EVALUATING THE OUTPUT QUALITY OF MACHINE TRANSLATION SYSTEMS: SYSTRAN 4.0

Noa Talaván Zanón
UNED

This paper presents a user-like black-box evaluation of SYSTRAN Premium 4.0., at the same time that it introduces an overall approach to the technique of Machine Translation evaluation. The evaluation of the output quality of a system such as SYSTRAN can give users an idea of the needs that can be covered by such a tool and of the type of uses that are more appropriate and can profit more from such a system.

1. OVERVIEW OF MACHINE TRANSLATION EVALUATION

The object of an evaluation is to determine whether a system responds adequately to given needs and constraints. Evaluating Machine Translation (henceforth, MT) systems is important for everyone involved in the field: linguists and computer scientists need to know if their theories make a difference, users have to decide which system to use and what to expect from it, and commercial developers want to please costumers, but firstly they need to know how well the system performs in real-time. It should be noted that nowadays, the main goal of MT is not to produce perfect high-quality translations, but useful and practical ones for a particular user in a particular context.

The history of MT evaluation is as old as MT itself, so there is a significant body of literature on MT evaluation and on the factors involved. One of the first major MT evaluations was carried out by Pierce and Carroll in 1966 and became known as the famous and pessimistic ALPAC Report, whose conclusions were popularly interpreted as "MT is hopeless"[1]. After a significant stop of several years, other reports, generally more optimistic, have been suggested and performed by experts like Lehrberger and Bourbeau in 1988 or King and Falkedal in 1990, to name just a couple of examples; Lehrberger and Bourbeau analyzed in full detail the functioning of TAUM (Traduction Automatique de l'Université de Montréal), and King and Falkedal utilized text suites in evaluating MT systems.

Despite forty years of research on MT, there is still not a generally accepted, satisfactory and comprehensive evaluation methodology, which would considerably help in enhancing knowledge on the general field of MT and MT evaluation and hence, on the essential issue of making expectations about MT realistic. This is often so because, on the one hand, companies do not want to make their flaws public and, on the other, many evaluation

[1] The main conclusions of the report about the viability of MT were that MT did not provide high quality translations and that it would not be desirable in the near term, since it would be more cost-effective hiring human translators. The worst consequence of the ALPAC report was the closing of most research groups on MT of the time, since economic support stopped due to the report's negative conclusions. In fact, the report recommended putting more effort in scientific research addressing more practical translator's tools, and leaving research on MT a little behind.

methodologies are in private hands. Besides, even if the limitations of MT are currently recognized, and an MT system it is not generally expected to produce a flawless translation without human intervention, some kind of evaluation will always be needed, either to improve a system, or simply to know if it suits the user's needs.

Typically, six types of evaluation are distinguished: **operational** evaluation (which assesses the economic benefits of using a particular system as part of a process in an organisation), **adequacy** evaluation (that allows users to be able to choose the system which best meets their specific needs), **declarative** evaluation (the standard method, so to speak, that looks at how systems perform according to criteria such as accuracy and intelligibility), **diagnostic** evaluation (when the state of a system is assessed to discover where it fails and why), **typological** evaluation (that provides information for developers on which linguistic constructions a system can handle, and typically employs test suites), and **progress** evaluation (when the actual state of a system is assessed with respect to some desired state, or when successive versions of a system are assessed to measure its progress in time).

Distinctions in the literature have also been drawn between **glass box** and **black box** evaluation, depending on whether or not the evaluator has access to the operational modules of the system. The first can be used to measure how well each component performs its specific function; usually it is the system builder or specialist that performs glass-box evaluation, having complete access to the system. In the latter, the evaluator has access only to the final results of the processing and hence can measure only how well the system as a whole performs its task; typically, black-box evaluations are performed by users.

Finally, different types of evaluations can also be distinguished depending on the person who carries them out: evaluation **by translators**, evaluation **by researchers**, evaluation **by developers**, evaluation **by potential users**, evaluation **by recipients**, and even evaluation **by system sponsors** (Saiz, 1995).

