Usefulness of MT output for comprehension — an analysis from the point of view of linguistic intercomprehension

Kenneth Jordan Núñez kjordan@usj.es
Universidad San Jorge, E-50830 Villanueva del Gállego, Spain

Mikel L. Forcada mlf@ua.es
Universitat d'Alacant, E-03690 Sant Vicent del Raspeig, Spain

Esteve Clua esteve.clua@upf.edu
Universitat Pompeu Fabra, E-08018 Barcelona, Spain

Abstract

The present paper describes and presents ongoing research on machine translation (MT) and linguistic intercomprehension. One main goal, although not the only one, is to evaluate three machine translation (MT) systems —Systran, Google Translate and Apertium— through an analysis of the readers' ability to understand the output generated. We compare the usefulness of MT output for comprehension to that of non-native writing in the readers' L_1 and that of native writing in languages similar to their L_1. The methodology used is based on *cloze* tests and the experiments are carried out using English, French and Italian as source languages and Spanish as the target language. The subjects involved are native Spanish first-year-undergraduate students and final-year secondary-school students. All of them have only very elementary knowledge, or in some cases no knowledge at all, of English, French and Italian (that is, a level equal to or lower than CEFRL B1[1]). Although the results suggest that MT output resulting from translating from English and French into Spanish is similar to natively-written Italian texts or texts written in Spanish by non-native speakers in terms of usefulness, that depends quite often on the level of specialty but also on the field and on the MT system used.

1 Introduction

The aim of the study, which is part of a broader research plan relating machine translation (MT) and linguistic intercomprehehsion, is to assess and compare three MT systems when used for assimilation[2] or *gisting* – Systran, a hybrid system[3] that combines statistical or corpus-based MT and rule-based MT; Google Translate,[4] a statistical corpus-based MT system at the time of testing for the language pairs tested,[5] and the Apertium rule-based system.[6] The aim is achieved

[1] Common European Framework of Reference for Languages: Learning, Teaching, Assessment (`http://www.coe.int/t/dg4/linguistic/Source/Framework\_EN.pdf`)

[2] MT systems can be divided into two groups: those aimed at *assimilation* or *gisting*, which allow the user to understand the content of the text, and those aimed at *dissemination*, which helps to translate a text to be published.

[3] `http://www.systran.es`

[4] `http://translate.google.com`

[5] As of May 15, 2017, the English–Spanish system and the English–Italian systems are no longer statistical, but rather neural; the French–Spanish system is still statistical.

[6] `http://www.apertium.org`

by comparing the usefulness of their output to that of non-native writing in the reader's first language L_1 and to that of native writings (or MT output) in a closely related language. This will enable us to determine which of the three MT systems is the most useful for gisting, that is, which MT system results in the highest level of usefulness when reading a text originally written in a language unknown to the reader.

Linguistic intercomprehension, the ability to understand one foreign language based on knowledge about another language (Meissner, 2004, 34), is an ability that the readers of a language naturally have and use unconsciously but that they can also develop in order to understand messages in another language without being able to produce them themselves (Martín-Peris, 2011, 247). Linguistic intercomprehension —in this work, *reading* or *written* intercomprehension, in contrast to *listening* or *spoken* intercomprehension— leverages on the similarity or identity of word forms and structures (Martín-Peris, 2011, 276).

In recent decades, although only in Europe (within the framework of Euro-comprehension or European intercomprehension), a new discipline focused on research into study methods has been developed, which is aimed at the simultaneous studying or learning of multiple languages. This discipline is centred on reading comprehension —as well as listening comprehension in some cases— and seeks to save time and effort when learning languages from the same linguistic family (Clua, 2003). Some of the methods designed have been adapted to Romance-language learning, such as Eurom4 and Eurom5,[7] Galatea,[8] Miriadi,[9] or EuroComRom[10] (within EuroCom, whose name refers straightforwardly to Euro-comprehension). EuroComRom shows the reader how to obtain information from texts in other languages —even if those languages are completely unknown to the reader— through the so-called *seven sieves* (Klein and Stegmann, 2000): international vocabulary or internationalisms, pan-Romance vocabulary, sound correspondences, spelling and pronunciation, basic structure of Romance sentence patterns, common morpho-syntactic structures developed by the Pan-Romance community, and the transfer of EuroFixes[11] (Martín Peris et al., 2005).

The work in this paper explores the quantitative aspects of a line of research that aims at exploring to what extent the differences between native text and MT output are similar to the differences between languages within a language family or the differences between native text and non-natively written text, and, then, explores whether spontaneous intercomprehension strategies (Klein and Stegmann, 2000; Martín Peris et al., 2005) play a role in the processing of raw machine-translated text by readers.

2 Research questions and hypotheses

We aim to answer four research questions: RQ1) To what extent are MT output and a text written by a native speaker in a language L' from the same language family as the reader's first language L_1 similar to the reader in terms of usefulness?; RQ2) Could the MT output in a language L' in the same language family as the reader's first language L_1 be useful for comprehension when translating texts originally written in a language from a language family different from that of their first language L_1?; RQ3) Are MT output and a text written in the reader's first language by non-native speakers similar to the reader in terms of usefulness?; and, lastly, RQ4) Which MT system is the most useful for comprehension when translating a text originally written in a language that is completely unknown to the reader?

