The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0022-0418.htm

Received 12 February 2009
Revised 2 September 2009
Accepted 3 September 2009

Analysis of errors in the automatic translation of questions for translingual QA systems

María-Dolores Olvera-Lobo
CSIC, Unidad Asociada Grupo SCImago, Madrid, España, and
Lola García-Santiago
Department of Library and Information Science, University of Granada, Granada, Spain

Abstract

Purpose – This study aims to focus on the evaluation of systems for the automatic translation of questions destined to translingual question-answer (QA) systems. The efficacy of online translators when performing as tools in QA systems is analysed using a collection of documents in the Spanish language.

Design/methodology/approach – Automatic translation is evaluated in terms of the functionality of actual translations produced by three online translators (Google Translator, Promt Translator, and Worldlingo) by means of objective and subjective evaluation measures, and the typology of errors produced was identified. For this purpose, a comparative study of the quality of the translation of factual questions of the CLEF collection of queries was carried out, from German and French to Spanish.

Findings – It was observed that the rates of error for the three systems evaluated here are greater in the translations pertaining to the language pair German-Spanish**.** Promt was identified as the most reliable translator of the three (on average) for the two linguistic combinations evaluated. However, for the Spanish-German pair, a good assessment of the Google online translator was obtained as well. Most errors (46.38 percent) tended to be of a lexical nature, followed by those due to a poor translation of the interrogative particle of the query (31.16 percent).

Originality/value – The evaluation methodology applied focuses above all on the finality of the translation. That is, does the resulting question serve as effective input into a translingual QA system? Thus, instead of searching for "perfection", the functionality of the question and its capacity to lead one to an adequate response are appraised. The results obtained contribute to the development of improved translingual QA systems.

Keywords Translation services, Computer applications, Knowledge management, Languages, Error analysis, Quality improvement

Paper type Research paper

1. Introduction

The tools known as question-answering systems or QA systems pursue the objective of supplying concrete data that respond to the queries formulated by users in natural language. The earliest QA systems came forth in the 1960s and used restricted domain databases with structured information. Classic examples are Baseball (Green *et al.*, 1961), responding to information such as: How many games did the Yankees play in July? Lunar (Woods *et al.*, 1972), a database for the chemical analysis of the Apollo lunar missions: What is the average concentration of aluminium in high alkali rocks? or Chat-80 (Warren, 1981), a geographical database (Which is the largest African

Journal of Documentation
Vol. 66 No. 3, 2010
pp. 434-455

0022-0418
DOI 10.1108/00220411011038485

country?) with an updated version that translates the question to the Structured Query Language (SQL). Another type of answer search process involves systems of dialog, like the classic Eliza (Weizenbaum, 1966). This system simulated a psychoanalysis, and can be considered the precursor of the current chatterbot software designed to emulate an intelligent conversation with one or more humans by means of text and/or audio. In the past decade, a noteworthy interest has evolved toward the QA systems that use diverse information sources to elaborate the response (anywhere from dictionaries and structured databases to Wikipedia or Google), which can be seen in operative examples on the Web such as Start (available at: http://start.csail.mit.edu/), an open domain system; or MedQA, specializing in biomedical information (available at: http://monkey.ims.uwm.edu:8080/MedQA/). This type of system took off after the TREC-8 Conference (Text REtrieval Conference, available at: http://trec.nist.gov) (Voorhees, 1999).

2. Translingual QA systems

The process that takes place in a QA system basically comprises the following stages: after analysis of a question put forth by the user, the system locates and extracts the response of different sources, afterwards eliminating the redundant information and elaborating and presenting one or several concrete responses that supposedly satisfy the query (Cui *et al.*, 2004) – that is, that are relevant in the context of the user's question. Therefore, once the question is expressed, usually via an interrogative adverb of the type: who, what, which, how, when, where – or in an imperative form ("tell me", "name", "indicate . . . "), the QA system proceeds to construct coherent responses.

The questions managed by QA systems tend to be of a factual sort, in which data referring to persons objects, dates, places, organizations, measurements, or so on are requested. In order to fulfil their objective, these systems apply increasingly sophisticated processing techniques, which permit them to elaborate responses on the basis of information stored in a repository of documents. Most systems carry out a detailed analysis of the question so as to extract and represent that information that might be of utility in the successive phases of the process (Vicedo, 2004), namely: the type of entity that each question expects as an answer (whether a date, a proper name, etc.); and restrictions and additional characteristics related with the type of response expected (terms within the question that give rise to retrieval of those extracts susceptible of containing the answer, and syntactic or semantic relations that should appear between the questioning entities and the response to be located).

The classification of the questions proves fundamental for the system, as this information will be used in the search phase and in the selection and extraction of the potential responses (Hermjakob, 2001; Kwok *et al.*, 2001; García Cumbreras *et al.*, 2005). Generally speaking, the questions are classified in terms of the expected response, assigning it to a category proposed by the system (time, place, person. . .). Then, during the final filtering, the question and its characteristics are contrasted with the candidate responses.

The techniques used in QA systems for processing and analysing information vary widely, and depending on the focus adopted by the system designers, they may involve the use of statistical methods, or the application of complex techniques for processing natural language. The different criteria and decisions adopted in the design and development of the QA system architecture give rise to a heterogeneous typology. The

language(s) of system operations, the subject matter of the database documents, the level or organization of the information (structured or not) contained therein or the degree of interactivity with the user, among other aspects, are determinant features of QA systems.

If the user of a QA system presents a question in one language in order to obtain the response from documents found in one or several other languages, we are dealing with a translingual or multilingual search system. These embrace the capacities of cross-language information retrieval or CLIR, plus those of a QA system per se (Airio, 2008). The evaluation of this type of system and the incorporation of novel techniques and proposals is carried out in the core of major international forums such as Text REtrieval Conference (TREC) and Cross-Language Evaluation Forum (CLEF, available at: www.clef-campaign.org/). Both have a section – Question-Answering Track – dedicated specifically to QA systems.

In the realm of multilingual information retrieval, we find different proposals directed to overcome the linguistic barriers arising when the questions posed and the relevant documents are in different languages (Oard *et al.*, 2004; Hansen and Karlgren, 2005). The various architectures that the CLIR system may adopt can be broadly classified as those focusing on the translation of the query, on the translation of the documents in the database, or else on interlinguistic methods (Oard and Diekema, 1998). The most frequent option is the translation of the query: being briefer than the text of the documents, the computational cost required for its translation is foreseeably less (Hull and Grefenstette, 1996). However, some research underlines the difficulty of resolving the ambiguity in the process of translation; questions are short and offer little context to aid in the semantic disambiguation of the query terms, although interaction with the user might contribute to enhanced results (Oard *et al.*, 2008). The translation processes involved rely on diverse linguistic resources, which include bilingual dictionaries, textual corpora, automatic translating software, thesauri, or semantic networks (López-Ostenero *et al.*, 2004; Abusalah *et al.*, 2005), sometimes used in conjunction (Jones *et al.*, 2008). Further mechanisms may be incorporated for the disambiguation or selection of the most adequate translation from amongst the available proposals (Kishida, 2005).

