that can be used to train the system.

Is TAPTA compatible with...format? * Xml.

Mark only one oval.

- Yes
- No

Is TAPTA compatible with...format? * Html. Mark only one oval.

- Yes
- No

Is TAPTA compatible with...format? * Doc.

Mark only one oval.

- Yes
- No

Is TAPTA compatible with...format? * Pdf.

Mark only one oval.

- Yes
- No

Is TAPTA compatible with...format? * Xliff.

Mark only one oval.

- Yes
- No

Is MTH compatible with...format? * Xml.

Mark only one oval.

- Yes
- No

Is MTH compatible with...format? * Html.

Mark only one oval.

- Yes
- No

Is MTH compatible with...format? * Doc.

Mark only one oval.

- Yes
- No

Is MTH compatible with...format? * Pdf.

Mark only one oval.

- Yes
- No

Is MTH compatible with...format? * Xliff.

Mark only one oval.

- Yes
- No

Security - TAPTA

The capability of the software product to protect information and data so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them.

11. Is the training corpus protected against external users*? *

*Protected against unauthorized access, modifications, etc. Mark only one oval.

- Yes
- No

12 Online translations* are not stored on the corpus *

*Online translations are those carried out by users bay means of the Association's Web Portal

Mark only one oval.

- Yes
- No

Corpora does not have to be uploaded to an external remote server* *

*Uploading information into external servers make it vulnerable. Mark only one oval.

- Yes
- No

Users' modifications* are not incorporated into future translations. *

*Users can edit translations directly on the Web Portal. This modifications might not be appropriate for the organisation. Mark only one oval.

- Yes
- No

Security - MTH

The capability of the software product to protect information and data so that unauthorized persons or systems cannot read or modify them and authorized persons or systems are not denied access to them.

Is the training corpus protected against external users*? *

*Protected against unauthorized access, modifications, etc. Mark only one oval.

- Yes
- No

Online translations* are not stored on the corpus *

*Online translations are those carried out by users by means of the Association's Web Portal

Mark only one oval.

- Yes
- No

17. Corpora does not have to be uploaded to an external remote server* *

*Uploading information into external servers make it vulnerable. Mark only one oval.

- Yes
- No

18 Users' modifications* are not incorporated into future translations. *

*Users can edit translations directly on the Web Portal. This modifications might not be appropriate for the organisation. Mark only one oval.

- Yes
- No

Usability - Learnability

Learnability: The capability of the software product to enable the user to learn its application.

TAPTA *

The system offers one or more of the following supports... Check all that apply.

- [ ] User's Manual
- [ ] Help Tips
- [ ] Tutorials
- [ ] Forum
- [ ] Technical Help
- [ ] MTH *

The system offers one or more of the following supports... Check all that apply.

- [ ] User's Manual
- [ ] Help Tips
- [ ] Tutorials
- [ ] Forum
- [ ] Technical Help
- [ ]

Usability - Understandability

Understandability: The capability of the software product to enable the user to understand whether the software is suitable, and how it can be used for particular tasks and conditions of use.

21. TAPTA *

Direct users would be able to train the system without using.

- [ ] Help Tips
- [ ] Tutorials
- [ ] Forums
- [ ] User's Manual
- [ ] Technical Help
- [ ] Other:

22. TAPTA *

Indirect users would be able to translate with the system without using.

- [ ] Help Tips
- [ ] Tutorials
- [ ] Forums
- [ ] User's Manual
- [ ] Technical Help
- [ ] Other:

23. MTH *

Direct users would be able to train the system without using.

- [ ] Help Tips
- [ ] Tutorials
- [ ] Forums
- [ ] User's Manual
- [ ] Technical Help
- [ ] Other:

Efficiency - Cost -TAPTA

Cost here covers all of the monetary costs of introducing MT, maintenance costs implied by operational use of the system and the potential costs of not introducing MT.

Can TAPTA be trained and installed* for free? *

*"Installed" refers to incorporating the translation engine to the client's web page or working environment Mark only one oval.

- ◯ Yes
- ◯ Yes, but up to a limit
- ◯ No

If there is a limit of data usage*, is it superior to what the client would need?