[7] http://www.eurom5.com

[8] http://galatea.u-grenoble3.fr

[9] https://www.miriadi.net/en

[10] http://www.eurocomresearch.net

[11] Eurofixes are lexical components used in word formation, such as prefixes and suffixes, and which are shared across European languages, Many of them are Latin- and Greek-based affixes (*pseudo-*, *-phobia*, *inter-*, etc.).

Regarding RQ1, one could argue that the relationship between the reader's L_1 (first language, *mother tongue*) and MT output into L_1 has similarities to that between L_1 and other languages in the family of L_1 (common international vocabulary, common morpho-syntactic structures), and that, therefore, the usefulness of MT output into L_1 is similar to that of the related language L'.

Regarding RQ2, by analysing the usefulness to Spanish (ES) native speakers of MT output when translating from English (EN) into Italian (IT), we can determine if MT could be used to read texts originally written in a language from a different language family when a certain language combination is not available; that is, if a speaker of (ES) wants to understand a text in EN but there is no EN–ES MT available, they could use the EN–IT system and their linguistic intercomprehension abilities.

3 Methodology

We explore the four research questions above using closure or *cloze* test methodology, a method that involves filling in gaps corresponding to single words that have been removed from an extract from a written text.[12] Our study is grounded on these three pillars: (a) reading-comprehension questionnaires with questions like *Who was the president of the Green Party in 2011?* have repeatedly been used to evaluate the usefulness of machine-translated text (Jones et al., 2005, 2009, 2007; Berka et al., 2011; Weiss and Ahrenberg, 2012); (b) cloze testing or gap-filling has extensively been used as an alternative way to measure reading comprehension (Rankin, 1959; Page, 1977); and (c) gap-filling may sometimes be considered to be roughly equivalent to question answering: *In 2011, _____ was the president of the Green Party.* Therefore, we work upon the assumption that gap-filling success measures the usefulness of machine-translated text for comprehension. Inspired in previous work (O'Regan and Forcada, 2013; Trosterud and Unhammer, 2012; Ageeva et al., 2015), the method was used for 65- to 75-word-long target texts previously translated by professionals with gaps every fifth word.[13]

Subjects (native ES speakers with an EN, FR and IT level equal to or lower than CEFRL B1 were provided with different kinds of hints to help them fill in the gaps in professionally translated ES text. The aim was to determine whether the hint was useful for comprehension, as measured by the rate of success in filling out the gaps (that is, the fraction of correctly filled gaps) in the incomplete professionally-translated ES text.

Forty-four test texts were taken from 4 different sources with different degrees of specialization, and were tested in four different hinting situations, distributed as follows:

a) *Using machine translation into L_1 as a hint*: EN–ES and FR–ES MT output[14] — one highly specialized text on natural sciences ('NAT'), one highly specialized text on human and social sciences ('SOC'), one journalistic or informative semi-specialized text ('INF'), and one non-specialized or general text ('NO') that had been translated by the three MT systems, both from EN to ES and from FR to ES (total, 24 texts);

b) *Using text in $L' \simeq L_1$ as a hint*: native or professionally-translated IT texts — four texts written by an IT native professional translator from four EN sources with the same degree of specialization;

[12] Gaps should be filled with either the exact word removed from the source text, or with a synonym or a functionally equivalent unit, that is, a lexical unit that creates a target text with the same meaning as the source in that specific context.

[13] Preliminary experiments showed that the readers' gap-filling performance showed a marked reduction when gaps occurred every five words. Note also that (O'Regan and Forcada, 2013) poke holes with a probability, not periodically as it is done here, but this may be considered be equivalent.

[14] These translations were retrieved in November–December 2014 and it is worth mentioning that online MT systems change over time as models get re-freshed or even when a technological change occurs (such as the recent switch from statistical MT to neural MT for EN–ES and EN–IT)

EN source text (not shown to subjects)	If no formal authorisation has been given by the host state, a third-country national's presence may be considered unlawful by that state. Both EU and ECHR law, however, set out circumstances in which a third-country national's presence must be considered lawful, even if unauthorised by the state concerned.
Hint: Machine-translated ES text	Si no se ha dado ninguna autorización formal por el estado de anfitrión, la presencia de un nacional de terceros países se puede considerar ilegal por ese estado. La ley de la UE y del ECHR, sin embargo, estableció las circunstancias en las cuales la presencia de un nacional de terceros países se debe considerar legal, incluso si es desautorizado por el estado trató.
Problem: Professionally-translated ES text with gaps	Si el Estado de acogida no le ha concedido una autorización formal, dicho Estado **[...]** considerar que la presencia **[...]** un nacional de un **[...]** país es irregular. Sin **[...]**, tanto el Derecho de **[...]** UE como el CEDH **[...]** circunstancias en las que **[...]** presencia de un nacional **[...]** un tercer país se **[...]** considerar legal, aunque el **[...]** miembro de que se trate no la haya autorizado.

Figure 1 – An example gap-filling problem: the subject has to fill with a single word the gaps (one every 5 words) introduced in the professionally-translated text, using the machine-translated text as a hint. The source text is not shown. The solutions in this case are *puede*, *de*, *tercer*, *embargo*, *la*, *establecen*, *la*, *de*, *debe*, *Estado*.

c) *Using machine translation into $L' \simeq L_1$ as a hint*: EN–IT MT output — eight texts from the same four sources that were translated from EN into IT by two of the MT systems (Google and Systran, as Apertium does not provide this combination);

d) *Using non-natively produced L_1 text as a hint*: either an EN text translated into ES by an EN native speaker or a FR text translated into ES by a FR native speaker (both with a ES B2 level according to the CEFRL) — four texts from the same four sources translated from EN into ES by a native speaker of EN and four texts from the same four sources translated from FR into ES by a native speaker of FR (8 texts in total).

