revista.tradumàtica
tecnologies de la traducció

118-136

Comparison of SYSTRAN and Google Translate for English→ Portuguese

Rodrigo Gomes de Oliveira
University of Bremen
rdeoliveira.nlp@gmail.com

Dimitra Anastasiou
SFB/TR8 Spatial Cognition
Computer Science & Languages and Literary Studies
University of Bremen, Germany
anastasiou@uni-bremen.de

ABSTRACT

Two machine translation (MT) systems, a statistical MT (SMT) system and a hybrid system (rule-based and SMT) were tested in order to compare various MT performances. The source language was English (EN) and the target language Portuguese (PT). The SMT tool gave much fewer errors than the hybrid system. Major problem areas of both systems concerned the transfer of verb systems from source to target language, and of the hybrid system the word-to-word translation, since its resources are mainly dictionaries and not corpora.

Keywords: statistical machine translation, rule-based machine translation, SYSTRAN, Google Translate.

RESUM *(Comparació de SYSTRAN i Google Tranlate per la combinació anglès→português)*

Dos sistemes de traducció automàtica (MT), un sistema estadístic (SMT) i un sistema híbrid (sistema basat en regles i SMT) es van posar a prova per tal de comparar divereses resultats en MT. La llengua d'orígen era l'anglès (EN) i a llengua de destí el português (PT). L'eina SMT va donar força menys errors que el sistema híbrid. La major part dels problemes d'ambdes eines se centraven en el sistema de transferència del verb de la llengua origen a la llengua destí, i en el cas del sistema híbrid en la traducció paraula a paraula, ja que els seus recursos són sobretot diccionaris i no pas corpus.

Paraules clau: traducció automàtica estadística, traducció automàtica basada en regles, SYSTRAN, Google Translate

RESUMEN (*Comparación de SYSTRAN y Google Translate para el par inglés → portugués)*

Se pusieron a prueba dos sistemas de traducción automática (MT), uno de de ellos estadístico (SMT) y el otro híbrido (basado en reglas y SMT) a fin de comparar diversos resultados en MT. La lengua de origen era el inglés (EN) y la lengua de destino el portugués (PT). La herramienta SMT generó bastantes menos errores que el sistema híbrido. La mayor parte de los problemas de ambos instrumentos se centraron en el método de transferencia del verbo de la lengua origen a la lengua destino, y en el caso del sistema híbrido, en la traducción palabra a palabra, ya que sus recursos son sobre todo diccionarios y no corpus.

Palabras clave: traducción automática estadística, traducción automática basada en reglas, SYSTRAN, Google Translate

1. Introduction

Machine Translation (MT) became one of the first real enterprises to computationally process human language, even before the term computational linguistics was coined. MT research started parallelly to the invention of computers with first attempts of automatic translation dating back to the 1940's (Hays, 1967). Only in the 1950s, in a collaborative work between Georgetown University (USA) and IBM, some sentences in Russian were translated into English. That system composed of about 250 words and 6 grammar rules. Source language strings were represented by punch-card codes. In that very early stage, what linguistics and the recently inaugurated computational science believed was that different human languages could indeed be translated via machines, because natural languages combine nothing more than a lexicon database and a set of rules, be those of basic levels such as morphology, or of higher level such as semantics, or even enclosed in the human encyclopedic, factual knowledge, which comprise the so-called "world knowledge" (Hays, 1967). Since computers were developed to store and manipulate large databases and function under logical rules, that task should have been feasible with the available technology in that moment. Nonetheless, that was the birth of MT and automatically the birth of Rule-Based Machine Translation (RBMT). For various decades, computational linguists attempted to improve system rules based on newer paradigms of linguistic theory. This technique became a new approach around the 1980's when Makoto Nagao proposed in Japan to look at bigger chunks of input in the source language that could be translated into the target language, if bilingual examples of those chunks were already at hand in a certain database (Nagao, 1984). That paradigm did not entirely rely on rules, but on previously existing translation cases, hence the coining of this new paradigm as Case-Based Machine Translation (CBMT). The sub-method that uses a bilingual example database got labeled Example-Based Machine Translation (EBMT). Later in that decade, research in EBMT began and researchers awoke to the possibility of using the output of ongoing research in corpus linguistics by using bilingual corpora as databases for EBMT systems.

In the meantime, computational linguistics had become a research field which gained ground, and work on both mono- and bilingual corpora was booming. CBMT started to look at even bigger input strings than its sub-branch EBMT did, so that not just words or phrases, but whole sentences now were paired within bilingual corpora. These new systems employed Statistical Machine Translation (SMT), because they contained a translation model that initially produced possible translations and an n-gram language model that evaluated those translations. Outputs were then passed by a statistical algorithm that evaluated the probability of those sentences being right, and would finally choose the best output as the final translation result. At present, SMT systems are considered state-of-the-art MT, but many commercial and open source MT tools are combining different methodologies, and thus get the label of hybrid systems.

In this study we tested two MT systems of various architectures: RBMT and SMT. The main idea is to closely inspect the target language output of the same source language input and draw conclusions on the general performance of different systems for the language pair English→Portuguese. In comparison are a commercial system, SYSTRAN, which is traditionally rule-based, but currently uses a hybrid RBMT-SMT approach and provides an online free edition, and the freely available pure SMT system Google Translate. The paper is laid out as follows: in section 2 we describe our methodology discussing our testing data, systems, languages, and the criteria we chose to evaluate MT performance. In section 3 we refer to the evaluation of MT outputs and categorization of mistakes based on morphology, syntax, semantics, pragmatics, and orthography. We summarize and conclude the paper with discussion in section 4.