The translation of the question served in machine translating (MT-based query translation) is possibly the most immediate approach, and it affords the best results, as it involves the use of linguistic analysis tools that help improve the precision of the translation (Jones *et al.*, 1999). The basic structure of the translingual QA systems that incorporate MT software for the translation of questions can be seen in Figure 1. The user introduces a question in the natural source language. The system analyses the question put forth by the user and classifies it according to the type of response anticipated. At this point, some sort of transformation of the question may (or may not) be applied by means of diverse techniques (keyword extraction, linguistic analysis, recognition of patterns in text, among others). The question is translated by the QA system with the help of tools such as bilingual dictionaries, systems of automatic translation, parallel texts or corpora that are statistically processed, ontologies or thesauri, that are more or less oriented toward the underlying concepts (Volk *et al.*, 2003); or else semantic networks like EuroWordNet (Vossen, 1998). Machine translating can be used with other methods to improve the quality of the phrases produced. The resulting search expression will be the input. Once the relevant

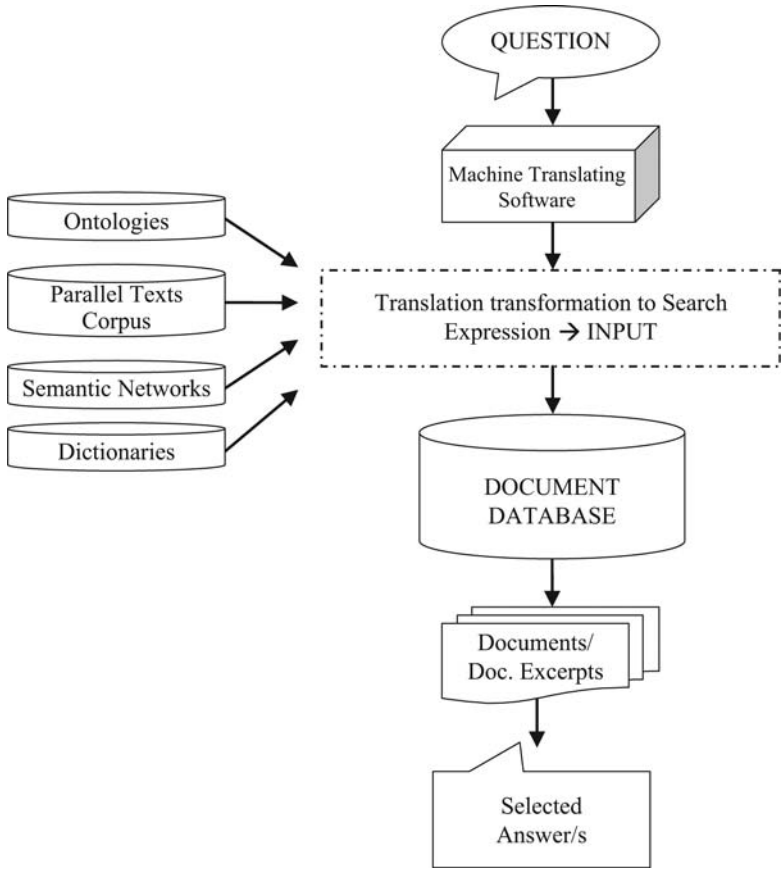

Figure 1.
Basic scheme of a translingual QA system featuring machine translation of the question (MT-based query translation)

documents for that query have been located, the system breaks them up into passages, chooses the fragments that include the candidate responses, and selects the definitive response. Finally, this response is supplied to the user along with its situation within the corresponding passage of text.

At the end of the 1990s, the MT-type devices were made available to web users, and at present internet can be used to translate texts from/to dozens of languages online. The online automatic translators are programs – in many cases free – available on the web to permit the practically instantaneous translation of web pages or more or less limited text fragments. The quality of the translation is poor on occasions, but even so the demand for such necessary services is great. Most QA systems currently rely on online translators (Larosa *et al.*, 2005).

One of the objectives of the present study is to identify how the online translator functions as an integrated tool in a translingual QA system entailing a collection of Spanish language documents. In this case, the questions were formulated in French and in German, and as part of the process; they had to be translated into Spanish to serve as system input. A second objective surrounded the application and, if so be the case, the demonstration of the validity of a number of measures for evaluating the

quality of the automatic translations obtained. Finally, a third objective was to identify the main problems that arise in the translation of questions as the expression of consultation in a QA system.

Given the framework of these proposals, our study involves an initial focus on the first module of these translingual QA systems, the one destined to translate the user's question. In subsequent sections we identify and analyse the types of errors committed by the three online translators used. To this end, a comparative study of the quality of the translation of factual questions from German and French into Spanish is carried out. A subjective manual assessment was performed and compared with automatic evaluation, to further explore the types of errors produced. This evaluation was proposed from a documental perspective: that is, to determine the functionality of the translator as a mediating instrument for the search of answers. The work was completed with the display of the results obtained, as we will show, and some pertinent conclusions.

3. Methodology for objective and subjective evaluation

This analysis was based on the CLEF 2008 collection of questions in German and in French. In order to carry out evaluation of the automatic translation, three separate online translators were used: Google Translator (available at: www.google.es/translate_t?hl = es), Promt Translator (available at: www.online-translator.com/) and Worldlingo (available at: www.worldlingo.com/en/products_services/worldlingo_translator.html). These allowed for the translation from German and French into Spanish. The translations obtained were assessed manually as well as automatically, applying well-established measures for the evaluation of automatic translation.

3.1 Online translators evaluated

The choice of Google Translator, Promt Translator and Worldlingo was based on the fact that they allow for comparisons involving the language pairs German-Spanish and French-Spanish. Moreover, they are free and widely diffused online services, they are fast, and the quality is quite reasonable at first glance, making them appropriate for a study of this sort.

The Google translator works with its own statistical translation system, and the languages offered as source language and as target language come to a total of 34 in its beta version (including German, Arabic, Bulgarian, Czech, Chinese, Portuguese, Rumanian, Russian, Swedish . . .). As the Google service affirms, the online translator is based on the introduction of many millions of words and texts within the system, both monolingual texts in the target language and aligned texts made up of translations elaborated by professional translators in the two languages involved. Then, the system applies statistical learning techniques to create a model of translation that progressively improves the efficiency of the system.