Figure 1 shows an example gap-filling problem in which a machine translated text from EN is shown as a hint, that is, an example of the first hinting situation.

All the texts were extracted from institutional publications, that is, works published or translated by staff for linguists in international institutions or organizations with the exception of the non-specialized or general texts, which came from different translations of the Bible in the languages involved. "NAT" texts were taken from the different language versions of WHO document http://www.who.int/pehemf/publications/en/EMF_Risk_ALL.pdf. "SOC" texts were taken from the FRA–EHCR–Council of Europe *Handbook on European law relating to asylum, borders and immigration* (http://fra.europa.eu/sites/default/files/handbook-law-asylum-migration-borders-2nded_en.pdf). "INF" texts were taken from the European Commission publication (http://ec.europa.eu/economy_finance/publications/general/pdf/the_road_to_euro_poster_en.pdf). "NO" texts were taken from four editions of the Bible (books of Job and Esther) in different languages —the Spanish *Dios habla hoy* (1996), the English *Easy-to-Read Version* (1987), the French *Bible Segond 21* (2007) and the Italian *Bibbia Diodati* (1991).

The test was taken by 71 native Spanish undergraduate students from either the first year of the Degrees in Journalism, Audiovisual Communication, Advertising and Public Relations, Translation and Intercultural Communication, Infant Education and Primary Education, or final-

year secondary students. All had only elementary knowledge or, in some cases, no knowledge at all, of EN, FR and IT (that is, a level equal to or lower than CEFRL B1). Students were asked to read the hint and, for the whole test, write a single word in each gap in all the texts. Each job contained 10 gaps; each student completed 22 jobs on average. The fraction of gaps that the students were able to fill successfully (either with the exact word removed or with a synonym) were considered as a measure of usefulness of the text used as a hint (in a scale from 0 to 1) and compared.

4 Results

The results of the test were compared in three blocks: 1) Systran and Google EN–IT and Systran and Google FR–ES MT output with IT texts written by a native speaker for reference; 2) Systran, Google and Apertium MT EN–ES output with texts written in ES by a native EN speaker for reference; 3) Systran, Google and Apertium FR–ES MT output with texts written in ES by a native FR speaker for reference. In each block, five comparisons were made: overall or general (that is, without considering the text type or its degree of specialization), NAT, SOC, INF and NO (see section 3).

In Tables 1 to 5, language combinations for each MT system or the other texts used as a hint were ordered according to the level of usefulness based on the results of the test undertaken by the students. Different data are given: ('m'), the average fraction of gaps correctly filled by the subjects on a scale of 0 to 1; its variance ('var'), the number of observations in each case ('n'), and the estimated probability that the established order for each group was due to chance and not to a real superiority. In other words, the arithmetic mean in each case is used to establish a ranking of hints as regards their usefulness for comprehension; the probability that the hypothesis behind that order or preference is false is also indicated.[15] As usual, when the latter probability shown on each table between two text squares is lower than 5%, we will interpret that the level of usefulness of the first text with respect to the second is clearly higher and not the result of luck.

4.1 Comparing MT output to text in a closely related language

The results in Table 1 answer the first two research questions, that is: RQ1) whether, in terms of usefulness, MT output and a text written by a native speaker in a language L' in the same language family as the reader's first language L_1 are similar; and RQ2) whether the MT output in a language L' in the same language family as the reader's first language L_1 could be useful for comprehension when translating texts originally written in a language from a different language family and much less connected to their first language.

According to these results, in general terms, a human-written IT text is less useful for comprehension than the MT output resulting from translating a text from a related language (that is, FR into ES; and it is basically as useful for comprehension as the MT output resulting from translating a text from a language belonging to a different language family (that is, EN) into IT.

Strangely, however, for NAT texts, a human-written IT text would be marginally more useful for comprehension to a native ES reader than Google FR–ES MT output; and Google EN–IT MT output would also seem marginally more useful for comprehension than Systran FR–ES MT output and clearly more useful than Google FR–ES MT output. Furthermore, for SOC texts, a human-written IT text is as useful for comprehension as Google FR–ES MT output but much more useful than Systran FR–ES MT output. AS to INF and NO texts, human-written IT texts are much less useful than any of the FR-ES MT outputs.

As stated by Jordan-Núñez (2015), the fact that usefulness of NAT IT texts is very similar to that of MT system output, but considerably higher for SOC texts could be due, perhaps, to the

[15] The probability of the null hypothesis (both averages being equal) has been computed using Welch's two-tail *t*-test, https://en.wikipedia.org/wiki/Welch's_t-test.