*Corpus size, translated characters or pages, etc. Mark only one oval.

- Yes, the client would not reach the limit
- No, the client would reach the limit

If the client would exceed the limit, is the price superior to the client's budget*?

*The exact amount (known by the evaluator) is private information and cannot be disclosed to the general public. Mark only one oval.

- Yes
- No

27 The Client can add more direct users (administrators, developers) for free. * Mark only one oval.

- Yes
- No

28. Are updates free?

Mark only one oval.

- Yes
- No
- No updates available

Efficiency - Cost - MTH

Cost here covers all of the monetary costs of introducing MT, maintenance costs implied by operational use of the system and the potential costs of not introducing MT.

Can MTH be trained and installed* for free? *

*"Installed" refers to incorporating the translation engine to the client's web page or working environment Mark only one oval.

- Yes
- Yes, but up to a limit
- No

If there is a limit of data usage*, is it superior to what the client would need?

*Corpus size, translated characters or pages, etc. Mark only one oval.

- Yes, the client would not reach the limit
- No, the client would reach the limit

If the client would exceed the limit, is the price superior to the client's budget*?

*The exact amount (known by the evaluator) is private information and cannot be disclosed to the general public. Mark only one oval.

- Yes
- No

Client can add more direct users (administrators, developers) for free. * Mark only one oval.

- Yes
- No

33 Are updates free? *

*If the system does not offer updates, just leave both boxes unchecked. Mark only one oval.

- Yes
- No
- No updates available

Efficiency - Time Behaviour

Time behaviour: The capability of the software product to provide appropriate response and processing time and throughput rates when performing its function under stated conditions.

Setting the System

---

34. Does the system have automatic language recognition? * Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

35. Does the system have a hot key (generally, Enter) to start translating? * Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

36. Can the system translate directly from the document*? *

*Without coping and pasting on the Web Portal Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

Translation Time

---

37. How long takes the system to translate a sentence of about 20 words? * Mark only one oval per row.

| | Less than 3 s. | 3 s. | More than 3 s, but less than 6 s. | 6 s. | More than 6 s. |
|---|---|---|---|---|---|
| MTH | | | | | |
| TAPTA | | | | | |

38 How long takes the system to translate a paragraph of about 80 words? * Mark only one oval per row.

| | Less than 3 s. | 3 s. | More than 3 s, but less than 6 s. | 6 s. | More than 6 s. |
|---|---|---|---|---|---|
| MTH | | | | | |
| TAPTA | | | | | |

39. How long takes the system to translate a whole text of about 300 words? * Mark only one oval per row.

| | Less than 4 s. | 4 s. | More than 4 s, but less than 8 s. | 8 s. | More than 8 s. |
|---|---|---|---|---|---|
| MTH | | | | | |
| TAPTA | | | | | |

Update Time: Training

---

40. How long takes the system to be trained*? *

*Considering the training corpus as reference. Mark only one oval per row.

| | Less than 2 days. | 2 days. | More than 2 days, but less than 4 days. | 4 days. | More than 4 days. |
|---|---|---|---|---|---|
| MTH | | | | | |
| TAPTA | | | | | |

Maintainability - Stability

Maintainability: The capability of the software product to be modified. Modifications may include corrections, improvements or adaptation of the software to changes in environment and in requirements and functional specifications. (ISO 9126: 2001, 6.5).

41. Indirect users cannot introduce changes to the corpus.

*Community Feedback might not always be appropriate for the organization Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

42. Indirect users cannot introduce changes to the Translation Model. Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

43 Indirect users cannot introduce changes to the Language Model.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

Maintainability - Changeability

Maintainability: The capability of the software product to be modified. Modifications may include corrections, improvements or adaptation of the software to changes in environment and in requirements and functional specifications. (ISO 9126: 2001, 6.5).

All of the questions in this section refer to direct users.

44. Regular available updates? Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

45. Is it possible to add more languages to the system? Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

46. Is it possible to introduce changes to the Language Model? Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

47. Is it possible to introduce changes to the Translation Model?

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

48. Is it possible to introduce changes to the translation interface? Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

Portability - Installability

The capability of the software product to be installed in a specified environment.