Many specific methods of MT evaluation have been developed over the years. The ISLE (International Standards for Language Engineering) project, funded by the European Union and the National Science Foundation of the USA, which started working a few years ago, continues its work to systematize these measures and procedures (ISLE Evaluation Working Group, 2003). The ISLE project builds up schemes that classify various aspects of import for MT, including user needs and system characteristics, with metrics associated with them, so as to measure the different aspects involved. This work is intended to be useful to MT users, evaluators, researchers, and system developers. The ISLE project has organized several workshops exclusively on MT evaluation, namely at LREC in Athens ("Workshop on the Evaluation of Machine Translation", http://www.lrec-conf.org/lrec2000/www.icp.inpg.fr/ELRA/lrec2000.html, 2000), at AMTA in Mexico ("Hands on MT Evaluation Workshop" http://www.isi.edu/natural-language/conferences/AMTA2000.htmt, 2000) at Geneva University (MT-Evaluation Workshop "An Invitation to get Your Hands Dirty", http://www.issco.unige.ch/projects/isle/mteval-april01, 2001) at NAACL in Pittsburgh (Workshop on MT Evaluation http://www.cs.cmu.edu/~ref/naacl2001.html, 2001), at MT Summit in Santiago de Compostela (MT Evaluation Workshop "Who Did What to Whom", http://www.issco.unige.ch/projects/isle/MT-Summit-wsp.html, 2001), at LREC in Las Palmas, Canary Islands (MT Evaluation: "Human Evaluators Meet Automated Metrics" http://www.issco.unige.ch/projects/isle/mteval-may02/, 2002), at USC/ISI, Marina del Rey ("MT Evaluation Workshop", 2003), and at MT Summit IX in New Orleans ("Towards

Systematizing MT Evaluation" http://www.issco.unige.ch/projects/isle/MTE-at-MTS9.html, 2003).

2. METHODS OF EVALUATING OUTPUT QUALITY

To assess the quality of the linguistic performance of any MT system, a number of methods have been proposed in the course of the last decades[2]. One of them is **error counting**, which consists of counting up the errors of a set of output sentences or a text, and then weighing them according to a specific scale; the disadvantage of this method is that what constitutes an error may depend on subjective judgement. Another possible proposal of evaluation is the one that rates output according to scales of **intelligibility**, **accuracy**, **style** and **fidelity**, through cloze techniques, judgements of reading processes, quantifying the time taken to post-edit a text in order to make it intelligible, etc. Also, recent attention has been paid to new tools for MT evaluation, particularly to **test suites** (organised sets of test inputs, especially constructed sentences and structures, used to test the syntactic coverage of the system) and **text corpora** (large quantities of real text containing any kind of linguistic phenomena). While good combinations of test suites may allow developers to assess in a controlled way how systems behave and how they can be improved, running corpora through MT may produce large amounts of data but cannot ensure that a particular linguistic phenomenon is tested (Wagner, 1998).

No matter which method for evaluating MT systems is used, it will have to assess the quality of a particular system. However, the MT quality and accuracy needed for translation purposes is relatively higher than for information purposes, and this fact must be considered before setting about evaluating MT outputs. Another very significant aspect to take into account in advance is that any evaluation of MT linguistic output, even if mathematically measurable, will involve a subjective factor (Lewis, 1997). Thus, each evaluation process has to establish a series of quality requirements that have to be matched to the methods the evaluation process is going to follow, so as to attempt to reduce that subjectivity factor to a minimum. All in all, what needs to be borne in mind is that even if the terminology and typology of MT evaluation varies slightly from one researcher group to another or from one system developer to another, the type of evaluation chosen will determine the methods employed and the type of information to be evaluated and reversely.

3. SYSTRAN AND ITS EVALUATION

From the growing number of MT systems on the software market and on the World Wide Web, a long-established and well-known one that has its own version available for free use on the web was selected: SYSTRAN and its version SYSTRAN Premium 4.0[3]. SYSTRAN was founded in 1968 by Peter Toma, an MT linguist researcher who established in 1957 a company in La Jolla, California, with a product called SYSTRAN, an acronym for System Translation. Soon afterwards, the company was hired to develop Russian to English MT for the US Air Force (USAF). The first SYSTRAN system was tested in early 1969, and since 1970, the system has continued to provide translation for the USAF Foreign Translation Division (Flanagan & McClure, 2002).

---

[2] Obviously, complex and time-consuming schemes for evaluation are more readily undertaken by large organizations, agencies or in-house developers than by end users or customers.