TOTAL	NATURAL SCIENCES ('NAT')	SOCIAL SCIENCES ('SOC')	JOURNALISTIC ('INF')	NOT SPECIALIZED ('NO')
FR→ES GOOGLE m = 0.7218 var = 0.0254 n = 142	**EN→IT GOOGLE** m = 0.8467 var = 0.0079 n = 30	**ITALIAN** m = 0.6667 var = 0.0354 n = 30	**FR→ES SYSTRAN** m = 0.8463 var = 0.0120 n = 41	**FR→ES GOOGLE** m = 0.8098 var = 0.0239 n = 41
74,30% ↓	10.38% ↓	54.39% ↓	11.86% ↓	41.18% ↓
FR→ES SYSTRAN m = 0.7141 var = 0.0537 n = 142	**FR→ES SYSTRAN** m = 0.7667 var = 0.0451 n = 30	**FR→ES GOOGLE** m = 0.6400 var = 0.0218 n = 30	**EN→IT SYSTRAN** m = 0.7951 var = 0.0310 n = 41	**FR→ES SYSTRAN** m = 0.7829 var = 0.0195 n = 41
4.43% ↓	21.91% ↓	40.23% ↓	17.32% ↓	**0.03%** ↓
EN→IT GOOGLE m = 0.6634 var = 0.0357 n = 142	**ITALIAN** m = 0.7067 var = 0.0248 n = 30	**EN→IT SYSTRAN** m = 0.6033 var = 0.0348 n = 30	**FR→ES GOOGLE** m = 0.7488 var = 0.0156 n = 41	**EN→IT GOOGLE** m = 0.6463 var = 0.0330 n = 41
3,39% ↓	13.64% ↓	15.05% ↓	**1.87%** ↓	**0.13%** ↓
ITALIAN m = 0.6120 var = 0.0469 n = 142	**FR→ES GOOGLE** m = 0.6467 var = 0.0226 n = 30	**EN→IT GOOGLE** m = 0.5400 var = 0.0318 n = 30	**ITALIAN** m = 0.6805 var = 0.0176 n = 41	**EN→IT SYSTRAN** m = 0.5146 var = 0.0308 n = 41
49.10% ↓	**0.00%** ↓	**0.04%** ↓	31.74% ↓	8.76% ↓
EN→IT SYSTRAN m = 0.5937 var = 0.0532 n = 142	**EN→IT SYSTRAN** m = 0.4167 var = 0.0401 n = 30	**FR→ES SYSTRAN** m = 0.3867 var = 0.0274 n = 30	**EN→IT GOOGLE** m = 0.6439 var = 0.0365 n = 41	**ITALIAN** m = 0.4341 var = 0.0578 n = 41

Table 1 – Gap-filling success rates for MT output compared with those for an Italian text written by a native Italian speaker (shaded boxes) and with those for MT output resulting from translating from English into Italian.

deficiencies in the glossaries in the MT systems given the appreciable conceptual and lexical variety within these fields (especially in law, due to partial or even false equivalence in legal terms or institution names).

Likewise, in general terms, the FR–ES MT output is more useful for comprehension than EN–IT MT output (see RQ2). However, depending on the degree and the field of specialization of the text and the MT system used, EN–IT MT output could be more useful to a native Spanish reader.

More specifically, for NAT texts, Google EN–IT MT output would appear to be more useful than the FR–ES MT output, and Systran EN–IT MT output is less useful. For SOC texts, any of the EN–IT outputs is slightly less or as useful than Google FR–ES MT output but far more useful than FR–ES Systran output.

For INF texts, Systran EN–IT MT output would be slightly less useful for comprehension than Systran FR–ES[16] but would be slightly more useful for comprehension than the output resulting from translating a text in the same language combination with Google. The output from translating a text from English into Italian with Google is less useful for comprehension than the output from the same MT system when translating from French into Spanish.

Lastly, for NO texts, all of the EN–IT MT outputs are less useful for comprehension than any of the FR–ES outputs.

[16]Note, however, that because two MT systems are made by the same company they do not have to be similar at all. For instance, while Apertium systems are structurally very similar to each other, their performance varies widely with the language pair or the level of development.

4.2 Comparing MT output with non-native writing in the target language

The results shown in Tables 2 and 3 aim at answering the third research question (RQ3), that is, to demonstrate whether MT output and a text written in the reader's first language L_1 by non-native advanced 'independent speakers' —with a CEFRL B2 EN or FR level— are similar to a native ES reader as regards intelligibility.

TOTAL	NATURAL SCIENCES ('NAT')	SOCIAL SCIENCES ('SOC')	JOURNALISTIC ('INF')	NOT SPECIALIZED ('NO')
EN→ES APERTIUM m = 0.7718 var = 0.0354 n = 142	**EN→ES APERTIUM** m = 0.7933 var = 0.0331 n = 30	**EN→ES GOOGLE** m = 0.6533 var = 0.0502 n = 30	**EN→ES APERTIUM** m = 0.8707 var = 0.0206 n = 41	**EN→ES APERTIUM** m = 0.8268 var = 0.0165 n = 41
2.13% ↓	62.32% ↓	95.13% ↓	34.70% ↓	**3.22%** ↓
EN→ES GOOGLE m = 0.7218 var = 0.0308 n = 142	**EN→ES GOOGLE** m = 0.7700 var = 0.0339 n = 30	**EN→ES SYSTRAN** m = 0.6500 var = 0.0384 n = 30	**EN→ES SYSTRAN** m = 0.8463 var = 0.0065 n = 41	**EN→ES GOOGLE** m = 0.7537 var = 0.0195 n = 41
0.19% ↓	19.64% ↓	**1.11%** ↓	**0.00%** ↓	**0.00%** ↓
EN→ES SYSTRAN m = 0.6493 var = 0.0454 n = 142	**ES ANGLO** m = 0.7167 var = 0.0159 n = 30	**EN→ES APERTIUM** m = 0.5400 var = 0.0135 n = 30	**EN→ES GOOGLE** m = 0.7049 var = 0.0115 n = 41	**ES ANGLO** m = 0.5537 var = 0.0385 n = 41
1.24% ↓	6.34% ↓	28.04% ↓	**0.02%** ↓	**2.51%** ↓
ES ANGLO m = 0.5930 var = 0.0258 n = 142	**EN→ES SYSTRAN** m = 0.6367 var = 0.0279 n = 30	**ES ANGLO** m = 0.5100 var = 0.0092 n = 30	**ES ANGLO** m = 0.6024 var = 0.0157 n = 41	**EN→ES SYSTRAN** m = 0.4610 var = 0.0289 n = 41