49. Can the system be trained on MS Window 2008*?

*OS chosen on the basis of client's available resources. Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

50. Can the system be trained using other MS Window* versions?

*OS chosen on the basis of client's available resources. Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

51. Can the system be trained on Linux* environments?

*OS chosen on the basis of client's available resources. Mark only one oval per row.

| | Yes | No |
|---|---|---|
| MTH | | |
| TAPTA | | |

## 1. Well-Formedness Test

**Instructions:** Compare the translations (**Machine Translation 1** and **Machine Translation 2**) in terms of **grammar correctness** and grade each segment in the following way: give 1 point to each segment using the cells in between both translations: 1MB (translation 1 **Much Better** than translation 2); 1SB (translation 1 **Slightly Better** than translation 2); 1 AS (**About the Same** quality); 2SB (translation 2 **Slightly Better** than translation 1); and 2MB (translation 2 **Much Better** than translation 1). Please, do not edit the translations. You can use the Human Reference Translation to help you decide. After grading the segments, you can add comments (common mistakes, interesting translation choices, and so on). These comments are optional. Do not forget to fill in the Consent Form (Consent_Form_participante_2) and the Q_gral Form. Thank you for your collaboration.

| Source | MT 1 | 1 MB | 1 SB | 1 AS | 2 SB | 2 MB | MT 2 | Comments | Ref. |
|---|---|---|---|---|---|---|---|---|---|
| Malaysia's social security scheme expands disability management based on ISSA Guidelines | Régimen de seguridad social de Malasia se expande gestión de la discapacidad basado en las Directrices de la AISS | | | | | | Malasia del esquema de seguridad social expande gestión de discapacidad basado en directrices de AISS. | | El régimen de seguridad social de Malasia ampliará la gestión de la discapacidad basándose en las Directrices de la AISS |
| SOCSO case management to integrate ISSA Guidelines on Return to Work and Reintegration | Manejo de casos SOCSO integrar Directrices de la AISS en volver al trabajo y Reintegración | | | | | | SOCSO de gestión de casos para integrar la Asociación Internacional de la seguridad social directrices sobre la reincorporación al trabajo y la reintegración. | | La gestión de casos de SOCSO integrará las Directrices de la AISS sobre el Regreso al Trabajo y la Reintegración |
| Malaysia's Social Security Organisation (SOCSO), an ISSA member responsible for the main national employee social protection scheme, has committed to further development of its pioneering return-to-work programme by integrating | Organización de Malasia Seguro Social (SOCSO), un miembro de la AISS responsable del esquema principal de la protección social de los empleados nacionales, se ha comprometido a fomentar el desarrollo de su | | | | | | Malasia la organización de seguridad social (SOCSO), un miembro responsable de la AISS nacional principal empleado plan de protección social. se ha comprometido a seguir desarrollando su actitud pionera regreso-- programa de | | La Organización de la Seguridad Social de Malasia (SOCSO), miembro de la AISS a cargo del principal régimen nacional de protección social para los empleados, se ha comprometido a seguir desarrollando su innovador programa de |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| the principles enshrined in the recently-published ISSA Guidelines. | pionero programa de regreso al trabajo mediante la integración de los principios consagrados en las Directrices de la AISS publicados recientemente. | | | | | | trabajo mediante la integración de los principios consagrados en el recientemente publicado directrices de la Asociación Internacional de la seguridad social | | regreso al trabajo gracias a la integración de los principios consagrados en las Directrices de la AISS recientemente publicadas. |
| The resolution setting out a comprehensive action plan was adopted by the Regional Return to Work Conference on Economic Empowerment and Societal Reintegration, held in Kuala Lumpur on 24 and 25 June. | El establecimiento de un plan de acción global de resolución fue aprobado por el retorno a la Conferencia de Trabajo Regionales para el Empoderamiento Económico de las sociedades, la Reintegración, celebrada en Kuala Lumpur el 24 y 25 de junio. | | | | | | La resolución en la que se establece un amplio plan de acción fue aprobado por la Conferencia de reincorporación al trabajo Regional sobre el empoderamiento económico y social reintegración, celebrada en Kuala Lumpur los días 24 y 25 de junio. | | La resolución que establece un plan de acción integral fue adoptada por la Conferencia Regional de Regreso al Trabajo sobre el Empoderamiento Económico y la Reintegración Social, celebrada en Kuala Lumpur los días 24 y 25 de junio. |
