Overcoming statistical machine translation limitations: error analysis and proposed solutions for the Catalan-Spanish language pair

Mireia Farrús*, Marta R. Costa-jussà©, José B. Mariño, Marc Poch, Adolfo Hernández, Carlos Henríquez, José A. R. Fonollosa

TALP Research Center, Department of Signal Theory and Communications
Universitat Politècnica de Catalunya
C/ Jordi Girona 1-3, 08034 Barcelona, Spain
{mfarrus,canton,mpoch,adolfohh,carloshq,adrian}@gps.tsc.upc.edu

©Barcelona Media Innovation Center, Speech and Language Department
Av Diagonal 177, 9th floor, 08018 Barcelona
marta.ruiz@barcelonamedia.org

** (corresponding autor): Tel: +34 93 401 0964, Fax: +34 93 401 6447*

Abstract

This work aims to improve an N-gram-based statistical machine translation system between the Catalan and Spanish languages, trained with an aligned Spanish-Catalan parallel corpus consisting of 1.7 million sentences taken from *El Periódico* newspaper. Starting from a linguistic error analysis above this baseline system, orthographic, morphological, lexical, semantic and syntactic problems are approached using a set of techniques. The proposed solutions include the development and application of additional statistical techniques, text pre- and post-processing tasks, and rules based on the use of grammatical categories, as well as lexical categorization. The performance of the improved system is clearly increased, as is shown in both human and automatic evaluations of the system, with a gain of about 1.1 points BLEU observed in the Spanish-to-Catalan direction of translation, and a gain of about 0.5 points in the reverse direction. The final system is freely available online as a linguistic resource.

Keywords: statistical machine translation, N-gram-based translation, linguistic knowledge, grammatical categories

Introduction

Currently, Statistical Machine Translation (SMT) is one of the most popular Machine Translation paradigms. The SMT approach allows programmers to build a translator with open-source tools as long as a parallel corpus is available. If the languages involved in the translation belong to the same language family, the translation quality can be surprisingly high. Furthermore, one of the most attractive reasons to build a statistical system instead of a standard rule-based system is that little human effort is required.

Theoretically, when using SMT, no linguistic knowledge is required. In practice, once the system is built, linguistic knowledge becomes necessary to achieve perfect translations at all grammatical levels, as it has been seen in recent works (Niessen and Ney 2000; Popović and Ney 2004; Popović and Ney 2006). In fact, the main question that arose at the beginning of this work was: *which are the steps to follow when the main goal is to improve an already high-quality statistical translation?*

We consider a translation system to be of high quality when almost no post-editing is needed. Actually, this is a relatively unusual situation in the field of MT, where most systems offer translations that have to be post-edited. This study is devoted to developing translation at this stage in the Catalan-Spanish pair in both directions. Therefore, the main objective here is to obtain a translation system that does not require any post-editing.

Starting with a high-quality baseline SMT system, a human analysis of the output is performed, which is then used to make further improvements via the introduction of statistical techniques and linguistic rules. One main consideration is the methodology of the study. Our study methodology is based primarily on the detection and classification of systematic errors, followed by the proposal of suitable solutions. When possible, solutions will be statistical; otherwise, they will be linguistic.

This paper is organized as follows. First, a brief description of the Catalan and Spanish languages and the N-gram-based statistical translation system used as the baseline system are presented. Next, we report the error analysis and classification performed on the system described in the previous section, and the proposed

solutions to solve these errors. Then, the improved system is carefully evaluated with respect to the baseline system and, finally, conclusions and future work are summarized in last section.

Catalan and Spanish languages

Catalan and Spanish belong to the family of Romanic languages, also referred to as Romance or Latin languages. This family also includes other languages, such as French, Italian, Portuguese and Romanian, as it is comprised of all the languages descending from Vulgar Latin, which is, in turn, a branch of the Indo-European language family. Although the evolution of Latin into such Romanic languages occurred at all levels (phonological, morphological, lexical, etc.), the main characteristic of these languages is found at the syntactic level, with the loss of the declension system, which has led to a generalized use of SVO sentence structures with a large amount of prepositions.

The Catalan language is spoken by more than nine million people distributed among Catalonia, the Valencian Community, the Balearic Islands, Andorra, and part of Aragon, Sardinia, Southern France and Murcia. On the other hand, Spanish is spoken by more than 400 million people as a native language in Spain and several countries in America, Asia and Africa. Spanish is the second most spoken language in the world after Mandarin Chinese.

The main sociolinguistic characteristic of Catalan is the fact that it is found in a socially bilingual environment in all the regions where it is spoken: with Spanish in Spain (Catalonia, Aragon, Murcia) and Andorra, with French in South France and Andorra, and with Italian in Sardinia. This fact highlights the need for bilingual communication in all of these regions, where helping systems and resources have become increasingly popular and necessary.

N-gram-based statistical translation system

The translation system used in the current study – called N-II[1] – is an N-gram-based SMT system developed at the Universitat Politècnica de Catalunya (UPC), trained with the aligned Spanish-Catalan parallel corpus taken from *El Periódico*

[1] http://www.n-ii.org

newspaper. This corpus contains 1.7 million sentences, a rich vocabulary and nearly $4.24 \cdot 10^7$ words in average, as shown in Table 1.

	Catalan	***Spanish***
Sentences	1.7 M	
Words	41.5 M	43.3 M
Vocabulary	397.4 k	390.2 k

Table 1. El Periódico corpus statistics.

An N-gram-based SMT system regards translation as a stochastic process. We are given a source string $s_1^J = s_1 \ldots s_j \ldots s_J$, which is to be translated into a target string $t_1^I = t_1 \ldots t_i \ldots t_I$. Among all possible target strings, we will choose the string with the highest probability:

$$\tilde{t}_1^I = \arg\max_{t_1^I} P\left(t_1^I \mid s_1^J\right) \quad (1)$$

The *argmax* operation denotes the search problem, i.e. the generation of the output sentence in the target language. In recent systems, such an approach will assume a general maximum entropy in which a log-linear combination of multiple-feature functions is implemented (Och 2003). This approach tends to maximize a linear combination of feature functions:

$$\tilde{t} = \arg\max_{t} \left\{ \sum_{m=1}^{M} \lambda_m h_m\left(t, s\right) \right\} \quad (2)$$

Translation feature functions

The main feature function is the N-gram-based translation model, which is trained with bilingual *n*-grams. This model constitutes a language model of a particular *bi-language* composed of bilingual units (translation units) which are referred to as *tuples*. In this way, translation model probabilities at the sentence level are approximated by using *n*-grams of tuples, such as described by the following equation:

$$\hat{t}_1^I = \arg\max_{t_1^I} \left\{ p\left(s_1^J, t_1^J\right) \right\} = \ldots = \arg\max_{t_1^I} \left\{ \prod_{n=1}^{N} p\left(\left(s,t\right)_n \mid \left(s,t\right)_{n-x+1}, \ldots, \left(s,t\right)_{n-1}\right) \right\} \quad (3)$$

where the *n-th* tuple of a sentence pair is referred to as $(s,t)_n$. As any standard *n*-gram language model, the bilingual translation model is estimated over a training

corpus composed of sentences in the language being modeled. In this case, we consider sentences in the *bilanguage* mentioned previously.