Table 2 – Gap-filling success rates for MT output compared with those for a Spanish text written by a native English speaker ("ES ANGLO", shaded boxes).

TOTAL	NATURAL SCIENCES ('NAT')	SOCIAL SCIENCES ('SOC')	JOURNALISTIC ('INF')	NOT SPECIALIZED ('NO')
ES FRANCO m = 0.7789 var = 0.0231 n = 142	**ES FRANCO** m = 0.8267 var = 0.0186 n = 30	**ES FRANCO** m = 0.6400 var = 0.0225 n = 30	**FR→ES SYSTRAN** m = 0.8463 var = 0.0120 n = 41	**ES FRANCO** m = 0.8293 var = 0.0246 n = 41
0.00% ↓	9.63% ↓	100.00% ↓	**2.31%** ↓	57.22% ↓
FR→ES GOOGLE m = 0.7218 var = 0.0254 n = 142	**FR→ES SYSTRAN** m = 0.7667 var = 0.0451 n = 30	**FR→ES GOOGLE** m = 0.6400 var = 0.0218 n = 30	**ES FRANCO** m = 0.7951 var = 0.0080 n = 41	**FR→ES GOOGLE** m = 0.8098 var = 0.0239 n = 41
74.30% ↓	**1.46%** ↓	**2.06%** ↓	5.70% ↓	41.18% ↓
FR→ES SYSTRAN m = 0.7141 var = 0.0537 n = 142	**FR→ES GOOGLE** m = 0.6467 var = 0.0226 n = 30	**FR→ES APERTIUM** m = 0.5333 var = 0.0382 n = 30	**FR→ES GOOGLE** m = 0.7488 var = 0.0156 n = 41	**FR→ES SYSTRAN** m = 0.7829 var = 0.0195 n = 41
0.00% ↓	72.94% ↓	**0.27%** ↓	**0.00%** ↓	**0.00%** ↓
FR→ES APERTIUM m = 0.5268 var = 0.0499 n = 142	**FR→ES APERTIUM** m = 0.6300 var = 0.0463 n = 30	**FR→ES SYSTRAN** m = 0.3867 var = 0.0274 n = 30	**FR→ES APERTIUM** m = 0.5854 var = 0.0218 n = 41	**FR→ES APERTIUM** m = 0.3878 var = 0.0616 n = 41

Table 3 – Gap-filling success rates for MT output compared with those for a Spanish text written by a native French speaker ("ES FRANCO", shaded boxes)

English into Spanish: According to these results, in general terms, for a native ES reader, the results show that EN–ES MT output is more useful for comprehension than a text written in ES by a native EN speaker ("ES ANGLO" in Table 2) with a CEFRL B2 level in ES. However, the text written by that native EN speaker would seem to be marginally more useful than Systran MT system output for NAT or NO texts.

French into Spanish: Overall, texts written in ES by a native FR speaker with a CEFRL B2 level in ES are more useful than any FR–ES MT output. However, the text written by the native French speaker ("ES FRANCO" in Table 3) is more useful for comprehension than most FR–ES MT output for NAT texts, as useful as Google FR–ES output for SOC texts, at least almost as useful as any FR–ES MT output for NO texts, and less useful than Systran MT FR–ES output for INF texts, but more useful than Google or Apertium FR–ES output. This may arise from the smaller differences between languages from the same family allowing a non-native speaker to write a text with fewer grammar and lexical mistakes, since the two languages involved share common vocabulary and structures. Alternatively, in any case, non-native mistakes do not prevent a native Spanish reader from understanding the text.

4.3 Comparing MT systems

When comparing data from the three MT systems, it has been shown, as already stated in Jordan-Núñez (2015), that, in general terms, the usefulness of any MT EN–ES output resulting from using the three MT systems is similar. However, the usefulness of the Apertium MT output is slightly larger. This is not the case for FR–ES: the usefulness of Apertium output is considerably lower.

More precisely, for highly specialized NAT texts (see table 4), Apertium seems to be the best EN–ES MT system for NAT texts, on par with Google, while Systran seems to be the best FR–ES system. Likewise, Google EN–ES and FR–ES are apparently the best MT systems for SOC texts.

For semi-specialized, INF texts (see Table 5), Apertium seems to be the best EN–ES MT system, on par with Systran, while Systran is the best FR–ES MT system. Equally, for NO texts (see Table 5), Apertium is again the best MT EN–ES system, and Google seems to be the best MT FR–ES system, on par with Systran.