2. Methodology

In the next sections we provide information about testing data (2.1), the systems (2.2) we compared, languages (2.2), and our evaluation criteria (2.3).

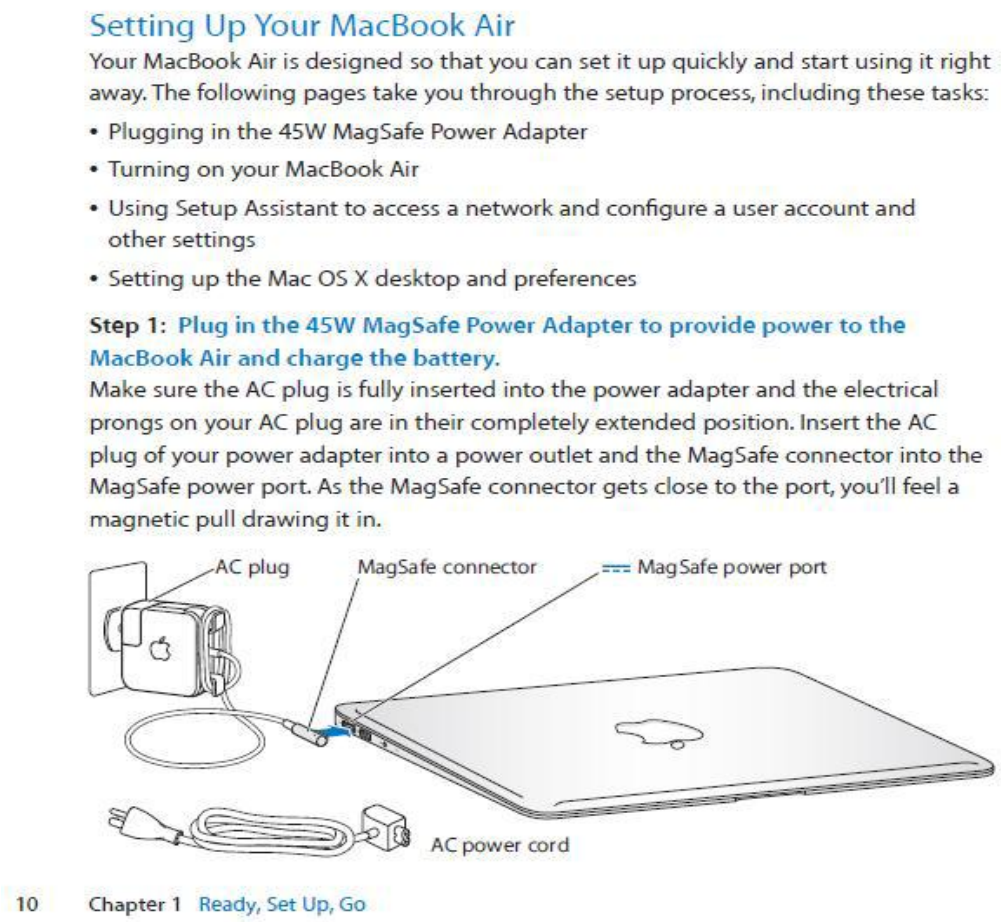
2.1 Testing data

The source data used for this research is a user-directed, instruction manual[1] for a laptop personal computer. Figure 1 shows a page extracted from the booklet/instruction manual, where we see the richness of layout formatting, including different font colors, bullet points, images, and footer details, such as page number and chapter title.

The preparation of the research data consisted basically of eliminating especially images, but also any visual elements added to the core text, including bullet points, colors, bold face, icons and logos. Also, any textual information outside the core text, which included footer information in the original book and text chunks within images were also excluded. Punctuation, or the lack of it, was preserved as in the English (henceforth: EN) original. The segmentation of the text relied on page breaks firstly, and then on line breaks. Line breaks comprise new paragraphs, bullet points (whereof just core text was maintained) and section and subsection titles. The sample data composed of 9 pages of the small booklet, starting at the actual instructive section, i.e. after cover, index, etc., so altogether word count was about 1100 words. Every segment was then translated in isolation with both MT systems under examination.

Test Data 1, following the Figure 1, shows the segmentation of the same page of the booklet. It is crucial here to mention that this text belongs to the technical genre and that being so, one may expect translation not to face artistic levels of language use, such as metaphors or poetical words and/or complex syntax. The text includes technical terms of very recent arrival, which may pose a lexical problem to the MT systems, since there is not a lot of data or dictionaries including the latest technical jargons.

[1] http://manuals.info.apple.com/en_US/MacBook_Air_13inch_Late2010_UG.pdf



Setting Up Your MacBook Air

Your MacBook Air is designed so that you can set it up quickly and start using it right away. The following pages take you through the setup process, including these tasks:

- Plugging in the 45W MagSafe Power Adapter
- Turning on your MacBook Air
- Using Setup Assistant to access a network and configure a user account and other settings
- Setting up the Mac OS X desktop and preferences

Step 1: Plug in the 45W MagSafe Power Adapter to provide power to the MacBook Air and charge the battery.