[3] Even if the one that is currently on the market is SYSTRAN Premium 5.0.

During the period 1974-1975, SYSTRAN was used by NASA for the joint US-USSR Apollo-Soyouz space project. In 1975, Toma demonstrated a prototype of English to French MT to representatives of the Commission of the European Communities (CEC), which resulted in a contract to develop MT systems to various European language pairs. Nowadays, the CEC uses more than 12 SYSTRAN MT systems for the translation of internal documents. Likewise, Xerox Corporation began using SYSTRAN in 1978 and it has continued to use the system for the translation of thousands of pages per year, allowing Xerox to launch multilingual products to the global marketplace. In 1981, SYSTRAN developed the Japanese-English pair, and in 1989, it created the utility "Customer Specific Dictionaries", dictionaries that are created by users with their own specific terminology. Other developments led to the bringing of the MT technology to the PC, and in 1995, SYSTRAN Professional for Windows was launched. In 1996, SYSTRAN received a contract from the US national Air Intelligence Centre to develop several Eastern European MT language pairs. Other companies, such as Seiko Instruments Inc. or Ford Motor Company, incorporated SYSTRAN MT to their work and in 1997, BabelFish, the first online translation system ever was launched powered by SYSTRAN's technology. From that moment on, MT usage has been reaching new heights, novel improvements have been gained, more and more corporations have tested the benefits of MT in today's multilingual society and today, 36 SYSTRAN MT pairs are commercially available (Flanagan & McClure, 2002).

As to the system characteristics, SYSTRAN can be described as a Fully Automatic Machine Translation (FAMT) system based on the direct approach[4], containing a certain degree of modularity[5] to the point that nowadays it could almost be classified as a *transfer* system (Hutchins & Somers, 1992). However, even if SYSTRAN appears to hold a clear separation between the phases of analysis, transfer and generation, it cannot be truly characterised as a transfer approach for several reasons, such as the evidence of inconsistency in the application of semantic features, or that, despite the labels, there is no clear separation between the phases of transfer and generation (Yuste-Rodrigo & Braun-Chen, 2001).

Most SYSTRAN language pairs, being widely and internationally used, have been evaluated several times in the course of the years; some of the evaluations have been public and many others have been private. Representative instances of these public evaluations are the one presented by Halliday and Briss (1977), the one commissioned by the C.C.E. and performed by Van Slype (1979), or the one presented by Heid (1988).

---

[4] A direct MT system simply translates source language texts to their corresponding target language texts in a word-for-word manner by means of bilingual dictionaries inserted in the system. Then the resulting TL words are reorganised according to the TL sentence conventions. In order to improve the output quality, some direct MT systems also perform some morphological analysis but they rarely analyse the sentence structure of the SL text. Direct MT differentiates itself from two more MT strategies known as the transfer approach and the interlingua method, the former translates using three stages known as analysis, transfer and synthesis, and the latter translates using one intermediate representation which happens to be universal for all languages and known as 'interlingua'.

[5] 'Modularity contributes to ease the maintenance and reusability of the sources and was thus an essential goal for SYSTRAN, whose linguistic resources are extensive. The redesign has modularised the code so that the output of each module is independent and can be used for external purposes as well as for input to the subsequent module' (Flanagan & McClure, 2002).

3.1. The evaluation

SYSTRAN Premium 4.0, the specific software that has been evaluated in the present study, is described at the SYSTRAN website (www.systransoft.com) as an MT system that '*allows you to manage sophisticated translation projects [...] extends productivity level to the maximum [...] and contains over two million words and twenty terminology-specific domains*'. It is a very practical system for personal and small-business use, since it incorporates itself to the Microsoft Word tool of the Microsoft Office software, and users can cut and paste any text they want to translate in their computers, without having to type it out. However, some users are usually soon discouraged due to some of the imperfections that most translations show, since the system's output is not perfect; details about the output quality and the utility of the system will be discussed in the next sections.

A non-expert user's role was played by the evaluator so as to carry out the evaluation of the system's English-Spanish pair, with the main goal of drawing out a number of conclusions that could guide potential non-specialized users of the system. As it was listed above, there are so many ways of evaluating, that many had to be left behind in the current evaluation; some of them because they would not fit the main goal of the evaluation, and others because it would have been impossible for the assumed user to embark on them, since he/she is supposed to have no access to the internal functioning of the system. The evaluation carried out is based on the selection of three texts of different genres (narrative, tourism, and legal) that contained arbitrary combinations of different phenomena. The reasoning behind this selection was having three representative texts of common standard varieties, useful for the average user, different from each other, and not markedly specific.