In Worldlingo translations can be produced between pairs from these languages: German, Arabic, Korean, Chinese, Spanish, French, Greek, English, Italian, Japanese, Dutch, Portuguese, Russian, and Swedish. It freely admits texts of up to 150 words. In turn, Promt Translator allows one to work with 25 language pairs, (similarly from English to Russian, German, Spanish, French and Portuguese, or from German to Russian, to English, to French and German, among others). In some of these linguistic combinations the user is allowed to indicate whether the text refers to a general lexicon

or to a specific jargon or specialized language (automobiles, banking, football, commercial correspondence, internet and computers, sports, travel, etc.), and texts of up to 3,000 characters are admitted. Promt translator applies a strategy of transference, although it presents the particularity that the systems' translating algorithms are not based on sequential procedures but rather on a hierarchical approximation that subdivides the translation processes and interconnects procedures for different units of analysis – lexical unit, nominal group, simple phrase or compound phrase (Sokolova, 2007). All the processes are interconnected, and they interact in trying to arrive at a quality translation.

Although, as we mentioned above, the free online translators in some cases limit the maximum length of text admitted for operations (this is not the case of Google translator, which admits very extensive texts), we must not forget that a QA system works with specific questions whose expression is generally succinct.

3.2 Sample and types of questions

The CLEF collection used here consists of 200 questions in German and the same 200 in French that were used in the CLEF Project so that the participants – research groups and private enterprises – could evaluate the techniques and systems they proposed for the translingual retrieval of information. In our case, to see how the translators work online as tools for a translingual QA system, one of the criteria assessed is the translators' capacity for maintaining the characteristics of the question. The questions in the CLEF collection pertain to three types: factual questions, of definition, and closed list. The factual questions refer to a very specific datum mainly related with names, amounts, dates, etc. These questions can be divided into subtypes, depending on the interrogative adverb used (who, what, when, where, why, how much, how many, which, whose), to evoke a response concerning person, object, time, place, etc. The definition questions solicit a somewhat more open informational item in terms of synonymy and formulation. They respond to the type "what/who is X?" in reference to an institution, person, thing or concept. Finally, in the closed list questions, the response calls for a limited set of data, and they are formulated either using an interrogative adverb (which, how) or in imperative form ("name all of London's airports").

Natural Language, that in which we habitually communicate, makes it possible for us to express feelings, narrate events, give orders, express reasoning, and much more. The great wealth and variety of registers that NL permits would be its main trait, but there lies the main problem it presents in automatic processing as well. One example for the case at hand: interrogative adverbs vary from one language to another in terms of characteristics that can make it difficult to clearly identify question types.

Table I shows the distribution of the questions we used according to the type of response anticipated (object, person, place, time, etc.). The 200 questions utilized were, for the most part (156), of the factual type; and in the Spanish language, they mainly present interrogative particles (adverbs or preposition plus adverb) as shown, in Table II. The rest of the questions (44) are divided fairly equally into definition questions (24) and closed list questions (20).

The combinations of preposition and interrogative adverb allow for a clearer identification of the intention of the question. For instance, in the case of the interrogative adverb in Spanish "¿qué?" (what?), the question might be of the definition

sort. If that particle is preceded by the proposition "en", however, the information requested would be factual. It might refer to a place ("¿en qué ciudad?") or be of a temporal character ("¿En qué año?"). Within the category "others" are the cases in which the interrogative particle is preceded by other elements such as a conjunction (e.g. "¿Y quién?") or some information delimiting the actual question (e.g. "¿En España quién. . .?").

3.3 Measures for the evaluation of the automatic translator of questions

Machine Translation evaluation is a persistent research problem, addressed by a number of recent studies. The measures most extensively used are of two major sorts: automatic methods (or objective ones) and subjective methods (Tomás *et al.*, 2003). Objective evaluation methods compare a set of correct translations or translations of reference against the set of translations produced by the software translator under evaluation. The best known metrics work at the lexical level, comparing chains or strings of text. Meanwhile, the subjective methods call for human intervention to carry out the evaluation.

In the present study, the online translators were evaluated applying both objective and subjective criteria. The automatic evaluation measures most widely used calculate the rate of error in the translation. Deserving specific mention among these are WER

Table I.
Categories of the questions

Category	Num. questions
Object	97
Person	38
Place	11
Time	2
Amount	28
Definition	21
Other	3

Table II.
Interrogative particles identified in Spanish

Type	Number
Cómo (How?)	7
Cuál (Which?)	14
Cuándo (When?)	2
Cuánto (How much?/How many? Masc.)	22
Cuánta (How much?/How many? Fem.)	6
Dónde (Where?)	11
Qué (What?)	62
A qué (To what?)	8
Con qué (With what?)	1
De qué (Of what?)	3
En qué (In what?)	23
Entre qué (Between/among what?)	1
Quién (Who?)	32
A quién (To whom?)	3
Contra quién (Against whom?)	1
Others	4

(Tillman *et al.*, 1997; Vidal, 1997) and SER, though they present the drawback of calculating from a single translation of reference, affording no possibility of considering several equivalent translations as correct alternatives. For this reason, our study also included consideration and application of aWER and aSER (Tomás *et al.*, 2003), briefly described below as well:

- WER (*Word Error Rate*) is based on the Levenshtein distance or edit distance (Levenshtein, 1966) between two strings of characters. It measures the minimum number of insertions, substitutions and erasures necessary to convert one string in the other. Yet unlike the Levensthein distance, which works at the level of characters, the WER calculates this distance with regard to words. WER, then, is the edit distance between the output of the online translator and a given string of reference (considered to be a valid translation). It is a pessimistic measure in that it focuses on errors (Pérez *et al.*, 2004): If the system's output does not coincide precisely with the chain of reference, the latter is penalized, even if the output provided were acceptable for a human translator:

$$WER[\%] = \frac{\sum_{i=1}^{n} d(t_i, t_i^r)}{\sum_{i=1}^{n} |t_i^r|} * 100$$

- aWER (*all references WER*) (Tomás *et al.*, 2003) takes the translation provided by the automatic translator and compares it with all the phrases or strings of reference contained in the system, which have also been included by the human translator as valid translations.
- SER (*Sentence Error Rate*) compares the string of output from the online translator and the string of reference globally, as units. It compares the phrase to be evaluated with a single phrase of reference and indicates the percentage of phrases whose translations do not precisely coincide with those of reference.
- aSER (*all references SER*) (Tomás *et al.*, 2003) is a measurement deriving from aWER. aSER entails the drawback of working with a single reference, and besides, it does not measure the number of erroneous phrases, but rather those that do not coincide exactly with the string of reference. On the other hand, aSER does indicate the percentage of phrases whose translation is incorrect.

Other widely used measures are the BLEU (Papineni *et al.*, 2002; Callison-Burch *et al.*, 2006), NIST (Doddington, 2002) and F-measure (Melamed *et al.*, 2003), based on precision; METEOR (Banerjee and Lavie, 2005), in turn based on F measurement; and PER (Leusch *et al.*, 2003) and TER (Snover *et al.*, 2005), which measure the rate of error.