| The Conference resolution aims to set standards of disability management in Malaysia and to provide guidance for stakeholders to integrate disability management in their respective organizational policies, human resource practices and programmes. | La resolución de la Conferencia tiene por objeto establecer las normas de gestión de la discapacidad en Malasia y proporcionar orientación a los interesados para integrar la gestión de la discapacidad en sus respectivas políticas de | | | | | | La Conferencia de resolución tiene por objeto establecer normas de gestión de la discapacidad en Malasia y proporcionar orientación para la gestión de las partes interesadas para integrar la discapacidad en sus respectivas políticas de | | El objetivo de la resolución de la Conferencia consiste en establecer normas de gestión de la discapacidad en Malaysia y en ofrecer una orientación a las partes interesadas para integrar la gestión de la discapacidad en sus respectivas políticas institucional |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | la organización, las prácticas de recursos humanos y programas. | | | | | | organización, las prácticas de recursos humanos y programas | | es, prácticas y programas de recursos humanos. |
| The Conference notably agreed to adopt the ISSA Guidelines on Return to Work and Reintegration as a basis for strengthening disability management, and committed to training case managers through the ISSA's Centre for Excellence. | La Conferencia acordó en particular la adopción de las Directrices de la AISS en volver al trabajo y Reintegración de base para el fortalecimiento de la gestión de la discapacidad, y el compromiso de los administradores de casos de formación a través del Centro de la AISS para la Excelencia. | | | | | | la Conferencia, en particular, convino en adoptar directrices AISS sobre la reincorporación al trabajo y la reintegración como base para fortalecer la gestión de la discapacidad, y se comprometieron a gestores de casos a través de la capacitación del centro para la excelencia de la AISS. | | Como resultado de la Conferencia, se acordó, en particular, adoptar las Directrices de la AISS sobre el Regreso al Trabajo y la Reintegración como base para mejorar la gestión de la discapacidad y formar a administradores de casos por intermedio del Centro para la Excelencia. |
| Held regularly since 2007, the SOCSO Conference is a platform to address issues and best practices related to disability, return to work and societal integration in Malaysia, bringing together a range of political, societal and economic actors from the country. | Celebrada regularmente desde 2007, la Conferencia SOCSO es una plataforma para abordar los problemas y las mejores prácticas relacionadas con la discapacidad, regresar al trabajo y la integración social en Malasia, que reúne a una serie de actores políticos, sociales y económicos del país. | | | | | | Mantenido regularmente desde 2007, la Conferencia SOCSO es una plataforma para abordar las cuestiones y las prácticas óptimas relacionadas con la discapacidad. el regreso al trabajo y la integración en la sociedad de Malasia, que reúne a una amplia gama de agentes políticos, sociales y económicos del país | | Celebrada regularmente desde 2007, la Conferencia de SOCSO es una plataforma que permite abordar problemas y examinar las mejores prácticas relacionadas con la discapacidad, el regreso al trabajo y la integración social en Malaysia. |

## Annex 4. Response Tables

### 1. Terminological Test

| Terminological Test: Results | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| **MTH** | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 2 | 2 | 2 | 0 | 2 | 1 | 15 |
| **TAPTA4ISSA** | 0 | 2 | 2 | 0 | 0 | 0 | 1 | 2 | 1 | 1 | 2 | 0 | 2 | 2 | 2 | 2 | 19 |
| **BING** | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 2 | 1 | 9 |

| Terminological Test: Results | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| MTH | 1 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 0 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 20 |
| TAPTA4ISSA | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 0 | 2 | 2 | 2 | 2 | 22 |
| BING | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 0 | 2 | 0 | 1 | 1 | 11 |

| Terminological Test: Results | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | |
| MTH | 1 | 2 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 2 | 2 | 2 | 1 | 0 | 0 | 14 |
| TAPTA4ISSA | 1 | 2 | 2 | 0 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 0 | 2 | 2 | 2 | 2 | 25 |
| BING | 2 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 2 | 0 | 0 | 1 | 9 |