A user is free to choose anything that seems important for him/her to be evaluated. However, acting as the average user, the author devised an evaluation of the most common factors this user would take into consideration when he/she is about to buy an MT system for personal or small-business use, such as cost-effectiveness, robustness, and user-friendliness. Besides, since this small study aims at helping users make their own evaluations, this one was a black box evaluation, given that the user envisaged was a non-specialized one, who accesses the system simply to acquire general information and comprehension of common and practical texts, as it was commented above. Hence, this potential user would want the MT system not to make flawless high-quality translations in terms of grammar and style of a very specific text type, but simply to obtain a general understanding of unrestricted texts originally written in a different language. To this end, the translations of a fragment of the beginning of the tale "Little Snow White" by the Grim Brothers (consisting on 380 words), a fragment of a tourist brochure about London (consisting on 348 words), and an extract of a working contract (consisting on 368 words) were undertaken, since they presented different linguistic features that an average person could need to translate for personal or small-business use: a domestic and more subjective text (the tale), one related to work (the contract), and another one that can be associated to both fields (the brochure)[6].

Following the double purpose of this paper, to evaluate the system and to help potential non-specialized users to make their own evaluations, and in order to determine the methods of evaluation used, it needs to be remembered that this is a black box evaluation and that in

[6] Fragments of the three original texts and their corresponding MT versions appear in the *General Findings*' section bellow.

this type of evaluations, it may be distinguished between an overall assessment of quality and a more detailed identification of errors. While the former tends to produce more subjective evaluations, the latter provides more objective practical data. Hence, after reviewing the ISLE MT Evaluation Taxonomy (ISLE Evaluation Working Group, 2003) it was decided to perform the evaluation of the texts on two different levels, sentence level and text level, taking into consideration intelligibility (or comprehensibility), readability, fidelity, error analysis (or post-editability), and accuracy (or classification of errors, considering punctuation, capital letters, morphology, lexis, syntax and style), following the ISLE taxonomy descriptors.

Intelligibility or comprehensibility is one of the most frequently used categories to measure the quality of output, and it expresses how intelligible the output of a translation device is under different conditions, and the ease with which a reader can understand the translation. The method used to assess intelligibility in the present evaluation involved asking three different people to read and evaluate each sentence of each text and each text in its entirety on a scale of 1 to 4 for intelligibility: 'unintelligible' (nothing or almost nothing of the message is comprehensible), 'barely intelligible' (only a part of the content is understandable, representing less than fifty per cent of the message), 'fairly intelligible' (the major part of the message passes as intelligible), 'very intelligible' (all the content of the message is comprehensible, even if there are errors of style and/or spelling, and if certain words are missing, or are badly translated but close to the target language.

Readability is a comparison of the time it takes to read a text translated by an MT system and the reading time spent in reading a human translation of the same text. This method was used as an indication of both the readability and the intelligibility of the translated texts in the present evaluation, since reading speed should increase along with intelligibility. This comparison was analysed through the three texts, as compared to human translations in terms of reading times of three people who knew nothing about the texts. Then, the amount of time necessary to achieve sufficient understanding to answer basic questions about the text was also measured, and the words read per minute (WPM) by the different individuals were calculated (WPM = number of words / reading time). Another test for readability performed was the Cloze Technique. This method is linked with textual cohesiveness and measures the success of a reader in replacing words that have been deleted from a translated text. Three subjects were asked to fill in the blanks and then the score was determined by the number of correct responses (either the exact word or any word that yields a paraphrase of the original text). The words deleted were chosen at random every 6 to 12 words. The readability of the translated text, as measured by this technique, is also assumed to be correlated with intelligibility.

In order to assess **fidelity**[7]**,** Lehrberber & Bourbeau (1988: 208) follow John B. Carrol, who devised an indirect method for measuring fidelity,

> 'based on the informativeness[8] of the original relative to the translated text: after digesting the meaning of the latter, the rater is then asked to read the original text and see how much information it adds to the translated version. If the original is very informative relative to the translated text, the fidelity is low, and if it adds little or no information, the fidelity is high.'