The N-gram-based approach is monotonic in that its model is based on the sequential order of tuples during training. Therefore, the baseline system may be especially appropriate for pairs of languages with relatively similar word order schemes.

Tuples are extracted from a word-to-word aligned corpus in such a manner that a unique segmentation of the bilingual corpus is achieved. Although in principle, any Viterbi alignment should allow for tuple extraction, the resulting tuple vocabulary strongly depends on the particular alignment set considered. According to our experience (Mariño, Banchs et al. 2006), in some specific tasks the best performance is achieved when the union of the source-to-target and target-to-source alignment sets is used for tuple extraction.

In this way, which is different from other implementations, where one-to-one (Bangalore and Riccardi 2000) or one-to-many (Casacuberta and Vidal 2004) alignments are used, tuples are extracted from many-to-many alignments. This implementation produces a monotonic segmentation of bilingual sentence pairs, which allows for the simultaneous capture of contextual information and the reordering of information into the bilingual translation unit structures. This segmentation also allows the estimation of the *n*-gram probabilities. To guarantee a unique segmentation of the corpus, tuple extraction is performed according to the following constraints:

- a monotonic segmentation of each bilingual sentence pair is produced,
- no word inside the tuple is aligned to words outside the tuple, and
- no smaller tuples can be extracted without violating the previous constraints.

Illustration 1 presents a simple example showing the unique tuple segmentation for a given pair of sentences, which mean *he left at half-past one*.

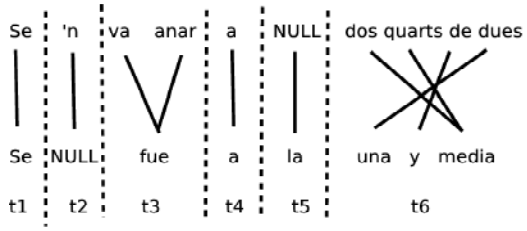


Illustration 1: Example of tuple extraction in the sentence *he left at half past one*.

Two important observations from *Illustration 1*, related to *null* and *embedded words*, must be considered:

NULL. In some cases, such as tuple 2, there are possible occurrences of tuples containing unaligned elements in its target side. This kind of tuple should be handled in an alternative manner in order for the system to be able to provide appropriate translations for such unaligned elements. The problem of how to handle this kind of situation has been discussed in detail (Mariño, Banchs et al. 2006). In short, since no NULL is actually expected to occur in translation inputs, this type of tuple is not allowed. Any target word that is linked to NULL is attached either to the word that precedes or to the word that follows it. To determine this, we use the IBM-1 probabilities. More specifically, the IBM-1 lexical parameters (Brown, della Pietra et al. 1993) are used for computing the translation probabilities of two possible new tuples: the one that results when the null-aligned-word is attached to the previous word, and the one that results when it is attached to the following one. Then, the attachment direction is selected according to the tuple with the highest translation probability.

Embedded words. Often, a large number of single-word translation probabilities are left out of the model. This happens for words that are always embedded in tuples containing two or more words. Consider for example the word *quarts* in *Illustration 1*. This word is embedded into tuple *6*. If a similar situation is encountered for all occurrences of *quarts* in the training corpus, then no translation probability for an independent occurrence of this word will exist. To overcome this problem, the tuple *n*-gram model is enhanced by incorporating *1*-gram translation probabilities for all the embedded words detected during the tuple extraction step. These *1*-gram translation probabilities are computed from the intersection of both the source-to-target and the target-to-source alignments.

Additional feature functions

Apart from the translation model, state-of-the-art systems include several component models are introduced in the maximum entropy approach: typically the target language model and the word bonus. For the Catalan-Spanish language pair, curiously, it is seen that the target language model does not help improving the translation quality (de Gispert, Mariño, 2006). However, the POS target language model, which introduces linguistic knowledge (Crego, de Gispert et al.

2006), may help. This POS target language model has to be combined together with the word bonus model for each produced target word, which compensates for the POS target language model preference for short sentences.

Error analysis of the baseline N-II system

The linguistic error analysis of the baseline N-II system was performed by analyzing, according to the standards of the Institute of Catalan Studies[2] and the Royal Spanish Academy[3], several press articles from the digitalized version of the following newspapers: *El Periódico*[4] (in both Spanish and Catalan), *La Vanguardia*[5] (Spanish), *Avui*[6] and *El Punt*[7] (Catalan). Since it is well-known that the same sentence can be translated in many different ways, the following criterion was applied in order to decide whether a sentence was correct or not: all the translations achieved can be considered as correct if they maintain the meaning of the original sentence and are grammatically correct.
The main errors found in the translation system were classified according to their corresponding linguistic level: orthographic, morphological, lexical, semantic, and syntactic. Below, the specific problems are described and exemplified in detail and summarized in Table 2, which specifies the direction of translation where the errors are found: Catalan to Spanish (ca2es), Spanish to Catalan (es2ca) or both.

Orthographic errors

Geminated l (l·l)

The geminated l is one of the characteristics of Catalan that differentiates it from the other Romanic languages. The correct way of writing this grapheme involves use of the middle dot (·); however, the incorrect use of the normal dot (.) is quite widespread among users. The system translates correctly the words containing a *geminated l* (*l·l*) provided that they are well-written – i.e. using the middle dot. Otherwise, the tokenization process is not performed properly, and erroneous translations are generated. An example related to this kind of error from Catalan

[2] http://www.iec.cat
[3] http://www.rae.es
[4] http://www.elperiodico.com
[5] http://www.lavanguardia.es
[6] http://www.avui.cat
[7] http://www.vilaweb.cat/www/elpunt

into Spanish is illustrated below, where the asterisk symbol (*) refers to the wrong translation.

***CA**: Aquesta normativa acaba de ser reformada a **Brussel.les** per liberalitzar la presència de publicitat de la televisió encara més del que ho està.*
***ES**: Esta normativa acaba de ser reformada en ***Bruselas. las** para liberalizar la presencia de publicidad en televisión aún más de lo que está.*
EN: These rules have been reformed in **Brussels** in order to liberalize still more the advertisements on TV.

Apostrophe

The apostrophe is used in Catalan to elide a sound. The singular articles (*el*, *la*), the preposition *de* and some pronouns are apostrophized in front of words beginning with a vowel or an *h* plus a vowel, though there are some exceptions. In some cases, the baseline N-II does not apostrophize some of the words that should be apostrophized:

***ES**: 'Cuando se acepta una defensa, **la acepta** hasta el final', dijo Plaza.*
***CA**: 'Quan s'accepta una defensa, ***la accepta** fins al final', va dir Plaza.*
EN: When a defense is accepted, **it is accepted** until the end', said Plaza.