### 2. Readability Test

| Fluency and Adequacy | | | | | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | | |
| | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | |
| MTH | 1 | 1 | 0 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 3 | 2 | 2 | 30 |
| TAPTA4ISSA | 1 | 1 | 3 | 3 | 2 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 3 | 0 | 0 | 2 | 3 | 3 | 3 | 42 |
| GT | 0 | 1 | 0 | 0 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 0 | 0 | 2 | 1 | 3 | 3 | 26 |
| TAPTA4UN | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 2 | 1 | 1 | 23 |

| Fluency and Adequacy | | | | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
| | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | |
| MTH | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 0 | 0 | 2 | 3 | 2 | 2 | 29 |
| TAPTA4ISSA | 1 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 3 | 3 | 2 | 2 | 0 | 0 | 2 | 2 | 2 | 2 | 35 |
| GT | 0 | 2 | 0 | 0 | 1 | 1 | 2 | 2 | 1 | 2 | 0 | 1 | 1 | 2 | 0 | 0 | 1 | 1 | 2 | 2 | 21 |
| TAPTA4UN | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 0 | 0 | 1 | 3 | 1 | 1 | 26 |

| Fluency and Adequacy | | | | | | | | | | | | | | | | | | | | Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | | 9 | | 10 | |
| | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | A | F | |
| MTH | 1 | 1 | 0 | 1 | 1 | 2 | 1 | 1 | 0 | 0 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 23 |
| TAPTA4ISSA | 0 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 2 | 0 | 0 | 2 | 2 | 2 | 2 | 20 |
| GT | 1 | 3 | 0 | 0 | 1 | 1 | 1 | 2 | 1 | 2 | 0 | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 1 | 2 | 20 |
| TAPTA4UN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 | 4 |

### 3. Well-Formedness Test

Comparative Test: Grammar Correctness

| Participant | MTH | | | TAPTA | |
|---|---|---|---|---|---|
| | MB | SB | AS | MB | SB |
| 1 | 5 | 13 | 2 | 6 | 3 |
| 3 | 5 | 11 | 2 | 1 | 10 |
| 4 | 4 | 9 | 6 | 5 | 5 |
| Points | 14 | 33 | 10 | 12 | 18 |
| Points . W | 28 | 33 | 0 | 24 | 18 |
| TOTAL | 61 | | | 42 | |

| Participant | TAPTA | | | MTH | | Control System | |
|---|---|---|---|---|---|---|---|
| | MB | SB | | MB | SB | MB | SB |
| 2 | 4 | 3 | | | | 7 | 10 |
| 5 | | | | 3 | 1 | 6 | 6 |

| | |
|---|---|
| MB | 2 |
| SB | 1 |
| AB | 0 |

## Annex 5: General Information Form: Participants

# CUESTIONARIO PARA PARTICIPANTES

# INFORMACIÓN GENERAL

| Participante nº: 1 |
|---|

Rellena el cuestionario que se presenta a continuación. Las preguntas con un asterisco (*) al lado del enunciado pueden tener más de una respuesta.

### SECCIÓN 1: INFORMACIÓN PERSONAL

[…]

Lengua(s) materna: Español

Variante(s) del español:*Castellano rioplatense

Lenguas de trabajo: Inglés - Español

¿Vives en un país hispanohablante actualmente? Sí

### SECCIÓN 2: FORMACIÓN

**Último nivel de estudios completado:** Diplomatura/Licenciatura/Grado; Otros: Formación en correctora de español (título no oficial)

### SECCIÓN 3: EXPERIENCIA LABORAL

**Experiencia como traductor(a)*:** Autónomo; Autónomo para una agencia de traducción; Empleado en una agencia de traducción

**Experiencia como revisor(a)***

Autónomo; Autónomo para una agencia de traducción; Empleado en una agencia de traducción