[7] Defined by the ISLE as 'the accurateness and completeness of the information conveyed'.

[8] Defined by the ISLE as 'semantic fidelity'; it questions whether the output reflects the content of the source text and whether distortions of meaning occur.

Based on this method, the fidelity of the texts used in the evaluation was measured on a scale from 1 to 4 in terms of the informativeness of the original relative to the translation: 1 for very informative, 2 for rather informative, 3 for not very informative, and 4 for not informative at all. As a final method to test fidelity, to see the extent to which the translated text contained the "same" information as the original, back translation was used, translating the output back into the original language, and comparing the result with the original text; at this stage, many of the shortcomings were magnified by the double process.

Error analysis or post-editability is a method of evaluation that has apparently not been used very much so far (Wagner, 1998). However, the quality of a translation should be reflected in the time needed for correcting errors. This method has been adapted to the purposes of the present evaluation, and this analysis has been measured by attempting to make the minimal number of corrections (deletions, substitutions, additions, rearrangements, etc.) necessary to render the MT "raw" translation output acceptable for information purposes. Errors were simply counted and not weighed, since 'a weighting according to improbability is useless if there is no cooperation with the developer' (Wagner, 1998). This metric is based on the intuition that the time required to produce an acceptable translation from a raw MT output is inversely proportional to the overall quality of the raw translation. In the present study, the following measurement was used to compare the three texts: (number of minutes spent in correction) / (total number of words in text) x 10= correction time. At the same time, each addition or deletion of a word was counted, as well as each substitution of one word by another, etc., and the percentage of corrected words in the whole text was calculated.

Finally, the evaluation of **accuracy** or classification of errors was performed according to the following categories: lexical errors, syntactic errors, untranslated words, morphological errors, wrong punctuation, wrong use of capital letters, and stylistic errors. Bearing in mind that a given translation error may be the manifestation of various interrelated linguistic phenomena, the most characteristic or most straightforward phenomenon was chosen, making the distinction, when necessary, between Source language (henceforth SL) phenomenon and Target Language (henceforth TL) phenomenon, so as to improve the understanding of the system's capability and its linguistic performance.

4. GENERAL FINDINGS

Having submitted the three representative texts to the system and having analysed the results of their corresponding evaluations, the first aspect to highlight is that the overall output of all three texts is rather comprehensible and intelligible, in spite of the existence of a certain number of errors of different types, with different degrees of significance in the three texts.

Thus, after undertaking a detailed analysis of the target texts produced under the different evaluation methods and techniques discussed above, a series of aspects related to several relevant findings need to be pointed out.

-First of all, the translation of the following types of linguistic structures between English and Spanish cause errors in terms of **accuracy**: TL negative structures, TL subjunctive structures, SL structures with "to-infinitive", SL complex noun phrases and noun + noun strings, SL multi-word verbs, anaphoric relationships of gender and number within sentences, and the maintenance of the SL text capital letters. There were other

specific errors, but these were the most repeated ones. Some representative instances of these accuracy errors are listed below:

TL negative structures:

Example taken from the tale:

> The Queen was horrified, and from that moment envy and pride grew in her heart like rank weeds, *until one day she called a huntsman and said "Take the child away into the woods and kill her, **for I can no longer bear the sight of her**. And when you return bring with you her heart,* that I may know you have obeyed my will."

This was translated as:

> Horrorizaron a la reina, y a partir de ese momento la envidia y el orgullo crecieron en su corazón como malas hierbas espesas, hasta un día ella llamó un huntsman y una "toma dicha el niño ausente en las maderas y le mata, porque no puedo ningún oso más largo la vista de ella. Y cuando usted de vuelta trae con usted su corazón, que puedo conocerle ha obedecido mi voluntad.".