And some words are apostrophized when they should not be:

***ES**: Ante una situación tan dramática, los mandos **del Luz** del Mar y Conde de Gondomar lanzaron un SOS a todos los barcos que pudieran navegar en la zona.*
***CA**: Davant d'una situación tan dramática, els comandaments ***del l'Llum** de Mar i Conde de Gondomar van llançar us SOS a tots els vaixells que poguessin navegar a la zona.*
EN: In view of such a dramatic situation, the command **of Luz** del Mar and Conde de Gondomar launched an SOS to all those ships that could be sailing in the area.

Coordinating conjunctions y *and* o

The Spanish coordinating conjunctions *y* and *o* change to *e* and *u* when they precede words starting by *i* or *o (*or *hi* or *ho)*, respectively. The baseline N-II system usually fails to perform such conversions, leading to an incorrect use of the conjunctions:

***CA**: Que en alguns casos, com Blanes **o** Olot, s'estenen interanualment.*
***ES**: Que en algunos casos, como Blanes ***o** Olot, se extienden interanualmente.*
EN: That in some cases, such as Blanes **or** Olot, they are extended interannually.

Morphological errors

Lack of gender concordance

Some words are given a different gender in Catalan and Spanish languages. For instance, the word *signal* is feminine in Spanish (*la señal*) and masculine in Catalan (*el senyal*). Therefore, it is common to find a lack of gender concordance in articles and adjectives with a noun that changes its gender from one language to the other:

***CA**: Ali tornarà a recuperar no només les cames i la mà, sinó també **el somriure**.*
***ES**: Ali recuperará de nuevo no solo las piernas y la mano, sino también ***el sonrisa**.*
EN: Ali will recover again; not only his legs and hand, but also his **smile**.

This lack of concordance affects also the Spanish-Catalan translation of some invariant possessives such as *mi*, *mis* (my), tu, tus (your), etc.

Lexical errors

Capital letters & Numbers

Other cases related to unknown words can be found at the lexical level; for instance, in numbers and words that appear only in capital letters in the training corpus. In this case, they will be recognized with capital letters:

***CA**: **Assisteixes** a tres actes consecutius.*
***ES**: **Asistes** a tres actos consecutivos.*
EN: **You are attending** three acts in a row.

and not recognized when written in lower case:

***CA**: No entenc per què no hi **assisteixes**.*
***ES**: No entiendo por qué no ***assisteixes**.*
EN: I don't understand why **you are** not **attending** to it.

Time

In general, time expressions in Spanish and Catalan differ considerably. The main difference is found in the use of the quarters: while in Spanish the time is expressed through the quarters that *pass* a specific hour, Catalan talks about the quarters that are approaching a specific hour. E.g.:

***ES (4:15h)**: Las cuatro y cuarto. (Literally: 'four and a quarter').*
***CA (4:15h)**: Un quart de cinc.* (Literally: 'a quarter of five').
EN (4:15h): It's a quarter past four.

These differences in the time expression lead to erroneous translations in the baseline N-II system:

***CA (7:45h)**: Són **tres quarts de vuit**.*
***ES (7:45h)**: Son * **tres cuartos de ocho**.*
EN (7:45h): It's **a quarter to eight**.

Semantic errors

The perquè *final/causal conjunction*

The conjunction *perquè* in Catalan is a polysemy-related case. When used to introduce a causal clause, it must be translated into *porque* (because). However, when it introduces a final subordinate clause, the meaning changes, and it should be translated into *para que* (in order to). Nevertheless, the baseline N-II system usually translates this conjunction into *porque*, irrespective of its meaning:

***CA**: Els agents van reduir al conductor per evitar que els fes explotar en cas de portar un detonador portátil, i van reclamar un equip d'artificiers **perquè** els desactivessin.*
***ES**: Los agentes redujeron al conductor para evitar que les hiciera explotar en caso de llevar un detonador portátil, y reclamaron un equipo de artificeros ***porque** los desactivessin.*
EN: The agents subdued the driver in order to prevent the explosion in case he had brought a portable detonator, and they called for bomb disposal experts **in order to** deactivate them.

Solo *adverb/adjective*

In Spanish, *solo* can be both an adjective and an adverb. Irrespective of its grammatical function, it must be used without a written accent, unless a possible semantic confusion exists. In this case, the adverb must be accented. However, the

training corpus does not always follow this rule: the adverb is usually accented, even when it is not necessary. Therefore, the non-accented *solo* is usually translated as if it was an adjective:

***ES**: Porque se quiera o no – y más en estos tiempos –, **solo** cuentan los que siguen vivos y no los que caen por descalificación o accidente.*
***CA**: Perquè es vulgui o no – i més en aquests temps - , ***sol** compten els que continúen vius i no els que cauen per desqualificació o accident.*
EN: Because like it or not – especially in the current days – **only** those who are still alive and not falling for disqualification or injury count.

The verb soler

Another homonymy-related case is found in the verb *soler* (to be used to). Usually, N-II translates the Catalan verb *sols* or *sol* (corresponding to the 2nd and 3rd singular persons of present tense in indicative mood, respectively, of the verb *soler*), into the Spanish adjectives *solos* and *solo* (alone), instead of the corresponding verb conjugations *sueles* and *suele:*

***CA**: La Creu Roja **sol** disposar de quatre lliteres per atendre els immigrants que arriben en piragua al port de los Cristianos, però ahir n'hi tenia vuit.*
***ES**: La Cruz Roja ***solo** disponer de cuatro literas para atender a los inmigrantes que llegan en piragua en el puerto de los Cristianos, pero ayer tenía ocho.*
EN: The Red Cross **usually has** four bunk beds, but yesterday it had eight.

Possessive pronouns and adjectives

A systematic error was observed in translations between possessive pronouns from Catalan to Spanish: the possessive pronouns *meu* (mine), *teu* (yours), *seu* (his/hers), etc. are translated as the possessive pronouns *mi* (my), *tu* (your), *su* (his/her), respectively, instead of *mío*, *tuyo*, *suyo*, etc., as it should be:

***ES**: No és fàcil portar la pàtria a escena perquè cadascú té **la seva**.*
***CA**: No es fácil llevar la patria a escena porque cada uno tiene ***su**.*
EN: It is not easy to bring the homeland on stage because everyone has **their own**.

Syntactic errors

Pronominal clitics

Usually, the pronominal clitics lead to translation errors in both Catalan-to-Spanish and Spanish-to-Catalan directions. The most frequent errors are due to an erroneous combination with the corresponding verb:

***ES**: No quiero ver**te** más por aquí.*
***CA**: No vull veure ***et** més per aquí.*
EN: I don't want to see **you** here anymore.

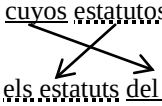
or the elision of the pronoun in the target language:

***ES**: Desde la operación el pequeño se mueve arrastrándo**se** por el suelo.*
***CA**: Des de l'operació el nen es mou arrossegant ***[]** per terra.*
EN: Since he underwent surgery, the little one crawls on the floor.