Make sure the AC plug is fully inserted into the power adapter and the electrical prongs on your AC plug are in their completely extended position. Insert the AC plug of your power adapter into a power outlet and the MagSafe connector into the MagSafe power port. As the MagSafe connector gets close to the port, you'll feel a magnetic pull drawing it in.

10 Chapter 1 Ready, Set Up, Go

Figure 1. A page of the instructions manual (Apple, 2010)

Setting Up Your MacBook Air

Your MacBook Air is designed so that you can set it up quickly and start using it right away. The following pages take you through the setup process, including these tasks:

Plugging in the 45W MagSafe Power Adapter

Turning on your MacBook Air

Using Setup Assistant to access a network and configure a user account and other settings

Setting up the Mac OS X desktop and preferences

Step 1: Plug in the 45W MagSafe Power Adapter to provide power to the MacBook Air and charge the battery.

Make sure the AC plug is fully inserted into the power adapter and the electrical prongs on your AC plug are in their completely extended position. Insert the AC plug of your power adapter into a power outlet and the MagSafe connector into the MagSafe power port. As the MagSafe connector gets close to the port, you'll feel a magnetic pull drawing it in.

Test data 1. Preparation and segmentation of the text for translation

2.2 Systems

SYSTRAN and Google Translate are the systems we tested in order to compare their performances. Some information about the two systems follows below.

SYSTRAN debuted in the market in the 1970s as a robust RBMT system and has been a successful commercial tool still to the present days. Currently SYSTRAN is a hybrid system, which is claimed to combine the predictability and consistency of RBMT systems with the fluency of SMT systems (Senellart, 2009). Another strong claim by the company for commercial targets but also of interesting academic value is that SYSTRAN possesses a learning module, which is used for system training. It extracts sentences from corpora, but rules may get adapted with repetitive use to fit translation domains, so people or parties using the system should gain speed and accuracy of translation with the long term use of SYSTRAN. Last but not least, SYSTRAN has become the translation engine behind the initial translation tool at Google.

Google Translate adopted a full statistical approach and thus abandoned work with SYSTRAN in 2006. The key argument is that translation based on rules has computationally high costs once exceptions of the rules need even more rules to be matched. Google Inc. claims to have adopted a pure SMT machinery to its free, online translation tool that does not rely on any linguistic rules and achieves better results than RBMT does (Google-Inc., 2011). The whole machinery in Google Translate is simultaneously based on and dependent of bilingual text data. This data provides paired corpora for specific language pairs, which are retrieved from large translated text sources, such as United Nations documents, electronic books, and websites. The system then is trained to search patterns of translation in this data and the repetition of this pattern search creates an enormous database of translated material.

The limitation of Google and SMT, in general, lies in the availability of source text and, most especially, in the availability of certain combinations of bi-texts for some language pairs. That is why translation involving certain language pairs will perform better than others. Google's SMT system needs to be fed with large amounts of text to work efficiently and these texts must correspond to the language pair chosen by the user, or the system will start using the interlingua approach, which first translates source language to interlingua (mostly English) and then into a target language.

2.3 Languages

In our study we translated a technical text from English into Portuguese; the first author is native speaker of Brazilian Portuguese. Noteworthy is that both Google Translate and SYSTRAN offer only Portuguese as a language, not European Portuguese or Brazilian Portuguese. The same holds for English; there is not an option to choose between American and British English. In this section we look at some variants of English and Portuguese, and the translation equivalence between them by means of morphology and syntax.

Table 1 discusses the “dichotomy” between American (AmE) versus British English (BrE), where is shown that orthographical peculiarities and word choices mark the identity of those variants.

Type of difference	**AmE**	**BrE**
Verb morphology	Learned, dreamed, spilled (strictly)	Learnt, dreamt, spillt (tendency)
Orthography	Color, theater, recognize	Colour, theatre, recognise
Lexicon	Truck, trunk, (shopping) cart	Lorry, boot, trolley

Table 1. Differences between AmE and BrE (Beare, 2011)

Other variants, such as African English(es), Australian English, Indian English and so forth should not be neglected either.

Portuguese also appeals for a solid dichotomy: European versus Brazilian; Table 2 sets a comparison between Brazilian Portuguese (BP) and European Portuguese (EP), where we see that not only orthography and lexicon, but also syntactical structures mark the identity of the variants. In addition, Portuguese is spoken in Africa and Asia too (Lewis, 2009).

Type of difference	**BP**	**EP**
Morpho-syntax	Eu te amo (preferably) *I you love*	Amo-te *[I-love]-you*
Orthography	Ação, batismo, fato *Action, baptism, fact*	Acção, baptismo, facto
Lexicon	Ônibus, trem, terno *Bus, train, suit*	Autocarro, combóio, fato
Syntax	Estou escrevendo *[I-am] writing*	Estou a escrever *[I-am] at write*

Table 2. Differences between BP and EP (Guimarães, 2005)

In our opinion, it is impossible that MT systems account for all existing language variants of the world, because that would boost the amount of language pairs and the corresponding workload enormously, once all variations at national and regional levels were taken into account. The point is that, when the two chosen MT systems deliberately concatenate all possible variants of English and Portuguese (henceforth: PT) into one language, the systems will have to make choices, since the translated output also needs to be unitary. That may cause natives of different variants to judge the translation quality lower than it deserves, because the system made a choice for a word or structure that natives do not recognize as their own.