The above measures, once the strings of reference have been provided, are applied automatically so that the translations and the phrases of reference are compared without allowing one to determine or specifically assess the type of error or discrepancy produced between the two strings. For this reason, another type of metrics has been developed, requiring human intervention to perform evaluation. These are known as the subjective measures. Possibly the most extensively used among them, and taken into account in the framework of our study, is SSER (Subjective Sentence Error Rate) (Nießen *et al.*, 2000), calculated from the assessment that the human

translator makes regarding the quality of the translation supplied by the system. For this index, the range of values goes from 0 to 1. A score of 0 means a perfect translation, while a score of 1 corresponds to a translation that is syntactically and semantically incorrect:

$$SSER(s_1^n, t_1^n)[\%] = \frac{100}{K \cdot n} \sum_{i=1}^{n} v(s_i, t_i)$$

Given the context of translingual QA systems that present an architecture of the type described here, the SSER measure of subjective evaluation was applied, as the objective was not so much a "perfect" translation but rather a translation capable of maintaining the characteristics of the questions and that would permit the system to locate appropriate responses. For this purpose, we designed and applied an evaluation grid to register the type of error most frequently produced by the machine translation of the questions.

For the human evaluation of the translations, we applied a Likert scale with six levels. Such evaluation holds that, for example, the position of the elements in the string or another type of error should not be penalized to the same extent as ambiguity or the loss of some characteristic of the question (interrogative adverb, entity of reference of the question, or similar). The quality of the translation of each question was assigned a value between 0 and 5. The score of 0 was given when the string was totally incorrect and made no sense as a translation, meaning it would be of no use as an input question in a QA system. A score of 5, in turn, was given when the translation was considered fully accurate.

The evaluation process was undertaken using EvalTrans software (Nießen *et al.*, 2000) in its graphic version for Windows (Tomás *et al.*, 2003), a tool that is freely available for machine translation evaluation (EvalTrans available at: www-i6.informatik.rwth-aachen.de/web/Software/EvalTrans/index.html).

3.4 Evaluation of translation errors

In itself, the application of evaluation measures does not serve to identify the most important source of error of a given system. It is necessary to carry out, in addition, a detailed analysis of the translations so as to determine the main problems generated and, a posteriori, to better focus research efforts. Unfortunately, few studies to date concentrate directly on this problem (Vilar *et al.*, 2006). Because we wished to specifically identify the main limitations or weaknesses of the automatic translators evaluated, the errors of translation they produced were determined, and we looked into them to see which were the most frequent. This called for detailed examination of each one of the phrases resulting from the translation of the questions. Likewise, we determined whether the evaluation score of the translations, according to the objective measures described above, was the consequence of the type of error made.

The translations that were given a score of 5 were the ones considered fully accurate, and therefore adequate for use in a translingual QA system. Those obtaining a score of 3 or 4 points presented errors that might be regarded as slight errors for the purpose of the translation. Those containing errors of medium importance received a score of 2. The translations getting the lowest score (from 0 to 1) contained serious errors of different types.

The errors identified were grouped into five categories: those referring to interrogative adverbs/particles, syntactical errors, lexical errors, errors involving the preposition, and errors related to punctuation.

4. Results and discussion

4.1 Efficacy of the online translators according to the measures used

The values of WER and of SER, as determined for Google Translator, Promt Translator and Worldlingo in automatic evaluations from German and French to the Spanish language (Tables III and IV), are based on the edit distance. The main difference between the two measures lies in the fact that the string of characters used by SER to perform the evaluation is the complete sentences, so that any minimal variation between the phrase of reference and that offered by the online translators is interpreted as an erroneous phrase and discarded. This leads to high percentages in the rate of error.

Of the 200 strings (questions) studied for each language, and regardless of their length, it was necessary to change, add, substitute or eliminate from four to six words of the string supplied by the translator in order to obtain the translation of reference or "fully accurate" translation, this value generally being lower in the case of translating from French to Spanish (Table V).

The aWER, and aSER indicators allowed us to consider more than one possible translation as correct. That is, a set of reference sentences are used (not just one) to compare with the proposed translation. This leads to a decrease into the error rates observed – though they are still high (See Tables VI and VII).

These automatic measurements were complemented with manual evaluation, in which a professional human translator which languages of work are these three, has assigned scores (from 0 to 5) to assess the quality of the translation of each one of the phrases in both language pairs. Besides it was possible to identify the different types of errors that were produced throughout the translation process. The manual evaluation is always over the phrases translated into Spanish and we must remember, a functional

Table III. Objective evaluation (WER and SER) of the German-Spanish translations

German-Spanish	Google	Promt	Worldlingo
WER (%)	41.9	54.4	57.6
SER (%)	95	99	98.5

Table IV. Objective evaluation (WER and SER) of the French-Spanish translations

French-Spanish	Google	Promt	Worldlingo
WER (%)	43.2	39.6	40.8
SER (%)	95.5	90	93

Table V. Average number of words that had to be modified to arrive at a fully accurate translation

	Google	Promt	Worldlingo
German-Spanish	4.29	5.57	5.85
French-Spanish	4.42	4.05	4.12

criterion is taken into account in this study to determine the translation quality. Figures 2 and 3 show the results of this human assessment.

The irregular behaviour of the online translators regarding translations from the French language (Figure 3) is seen in greater differences from one to the next. The translations generated were either very adequate or else completely inadequate, with practically no median (over 50 percent were given a score of 0). Thus, the mean score is very low. There is one exception, the case of Worldlingo (Figure 3), which obtained higher scores for its translations, with a median of two points – that is, a nearly correct translation. This gives rise to the somewhat asymmetrical distribution of results to the left, where we see the lowest values according to the scale used. In this case, a high average score was not obtained (meaning there were not many perfect translations), yet over 50 percent proved nearly acceptable as system input. Meanwhile, the normal distribution of the Google Translator reflects a normal and almost symmetric distribution, with more heterogeneous values, some very high and others very low (Figures 2 and 3).