The relative pronoun cuyo

The translation of Spanish relative pronoun *cuyo* (whose) involves a reordering of elements when it is to be translated into Catalan:

Es un organismo cuyos estatutos están pendientes de aprobación.



És un organismo els estatuts del qual están pendents d'aprovació.

If this reordering is not considered, the N-II will lead to erroneous translations:

***ES**: Es un organismo **cuyos** estatutos están pendientes de aprobación.*
***CA**: És un organisme ***que té** els seus estatuts ***están** pendents d'aprovació.*
EN: It is an organism **whose** statutes are pending approval.

Obligation

The obligation in Catalan and Spanish are expressed in different way: where Spanish uses the verb *tener que* (to must) + infinitive, Catalan utilizes the form *haver de* + infinitive. The verbal periphrasis *tenir que* + infinitive, which is not correct, is usually obtained in the Spanish-Catalan translation:

***ES**: Nos los **tenemos que** creer, es lo único que nos falta.*
***CA**: Ens ho ***tenim que** creure, és l'únic que ens falta.*
EN: We **must** believe it, it is the only thing we need.

Elision/insertion of the preposition de

It has been seen previously that the verbal periphrasis *tener que* was translated as *tenir que* instead of *haver de*. However, when translating the Spanish form *deber*, which has the same meaning, the correct expression, *haver de*, is obtained, but the preposition *de* is usually missing:

***ES**: Las islas Británicas (…) están teniendo unos meses de junio y julio extraordinariamente lluviosos y frescos por la continua entrada de borrascas atlánticas que, por estas fechas, **deberían** discurrir por latitudes mayores.*
***CA**: Les illes Britàniques (…) estan tenint uns mesos de juny i juliol extraordinàriament plujosos i frescos per la contínua entrada de borrasques atlàntiques que, per aquestes dates, *** haurien []** transcórrer per latituds més grans.*
EN: The British Isles (…) are having unusually rainy and cool June and July months due to the continuous income of Atlantic squalls which, by these days, **should be** running at higher latitudes.

Moreover, the preposition *de* is also elided when translating *desde* (from, since) to Catalan, obtaining *des* instead of *des de*:

***ES**: **Desde** el neolítico hasta la actualidad hemos modificado profundamente las plantas y los animales silvestres.*
***CA**: ***Des []** el període del neolític fins a l'actualitat hem modificat profundament les plantes i els animals silvestres.*
EN: **Since** the Neolithic Age era until now we have strongly modified the plants and wildlife.

On the other hand, the preposition *de* falls in front of the conjunction *que* in Catalan, and this is normally not taken into account by the translator:

***ES**: Tenías que acordarte **de que** hoy era el día.*
***CA**: Havies de recordar-te ***de que** avui era el día.*
EN: You should have remembered **that** today was the day.

Level	*Error type*	*Specific problem*	*ca2es*	*es2ca*
Orthography		geminated *l*	x	
		apostrophe		x
		y/o conjunctions	x	
Morphology	concordance	gender concordance	x	x
Lexicon	unknown words	capital letters	x	x
		numbers	x	x
	expression	time	x	x
Semantics	polysemy	*perquè*	x	
	homonymy	*solo*		x
		soler	x	
		possessives	x	
Syntax	pronouns	clitics	x	x
		cuyo	x	x
	verbal periphrasis	obligation		x
	prepositions	elision *de*		x

Table 2. Error classification and their corresponding direction of translation.

Proposed solutions

To resolve the errors described in the previous section, several techniques need to be applied according to the idiosyncrasy of the problem. These can be classified into several types: direct text processing–which will be mainly used to solve orthographic errors–as well as the use of rules based on grammatical categories, the statistical model and word categorization. Below, we present a detailed description of such techniques and the errors they can solve.

Text edition

Some of the errors need to be solved by processing the text before or after the translation. The *geminated l,* for instance, was corrected before translation by normalizing the writing of the middle dot. Other cases, such as the obligation *tener que* and the conjunctions *y* and *o*, have been corrected through post-processing after the translation. Some examples of correction by text edition can be found in Table 3.

geminated l	(S) S'ha reformat a **Brussel.les**. *(It has been reformed in Brussels).* (T1) Se ha reformado en ***Bruselas. las**. (T2) Se ha reformado en **Bruselas**.
obligation	(S) Nos lo **tenemos que** creer. (*We have to believe it*). (T1) Ens ho ***tenim que** creure. (T2) Ens ho **hem de** creure.
conjunctions y/o	(S) Que en alguns casos, com Blanes **o** Olot, s'estenen interanualment. (T1) Que en algunos casos, como Blanes ***o** Olot, se extienden interanualmente. (T2) Que en algunos casos, como Blanes **u** Olot, se extienden interanualmente.

Table 3. Examples of text edition correction.

Grammatical category-based approach

Grammatical categories have been successfully used in statistical translation in order to deal with several problems like reordering (Crego and Mariño 2007) and automatic error analysis (Popović, de Gispert et al. 2006). The aim is to add the grammatical category, via a tag, corresponding to the word to translate, so that the statistical model is capable of distinguishing the word depending on its category and to learn from context.

In the current improved system, the grammatical category was provided by the *Freeling* tool (Carreras, Chao et al. 2004), and this information was then used to solve some of the problems found in the baseline system. The grammatical category can be used either in *pre- or post-processing rules*, or in the translation model as a *decoder-based solution*. While the former includes solutions for apostrophes, clitics, capital letters at the beginnings of sentences, the relative pronoun *cuyo* and polysemy disambiguation, the latter becomes useful for the homonymy disambiguation and the lack of gender concordance. A detailed description of the problems solved by this grammatical category-based approach is presented below.

Apostrophe

Use of the apostrophe in Catalan relies mostly on basic rules in which the article *el*, *la* and the preposition *de* are elided when preceding words that begin with a vowel or a silent *h:*

***el** arbre*	***l'**arbre* (the tree)
***la** hora*	***l'**hora* (the time)
***de** eines*	***d'**eines* (of tools)

Nevertheless, there are a considerable number of exceptions related to these rules, for instance:

exception rule	**example**
1. *el, la* and *de* when preceding words beginning with a semiconsonantic vowel	*el uombat* (the wombat), *la hiena* (the hyena), *de iogurt* (of yogurt)
2. *la* in front of words beginning with unstressed vowels *i* and *u*	*la universitat* (the university), *la Irene* (Irene)
3. *la* and *de* in front of the negative prefix *a*	*la anormalitat* (the anormality), *de asimètric* (of asimetric)
4. names of letters	*la e*, *la erra*, *la hac*, etc. (the e, the r, the h)
5. exception words	*la una* (one o'clock), *la ira* (the wrath), *la host* (the host)

Pronominal clitics

Pronominal clitics are first detected by using the *Freeling* tool and are separated from the corresponding verb. Then, the isolated pronominal clitic is translated and, finally, a rules-based text correction is performed as a post-processing of the translation. These correction rules consider the following factors:

1. The Spanish accentuation rules, since the position of the stressed syllable changes when adding an enclitic pronoun to the verb:

   vende+***lo*** | *véndelo* (sell **it**)

2. The combination of the Catalan pronouns, which, unlike Spanish pronouns, use hyphens or apostrophes, but the accentuation rules are not altered.

seguir+***lo***	*seguir-**lo*** (follow **it**)
***la** hora*	*compra'**l*** (buy **it**)
***de** eines*	***d**'eines* (**of** tools)
el+*aixecava*	***l**'aixecava* (lifted **it**)

Some examples of clitics and apostrophe correction can be found in Table 4.

clitics	(S) No quiero ver**te** más por aquí. *(I don't want to see you here anymore).* (T1) No vull veure ***et** més per aquí. (T2) No vull veure**'t** més per aquí.
apostrophe	(S) **La acepta** hasta el final. (*He accepts it until the end*). (T1) ***La accepta** fins al final. (T2) **L'accepta** fins al final.

Table 4. Examples of clitics and apostrophe correction.

Capital letters at the beginning of the sentence

To solve one of the causes of unknown words, a technique that uses morphological information is used: all the words that can be found at the sentence beginning are changed into lower case words, except for common and proper nouns and adjectives, which tend also to be proper nouns and are usually not found at the beginning of sentences. Therefore, all those words that were found with capital letters in the training corpus will no longer seem to be unknown words when lower case letters are used during translation. Table 5 shows an example in which, given a source sentence (S), we see the corresponding translations before (T1) and after (T2) applying the capital letters processing.

Cuyo *pronoun*

Problems related to the relative pronoun *cuyo* mainly arise because few sentences including this pronoun are contained in the training corpus, and so the system is not able to learn the specific structure properly. To solve this problem, a preprocessing rule was applied to transform the Spanish structure into a literal translation of the Catalan structure *del qual*; in other words, the sentences containing *cuyo* or some of its other forms (*cuya, cuyos, cuyas*) were transformed into sentences containing *del cual* or its corresponding forms (*de la cual, de los cuales, de las cuales)*, so that alignment was easier. Some translation errors related to this pronoun were avoided in this way (see Table 5).
A similar process was then performed in the inverse direction --Catalan-to-Spanish-- as a post-processing procedure: when the Catalan grammatical structure that corresponds to the Spanish structure that makes uses of *cuyo* is found, *del qual* is translated into *cuyo*, and the proper word reordering is performed. Otherwise, it is literally translated into *del cual*.

capital letters	(S) No entenc per què no hi **assisteixes**. (T1) No entiendo por qué no ***assisteixes**. (T2) No entiendo por qué no **asistes**.
cuyo	(S) Un pueblo **cuyo** nombre es largo. (*A village whose name is long*). (T1) Un poble ***amb un nom** és llarg.. (T2) Un poble el nom **del qual** és llarg.

Table 5. Examples of translation before and after applying capital letters processing and 'cuyo' correction.

Polysemy disambiguation

As noted in the previous section, the conjunction *perquè* is translated into *para que* when followed by a final subordinate clause, which is expressed by a verb in the subjunctive mood. Thus, when a verb in the subjunctive is detected after the conjunction *perquè*--even if there are other words in between--the translation into Spanish will be *para que*. In the same way, another rule will fix errors that occur if the verb is expressed in the indicative mood, so that the translation will be *porque.* Table 6 shows an example of correcting the use of this conjunction. This rule is not 100% fulfilled, since it is possible to find a verb in the indicative which is preceded by the conjunction *porque*. However, an estimation made from the training corpus showed that this structure is found in less than the 0.5% of the cases in which the conjunction *porque* appears.

Homonymy disambiguation

In the translation task, it is quite common to find two words spelled the same in the source language that differ in the target language. When both words are homonymous and they are characterized by a different grammatical category, such a category will be useful to disambiguate their meaning. The verb *soler*, the adjective/adverb *solo* and the possessives are examples of homonymous words that lead to problematic ambiguity. In some cases, the tag given by *Freeling* is simply added to the homonymous words to disambiguate them. In other cases, the output tag given by *Freeling* is not correct – as in the case of the adjective/adverb *solo* and the possessives – so it becomes necessary to establish a set of rules that will provide the correct tag. This set of rules is described below.

The verb *soler*

In order to disambiguate the homonymous words *sol* and *sols*, the corresponding adjective or verb tag given by the *Freeling* was added to the words in question.

The *solo* case

In order to solve the *solo* problem, the following rules were implemented in order to identify, in doubtful cases, whether the *solo* term was an adverb or an adjective, since the *Freeling* tool would output erroneous tags.

category	context	example
adjective	at the beginning of a sentence or after a semicolon, followed by conjunction or comma	*Solo, cansado y desesperado, se marchó.* (Alone, tired and desperate, he left)
	when followed by one of these punctuation marks: *. , ; ! ? " ( )* –	*Estoy solo*. (I am alone).
	before the conjunctions *y* or *o*	*Estaba solo y triste.* (He was alone and sad).
adverb	at the beginning of a sentence or after a semicolon and not followed by a conjunction or comma	*Solo quiero manzanas.* (I just want apples).
	after the verbs *ser* and *haber*, followed optionally by the adverb *tan*	*Esto es (tan) solo es principio del partido.* (This is just the beginning of the match.
	after a preposition	*Se ha producido a solo tres meses de los Juegos Olímpicos.* (It happened only three months before the Olympic Games)
	at the beginning of a sentence and followed by a verb	*Solo quiero escuchar un poco de música.* (I only want to listen to some music).
	after a verb with which there is no number concordance	*Vinieron solo para comer*. (They came only for lunch).

These rules were then applied to the source language and the corresponding tag was added to the word. Thus, a source sentence like *Venía solo* (he was coming alone) was changed to *Venía solo*_<ADJ>, and this allowed the model to distinguish both possibilities and to learn from context.

Possessives

A similar task was performed on Catalan possessive adjectives and pronouns. A set of rules (listed below) was designed in order to tag a specific word as an adjective or pronoun. These tags are then added to the Catalan corpus.

category	context	example
adjective	before noun or adjective	*El meu cotxe.* (My car).
pronoun	otherwise	*Aquest cotxe és meu.* (This car is mine).

Table 6 shows some examples of correction by using homonymy disambiguation in the three cases described in this section: *soler, solo* and possessives. Given a source sentence (S), the resulting translations before (T1) and after (T2) applying the obtained tags are shown. The errors produced in T1 are solved in T2.

perquè	(S) T'ho pregunto perquè m'ho **diguis**. (T1) Te lo pregunto ***porque** me lo digas. (T2) Te lo pregunto **para que** me lo digas.
soler	(S) La Creu Roja **sol** disposar de quatre lliteres. (T1) La Cruz Roja ***solo** disponer de cuatro literas. (T2) La Cruz Roja **suele** disponer de cuatro literas.
solo	(S) Era **solo** un niño. (T1) Era ***sol** un nen. *(T2) **Només** era un nen.*
possessives	(S) Els meus amics no són **els teus**. (T1) Mis amigos no están ***tus**. (T2) Mis amigos no son **los tuyos**.