**Experiencia en servicios de apoyo a la traducción*:** Gestión de proyectos; Terminología

### SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO?** Sí

**¿Con cuál(es)?*** SDL Trados; Wordfast; OmegaT

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta?** Avanzado: las tres mencionadas

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** Sí

**¿Cuál(es)? *** Otros

**¿En qué forma utilizas estos sistemas?** Producto para posedición

**¿Cuál es tu opinión sobre este tipo de sistemas?**

En las áreas de conocimiento en las que trabajo (Recursos humanos, medicina, turismo, comercialización, educación), estos tipos de sistemas no nos han sido eficaces, nos llevan mucho trabajo de posedición y retraducción.

**¿Alguna vez escuchaste la expresión "acelerador de traducción", referido a los sistemas de TA?**

No, pero suena interesante y lógico, ya que el uso de sistemas de traducción automática pueden ahorrar tiempo de escritura.

**Comentarios personales:**

Por falta de conocimiento y de uso de otras opciones que pueden llegar a ser mejores de los que he utilizado, no soy muy defensora del uso de la traducción automática, pero sé que si se desarrolla un buen sistema cuando hay memorias de traducción y bases de terminologías específicas a algún área de conocimiento o cliente en particular, puede ahorrarnos tiempo de producción y pasar directamente a la etapa de edición. Lamentablemente, en mis trabajos, no he visto ese tipo de desarrollo aún como para aprobar el uso de este tipo de traducción. Esperemos que a futuro, si se sigue con esta idea, en Argentina se aplique como corresponde para beneficiarnos y no obstaculizarnos.

| Participante nº: |
|---|

## SECCIÓN 1: INFORMACIÓN PERSONAL

Lengua(s) materna: español

Variante(s) del español:*España

Lenguas de trabajo: inglés y francés al español

¿Vives en un país hispanohablante actualmente? No

**Si la respuesta es "no": ¿en qué país? ¿cuánto tiempo has vivido allí?** Suiza, 2 años

## SECCIÓN 2: FORMACIÓN

## Último nivel de estudios completado

Diplomatura/Licenciatura/Grado

## Formación en curso

Máster

## SECCIÓN 3: EXPERIENCIA LABORAL

**Experiencia como traductor(a)*** Autónomo para organizaciones o empresas internacionales

**Sin experiencia profesional**

**Experiencia como revisor(a)*** Empleado para organizaciones o empresas internacionales

**Experiencia en servicios de apoyo a la traducción* No**

## SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO? Sí**

**¿Con cuál(es)? *** SDL Trados; Wordfast ; MemoQ ; Multitrans ; Otras : Passolo

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? ***

Avanzado: **SDL Trados/MemoQ**

Medio: **Wordfast/Multitrans/Passolo**

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** Sí

**¿Cuál(es)? * Google Translate; SYSTRANet; Reverso**

**¿En qué forma utilizas estos sistemas?**

**¿Cuál es tu opinión sobre este tipo de sistemas?**

A mí personalmente no me sirven, hacen que vaya más despacio.

**¿Alguna vez escuchaste la expresión "acelerador de traducción", referido a los sistemas de TA?** No.

**Comentarios personales:**

Oí hablar de la post-edición que me parece que es el mejor uso que se podría dar a estos programas, pero por el momento son demasiado básicos y no ayudan a los traductores.

| Participante nº: 3 |
|---|

## SECCIÓN 1: INFORMACIÓN PERSONAL

[…]

Lengua(s) materna: español

Variante(s) del español:*ríoplatense

Lenguas de trabajo: español, inglés, francés

¿Vives en un país hispanohablante actualmente? Sí

## SECCIÓN 2: FORMACIÓN

**Último nivel de estudios completado:** grado

**Formación en curso:** MA

## SECCIÓN 3: EXPERIENCIA LABORAL

**Experiencia como traductor(a)*** Autónomo para una agencia de traducción; Autónomo para organizaciones o empresas internacionales

**Experiencia como revisor(a)*** Autónomo para una agencia de traducción.

**Experiencia en servicios de apoyo a la traducción*** Maquetado

## SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO?** Sí

**¿Con cuál(es)? *** SDL Trados

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? *** Avanzado.