TL subjunctive structures:

Example taken from the tale:

> Long, long ago, in the winter-time, when the snowflakes were falling like little white feathers from the sky, a beautiful Queen sat beside her window, *which was framed in black ebony,* and stitched. As she worked, she looked sometimes at the falling snow, and so it happened that she pricked her finger with her needle, so that three drops of blood fell upon the snow. How pretty the red blood looked upon the dazzling white! The Queen said to herself as she saw it, "Ah me! If only I had a dear little child as white as the snow, as rosy as the blood, and *with hair as black as the ebony window-frame."*

This was translated as:

> *Larga, largo hace, en el invierno, cuando los copos de nieve caían como* pequeñas plumas blancas del cielo, una reina hermosa se sentó al lado de su ventana, cuál fue enmarcado en ébano negro, y cosido. Mientras que ella trabajó, ella miraba a veces la nieve que caía, y así que sucedió que ella pinchó su dedo con su aguja, de modo que tres gotas de la sangre bajaran sobre la nieve. *¡Cómo es bonito la sangre roja miraba sobre el blanco del deslumbramiento! ¡La reina dijo a se como ella la vio, "amperio hora yo! **Si*** solamente tenía un pequeño niño querido tan blanco como la nieve, tan atractiva como la sangre, y con del pelo negro tan como el ventana-marco del ébano."

SL complex noun phrases and SL structures with to-infinitive:

Example taken from the tourist brochure:

> London is made up of many varied and quite distinct districts, all offering a unique selection of attractions, places to stay and numerous places to eat: Bankside, stretches from Southwark Bridge to just beyond Tower Bridge, enhancing both old and new London. Look out for *The Tower of London*, the new Millennium Bridge, *Shakespeare's Globe Theater*, the Tate Modern, Southwark Cathedral, and plenty of fashionable shops and restaurants at Hay's Galleria.

This was rendered as:

> Londres se compone de muchos variados y de districtos absolutamente distintos, todo ofreciendo una selección única de atracciones, lugares para permanecer y los lugares

> numerosos a comer: Bankside, estiramientos del puente de Southwark a justo más allá del puente de la torre, realzando Londres viejo y nuevo. Mire hacia fuera para La torre de Londres, el puente nuevo del milenio, Teatro del globo de Shakespeare, el Tate moderno, catedral de Southwark, y un montón de tiendas y de restaurantes de moda en Galleria del heno.

Noun+noun strings:

Example taken from the contract:

> Between CATERING & INDUSTRIAL PERSONNEL LIMITED, an Employment Business (HEREINAFTER REFERRED TO AS "WE) AND THE TEMPORAL WORKER (HERIENAFTER REFERRED TO AS "YOU")

This was translated as:

> Entre el ABASTECIMIENTO y PERSONAL INDUSTRIAL LIMITADOS, un negocio del empleo (MÁS ABAJO DESIGNADO "NOSOTROS") Y EL TRABAJADOR TEMPORAL (HERIENAFTER REFERIDO COMO "USTED")

SL multi-word verbs:

Example taken from the tale:

> But as time passed on, Little Snow-White grew more and more beautiful, until when she was seven years old, she was as lovely as the bright day, and still more lovely than the Queen herself, so that when the lady one day asked her mirror-

This was translated as:

> Sino como el tiempo pasado encendido, poco Nieve-Blanco creció más y más hermoso, hasta cuando ella era siete años de viejo, ella era tan encantador como el día brillante, y aún más encantador que la reina misma, de modo que cuando la señora un día pidió su espejo.

Anaphoric references:

Instance taken from the contract:

> *We shall pay to you remuneration calculated at not less than a minimum* hourly rate of £3.60 per hour, which shall be notified to you on a per assignment basis for each hour worked to be paid weekly subject to such deductions relating to PAYE as are required by Section 134 of the Taxes Act 1988 and all other such deductions required by law to make.

This was rendered as:

> Le pagaremos la remuneración calculada en no menos que un precio por hora mínimo de £3.60 por la hora, que será notificada a usted en a por la base de la asignación para cada hora trabajada para ser pagado semanalmente conforme a tales deducciones referente a PAYE como son requeridos por Section 134 del acto 1988 de los impuestos y de el resto de las tales deducciones requeridas por la ley para hacer. (Instead of 'referentes')

Capital letters:

Example taken from the brochure:

> The East End has a vibrant artistic scene, one of London's hippest districts with contemporary bars, restaurants, shops and markets. **Greenwich**, to the south and east is home to the Meridian Line, at the Royal Observatory, the National Maritime Museum and the infamous Dome. In **Holborn**, buildings date back to the 15th century. This is London's legal epicenter.

> ***Islington*** is traditionally the home to non-conformist, actors, artists, journalist and politicians.