Table VI.
Objective evaluation (aWER and aSER) of the German-Spanish translations

German-Spanish	Google	Promt	Worldlingo
aWER (%)	57.6	50.4	54.6
aSER (%)	88	91	94

Table VII.
Objective evaluation (aWER and aSER) of the French-Spanish translations

French-Spanish	Google	Promt	Worldlingo
aWER (%)	36.7	27.5	29.7
aSER (%)	87.5	75	78.5

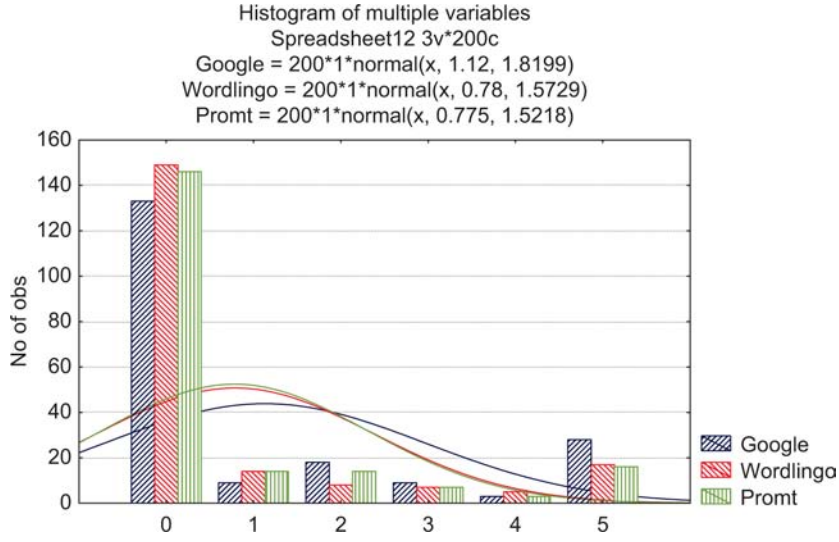

Figure 2.
Distribution of the manual assessments of the German-Spanish translations

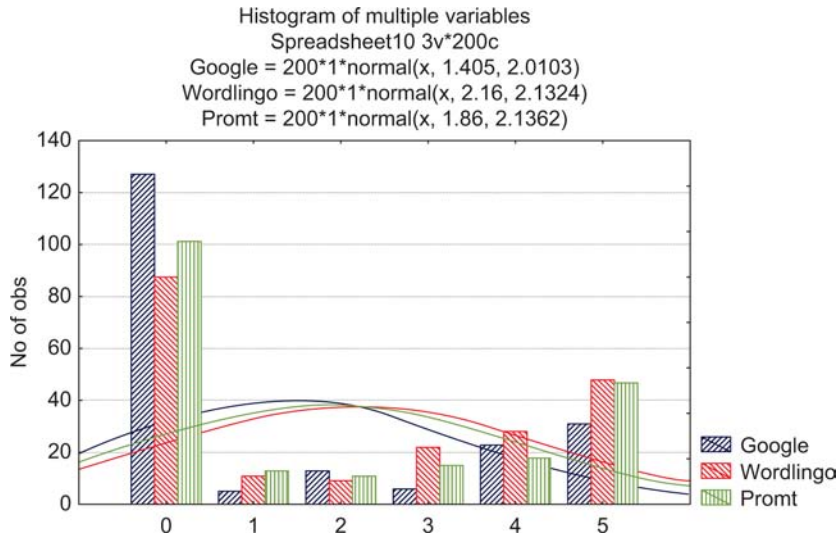

Figure 3.
Distributions of the manual assessments for the French-Spanish translations

Table VIII shows the standard deviation of the scores established for the translations on the basis of the manual assessments made. We find that, while in general the scores assigned to the translation of the German-Spanish linguistic pair were notably lower, they were also more homogeneous than for the other language pair analysed.

The scores resulting from the human assessments were taken in applying the subjective measure of evaluation sSER (Table IX).

The results corresponding to the measures WER, aWER, SER and aSER show substantial differences in the rates of error reflected by these automatic measurements (Tables III to VII) and in the rates of error of the subjective assessments (Table IX). According to the latter, for German-Spanish machine translations the best results are obtained with Promt (77 percent error in sSER), followed by Google Translator (here with 90.2 percent); whereas for online translations from the French, Worldlingo (53.7 percent) is the most adequate even though Promt (55.5 percent) presents a very similar sSER index.

Table VIII.
Standard deviation of the human assessments

	Google	Promt	Worldlingo
German-Spanish	1.820	1.522	1.573
French-Spanish	2.010	2.136	2.132

Table IX.
Subjective evaluation (sSER) of the translations

SSER	Google	Promt	Worldlingo
German-Spanish (%)	90.2	77	91.3
French-Spanish (%)	70.4	55.5	53.7

4.2 Analysis of the translation errors

Each question was introduced individually in the online translator, this being the manner for the user to enter a query in a translingual QA system. However, some questions from the collection presented co-references, the most common form of which being the anaphora, where the subject is declared explicitly in the first question and is implicit in the second (Q1: ¿De qué planeta viene Superlópez?, Q2: ¿Quienes son sus enemigos? = What planet is Superlopez from? Who are his enemies?), or in the response to the question formulated previously (Q1: Q'était le Max bleu?, Q2: En quelle année le Baron Rouge l'a-t-il obtenu? = What was the "Blue Max"? In what year did the Red Baron obtain it?). As each question is translated individually by the online translator, the anaphoric sequels are semantically de-linked from their predecessors. In these cases, it is quite habitual for the translator to automatically transform the question to the passive voice, in view of the elliptic subject. This occurs in the majority of the 60 cases of anaphoric questions we encountered. Notwithstanding, the second question in such a question chain, being independent, may be correctly translated even when the first is not.

We also observed that, in some cases, the online translators managed to produce questions with a better quality than the questions of reference included in the multilingual collection used, whose proposals on some occasions were not exactly correct. One example was "*Welches sind die Präfekturen Griechenlands*?" (= What are the prefectures of Greece?), whose equivalent in Spanish according to the collection was "*¿Cuáles son las prefecturas de Creta [sic]?",* confusing "Crete" and "Greece"; yet it was translated more correctly by Google Translator to "*¿Cuáles son las prefecturas de Grecia?"*

In the case of the translation of names of organizations/public entities, the automatic translator sometimes translated them to English as a pivot language not included in our study, and other times to the Spanish target language. This may be due to the fact that the name is well known in both its Spanish version and the English denomination. For instance, the Organización Europea para la Investigación Astronómica en el Hemisferio Austral (ESO) is the translation of ESO (European Southern Observatory) whose full name is European Organization for Astronomical Research in the Southern Hemisphere**.** However, the reference sentence that CLEF proposes, "¿Qué países forman parte del European Southern Observatory?" does not translate the name of the institution into Spanish. Contrariwise, all the online translators, in both languages, did translate the organization name:

- *In German*: Welche Länder sind Mitglied bei der Europäischen Südsternwarte?
- *Google*: ¿Qué países son miembros del Observatorio Europeo Austral?
- *Worldlingo*: ¿Qué países son miembro con el observatorio del sur europeo?
- *Promt*: ¿Qué país son miembro en el observatorio de Sur europeo?
- *In French*: Quels sont les pays membres de l'Observatoire européen austral?
- *Google*: ¿Qué países son miembros del Observatorio Europeo Austral?
- *Worldlingo*: ¿Cuáles son los países miembros del Observatorio europeo meridional?
- *Promt*: ¿Qué son los países miembros del Observatorio europeo austral?