This all might be just a small detail of the whole MT complex of problems. However, what happens when different variants present differences in syntax? Will systems account for this peculiarity and remain consistent with translation outputs or will different structures of different variants be at some point mixed in the same output? Even at lower levels, such as word choice for certain things, how consistent can systems be and would inconsistency harm understanding? And even if understanding is not excessively harmed, would professionals publish texts with these translations in their geographical area? The most important factor that may influence translation is perhaps the level of similarity within the language pair chosen. EN and PT are two western European languages that share many similarities due the romanization of the English language and because PT is a Romance language. They are nonetheless two different languages, belonging to two different language families, and thus difficulties in translatability vary.

Below are some examples given by Shoebottom (1996) about the EN-PT equivalence of verb systems, double negation, and syntactic word order, where MT systems often encounter obstacles.

a) The verb systems of EN and PT share many commonalities; however, there are some differences which we highlight below:

- Interrogative sentences in PT are given only by intonation in spoken language. Instead EN uses auxiliary verbs, such as ‘do’, does’, or ‘did’. The equivalent of the EN

question "Do you like me?" is in PT "Você gosta de mim?" (Gloss: *You like of me?*) without any auxiliary verb. Another example is "Did he come yesterday?" which is translated as "Ele veio ontem?" (*He came yesterday?*). In translation from EN to PT (vice versa), the auxiliary verbs may be interpreted as independent lexicon.

Você gosta de mim?	Ele veio ontem?
You like of me?	*He came yesterday?*
Do you like me?	Did he come yesterday?

- EN and PT may sometimes disagree in the choice of the Simple Present and Present Continuous tense. While the sentences "Ela não está sabendo de nada disso" (*She not is knowing of nothing of this*)" would be correct in PT (verb in Present Continuous), in English though the translation should be "She doesn't know about (any of) this" (verb in Simple Present) instead of "She is not knowing anything".

Ela não está sabendo de nada disso

She not is knowing of nothing of this

She doesn't know about (any of) this

- Even though the tenses Simple Past and Present Perfect exist in both languages, the use of the Present Perfect tense may differ across them. To give an example, a PT construction like "Faz tempo que eu não vejo um filme" (*It-does time that I not see a movie*) could generate a translation (through an MT system) in EN like "It's ages since I don't see a movie" instead of the correct translation "It has been some time since I last saw a movie". PT uses the verb "fazer" (to do) to covey duration of actions that started in the past and continue until the present. Another example is "Faz quanto tempo que você está aqui?" the translation of which is "How long are you here already", instead of "How long have you been here already?".

Faz tempo que eu não vejo um filme

It-does time that I not see a movie

It has been some time since I last saw a movie

Faz quanto tempo que você está aqui?

It-does how-much time that you are here?

How long have you been here already?

- Double negatives are permitted in PT, i.e. "Eu não sei de nada" (*I not know of nothing*). However, in EN the equivalent sentence should be "I know nothing/I don't know anything" with only one negation. That may be a problem for an MT systems, if systems maintain the double-negation constraint from EN in the PT version, when double-negation would actually be expected;

Eu não sei de nada

I not know of nothing

I know nothing/I don't know anything

To summarize the above points, when the MT system translates from EN into PT, the i) auxiliary verb in the interrogative sentence has to be omitted, ii) the correct tense has to chosen (Present Continuous vs. S. Present and S.Present vs. Present Perfect), and iii) double negation should be created.

b) Syntax is also an area where natural languages differ in and many translation flaws can happen. NPs' word order as well as the placing of adverbs and pronouns are some characteristic elements of syntactic structures. Like most Romance languages, PT preferably uses NPs with N + ADJ formation, whereas that would be ungrammatical in EN: sentences like "they still are eating" or "that you him give the book" are grammatical in PT, but not in EN;

Eles ainda estão comendo	que você lhe dê o livro
They still are eating	*that you him give the book*
They are still eating	that you give him the book

c) One special feature of EN is the capability of omitting the conjunction (e.g. "that") in reported speech when the reported clause already possesses a subject. This is ungrammatical in PT, but what is the reaction of the MT system, when EN provides a null input in that respect?

d) PT may omit certain object pronouns, so answers like "yes, I like" are accepted, for PT but not in EN. This might be an issue for an MT system that aims at translating into PT and includes these pronouns when it would be stylistically better not to;

Sim, eu gosto

Yes, I like

Yes, I like it/that/her…

e) In rather formal contexts, technical texts for instance, PT often uses one single adjective pronoun that ambiguously designates 3rd or 2nd person (masculine or feminine), what EN distinctively calls "his/her" or "your" respectively. Since PT maintains this "ambiguity", it might be an issue for a MT system that attempts to maintain the genus marking and use some other pronoun that does not belong there at all.