Table 6. Examples of correction after homonymy and polysemy disambiguation.

Gender concordance

The problem of gender concordance is solved by using a POS language model in the target language. This allows us to assign higher value to those word sequences that maintain gender coherence; therefore, the likelihood of a sequence like *pilota*_FN *vermella*_FAdj (where FN is a feminine noun and FAdj a feminine adjective) will be higher that *pilota*_FN *vermell*_MAdj (where MAdj is a masculine adjective), since the POS language model will have seen the sequence FN-FAdj more times than the sequence FN-MAdj (see an example of correction in Table 7). Nevertheless, the POS language model will be useful only if the bilingual unit exists. Thus, the translation of *senyal*_MN *blanc*_MAdj will remain as *señal*_FN *blanco*_MAdj instead of *señal*_FN *blanca*_FAdj, since the tuple *blanc#blanca* is not contained in the translation model.

Remember that the POS language model requires the introduction of the word bonus feature to compensate for short outputs. To optimize the POS language model and the word bonus with the minimum error rate procedure (Och 2003), a specific development corpus was created. This development corpus, which consists of 476 sentences and one reference, was manually created to ensure that the corpus was grammatically correct and that it contained all the problems we had encountered.

Given that we have tuned these features on a special corpus, we consider them as part of the improved system. Therefore, notice that the baseline system as in previous Catalan-Spanish works (de Gispert, Mariño, 2006) does only include the translation model.

gender	(S) Me encantan **las espinacas**. *(I love spinachs).* (T1) M'encante ***les espinacs**. (T2) M'encanten **els espinacs**.

Table 7. Examples of gender concordance.

Numbers and time categorization

Numbers

As shown in the previous section, some numbers that were not in the training corpus might appear in the source sentence to be translated, so such numbers are marked as unknown and thus not translated. To avoid this problem, a set of rules was implemented to detect the numbers in the source language–except for *un/una* (one), *dos/dues* (two), *nou* (nine) and *deu* (ten), which can also correspond to other terms than numbers–and then to codify and generate them in the target language.

In order to detect the numbers, it is necessary to consider their formal structure--i.e. whether the number is a compound noun, using or not using dashes, etc. It is also necessary to take into account that numeral adjectives can have an associated gender. A specific codification is thus defined so that, in the generation process, the final number is coherent with the detected number. For instance, the numeral adjective *quaranta-una* would have the corresponding codification:

Z|41|F|2|Z

The first and last fields (Z) indicate a numerical expression. The second, third and fourth fields indicate: the numerical expression in figures, a feminine adjective, and the number of words replaced in the codification. The feminine mark is used to establish gender coherence with the following term.

Time

Catalan time expression is formally different from that in Spanish, as discussed above. Since the corpus contains few examples related to time expressions, time translation errors were frequent. In order to solve this problem, the same procedure performed with the numbers is performed here: the time expressions are detected, codified, and generated again in the target language.

Time expressions are identified using three different structures, following the most formal notations presented by J.M. Mestres (2004)[8]: the precise international notation, the formal general notation, and the mixed formal notation. In the Catalan to Spanish case, only the third one was activated, since the other ones can be literally translated between these pair of languages and do not need any special codification. However, the code was thought for all the notations in order to be able to be used in other language pairs.

The codification is performed by using a ten-field code, where the first and the last fields express the fact that we are dealing with a time expression (HORA). Thus, the time phrase *it's 4.45pm*, expressed in Catalan as *Són tres quarts de cinc de la tarda*, would be codified as follows:

HORA|4|Ø|45|Ø|C|T|8|S|HORA

where the inside fields express the following:

4	numerical hour
Ø	the term *hores* (hours) is not included
45	numerical minutes
Ø	the term *minuts* (minutes) is not included
C	type of time notation (mixed formal)
T	timeslot: *tarda* (afternoon)
8	number of words replaced in the codification
S	the verb *ser* (to be) is included

Table 8 shows an example in which, given a source sentence (S) with an unknown numerical expression leading to a non-correct translation (T1), the translation is corrected (T2) by applying the corresponding categorization. Examples for time expressions are also included.

numbers	(S) L'alliberament de **quatre-cents quaranta-un** presoners. *(The liberation of four hundred and forty-four prisoners).* (T1) La liberación de ***quatre-cents quarant-un** prisioneros. (T2) La liberación de **cuatrocientos cuarenta y un** prisioneros.
time	(S) Són **tres quarts de vuit**. (*It's a quarter to eight*). (T1) Son ***tres cuartos de ocho**. (T2) Son **las ocho menos cuarto**.

Table 8. Examples of correction after the number and time categorization.

[8] http://www.comunicaciodigital.com/rellotge_catala/pdf/notacions_horaries.pdf

Table 9 summarizes the errors encountered in this study and classified according to the type of solution proposed for overcoming them.

Evaluation

The SMT scientific community tends to agree on using the BLEU automatic measure to evaluate improvement in a translation system. Additionally, human evaluation is used when enough resources are available.

In this study, we introduced linguistic knowledge into an SMT system, mainly by using linguistic rules in order to improve a baseline system. Both automatic and human evaluations were used in the final system in order to make the final evaluation complete and objective. The baseline system–an N-gram-based SMT system without linguistic information and only including the translation model as in previous Catalan-Spanish works (de Gispert, Mariño, 2006)–was thus compared with the same system after adding the improvements described in this paper.

Error	***Text edition***	***Grammatical information***		***Categorization***
		pre-/postprocess.	***translation model***	
geminated 'l'	x			
apostrophe		x		
y/o conjunctions	x			
gender concordance			x	
capital letters		x		
numbers				x
time				x
perquè		x		
solo			x	
soler			x	
clitics		x	x	
cuyo		x		
possessives			x	
obligation	x			
elisión *de*	x			

Table 9. Errors treated by types of solution.

Test corpora

The Spanish source test corpus consists of 711 sentences extracted from *El País* and *La Vanguardia*, while the Catalan source test corpus consists of 813 sentences extracted from the *Avui* newspaper and from transcriptions from the TV program *Àgora*. Each direction set contains two manual references. Table 10 shows the statistics on the number of sentences, words and vocabulary for each language.

	Sentences	***Words***	***Vocabulary***
Spanish	711	15974	5702
Catalan	813	17099	5540

Table 10. Test corpora statistics.

Human Evaluation

To better analyze the output translations of both systems, a human evaluation is provided. First, the test sentences are classified into different categories according to the errors that were outlined in the error analysis section. This classification allowed us directly to compare the success of the specific techniques used to solve the corresponding problem. Second, a generic comparison was performed in order to evaluation how much better the improved system was, in general terms, with respect to the baseline system.