Other times, the program does not limit itself to offering a mere translation of the question but even adds or offers other equivalents that may be more informative. This

may occur when substituting abbreviated forms or initials with their full version as in socio-political concepts. For example, "*Wer war US-Präsident während der Kubakrise?*", was translated by Google Translator as "*Que [sic] fue presidente de EE.UU. durante la crisis de los misiles de Cuba?*", with the original "Cuban crisis" translated as "Cuban missile crisis". This shows us that Google, in addition to translating the string, may also include metalinguistic information about the context of the expression that is not included in the sentence.

There were three questions for which no correct translation was obtained in any case. These questions made reference to a concept of popular culture, "*el ratoncito Pérez*" (the little Mouse who is the cultural equivalent of the "Tooth Fairy"); to the title of the work "*El ingenioso hidalgo Don Quijote de la Mancha*" (the complete title of "Don Quixote"); and to a specific request, "Dame ejemplos de desiertos originados en regiones pluviométricas" (= "Give me examples of deserts originating in pluviometric regions").

The errors were considered slight, medium or serious on the basis as how much they affected the functionality of the question as translated. The most frequent cases are shown in Table X.

The errors identified above were grouped into five categories, corresponding to errors involving pronouns and interrogative adverbs, errors of a syntactic nature, lexical errors, errors involving the preposition, and punctuation errors (Table XI). In the translations coming from the German, the errors were mostly due to lexical matters (40.35 percent) and the use of the interrogative particle (39.77 percent). This behaviour was similar for the translations from the French to the Spanish in the case of lexical errors (56.19

Score of error	Case
Slight (3-4 points)	Lack of agreement in gender and/or number Change in interrogative particle or verb (but obtaining a similar result) Change in the order of the words without altering the meaning of the question
Medium (2 points)	Loss of interrogative particle or elimination of accent mark (making it a pronoun instead of a relative adverb) Change in personal pronoun (in gender or in case, as in German) Change of number (singular/plural) of the noun Change of order of the words in a proper noun
Serious (3-4 points)	Change of gender of the article (when preceding nouns with different meanings depending on gender; or as part of a superlative form in an anaphoric question) Elimination of the verb in the phrase, or confusion between Spanish verbs "ser" and "estar", or change in the number of the verb (singular/plural) No translation of the interrogative particle of the target language, or changing the interrogative particle to another, of change in gender of the interrogative particle No translation of the proper noun (despite its adequate equivalent in target language), or translation of a proper name that cannot be translated, translation but erroneous in form Elimination of the preposition, or a change in it, or the unnecessary inclusion of a preposition, which distorts the meaning of the phrase Literal translation of certain expressions

Table X.
Degrees of error in the translation

percent), again the majority, although for this language pair there were considerably fewer errors relating to adverbs and interrogative pronouns (17.14 percent).

In our analysis of the errors made by each one of the online translators, it is interesting to observe the homogeneity of the three systems insofar as the most frequent errors (Table XII), with errors in vocabulary predominating in all cases (though with Promt Translator, the errors due to a poor translation of the interrogative particle were equally numerous).

As seen in Table XIII, Google is the translator with the greatest number of lexical errors for the combination German-Spanish (but only 46.88 percent of the errors made by this system), whereas Promt leads in the French-Spanish language pair (where 60.98 percent of the errors were of this type); yet curiously enough, Promt presents fewer lexical errors (23.94 percent) than the other two programs (Google 46.88 percent and Worldlingo 61.11 percent when working with the combination German-Spanish. Worldlingo stands out in the reduced percentage of errors referring to the adverb or interrogative pronoun, while in this case Promt, together with Google, give the poorest results for both language pairs analysed. At the same time, Worldlingo produces an error of generation and/or elimination of punctuation signs (question marks and quotation marks) not habitual in the other two systems.

As said above, from the standpoint of formal construction, the translated sentences were definition, closed list and factual questions. It must be remembered that the amount of questions of the first two types is considerably lower than that of the third type. These factual questions are 156 out of a total of 200 sentences. And this proportion is understandable due to the fact that the usual natural way of asking is by constructing factual questions. Still, the subjective scores were grouped according to the question type and then the average and the standard deviation for each automatic translation for both pair of languages was calculated (Table XIV).

Table XI.
Types of translation errors by language pair

Type of error	Question (German-Spanish) *n*	(%)	Question (French-Spanish) *n*	(%)	Average *n*	(%)
Lexical	69	(40.35)	59	(56.19)	64	(46.38)
Interrogative particle	68	(39.77)	18	(17.14)	43	(31.16)
Syntactical	15	(8.77)	15	(14.29)	15	(10.87)
Preposition	17	(9.94)	7	(6.67)	12	(8.7)
Punctuation	2	(1.17)	6	(5.71)	4	(2.9)
	171	(100)	105	(100)	138	(100)

Table XII.
Types of translation errors by online translator

Type of error	Google (%)	Promt (%)	Worldlingo (%)
Lexical	58.73	42.46	53.96
Interrogative particle	26.195	42.1	8.42
Syntactic	9.785	8.075	14.48
Preposition	5.285	8.93	13.975
Punctuation	0	0	9.165
	100	100	100

Types of errors	Google				Promt				Worldlingo			
	German-Spanish		French-Spanish		German-Spanish		French-Spanish		German-Spanish		French-Spanish	
	n	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)	*n*	(%)
Lexical	30	(46.88)	12	(70.59)	17	(23.94)	25	(60.98)	22	(61.11)	22	(46.81)
Interrogative particle	26	(40.63)	2	(11.76)	39	(54.93)	12	(29.27)	3	(8.33)	4	(8.51)
Syntactic	5	(7.81)	2	(11.76)	8	(11.27)	2	(4.88)	2	(5.56)	11	(23.40)
Preposition	3	(4.69)	1	(5.88)	7	(9.86)	4.8	(8)	7	(19.44)	4	(8.51)
Punctuation	0	(0.00)	0	(0.00)	0		0	(0.00)	2	(5.56)	6	(12.77)
	64	(100)	17	(100)	71	(100)	41	(100)	36	(100)	47	(100)

Table XIII.
Types of translation errors by translator and language pair

Table XIV.
Average and standard deviation of human assessments by question type

Question types	Google				Promt.				Worldlingo			
	German-Spanish		French-Spanish		German-Spanish		French-Spanish		German-Spanish		French-Spanish	
	Aver.	Stand. Dev.	Aver.	Stand. Dev.	Aver.	Stand. Dev.	Aver.	Stand. Dev.	Aver.	Stand. Dev.	Aver.	Stand. Dev.
Definition	0.958	1.899	0.833	1.736	0.583	1.412	1.375	2.143	0.708	1.681	0.833	1.685
Closed list	1.300	2.003	1.800	2.353	0.500	1.051	1.750	2.173	1.450	2.114	1.750	2.173
Factual	1.122	1.794	1.442	1.998	0.840	1.588	1.949	2.134	0.705	1.464	2.417	2.116

On the basis of the data obtained, it can be confirmed the effectiveness of the selected automatic translators, on the whole, for each language. This in-depth study brings some nuances about these translators to light. Thus, it was observed that the most suitable automatic translator varies depending on the language and on the type of question used.