Dimitra e seu carro

Dimitra and her/your? car

Dimitra and her car

Você e seu carro

You and your/his/her? car

You and your car

2.4 Evaluation criteria

In order to evaluate the performance of different MT systems, we choose criteria based on the five[2] distinct levels of language knowledge according to Akmajian et al. (2001):

a) **Morphology:** Morphology includes the correct marking of plurals, case marking, anaphoric pronominalizations, in special the well formation of verbal systems, including right conjugations, right aspect choice, tense choice etc.

b) **Syntax:** Syntax includes, correct structure transformations (e.g. passive x active, question x affirmative x negative, etc.), correct word order per language and context (SVO, SOV), well formation of smaller phrases (are NPs = Adj + N or N + Adj) and other structure-related issues such as the use of little particles between sentence chunks, like prepositions and conjunctions.

c) **Semantics:** Semantics is actually the field of linguistics that studies meaning, so a lot of problems could fall under this category. We will focus, however, on the idea of semantics as assumed by constructions grammars (Bergen & Chang, 2005), which address the connection of certain forms with certain meanings. So, semantic problems here will be seen as the capacity of a system to give/retrieve a meaning or multiple meanings to words or phrases. This includes, for instance, the correct disambiguation of word-intrinsic ambiguities. Under this category, we will also be looking at how named entities (Apple, Windows, Brazil, etc.) are translated. All lexical choices, insertions and omissions in the translated output fall also under this category. Also, if the system chooses another syntactic form to convey some meaning and generates different meanings with that, this error will be assumed to be a semantic error rather than a syntactic one.

d) **Pragmatics**: Pragmatics involves for this research mainly social formalisms conveyed by language, such as treatment pronouns. Further, we included under pragmatic problems, phrases or sentences that are totally grammatical in the target language PT, but are unlikely to be read in natively written texts, so these phrases/sentences would sound unnatural to a native speaker.

e) **Orthography**: This category encapsulates all phenomena which have strictly to do with written texts, such as capitalization and punctuation. Another phenomenon of the phonetic realm was also taken into account pertaining to the orthographic order of errors. It concerns official written conventions of offcial natural languages, which encapsulate certain phonetic processes in their spoken form. EN, for instance, the subject "I" contracts with the verb "am" in a one single written form "I'm". PT, for example, merges certain verbs with certain endings, e.g. all infinitives (-r ending), with an accusative, masculine clitic pronoun "o" by means of a linking consonant "l", that means that when the verb "comer" (eat) is joined with the pronoun "o" (it), it becomes "comê-lo." Hence, systems could produce flawed outputs of the orthographic order, if they return translations such as "I'am" or "comer-o."

3. Evaluation

In order to provide a fair evaluation of the systems under study, the outputs where eye-examined and every error encountered was added to its corresponding category (in accordance to the categories explicated in 2.4). The amount of errors encountered was manually counted and stored. Then, a second analysis of the errors was performed in order to group errors in more detailed subcategories. The idea was to identify very specific linguistic

[2] Phonology would be a sixth category, but since we are dealing with written electronic texts, this category falls outside the chosen criteria.

areas where systems performed faultily, so as to make use of this information for improvement suggestions for the systems.

Table 4 demonstrates what sort of categories were annotated based on errors encountered. In the above table we see how Google outperformed SYSTRAN. Below are some of the subareas where Google performed clearly better than SYSTRAN.

Category	**Error Description**		**Google**	**Systran**
Morphology	Verb system fidelity		6	17
	Morphological agreement		1	3
	Case marking		-	1
	Subtotal		**7**	**21**
Syntax	Particles in general that affect syntax		3	16
	Within-phrase word order		3	4
	Subtotal		**6**	**20**
Semantics	Omission		10	4
	Word choice		8	38
	Proper name translation		-	6
	Insertion		2	3
	Syntactic changes that altered meaning	NP for VP	3	2
		VP for NP	1	1
	Subtotal		**24**	**54**
Pragmatics	Word choice		1	4
	Word order		1	1
	Cohesion		2	-
	Definiteness marking		1	6
	Plural Marking		-	1
	Subtotal		**5**	**11**
Orthography	Punctuation, spacing		3	-
	Clitic binding		1	-
	Subtotal		**4**	**-**
Total			**46**	**107**

Table 3. Preparation and segmentation of the text for translation

PT is a more demanding language than EN in respect to verbal and nominal flexion, so differences between verbal aspect, tense, verb-person agreement are much more salient in PT. The systems had to guess the designated verb form to give in PT as output, and SYSTRAN often failed at guessing more than Google did. That happens perhaps due to the fact that Google does not guess, but looks into real bilingual cases in its database.

In the following sections we give some examples regarding the mistakes in morphology (4.1), syntax (4.1), semantics and specifically proper names (4.3), pragmatics (4.4), and orthography (4.5).

3.1 Morphological agreements in verbal and nominal systems

Systran failed more often than Google when trying to form well many verbal and certain nominal agreements in the target language (PT). Below are some of SYSTRAN's errors:

Original: if you want

SYSTRAN: se você quer(indicative)

if you want

Translation: se você quiser(subjunctive)

EN does not distinguish the present indicative and present subjunctive forms of verbs for the 2nd person singular *you* (in PT, *você*), whereas PT does. The system returned an indicative flexion of the equivalent verb in PT, *quer* (indicative) instead of *quiser* (subjunctive), due to the fact that the flexion *want* in the original text lacks a specific morphological marking for subjunctive.

Original: your… package includes the… drive that contains

SYSTRAN: seu… pacote inclui o… drive que contenha(subjunctive)

your… package includes the… drive that contain

Translation: seu… pacote inclui o… drive que contem(indicative)

Contrary to the first example, the indicative flexion *contains* was translated as the subjunctive flexion *contenha* (instead of its indicative correspondent *contem*). This was caused by the generation algorithm that simply suggests the generation of a subjunctive flexion of verbs after the PT relative pronoun *que* (that).