5 Discussion

In view of the cloze-test results reported, one can answer the research questions posed and prove the hypotheses described:

RQ1) As already indicated in Jordan-Núñez (2015), to a native ES reader, the usefulness of a text written in a language L' from the same language family as the reader's L_1 (in this case, L' =IT) is higher for highly specialized texts than for general or non-specialized texts, as a consequence of the higher density of international vocabulary in NAT, and to a lesser extent, SOC texts. However, although the usefulness of a human-written IT text is very similar to MT output for highly specialized NAT texts, the IT text would appear to be more useful than any of the MT outputs studied for SOC texts. This may be the result, as mentioned above, of the deficiencies of the glossaries in the MT systems given the appreciable conceptual and lexical variety within these fields.

RQ2) Although, generally speaking, any of the FR–ES MT outputs studied is more useful for comprehension to a native ES speaker than EN–IT MT output, usefulness depends on the level and the field of specialty, and on the MT system. In many cases, especially when translating highly specialized material, EN–IT MT output may be more useful than FR–ES MT output– for example, when using Google to translate NAT texts. This leads to the conclusion that MT output in a language L' in the same family as the reader's L_1 –e.g., IT for a native ES speaker– could be used to facilitate their comprehension of texts originally written in a language from a different family and, evidently, far removed from their first language (that is, EN). Some findings, however, as already stated, do not apply to MT in general, but to limitations in the MT system that lead to differences in performance across text genres.

RQ3) As stated above, in general terms and to a native Spanish reader, EN–ES MT output is

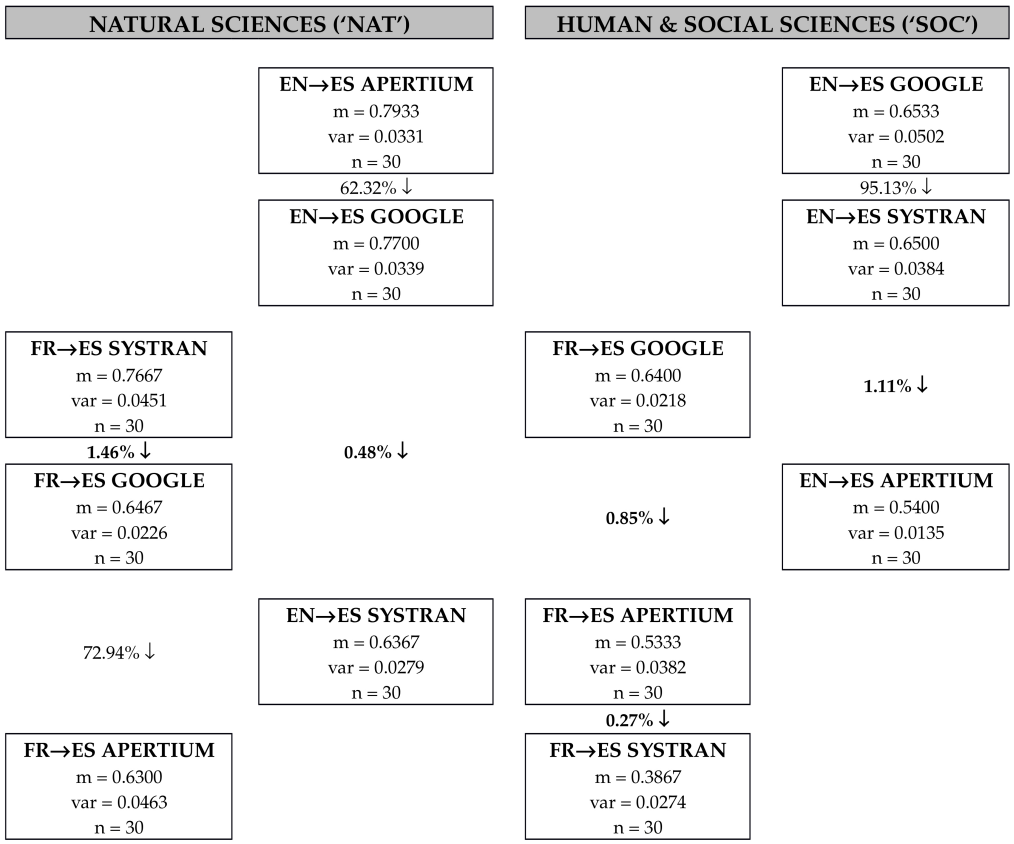

Table 4 – Comparing the gap-filling success rate for the output of different MT systems, for highly specialized texts.

more useful for comprehension than a text written in ES by a native EN speaker. However, texts written in ES by a native FR speaker are more useful than any of the FR–ES MT outputs. In fact, the text written in ES by the native FR speaker is much more useful than the text written by the native EN speaker (considering that both speakers have the same level of ES; that is, CEFRL B2). This could be due to that there is less interference or carry-over from FR when writing in ES (especially in syntax) given that both languages belong to the same language family.

RQ4) Although, in general terms, Apertium seems to be better for EN–ES than the other two MT systems and considerably worse for FR–ES, this also depends, to a greater or lesser extent, on the degree and field of specialty, and on the MT system used. In these experiments, Apertium is the best EN–ES MT system for any text type except SOC, which are best translated by Google and Systran. Google is the best FR–ES MT system for SOC and NO texts, while Systran is the best FR–ES MT system when dealing for NAT and INF texts.