For translations from German to Spanish, Google is the first in the ranking of online translators for definition and factual questions; whereas Wordlingo surpasses it for closed list questions.

Certain facts can be observed when only one of the three translation tools used in our research produces a right translation. When either Wordlingo or Google produce the only valid translation of a question; the mistakes are related to lexical issues, while the mistakes produced by the other translation tools are related to the interrogative adverb used. These errors are even produced in accepted sentences as valid inputs in an information retrieval system. An example of this is found in Google. It misinterprets the verb "sein" in German or "être" in French ("to be" in English) when translating it into Spanish because this verb has two possible meanings: "ser" or "estar". As regards any type of sentence, Promt is the worst candidate to translate questions into German.

In the last analysis, the interrogative adverbs in definition questions are usually mistranslated no mater what language. The mistakes produced in closed list questions are lexical errors both in French and German. But in the latter language there are also mistakes related to interrogative adverb identification and translation. Finally, in factual questions, on the one hand there are more lexical errors (verbs and prepositions) in the German language. On the other hand, in French, the wide range of types of mistakes increases with additional syntactical and formal (punctuation) errors and also other lexical errors related to the interrogative adverbs used.

5. Conclusions and future lines of research

Despite their short history of use, QA systems constitute an interesting option for the retrieval of information and the satisfaction of user needs. These systems aspire to reach greater realms of usability and interaction by improving their informational search procedures, yet there are many obstacles that must be overcome, particularly in relation with natural language processing. The QA system based on the translation of questions must also find an optimal solution for the translation stage of the process in order to guarantee adequate levels of precision overall.

Here we studied the workings of Google Translator, Promt Translator and Worldlingo when working from German and French to produce questions in Spanish destined for a QA system. Both objective and subjective evaluation measures were applied. The indexes of error for these systems were seen to be greater in the translations from German to Spanish, making manifest the greater difficulties for automatic translating tools in working with languages that are comparatively dissimilar. This suggests a long road ahead for automatic translating into Spanish when we are dealing with languages distant from English or the Romance languages. In this sense, we determined that the linguistic resources used by automatic translators for the German-Spanish language pair are less effective than for French, as shown by the results of translation evaluation. We identified Promt as the most reliable translator, on the average, for the two linguistic pairs studied. However, for German-Spanish online translating, a good evaluation was obtained by the Google Translator.

We are also able to underline the need to complement evaluation based on the determination of error with more detailed analysis of the behaviour of the translators involved. While indeed the objective and subjective translation evaluation measures give us the rates of error, they do not point to the cause thereof. Our approach serves to typify the translation errors and problems most often encountered, and it could therefore contribute to improving the techniques underlying online translation. It is noteworthy that most of the errors were of the lexical sort, or originating from a poor translation of the interrogative adverb or particle of the question, when these would seem to be matters easily resolved by automatic translating systems. The present study focused only on the most representative errors. More types of errors were not itemized because the number of cases was so small that it would have made it difficult to typify representative categorization. Anyhow, lexical errors have been identified, all of them concerning vocabulary which varies with the semantic context of each sentence. The translation mistakes of the interrogative particles identified in Table II vary according to the question and the online automatic translator used. The findings show that there is not a direct relationship between the type of interrogative particle and the error produced. It is worth noting, however, that the mismatch among automatic translators when they make an error proves that the information stored into each online automatic translator could complement one another so as to improve their efficiency.

As was pointed out previously, it has been observed that no online automatic translator is better than the rest to translate a type of question. It is encouraging to see that the findings of this study have provided some support to the assumption that the combination of factors such as language and formal construction of the question determines the translation success rate by each of the tool analysed.

In the near future our work along these research lines will take us to explore the use of other objective and subjective evaluation measures, such as those based on automatic learning, which attempt to reconcile human evaluation with automatic evaluation.

On the basis of the data obtained, one gets the impression that at present, translingual QA is not a very effective method. Nevertheless, it can be said that:

- First, research has found that free online automatic translators nowadays are not really effective for these types of sentences. Consequently, these tools must be improved to achieve an efficient specialized and contextualized translation. On the other hand, our analysis not only reveals automatic translators' shortcomings but also the notion that the formal aspect of the question or input determines the success of the translation. The homogeneity of the construction of the definition and closed list questions reduces their translation errors on the interrogative adverbs in general and on the lexicon in the case of the French language. Moreover, factual questions have generated a wider range of errors, above all in French. This fact obviously produces a lower number of cases for each type of question.
- Furthermore, this study reveals the main types of formal and syntactic errors which can be used to develop syntax rules to be added to automatic translator. And, as a result of that, they can help to correct the translation process and improve the resultant sentence too. When an automatic translator produces a good translation of an interrogative adverb over the other automatic translators

which are unable to do it, some interesting observations can be made. The improvement of QA translation effectiveness depends on the automatic translator's stored information as well as on the vocabulary interchange, sentence construction algorithms, context identification algorithms and so on.

Likewise, we plan to assess other tools and strategies for the automatic translation of questions for the final purpose of helping design efficient QA systems.

References

Abusalah, M., Tait, J. and Oakes, M. (2005), "Literature review of cross-language information retrieval", *Proceedings of World Academy of Science, Engineering and Technology*, Vol. 4, pp. 175-7.

Airio, E. (2008), "Who benefits from CLIR in web retrieval?", *Journal of Documentation*, Vol. 64 No. 5, pp. 760-78.

Banerjee, S. and Lavie, A. (2005), "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments", *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, MI*.

Callison-Burch, C., Osborne, M. and Koehn, P. (2006), "Re-evaluating the role of BLEU in machine translation research", *Proceedings of the EACL 2006: 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy*, pp. 249-56.

Cui, H., Kan, M.-Y., Chua, T.-S. and Xiao, J. (2004), "A comparative study on sentence retrieval for definitional question answering", *Proceedings of the SIGIR Workshop on Information Retrieval for Question Answering (IR4QA), Sheffield, 29 July*.

Doddington, G. (2002), "Automatic evaluation of machine translation quality using *n*-gram co-occurrence statistics", *Proceedings of the 2nd International Conference on Human Language Technology Research, San Diego, CA*, pp. 128-32.