Original: software for reinstalling… resetting… and using…

SYSTRAN: software para reinstalar(infinitive)… restaurando(gerund)… e usando(gerund)…

software for reinstalling… resetting… and using…

Translation: software para reinstalar(infinitive)… restaurar(infinitive)… e usar(infinitive)…

The original text in EN displays an example of the usage of the gerundive flexion, *-ing* form as in *reinstalling*, after prepositions in prepositional phrases. This concept is conveyed in PT by the infinitive form, such as *reinstalar*, and not by gerund, in PT *–ndo* as in *reinstalando*. The translation was only successful in the first item of the listing, but it then returned literal translations, i.e. from EN gerund to PT gerund, for the other phrases.

Original: on the same

SYSTRAN: no(singular) mesmos(plural)

on-the same

Translation: nos(plural) mesmos(plural)

In the example above, *em* (on) was correctly fused with *os* (the, plural masculine) to produce *nos*. However, the number/gender agreement between the words within the nominal phrase must remain. The output eliminated the plural marking of the fusion *nos* and returned the singular *no*.

PT is a morphologically richer language than EN in respect to verbal and nominal flexion, so differences between verbal aspect, tense, verb-person agreement are much more salient in PT. The systems had to guess the designated verb form to give in PT as output, and Systran often failed at

guessing more than Google did. That happens perhaps exactly due the fact that Google does not guess, but looks into real bilingual cases in its databases.

3.2 Syntactic particles

Small particles in sentence construction like prepositions or conjunctions that are required by previous words in the sentence were just not included, probably because the grammar of the source language (EN) did not include them, so the translation likewise did not either. Moreover, the opposite case happened as well, i.e. particles that should be omitted, were added by the system.

Original: If you know you won't

SYSTRAN: Se você sabe você não

If you know you not

Translation: Se você sabe que (você) não

If you know that (you) not

EN often omits the conjunction *that* between clauses, when the second clause's own subject is declared. PT does not allow this. Systran employed in this case a literal translation, in which the translated sentence lacks the conjunction *que* (that) the verb *sabe* (know) and the subject of the next clause *você* (you).

Original: reinstalling Mac OS X

SYSTRAN: reinstalar Mac OS X

reinstalling Mac OS X

Translation: reinstalar o Mac OS X

reinstalling the Mac OS X

EN may convey a determination by means of other determiners other than the article. PT is very strict in that matter and requires an article determination or a generalization through a plural form. The nominal phrase *MAC OS X* is an incomplete object to the verb *reinstalar* (reinstall), since the determiner in form of the masculine, singular definite article *o* is missing.

Original: see “Migrating…

SYSTRAN: veja da “Migrando…

see of “Migrating...

Translation: veja “Migrando …

see “Migrating…

Here the translation included the preposition *de* (of), fused with a feminine, singular definite article *a*, in the nominal phrase initiated by quotation marks, when in fact the construction should be a simple literal translation from EN, i.e an direct object after the subjunctive verb *veja* (see) without any preposition.

Original: making sure it

SYSTRAN: certificando-se que

certifying-yourself that

Translation: certificando-se de que

certifying-yourself of that

Even though the colloquial usage of the reflexive verb *certificar-se* (make sure) confirms a decline of the preposition *de* (of), standard formal PT remains strict to the mandatory use of *de* after *certificar*. Systran’s output thus would be acceptable in informal texts but it is incorrect in formal texts, which is the style of the text used in this study.

Original: the included AC power cord

SYSTRAN: o cabo incluído da alimentação CA

the cord included of power-supply AC

Translation: o cabo de alimentação CA incluso

the cord of power-supply AC included

Here we face a typical crosslinguistic problem between Germanic and Romance languages, where rearranging adjectives and nouns within nominal phrases often causes confusion to both machine and humans. In the EN phrase, *cable* is the nucleus of the nominal phrase and *AC power* and *included* are attributes between the nucleus and the determiner *the*. PT will contrarily place attributes to the right of the noun. The problem is that *incluso*, being an adjectival flexion of a verb, conventionally takes the final position of the nominal phrase in PT, and not between nucleus and complement as in this output.

3.3 Semantics in respect to word choice

Many of the referents that Systran disposed of in its database did not really correspond to the meaning of the referent in the original text. Table 5 illustrates some of the many errors included in Systran's output.

Original	**Original meaning**	**SYSTRAN**	**Meaning**	**Corrected**
Content	Contents of a box; included parts	Índice	Contents (index) in a book	Conteúdo
be slightly different	Not much; a little; not overall but in a detailed level	ser leve diferentes	Light-weighted; not heavy	ser ligeiramente diferentes
Power Adapter	Electric energy	adaptador do poder	Political or supernatural force	energia
plug in	Insert a pointed device into its matching dock	obstrua dentro	Obstruct something inside something else	insira
It takes	Used to convey time durations of actions	toma	Take something for one's possession	Leva
make the desktop look	To have a certain visual configuration or appearence	fazer o desktop olhar	To look at something with the eyes	fazer o desktop ter a aparência

Table 4. Wrong meanings caused by word choice

Google seems to have access to bilingual data, and so its database is complete enough to provide much fewer word choice errors. Despite Google's performance being again better in this category, Google's major semantic problem regarded translation omissions. Words such as *power adapter*, *plug AC*, *midprocess* were simply not translated but directly transferred to the output as in the original (EN). In other sentences, Google just chose not to include anything in the output, when the original text had a meaningful token. Below are some of the examples:

Original: Working power outlet

Google: tomada elétrica (?)

outlet electric (?)