Specific evaluation

Although the proposed solutions were applied to improve detected errors, a specific technique may not be completely successful in solving a problem. Thus, given a specific problem and a specific technique to solve it, the aim here is to know how many cases were

(a) *already correct*: baseline outputs that were not modified by the improved system,

(b) *not corrected*: baseline outputs that were incorrect and that the improved system could not solve,

(c) *improved*: baseline outputs that were incorrect and that the improved system did solve, and

(d) *damaged*: baseline outputs that were correct and that the improved system translated incorrectly.

This classification can be seen in Table 11 for both directions of translation, together with the overall percentage of *already correct*, *not corrected*, *improved* and *damaged* cases. It can clearly be seen that, although most of the cases did not present a problem, a considerable number of cases (18.9% in *ca2es* and 13.9% in *es2ca*) were initially problematic and solved using the presented techniques. Thus, the analysis shows that the proposed techniques are useful for handling the detected problems. Regarding clitics, some cases were still not solved, mainly because the decoder chose the wrong pronoun or a different place: *l'últim any* ****ho*** *va passar* instead of *l'últim any va passar-**lo***. In this case, for instance, the place is interchangeable, but the wrong pronoun selection results from a case of polysemy, since both pronouns are related to the same source (Spanish) pronoun. The clitics solution takes for granted that the translation is correct and then processes a grammatically correct sequence; if the translation is not correct, the solution is not able to solve the problem.

Since the human evaluation was performed from the journalism corpus, not all the error types were present in the evaluated sentences. This is the reason why not all the analyzed error types are contained in Table 11.

	ca2es					***es2ca***				
	(a) correct	***(b) not corrected***	***(c) improved***	***(d) damaged***	**total**	***(a) correct***	***(b) not corrected***	***(c) improved***	***(d) damaged***	**total**
clitics	48	1	**25**	3	77	70	3	**27**	5	105
y/o	5	0	**3**	0	8	-	-	-	-	-
numbers	53	0	**5**	2	60	49	0	**2**	0	51
possessives	55	0	**10**	0	65	-	-	-	-	-
soler	1	0	**1**	0	2	-	-	-	-	-
l·l	34	0	**3**	0	37	-	-	-	-	-
apostrophe	-	-	-	-	-	315	0	**39**	3	357
solo	-	-	-	-	-	17	0	**4**	0	21
obligation	-	-	-	-	-	5	0	**3**	0	8
%	78.7	0.4	**18.9**	2.0	100	84.1	0.5	**13.9**	1.5	100

Table 11. Classification of error cases.

The damaged cases appear for several reasons. Clitics such as *estar-**se***, for instance, were damaged because in some contexts, *estar* is not recognized as a verb by the tagger. Therefore, the proposed solution damages the baseline system, which was able to learn the correct translation from the training corpus.
In two other situations, translations in the number category were not able to maintain gender concordance, so when a sequence had been seen in the training corpus, the translation of numbers was better performed by the baseline than by the improved system. Finally, the solution applied to handle the apostrophe problem did not cover all the exceptions, such as the apostrophe of some Catalan acronyms that do not follow the basic apostrophe rules. This led to damaged cases, for which an additional dictionary would be required to build a proper apostrophe rule.
Figure 1 shows the percentage of correct cases over the total number of problematic cases treated above. The percentage of correct cases increases clearly in both directions of translation: in Catalan to Spanish (es2es) correct cases go from 78.7% to 95.6%, and in the inverse direction (Spanish to Catalan) the percentage increases from 84.2% to 94.6%.

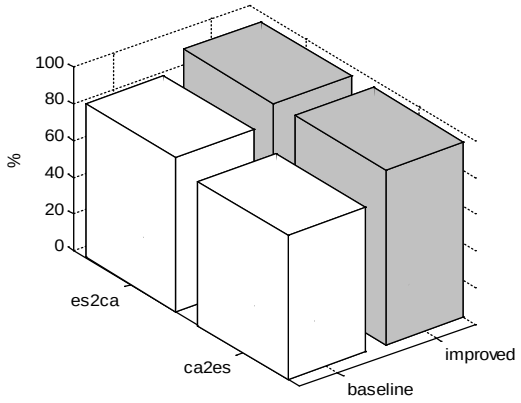


Figure 1. Percentage of *correct cases* for both baseline and improved systems in each direction of translation.

General evaluation

Second, a generic comparison of both systems was performed using human criteria. Ten native evaluators made a blind comparison of 200 different sentences randomly selected from each system. Five evaluators are native in Catalan and Spanish and they evaluated Spanish-to-Catalan. Five different evaluators are

native in Spanish and they evaluated Catalan-to-Spanish. Score criterion was simple: given the source sentence, they were asked to choose the best translation from the two outputs. Tails were accepted. The results obtained in this evaluation are shown in Table 12, ordered by evaluator and system, and the overall resulting percentages of the results are illustrated in Figure 2.

Catalan to Spanish				***Spanish to Catalan***			
evaluator	***baseline***	***improved***	***equal***	***evaluator***	***baseline***	***improved***	***equal***
1	10	21	169	6	22	51	127
2	44	52	104	7	34	52	114
3	32	35	133	8	31	51	118
4	35	36	129	9	12	69	119
5	4	16	180	10	6	31	163
%	**12.5**	**16.0**	**71.5**	**%**	**10.5**	**25.4**	**64.1**

Table 12. Human evaluation results.

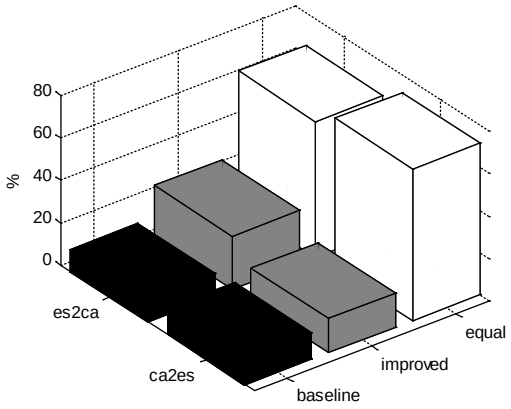


Figure 2. Percentage of better and equal cases in the human evaluation.

Automatic evaluation

The test corpus and two available references were used to compute the BLEU score for the translations provided by the baseline and the improved systems. The results are presented in Table 13. It can clearly be seen that the improved system performs significantly better in terms of BLEU. In the Catalan-to-Spanish direction the gain is 0.47 points BLEU, whereas in the Spanish-to-Catalan translation the improvement rises to 1.09 points BLEU. The margin of error was calculated with a 90% Wald confidence interval (Newcombe 1998).

	Catalan to Spanish (ca2es)	***Spanish to Catalan (es2ca)***
Baseline	86.57 ± 0.44	84.91 ± 0.45
Improved	87.04 ± 0.44	86.00 ± 0.44

Table 13. BLEU results obtained in the baseline and improved systems.