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** No

**¿Cuál(es)? ***

[…]

**¿En qué forma utilizas estos sistemas?**

**[…]**

| Participante nº: 4 |
|---|

## SECCIÓN 1: INFORMACIÓN PERSONAL

[…]

Lengua(s) materna: Español

Variante(s) del español:* Español peninsular

Lenguas de trabajo: Inglés, francés y español

¿Vives en un país hispanohablante actualmente? No

**Si la respuesta es "no": ¿en qué país? ¿cuánto tiempo has vivido allí?**

Inglaterra, durante varios meses y Suiza casi tres años.

## SECCIÓN 2: FORMACIÓN

**Último nivel de estudios completado:** Diplomatura/Licenciatura/Grado

**Formación en curso:** Máster

## SECCIÓN 3: EXPERIENCIA LABORAL

**Experiencia como traductor(a)*:** Autónomo; Autónomo para organizaciones o empresas internacionales

**Experiencia como revisor(a):** Autónomo

**Experiencia en servicios de apoyo a la traducción*:** Maquetado; Otras

## SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO?** Sí

**¿Con cuál(es)? *** SDL Trados

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? *** Experto:

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** No

**¿Cuál(es)? * [...]**

**¿En qué forma utilizas estos sistemas?**

**[...]**

| Participante nº: 5 |
|---|

## SECCIÓN 1: INFORMACIÓN PERSONAL

Lengua(s) materna: Español

Variante(s) del español:* Español Argentina

Lenguas de trabajo: Inglés y español

¿Vives en un país hispanohablante actualmente? **Sí**

**Si la respuesta es "no": ¿en qué país? ¿cuánto tiempo has vivido allí?**

## SECCIÓN 2: FORMACIÓN

**Último nivel de estudios completado:** Máster

**Formación en curso: [...]**

## SECCIÓN 3: EXPERIENCIA LABORAL

**Experiencia como traductor(a)*:** Autónomo; Autónomo para una agencia de traducción; Autónomo para organizaciones o empresas internacionales

**Experiencia como revisor(a)*** Autónomo

**Experiencia en servicios de apoyo a la traducción*** Mantenimiento de las herramientas de TAO; Maquetado; Otras

## SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO?** Sí

**¿Con cuál(es)? *** SDL Trados

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? *** Avanzado

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** No

**¿Cuál(es)? ***

**¿En qué forma utilizas estos sistemas?**

[…]

**¿Cuál es tu opinión sobre este tipo de sistemas?**

[…]

**¿Alguna vez escuchaste la expresión "acelerador de traducción", referido a los sistemas de TA?** NO

**Comentarios personales:**

[…]

| Participante nº: 6 |
|---|

## SECCIÓN 1: INFORMACIÓN PERSONAL

[…]

Lengua(s) materna: español

Variante(s) del español:* rioplatense

Lenguas de trabajo: inglés, francés y español

¿Vives en un país hispanohablante actualmente? No

**Si la respuesta es "no": ¿en qué país? ¿cuánto tiempo has vivido allí?** Vivo en Suiza desde hace dos años.

## SECCIÓN 2: FORMACIÓN

**Último nivel de estudios completado:** Máster

**Formación en curso:** Ninguno

## SECCIÓN 3: EXPERIENCIA LABORAL

**Experiencia como traductor(a)*** Autónomo; Autónomo para una agencia de traducción

**Experiencia como revisor(a)*** Autónomo; Autónomo para una agencia de traducción

**Experiencia en servicios de apoyo a la traducción*** Terminología

## SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO?** Sí

**¿Con cuál(es)? *** SDL Trados; Wordfast; OmegaT

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? * [...]**

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** Sí

**¿Cuál(es)? *** Google Translate; SYSTRANet; Reverso

**¿En qué forma utilizas estos sistemas?** Búsquedas terminológicas; Herramienta integrada a la memoria de traducción

**¿Cuál es tu opinión sobre este tipo de sistemas?**
Cuando están integrados a memorias de traducción, así como a otras plataformas que el traductor use (como, en mi caso, de subtitulado) los sistemas de traducción automática pueden ser útiles para acelerar el proceso de traducción. Sin embargo, en mi experiencia, he notado que una vez que uno cuenta con la opción del sistema de traducción automática es más difícil idear una solución propia que se aleje más del original o de la opción propuesta por el sistema.
**¿Alguna vez escuchaste la expresión "acelerador de traducción", referido a los sistemas de TA?** No.
**Comentarios personales: [...]**