This was translated as:

> El extremo del este tiene una escena artística vibrant, uno de los districtos ma's hippest de Londres con las barras contemporáneas, restaurantes, tiendas y mercados. **Greenwich**, al sur y al este es casero a la línea meridiana, en el observatorio real, al museo marítimo nacional y a la bóveda infame. En **Holborn**, los edificios datan del décimo quinto siglo. Éste es epicenter legal de Londres. **Islington** es tradicionalmente el hogar al disidente, a los agentes, a los artistas, al periodista y a los políticos.

-Secondly, it is also quite evident that the system does not have sophisticated semantic capability (perhaps only semantic markers) as it can be derived from the frequency of **lexical errors** (mainly due, apparently, to the system's inability to deal with inner-categorical homography, its limited treatment of cross-categorical homography, and the problems related with preposition attachment[9]) in all three texts. Instances of these lexical errors are enumerated in the list below:

Inner-categorical homography: *bars* repetitively rendered as *barras* instead of *bares* in the brochure, or *the right* translated by *la derecha* instead of *el derecho* in the contract.

Cross-categorical homography: *the dazzling white* as *el blanco del deslumbramiento*, in the tale, although it goes together with a syntactic problem of unrecognition of the internal structure of the Noun Phrase.

Preposition attachment: *as time passed on* as *como el tiempo pasado encendido* in the tale, as it has already been commented.

-Thirdly, words unrecognised by the system are left **untranslated** in the TL text, in their SL form, for example, *Queen* in the tale, *vibrant* in the brochure, or *section* in the contract.

-Fourthly and finally, in terms of intelligibility, readability, fidelity and error analysis, the evaluations undertaken were rather successful. To have an idea of the results obtained with all the types of measurements exposed before, the following summary list is presented below:

In terms of **intelligibility**, in the tale, at least fifty per cent of the message can be considered intelligible, the main part of the brochure text passes as intelligible, and the major part of the message contained in the contract passes as intelligible. Hence, the system is much better prepared for translating close texts with specific vocabulary, with a consistent use of denotative lexis and clear syntax, such as legal texts.

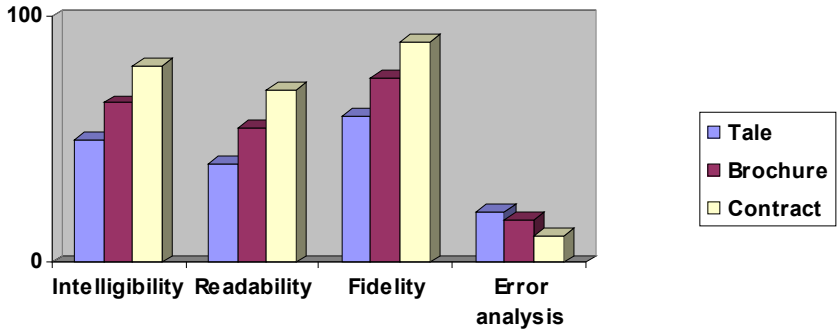
As far as **readability** is concerned, the different tests performed (with all the data related to the reading times of each evaluator and compiling all the evaluators' results for every single text) lead to the conclusion that, as happens with intelligibility, the highest degree of readability corresponds to the contract, going down to the brochure, and then to the tale, which is reasonable due to the various aspects explained up to here, and taking into account that the only text of the three for which the system contains a

[9] The ability to identify a part of a given preposition as the complement of a noun, the argument of a verb, a sentence adverbial, a multi-word verb or as part of a prepositional phrase.

specific dictionary (a business dictionary) is the last one—the others used the so-called "general dictionary" contained in the system.

As regards **fidelity**, the marks proposed for each text were 2.5 for the tale, 3 for the brochure, and 3.5 for the contract, according to the scale explained in the previous section.

Finally, taking into account **error analysis**, the percentage of errors analysed in order to provide a reasonable output with the minimal number of corrections was 15.63 per cent in the tale, 11.38 per cent in the brochure, and 7.8 percent in the contract (not counting the repetitions in any of the texts). And the percentage of the total number of corrected words of the text was 20.43 per cent in the tale, 17.41 per cent in the brochure, and 11.29 per cent in the contract.