Translation: tomada elétrica com energia

outlet electric with energy/power supply

Original: the remaining prompts to

Google: os prompts (?) para

the prompts (?) to

Translation: os prompts restantes para

the prompmts remaining to

Original: after you first start up your MacBook Air

Google: depois que você iniciar o seu MacBook Air (?)

after that you start-up the your MacBook Air (?)

Translation: depois que você iniciar o seu MacBook Air pela primeira vez

after that you start-up the your MacBook Air for-the first time

3.4 Proper name translation

Google recognized all entity names of the sample text in this experiment, whereas Systran translated some of them. It is important to mention that names of entities in the text examined were not names of animated beings (people, animals, characters, etc.), but the name of the personal computer about which the instructional manual was written, and names of in-system labels, such as menus and programs. This last sort of proper names, like some geographical sites, do get translated into a correspondent variant across different languages, but their minimal parts should not be treated as separate syntactic terms. For instance, when one wishes to translate "the New Yorker" into PT, one isolates the proper name of the geographical site (New York), applies its corresponding term in the target language (Nova Iorque), and translates the remaining morphemes (-er, -ino in PT), so one has "o novaiorquino" at the end. A faulty translation would, on the other hand, perceive the token "new" in the proper, geographical name as an independent adjective and provide the mistaken translation "o iorquino novo." This last kind of error was more common within SYSTRAN's output.

Original: MacBook Air

SYSTRAN: Ar de MacBook

Air of MacBook

Original: Setup Assistant helps you

SYSTRAN: As ajudas assistentes da instalação

The [assistant helps(plural noun)] of-the installation

Corrected: O Assistente de Instalação ajuda você

The Assistant of Installation helps you

Original: DVD or CD Sharing

SYSTRAN: DVD ou de CD que compartilha

DVD or of CD that shares

Corrected: Compartilhamento de DVD ou CD

Sharing of DVD or CD

Google was efficient enough to perceive when the text referred to the name of one of those as a proper name, given by the company, or when the words where regular nouns and adjectives.

3.5 Pragmatics

No system seems to have performed badly in respect to this category. Google's pragmatic capacity was slightly better than SYSTRAN's, especially in what concerns word choices that are not strange to native readers or noun definiteness marking, a phenomenon which can be very peculiar across languages.

Original: for help

SYSTRAN: para a ajuda

for the help

Translation: para ajuda

for help

Contrarily to what has discussed before, here is an example where PT, to convey certain generalizations, prefers a null determination for a given noun, in this case, *ajuda* (help). Systran chose to include the determiner *a*, when it should have just omitted it.

Original: Setup Assistant

SYSTRAN: assistente da instalação

assistant of-the installation

Translation: Assistente de Instalação

Assistant of Installation

Similar to the last example, the determination by means of *a*, fused with the preposition *de* could have been omitted, leaving the preposition *de* in its original form.

Original: before you first use

SYSTRAN: antes que você use primeiramente

before that you use firstly

Translation: antes de usar pela primeira vez

before of using for-the first time

Systran's output *antes que você use* (before you use) is not the most frequent use in instructive texts and could even suggest something different than the translation suggestion *antes de usar* (before using). PT normally uses an infinitive construction in this context and EN, the gerund.

Original: glow

SYSTRAN: incandescer

glow (can be used for lights and leds, but mostly referred to metal when heated)

Translation: brilhar

glow (used for things that are meant to be light sources or reflective things)

3.6 Orthography

Although problems in this category were too few to be of any statistical significance, the evaluation of system performance gets inverted. Systran's output did not present any error of the orthographic order, whereas Google's output presented a few.

Original: drawing it

Google: puxando-lo

Translation: puxando-o

Original: one or more Mac or Windows computers to partner with your MacBook Air

Google: um ou mais computadores Mac ou Windows, em parceria com o MacBook Air

one or more computer Mac or Windows, in partnership with the MacBook Air.

Translation: (remove comma)

4. Summary and Discussion

The task of fully automated and error-free MT is far from being achieved. Hence we must concentrate at the overall performance of systems to judge their quality, using as main criterion how far from perfection they perform and not whether they perform at a perfect level or not.

The results encountered speak primarily in favor of the SMT system used by Google in comparison with the RB/SMT-hybrid tool provided on the web cost-free by SYSTRAN. This online version allows 10.000 characters per query. Since we sectioned the input text into pages and paragraphs, the word limit was not a problem. It is also important to remember that SYSTRAN provides only a free online version, but paid version are also produced and are claimed by the company to be equipped with more resources that enhance translation outputs). SYSTRAN not only produced a larger number of different mistakes in the automatic translation process, but some of the mistakes were repeated throughout the output. The results at this experiment suggest then that an SMT system performs better than an RB/SMT-hybrid system, because they produce less language-specific mistakes and consequently achieve better levels of text fluency.