5.1 Critical appraisal of the methodology

It is important to formulate critical comments regarding the methodology used in this pilot study, to be taken into account when pursuing further research.

It has been noticed that the fraction of gaps corresponding to function, structure or "stop" words — that is, articles, pronouns, prepositions or conjunctions— varies from one text type to

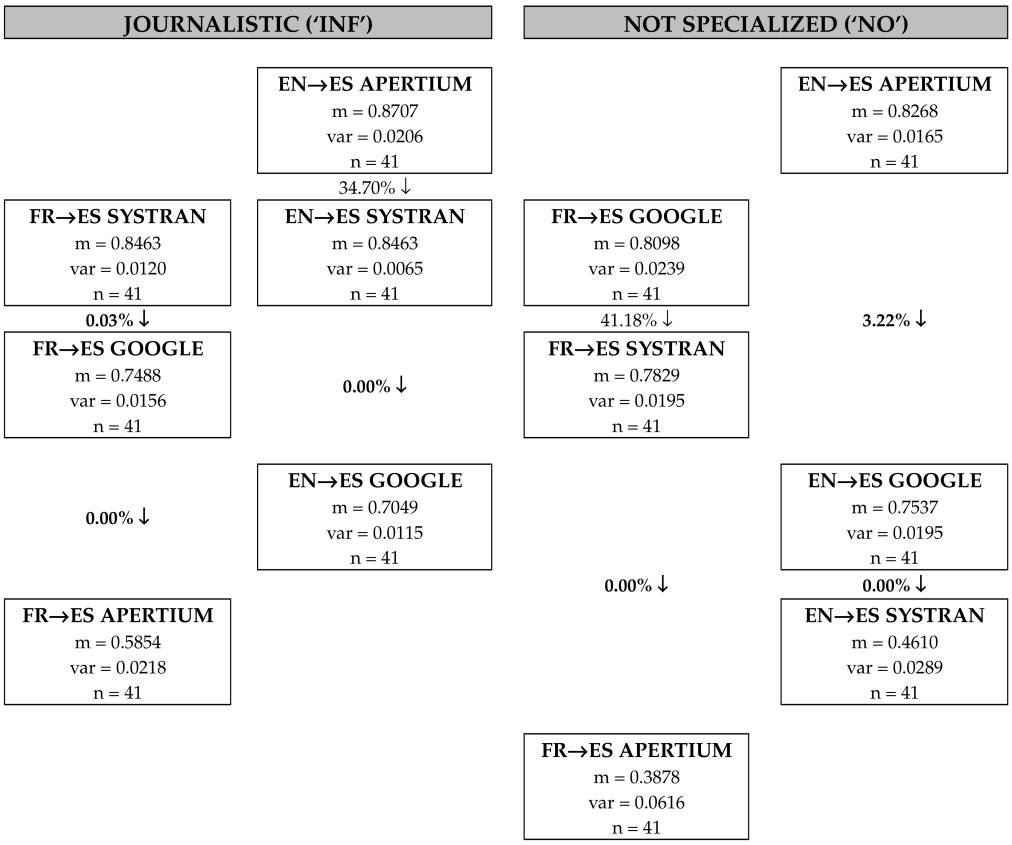

Table 5 – Comparing the gap-filling success rate for the output of different MT systems, for informative or journalistic texts and non-specialized texts

another. This could have a considerable effect on the results. In order to avoid this, in the test for the final project, the gaps should be done just on content words, that is, avoiding gaps at "stop" words (as was done by O'Regan and Forcada (2013)).

The results may also have been affected by the source of texts chosen when designing the test. It has been noticed that the Spanish version of the NAT text used a Latin-American variety of Spanish, which may have been less recognizable to a group of students familiar with Castilian Spanish; this should have been avoided when designing the research plan, as language varieties introduce an uncontrolled variable within the experimental design. Likewise, the Bible may have not been a good choice, since the different versions used differ quite noticeably from each other and, actually, they cannot strictly be considered mutual translations.[17] In order to avoid this, in the final project, all the texts should be taken from publications by institutional organizations using Castilian Spanish and where writing/editing and/or translation processes are followed by proofreading or quality assurance processes. In the case of non-specialized texts, they could also be taken from lesser-known novels that have been translated into all the languages involved in the study, to avoid the risk that the student recognizes the text and effortlessly fills the gaps without using information from the hint (as found by O'Regan and Forcada (2013) when no hint

[17] In fact, some passages have been translated so differently in each language version that the students could have found it difficult to find the information to help them fill the gaps from the hint given.

was given).

6 Conclusions and future work

As previously indicated, the broader research within which this study is framed seeks to identify whether the differences between a professional translation and MT output are similar to the differences between languages within a language family and, then, explore whether intercomprehension strategies (Klein and Stegmann, 2000; Martín Peris et al., 2005) are useful to avoid those non-human traits and understand the main message of the text. However, it is fair to say that the subjects participating in this study could have made use of spontaneous intercomprehension abilities not necessarily corresponding to the strategies described by the above authors and that some of those strategies may not apply to machine-translated texts.

As shown above, it should be admitted that the usefulness of a text written in a language from the same language family as the reader's L_1 is higher for highly specialized than for non-specialized texts. However, regarding MT output, this level of usefulness depends quite often on the level of specialty but also on the field and on the MT system used.