The results show a higher improvement in the Spanish-to-Catalan direction than in the reverse direction. The first reason for such a difference may be that in the Catalan-to-Spanish translation, we are starting from a better baseline system, which is more difficult to outperform. The second reason may be that, as was seen before, the test does not cover many of the problems that were addressed in the present work. To evaluate the improvement of the system when using the specific problems addressed, an *ad hoc* test was proposed, consisting of 636 sentences for both directions of translation. Each direction set had one manual reference. The number of sentences, words and vocabulary for each language are found in Table 14. This test was built to evaluate exclusively the proposed methods. Therefore, a Catalan-Spanish linguist wrote this test set from scratch focusing on the points that we were trying to solve.

	Sentences	***Words***	***Vocabulary***
Spanish	636	6711	2505
Catalan	636	6734	2516

Table 14. *Ad hoc* test corpora statistics.

Subsequently, a second automatic evaluation was performed with this test, which tried to represent the challenges the present study was trying to overcome, so that it helped to amplify and to better analyze the influence of the proposed solutions. Moreover, unlike the previous corpus, all types of errors were contained in the tested sentences. Table 15 shows the results of this evaluation, where the improvement is about 5.5 points BLEU in both directions of translation.

	Catalan to Spanish (ca2es)	***Spanish to Catalan (es2ca)***
Baseline	73.50 ± 0.89	75.91 ± 0.86
Improved	79.02 ± 0.82	81.35 ± 0.78

Table 15. BLEU results obtained in the baseline and improved systems.

Conclusions and future work

In this paper, a set of strategies was applied with the aim of overcoming the problems found in a preliminary statistical machine translation between the Catalan-Spanish language pair. The translation errors addressed were related to different linguistic dimensions: orthographic, morphological, lexical, semantic and syntactic. The proposed solutions made use of several strategies, such as text processing, grammatical category-based rules, statistical modeling, and word and phrase categorization. These techniques resulted in an improvement of the baseline translation system, especially in the Spanish-to-Catalan direction, where a gain of up to 1.1 points BLEU was obtained. Human evaluation also showed the employed strategies to have a positive effect in both directions of translation, when considering both the specific problems addressed in this work and the overall performance of the system.

Nevertheless, the list of problems to overcome is still not finished, and they have been left for future work. These are, for instance, the verbal forms, at the morphological level, and the use of specific prepositions, at the syntactic level. The problem associated with verbal forms resides in the fact that only the lemma or other conjugated forms of the verb in question are included in the training corpus, so that the conjugated word is returned as unknown. In the following example, the Spanish first person simple future of the verb *to pass* is translated, while the second person is output as an unknown word:

***ES**: ¿Crees que **aprobarás** la asignatura? Por supuesto que **aprobaré**.*
***ES**: ¿Creus que ***aprobarás** l'assignatura? Per descomptat que **aprovaré**.*
EN: Do you think **you will pass** the exam? Of course **I will pass** it.

Other common errors are related to the different use of prepositions in both languages. In Catalan, for instance, the preposition *a* is normally not used in front of a direct object, so this preposition will not appear in the target (Spanish) language when it is, in fact, required:

***CA**: Va dir que el taxista s'havia alegrat molt quan va agafar **el** viatger.*
***ES**: Dijo que el taxista se había alegrado mucho cuando cogió ***el** viajero.*
EN: He said that the taxi driver was very happy to take **the** passenger.

On the other hand, the *a* and *en* prepositions are commonly used in Catalan and Spanish languages to introduce both special and temporal elements. However, they are used in very distinct ways in these two languages, so translation errors are frequently observed in such cases due to these usage differences.

***ES**: El sábado, **en** las ocho horas de jornada laboral, solo 30 personas se pusieron en contacto con ella.*
***CA**: Dissabte, ***a** les vuit hores de jornada laboral, només 30 persones es van posar en contacte amb ella.*
EN: On Saturday, **during** eight hours of work, only 30 persons contacted her.

Future work should include the detection of additional errors. Polysemy and homonymy, for instance, have been detected in the current work by means of human knowledge. A new challenge would be to generalize homonymous and polysemic errors and to detect them using machine learning techniques. Finally, apart from solving remaining errors, future work should also work to avoid damaged cases: those cases already correct and included in the training corpus should remain unchanged in the translation process.

However, the current work has demonstrated that, despite the list of errors that are still to be solved, the presented N-II system provides high-quality translations between Catalan and Spanish languages and is freely available online as a linguistic resource.

Acknowledgements

The authors would like to thank TALP Research Center and Barcelona Media Innovation Center for its support and permission to publish this research. We would like to give credit to the anonymous reviewers of this paper for their valuable suggestions.

This work has been partially funded by the Spanish Department of Science and Innovation through the Juan de la Cierva fellowship program and the Spanish Government under the BUCEADOR project (TEC2009-14094-C04-01).

References

Bangalore, S. and G. Riccardi (2000). Stochastic finite-state models for spoken language machine translation. Workshop on Embedded Machine Translation Systems, Seattle, WA.

Brown, P., S. della Pietra, et al. (1993). "The mathematics of statistical machine translation." Computational Linguistics **19**(2): 263-311.

Carreras, X., I. Chao, et al. (2004). FreeLing: An Open-Source Suite of Language Analyzers. Conference on Language Resources and Evaluation, Lisbon, Portugal.
Casacuberta, F. and E. Vidal (2004). "Machine translation with inferred stochastic finite-state transducers." Computational Linguistics **30**(2): 205-225.
Crego, J. M., A. de Gispert, et al. (2006). N-gram-based SMT System Enhanced with Reordering. Human Language Technology Conference (HLT-NAACL'06): Proceedings of the Workshop on Statistical Machine Translation, New York.
Crego, J. M. and J. B. Mariño (2007). "Improving SMT by coupling reordering and decoding." Machine Translation **20**(3): 199-215.
de Gispert, A. and Mariño, J. (2006). Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. In Proc. of LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages (SALTMIL'06). Genova, 65-68.
Mariño, J. B., R. E. Banchs, et al. (2006). "N-gram based Machine Translation." Computational Linguistics **32**(4): 527-549.
Newcombe, R. G. (1998). "Two-sided confidence intervals for the single proportion: Comparison of seven methods." Statistics in Medicine **17**(8): 857-872.
Niessen, S. and H. Ney (2000). Improving SMT quality with morpho-syntactic analysis. International Conference on Computational Linguistics, Saarbrücken, Germany.
Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. 41st Meeting of the Association for Computational Linguistics, Sapporo, Japan.
Popović, M., A. de Gispert, et al. (2006). Morpho-syntactic information for automatic error analysis of statistical machine translation output. HLT/NAACL Workshop on Statistical Machine Translation, New York.
Popović, M. and H. Ney (2004). Towards the use of word stems and suffixes for statistical machine translation. International Conference on Language Resources and Evaluation, Lisbon, Portugal.
Popović, M. and H. Ney (2006). POS-based Word Reorderings for Statistical Machine Translation. International Conference on Language Resources and Evaluation, Genoa, Italy.