5. CONCLUSION

Thus, after considering the ease and speed of the translation process and undertaking a detailed analysis of the system's output for the three texts following the various evaluation techniques, the general conclusion that an average external user would draw from this study is that SYSTRAN is suitable for information purposes for personal and small-business use, given the sufficient degree of intelligibility, readability, and linguistic quality of the target texts. However, it needs to be added that the translations of texts of a type for which the system does not contain specialized dictionaries will include more errors, and the overall quality of the output will generally be lower.

Given the results of the analysis, a series of hypotheses on the improvement of the system by the final user or by the developer can be formulated, when trying to account for some of the errors produced:

> Some of the lexical errors found would probably be easily rectified through direct changes in the dictionary entries, in a system's version with open dictionary modules.
>
> Solving most grammatical errors by the developer could cause a "ripple effect" (Hutchins & Somers, 1992), due to the fact that the resolution of various grammatical errors may open the possibility for other errors to be solved, while others may require adjustment of the basic design of the system.

> Anaphora and cataphora errors can be easy to correct by the user in the majority of cases since it is possible to deduce the referents and correspondents, so these errors are not as relevant as they may seem at first sight, since they are very difficult for the system to "understand" and very easy for the user to change.

In any applied science it is normal to find some problems for which no solution is known at a particular time and, as new systems are developed, the assessment of their ability to cope with specific phenomena is a useful guide for further research and development. For the time being, SYSTRAN can offer the average user rather acceptable translations of different types of texts, as far as informative purposes are concerned. However, specific translations of technical texts for which the system already has specific dictionaries will provide a much better quality of translation output, as it has been observed in the case of the contract. Finally, and taking into consideration the lack of semantic and pragmatic information that exists in the system, we can never expect perfect translations; nonetheless, the informative quality has proved at times unexpected, and most importantly, the informative purposes of the average user will definitely be covered for almost any text with varying degrees of intelligibility, readability, fidelity and accuracy, ranging from standard to high.

References

ARNOLD, D., *Machine Translation: An Introductory Guide,* Oxford, NCC Blackwell, 1993.

FLANAGAN, M./& MCCLURE, S., "SYSTRAN and the Reinvention of MT", *IDCBulletin*, 26459 (2002), http://www.systransoft.com/IDC/26459.html.

HALLIDAY, T./& BRISS, E., "The Evaluation and Systems Analysis of the SYSTRAN Machine Translation system", *RADC-TR-76-399 Final Technical Report,* New York, Griffiss Air Force Base (1977).

HEID, H., *Evaluation der französisch-deutschen SYSTRAN-übersetzung*, Stuttgart, Vorhabenskizze, IMS, 1988.

HUTCHINS, W. & SOMERS, H., *An Introduction to Machine Translation,* London, Academic Press, 1992.

ISLE, Evaluation Working Group, "FEMTI- A Framework for the Evaluation of Machine Translation in ISLE", California, Information Sciences Institute, University of Southern California (2003), http://www.isi.edu/natural-language/mteval/

LEHRBERGER, J. & Bourbeau, L., *Machine Translation. Linguistic characteristics of MT systems and general methodology of evaluation,* Amsterdam, John Benjamin's, 1988.

LEWIS, D., "MT Evaluation: Science or Art?", *Machine Translation Review,* 6 (1997), pp. 25-36.

SAIZ, M., "Issues and approaches in NLP evaluation", *Procesamiento del Lenguaje Natural*, 17 (1995), pp. 289-300.

VAN SLYPE, G. & Pigott, I., "Description du système de traduction automatique SYSTRAN de la Commission des Communautés Européennes", *Documentaliste*, 16 (1979), pp. 150-159.

WAGNER, S., "Small Scale Evaluation Methods", Proceedings of the Workshops *Evaluation von maschinellen Übersetzungssystemen* at the KONVENS-98, Bonn (1998), http://www.ifi.unizh.ch/CL/swagner/SmallScaleAbstract.html.

YUSTE-RODRIGO, E. & BRAUN-CHEN, F., "Comparative Evaluation of the Linguistic Output of MT Systems for Translation and Information Purposes", CD-ROM *Proceedings of the Machine Translation Summit VIII*, *MT Evaluation Workshop,* Santiago de Compostela (2001), http://www.eamt.org/summitVIII/papers/yuste-1.pdf.