In view of the results, it seems also reasonable to postulate that MT output into a language L' in the same family as the reader's first language L_1 could be used to facilitate their comprehension of texts originally written in a language from a different family. Likewise, EN–ES MT output is more useful for comprehension than a text written in ES by a native EN speaker, while texts written in ES by a native FR speaker are more useful than FR–ES MT output.

Finally, in the future, it will be interesting to assess to what extent the skills used to understand MT output and the linguistic intercomprehension skills used by the methods designed are similar. If EN–ES MT output is, in general terms, as useful for comprehension as an IT text (depending, however, on the MT system used, the language combination and the level of specialization) and that MT output has a common vocabulary and some common morpho-syntactic structures (despite containing some mistakes that distance it from a native human-written text), one could argue that (generally unconscious) linguistic skills used in linguistic intercomprehension may be similar to those used to understand MT output. To shed some light on this, we will classify, label and quantify the types of machine translation errors or disfluencies and study their effect in the comprehension process, distinguishing errors or disfluencies that may be taken to be similar to the divergences observed between languages in the same family from those that are not. For instance, when a source word is out of the vocabulary of the MT system and left untranslated, but it is however similar to the target word, its negative impact should be less severe than in the case in which the word is very different.

Acknowledgements: KJN thanks Prof. Joan Fonollosa for his help in determining the statistical significance, and San Jorge University and Pompeu Fabra University for allowing him to have a research stay in the latter in order to work on this project. MLF thanks the Universitat d'Alacant for support during a sabbatical stay at the Universities of Sheffield and Edinburgh and acknowledges the support of the Spanish Ministry of Economy and Competitiveness through grant TIN2015-69632-R and the Spanish Ministry of Education, Culture and Sport through grant PRX16/00043. EC thanks the Spanish Ministry of Economy and Competitiveness for support through grant FFI2016-76245-C3-3-P. We also thank Dr. César Rodríguez, Dr. Naswha Nashaat, Dr. Manuel Gómez, and lecturers Santiago Lamas, María Pilar Cardos and Rachel Harris from San Jorge University and Dr. Elena Alcalde from the University of Granada for allowing us to test their students.

References

Ageeva, E., Tyers, F. M., Forcada, M. L., and Pérez-Ortiz, J. A. (2015). Evaluating machine translation for assimilation via a gap-filling task. In *Proceedings of EAMT*, pages 137–144.

Berka, J., Černý, M., and Bojar, O. (2011). Quiz-based evaluation of machine translation. *The Prague Bulletin of Mathematical Linguistics*, 95:77–86.

Clua, E. (2003). Eurocom: ciutadans plurilingües per a una europa multilingüe. *Enxarxa't: revista de la Xarxa de Dinamització Lingüística de la UB*, 2:4–5.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., and Weinstein, C. (2005). Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 5, pages v:1009–v:1012. IEEE.

Jones, D., Herzog, M., Ibrahim, H., Jairam, A., Shen, W., Gibson, E., and Emonts, M. (2007). Ilr-based mt comprehension test with multi-level questions. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pages 77–80. Association for Computational Linguistics.

Jones, D., Shen, W., and Herzog, M. (2009). Machine translation for government applications. *Lincoln Laboratory Journal*, 18(1).

Jordan-Núñez, K. (2015). ¿es igual de comprensible la salida de un sistema de ta que un texto escrito en otra lengua de la misma familia de lenguas? In LEXITRAD, editor, *New Horizons in Translation and Interpreting*. Tradulex.

Klein, H. G. and Stegmann, T. D. (2000). *EuroComRom - Die sieben Siebe: Romanische Sprachen sofort lesen können*. Shaker.

Martín-Peris, E. (2011). La intercomprensión: concepto y procedimientos para su desarrollo en las lenguas románicas. In de Zarobe, Y. R. and de Zarobe, L. R., editors, *La lectura en lengua extranjera*, pages 246–271. Portal Education.

Martín Peris, E., Clua, E., Klein, H., and Stegmann, T. (2005). *EuroComRom — Los siete tamices: Un fácil aprendizaje de la lectura en todas las lenguas románicas*.

Meissner, F. (2004). Esquisse d'une didactique de l'eurocompréhension. In *EuroComRom: Les sept tamis: lire les langues romanes dès le départ; avec une esquisse de la didactique de l´eurocompréhension*, pages 19–83.

O'Regan, J. and Forcada, M. L. (2013). Peeking through the language barrier: the development of a free/open-source gisting system for Basque to English based on apertium.org. *Procesamiento del Lenguaje Natural*, pages 15–22.

Page, W. D. (1977). Comprehending and cloze performance. *Literacy Research and Instruction*, 17(1):17–21.

Rankin, E. F. (1959). The cloze procedure: its validity and utility. In *Eighth yearbook of the National Reading Conference*, volume 8, pages 131–144. Milwaukee: National Reading Conference.

Trosterud, T. and Unhammer, K. B. (2012). Evaluating North Sámi to Norwegian assimilation RBMT. In *Proceedings of the Third International Workshop on Free/Open-Source Rule-Based Machine Translation (FreeRBMT 2012)*.

Weiss, S. and Ahrenberg, L. (2012). Error profiling for evaluation of machine-translated text: a Polish-English case study. In *LREC*, pages 1764–1770.