García Cumbreras, M.Á., Ángel, M. and Santiago, M. (2005), "Búsqueda de respuestas multilingüe: clasificación de preguntas en español basadas en aprendizaje", *Procesamiento del lenguaje natural*, Vol. 34, pp. 31-40.

Green, A., Wolf, A.K., Chomsky, C. and Laughery, K. (1961), "Baseball: an automatic question answerer", *Proceedings of the Western Joint Computer Conference, Los Angeles, CA*, pp. 219-24.

Hansen, P. and Karlgren, J. (2005), "Effects of foreign language and task scenario on relevance assessment", *Journal of Documentation*, Vol. 61 No. 5, pp. 623-38.

Hermjakob, U. (2001), "Parsing and question classification for question answering", *Annual Meeting of the ACL: Proceedings of the Workshop on Open-Domain Question Answering. Toulouse, France*, pp. 1-6.

Hull, D.A. and Grefenstette, G. (1996), "Querying across languages: a dictionary-based approach to multilingual information retrieval", *Proceedings of the 19th International Conference on Research and Development in Information Retrieval, Zurich*, pp. 49-57.

Jones, G.J.F., Fantino, F., Newman, E. and Zhang, Y. (2008), "Domain-specific query translation for multilingual information access using machine translation augmented with dictionaries mined from Wikipedia", *Proceedings of the 2nd International Workshop on Cross-Lingual Information Access Addressing the Information Need of Multilingual Societies, Chicago, IL, July*.

Jones, G.J.F., Sakai, T., Collier, N., Kumano, A. and Sumita, K. (1999), "A comparison of query translation methods for English-Japanese cross-language information retrieval (poster abstract)", *Annual ACM Conference on Research and Development in Information Retrieval Archive, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA*, pp. 269-70.

Kishida, K. (2005), "Technical issues of cross-language information retrieval: a review", *Information Processing & Management*, Vol. 41 No. 3, pp. 433-55.

Kwok, C., Etzioni, O. and Weld, D.S. (2001), "Scaling question answering to the web", *ACM Transactions on Information Systems*, Vol. 19 No. 3, pp. 242-62.

Larosa, S. (2005), "Best translation for an Italian-Spanish question-answering system", *Proceedings Information and Communication Technologies International Symposium, ICTIS'2005, Tetuan, Morocco, June 3-6*.

Leusch, G., Ueffing, N. and Ney, H. (2003), "A novel string-to-string distance measure with applications to machine translation evaluation", *Proceedings of MT Summit IX, New Orleans, LA*.

Levenshtein, V.I. (1966), "Binary codes capable of correcting deletions, insertions and reversals", *Soviet Physics Doklady*, Vol. 10, pp. 707-10.

López-Ostenero, F., Gonzalo, J. and Verdejo, F. (2004), "Búsqueda de información multilingüe: estado del arte", *Revista Iberoamericana de Inteligencia Artificial*, Vol. 8 No. 22, pp. 11-35.

Melamed, D., Green, R. and Turian, J. (2003), "Precision and recall of machine translation", *Proceedings of the HLT-NAACL, Edmonton*.

Nießen, S., Och, F.J., Leusch, G. and Ney, H. (2000), "An evaluation tool for machine translation: fast evaluation for MT research", *Proceedings of the 2nd International Conference on Language Resources and Evaluation, Athens*, pp. 39-45.

Oard, D.W. and Diekema, A. (1998), "Cross-language information retrieval", *Annual Review of Information Science and Technology*, Vol. 33, pp. 223-56.

Oard, D.W., He, D. and Wang, J. (2008), "User-assisted query translation for interactive cross-language information retrieval", *Information Processing & Management*, Vol. 44 No. 1, pp. 181-211.

Oard, D.W., Gonzalo, J., Sanderson, M., Lopez-Ostenero, F. and Wang, J. (2004), "Interactive cross-language document selection", *Information Retrieval*, Vol. 7 No. 1 & 2, pp. 205-28.

Papineni, K., Roukos, S., Ward, T. and Zhu, W.J. (2002), "BLEU: a method for automatic evaluation of machine translation", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, PA*, pp. 311-18.

Pérez, A., González, J., Casacuberta, F. and Torres, I. (2004), "Traducción automática mediante transductores estocásticos de estados finitos basados en gramáticas k-explorables", *Actas de las III Jornadas Techabla*, Valencia, pp. 207-12.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J. (2005), "Study of translation error rate with targeted human annotation", paper presented at Machine Translation Workshop, NIST, North Bethesda, MD.

Sokolova, S. (2007), *How the Computer Translates*, available at: www.promt.com/company/technology/pdf/e_how_computer_translates_sokolova.pdf (accessed 12 June 2008).

Tillman, C., Vogel, S., Ney, H., Sawaf, H. and Zubiaga, A. (1997), "Accelerated DP based search for statistical translation", *Proceedings of the 5th European Conference on Speech Communication and Technology, Rhodes, Greece*, pp. 2667-70.

Tomás, J. (2003), "A quantitative method for machine translation evaluation", *Proceedings of the of EACL 2003 Workshop on Evaluation Initiatives in Natural Language Processing, Budapest*.

Vicedo, J.L. (2004), "La Búsqueda de Respuestas: Estado Actual y Perspectivas de Futuro", *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial*, Vol. 8 No. 22, pp. 37-56.

Vidal, E. (1997), "Finite-state speech-to-speech translation", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich*.

Vilar, D., Xu, J., D'Haro, L.F. and Ney, N. (2006), "Error analysis of statistical machine translation output", *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genova, Italy, May*, pp. 697-702.

Volk, M., Vintar, S. and Buitelaar, P. (2003), "Ontologies in cross-language information retrieval", *Proceedings of 2nd Conference on Professional Knowledge Management, Lucerne*.

Voorhees, E.M. (1999), "The TREC 8 question-answering track report", *Proceedings of the 8th Text REtrieval Conference, College Park, MD, November*.

Vossen, P. (Ed.) (1998), "Introduction to EuroWordNet", *Computers and the Humanities*, Vol. 32 Nos 2-3, pp. 73-89.

Warren, D. (1981), "Efficient processing of interactive relational database queries expressed in logic", *Proceedings of the 7th International Conference on Very Large Databases, Cannes*, pp. 272-83.

Weizenbaum, J. (1966), "ELIZA: a computer program for the study of natural language communication between man and machine", *Communications of the ACM*, Vol. 9 No. 1, pp. 36-45.

Woods, W., Kaplan, R.M. and Nash-Webber, B. (1972), *The Lunar Sciences Natural Language Information System, BBN Final Report 2378*, Bolt, Beranek and Newman, Cambridge, MA.

Corresponding author
Lola García-Santiago can be contacted at: mdolo@ugr.es