Assessing a system's performance with such a birds-eye-view, however, means also to read the translation output as a native, educated adult speaker of the target language and try to retrieve the message from, firstly, the whole text, and secondly, from smaller chunks of it. Adult, educated humans are capable of making out meaning out of certain linguistic inputs that for any imaginable reason are not clear enough. Google's output provided a more fluent reading, thus quicker and less demanding or tiring, but still took the reader to garden path constructions, where the reader needs to go back to the beginning of the sentences and try to

make out some meaning out of it. Consequently, SYSTRAN's output required far more cognitive effort from the reader. Some translated parts by both tools were unintelligible.

The reading of both texts, thus, confirms two facts: (1) that regardless of the amount of errors, texts with construction errors, like both outputs in this MT experiment, will cause some reading problem; and (2) that there is a direct relationship between the amount of errors in a text and how much cognitive effort is demanded from human readers. Nonetheless, reading a text like the output translation by Google is overall understandable enough by native, educated adult speakers of the target language.

Perhaps the borderline between acceptable outputs and desirable outputs is defined in the human choice of publishing that text as official texts. Companies and professionals in general rely on their reputation in all areas of their business to continue surviving in the market. A text that is published is automatically official and also a product of a company's work, so its quality may reflect the quality of the products manufactured by the same. Hence, it is to doubt that big companies, with a large income volume, or professionals that live basically to write texts, would publish any of the outputs produced by both systems experimented in this study.

In order to reverse the situation and make MT systems produce publishable texts, one has to think of the areas annotated by this experiment (see Table 4), as limitations of those MT systems. The free, online SYSTRAN tool clearly holds more limitations than Google's translation tool, but some key areas where SYSTRAN needs to be improved, in our opinion, are the following:

(1) maintaining fidelity of verbal systems with the target language and not with the source language;

(2) improving sentence formation and the use of particles such as prepositions, articles and conjunctions again as required by the target language and not as it is in the source language;

(3) most critically, SYSTRAN's database in its online, free tool returns too many false word-word translations.

According to the results of this experiment, Google's second biggest limitation (but still not as severe as in SYSTRAN) concerned also verbal system fidelity. Its most recurrent error area, however, concerned translation omissions. SYSTRAN also presented some omission problems, but because Systran's online tool was practically overall less effective than Google Translate, this difference is important.

Keeping in mind that the scope of this experiment concerns the language pair EN-PT (in this order), one is inclined to acknowledging the good performance of SMT systems such as Google Translate. Both systems had, however, an acceptable level of output for reading.

Still, Google Translate demonstrates an almost desirable level of output for publishing, which is a remarkable performance for an MT system. Hence, Google Translate can be seen as a powerful tool for domestic use as well as a powerful tool to be used for human translators. A human who wishes to translate texts in EN into PT being aided by this SMT tool will save typing work and will be able to handle mistakes more easily because they occur much less than with other MT systems. Also, because of its huge database, SMT systems may even be more efficient than humans in remembering or suggesting certain translations.

Lastly, it is desirable then that research with MT systems continue so that performance improves. SMT systems seem to be indeed state of the art MT, but when one looks at problems with language-intrinsic things such as verbal systems, one becomes skeptical that pure statistical models will accomplish the task of fully automatically translating texts. Language models shall play a key role in helping statistical models decide between concurring hypothetical structures o translate a given phrase.

For this study, we applied an SMT system and an RB/SMT-hybrid system to automatically translate a technical text to compare MT performances. No translation output was fluent enough to be suitable for publishing, but in general they were intelligible. The SMT tool used, provided by Google, returned much fewer errors than the hybrid system by Systran.

Major problem areas for both systems were the transfer of verbal systems from source to target language and the expected worse performance of hybrid systems to translate referents efficiently, since their databases are dictionaries and not corpora. It is hence admitted that SMT corresponds to the state-of-the-art MT and it is to believe that language models will help SMT systems to give better outputs, where statistical models do not suffice.

5. References

Akmajian, A., Demers, R. A., Farmer, A. K. & Harnish R. M. (2001) *Linguistics: An introduction to language and communication*.The MIT press.

Beare, K. (2011) Differences Between American and British English. http://esl.about.com/od/toeflieltscambridge/a/dif_ambrit.htm

Bergen, B.K. & Chang N. (2005) Embodied construction grammar in simulation-based language understanding. Construction grammars: Cognitive grounding and theoretical extensions,147-190.

Google-Inc. (2011) Inside Google Translate. URL http://translate.google.com/about/intl/en_ALL/.

Guimarães, E. (2005) A Língua Portuguesa no Brasil, http://cienciaecultura.bvs.br/scielo.php?pid=S0009-67252005000200015&script=sci_arttext.

Hays, D.G. (1967) *Introduction to computational linguistics*. American Elsevier, 206-207.

Lewis, M. P. (2009) Ethnologue: Languages of the World, http://www.ethnologue.com/show_language.asp?code=por .

Nagao, M., (1984), “A Framework of a Mechanical Translation between Japanese and English by Analogy Principle”. In Elithorn, A.; Banerji, R. (Eds.), *Artificial and Human Intelligence*, Amsterdam, North-Holland, 173-180.

Senellart, J. (2009) Systran, http://www.b-eye-network.com/watch/11519.

Shoebottom, P. (1996) The differences between English and Portuguese, http://esl.fis.edu/grammar/langdiff/portuguese.htm.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, & J. Makhoul (2006) A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas,* 223-231.