| Usuario nº: 7 |
|---|

**SECCIÓN 1: INFORMACIÓN PERSONAL**

Lengua(s) materna: español

Variante(s) del español:*ríoplatense

Lenguas de trabajo: español, inglés

¿Vives en un país hispanohablante actualmente? Sí

**SECCIÓN 2: FORMACIÓN**

**Último nivel de estudios completado:** Educación secundaria; Diplomatura/Licenciatura/Grado

**Formación en curso [...]**
**SECCIÓN 3: EXPERIENCIA LABORAL**

**Experiencia como traductor(a)***Autónomo para una agencia de traducción**;** Empleado para organizaciones o empresas internacionales

**Experiencia como revisor(a)*** Autónomo para una agencia de traducción

**Experiencia en servicios de apoyo a la traducción*:** Terminología; Maquetado

## SECCIÓN 4: CONOCIMIENTOS DE HERRAMIENTAS DE TRADUCCIÓN ASISTIDA POR ORDENADOR (TAO)

**¿Has trabajado con alguna herramienta de TAO?** Sí

**¿Con cuál(es)? *** Wordfast

**¿Cuál es tu nivel de experiencia? ¿Con qué herramienta? *** Avanzado:

**¿Sueles utilizar, o has utilizado, algún sistema de traducción automática (TA)?** No.

## Annex 6: Corpus for Time Behaviour Test

Translation Time:
These samples have been extracted from the Corpus for Testing of the Association.

PHRASE

The track record of the Americas in driving the innovative design and delivery of social security programmes is widely acknowledged.
Word count: 20 words.

Paragraph:

In spite of the financial challenges to coverage extension, the recent evidence of social security's impact in the Americas is a positive one, witnessed via a reduction in poverty levels and inequalities, in particular for primary health care indicators. Historically, income distribution in Latin American and Caribbean countries was one of the most unequal in the world. However, over the last ten years, the situation has generally improved in most, but not all, countries. This improvement has often been accompanied by significant increases in public social spending.

Word count: 87.
TEXT:

SOCIAL JUSTICE
No social justice without social security. Message of the ISSA Secretary General on the UN World Day of Social Justice. On 20 February, the United Nations will observe the World Day of Social Justice. To mark this global commemoration, the ISSA Secretary General has reasserted that there can be no social justice without social security. Message of the ISSA Secretary General on the UN World Day of Social Justice. For the United Nations, the pursuit of social justice for all is at the core of its global mission to promote development and human dignity. In a practical sense for social security systems, the World Day of Social Justice draws global attention to the need to promote efforts to tackle important issues such as poverty, exclusion and unemployment. In tackling such issues, endeavours to build an inclusive society for all should be founded on a more equitable access to adequate income protection, resources, and economic development that promotes equity and social justice, while respecting human rights and fundamental freedoms. The advancement of social justice demands the removal of obstacles to realizing human potential. In important and precise ways these objectives convey core elements of the ISSA's strategic vision of Dynamic Social Security: to promote social security systems that are accessible, sustainable, adequate, socially inclusive and economically productive, and that are based on high-performing, well-governed, proactive and innovative social security institutions. Social security is a fundamental right, and just as there can be no social justice without social security, sustainable and effective social security can be made possible by strengthening the capacity of social security institutions in their pursuit of excellence in social security administration The World Day of Social Justice is a call to all those responsible for social security programmes in all regions to commit to the realization of excellence in social security administration, and to uphold the social dimension of globalization and the vision of social justice for all throughout the world. Hans-Horst Konkolewsky. ISSA Secretary General

Word count: 330.