Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 187-195 recibido 1-05-2009; aceptado 5-06-2009

Development of a free Basque to Spanish machine translation system[*]

Desarrollo de un sistema libre de traducción automática del euskera al castellano

Mireia Ginestí-Rosell,
Gema Ramírez-Sánchez,
Sergio Ortiz-Rojas,
Prompsit Language Engineering
E-03195 l'Altet (Spain)
{mginesti,gema,sergio}@prompsit.com

Francis M. Tyers
Mikel L. Forcada
Dept. Lleng. i Sist. Informàtics
Universitat d'Alacant
E-03071 Alacant (Spain)
{ftyers,mlf}@dlsi.ua.es

Resumen: Este artículo presenta un sistema de traducción automática libre (de código abierto) basado en reglas entre euskera y castellano, construido sobre la plataforma de traducción automática Apertium y pensado para la asimilación, es decir, como ayuda a la comprensión de textos escritos en euskera. Se describe el desarrollo y la situación actual y se muestra una evaluación de la calidad de las traducciones.
Palabras clave: traducción automática, euskera, castellano, software libre o de código fuente abierto

Abstract: This paper presents a free (or open-source) rule-based machine translation system between Basque and Spanish, based on the Apertium machine translation platform aimed at assimilation, that is, as a help for the understanding of texts written in Basque. The development process and current status are described and an evaluation is given of the utility of the output.
Keywords: machine translation, Basque, Spanish, free/open-source software

1 Introduction

This paper describes the development of a "gisting" machine translation system between Basque, or *Euskara* and Spanish.[1] The first section will give a general overview of the two languages in question and describe our aims for the current system. The subsequent sections will describe the existing resources which were re-used, some of the development work, the current status, an evaluation, and some prospects for future development.

Basque is a *language isolate*[2] spoken in areas of the the Basque Autonomous Community (*Euskal Autonomia Erkidegoa*) and the Foral Community of Navarre (*Nafarroako Foru Komunitatea*) in Spain and in the southwestern area of Department of the Atlantic Pyrenees (*Pirinio Atlantikoen Departamentua*) in France. These areas have been traditionally referred to as *Euskal Herria* ("the people of the Basque language") by Basque speakers. Approximately 30% of people in Euskal Herria speak Basque, and around 20% have Basque as their first language, making it a minority language. South of the Pyrenees, where most of the Basque speakers live, the majority language is Spanish.

Basque is very different from the Romance languages surrounding it (Spanish, French, Occitan). Their lexica do not have much in common, except for many modern loanwords that may be recognized rather easily, and older ones, which are much more difficult to recognize. This is one of the main obstacles to mutual understanding, but not the only one. Syntax is also very different: Basque is left-branching and as such uses postpositions where right-branching Romance languages use prepositions, relative clauses come before the noun they modify; also, it has a subject–object–verb word order where Romance languages rather use subject–verb–object, and has a special case (ergative) for

[*] Development was supported and funded by Prompsit Language Engineering S.L. and the Universitat d'Alacant.

[1] The system can be tried out online at http://www.erdaratu.eu/

[2] "**Language isolate**: A language that cannot to our knowledge be assigned to any larger family. Basque is a classic example." (Matthews, 1997).

ISSN: 1135-5948

the subject in transitive sentences.

Minority language speakers typically differ from the majority in being bilingual, speaking both their language, and the language of the majority. In contrast, a majority language speaker does not usually speak the language of the minority. This has some implications for the requirements society will put on machine translation systems.

Applications of machine translation system can be divided in two main groups: *assimilation*, that is, to enable a user to understand what the text is about; and *dissemination*, that is, to help in the task of translating a text to be published. The requirements of either group of applications is different.

Assimilation may be possible even when the text is far from being grammatically correct; however, for dissemination, the effort needed to correct (*post-edit*) the text must not be higher than the effort needed to translate it manually from scratch.

A majority to minority language system will mainly be used for dissemination purposes; it must therefore be such that post-editing the output is faster than translating from scratch. Intelligibility is secondary, and only important if it helps the post-editor. A minority to majority language will however be mainly used for assimilation, for instance, to answer vital questions such as "what are they writing about me in the minority language newspaper?". Therefore, the main goal is intelligibility.

The system in this paper was indeed developed with this second objective in mind, to be able to provide intelligible translations into Spanish of text published in the Basque language media. As a result of this, certain design decisions were made in the course of development, for example, it was decided that word order within small constituents (or *chunks*) was much more important for intelligibility than the ordering of these constituents, so for the moment we have only written rules performing frequent reorderings in the range of roughly 2–10 words. In the same sense, we gave priority to those aspects which most affected the intelligibility of translation, leaving unsolved for the moment not so crucial aspects like the overgeneration of definite articles in some sentences.

For instance, the Basque sentence in Figure 1 is currently translated by Apertium as shown in 2. As may be seen, short-range reorderings are performed, but not longer-range (sentence-level) reorderings.

When there are no unknown words, the reordering of chunks may include sequences of 7, 8 or more words, although the most frequent combinations have a scope of 4 to 6. See figure 4 for more detailed examples of chunk reorderings.

2 Development

2.1 Existing data

We were fortunate in the development of the system to be able to take advantage of some existing data for Spanish and Basque from the Matxin (Alegria et al., 2005) system for Spanish to Basque translation. Although only a fraction of the linguistic data is available under a free licence, we were able to build upon this to create a Basque morphological analyser and bilingual dictionary with acceptable coverage for our purposes. In total, approx. 5,800 entries from the Basque morphological dictionary in Matxin were re-used, although changes were made to the tagset (see section 2.4).[3] We also re-used the bilingual dictionary to obtain the Spanish translations of these entries, although in some cases it was necessary to choose a single translation from the multiple equivalences that the dictionary contained. The Spanish monolingual dictionary was taken from the Apertium Spanish–Catalan translator. No data for transfer rules was re-used.

2.2 The Apertium platform

The system is based on the Apertium machine translation platform.[4] The platform was originally aimed at the Romance languages of the Iberian peninsula, but has also been adapted for other language pairs, such as languages from the Celtic group, e.g. Welsh (Tyers and Donnelly, 2009), with much of the work on new languages being pursued by volunteers, using the increasingly common collaborative development model used for free[5] and open-source software. Apertium is licensed under the Free Software Founda-

[3] Note that we had to use data for the *generation* of Basque to build our morphological analyser. There is a high-coverage analyser for Basque (Alegria et al., 2004), but it is not free/open-source and therefore cannot be re-used.

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

[5] We follow the definition of "free" used by the Free Software Foundation (`http://www.fsf.org`)

```
Ertzainek  biktimen       etxean   atxilotu zuten gizonezkoa
Police-the victims-the-of house-in arrested had   man-the
```
“The police arrested the man at the victims’ house”

Figure 1: Word order in a Basque sentence

```
Los policías en la  casa  de las víctimas detuvieron el  hombre
The police   in the house of the victims  arrested   the man
```
“The police arrested the man at the victims’ house”

Standard word order: *“Los policías detuvieron al hombre en la casa de las víctimas”*

Figure 2: Word order in the Spanish output of Apertium for the sentence in figure 1

tion’s General Public Licence[6] (GPL) and all the software and data for the 17 supported language pairs (and the other pairs being worked on) is available for download from the project website.

Apertium uses a shallow-transfer engine. Finite-state transducers processing up to 40,000 words per second (Ortiz-Rojas, Forcada, and Ramírez-Sánchez, 2005) are used for lexical processing, hidden Markov models are used for part-of-speech tagging, and multi-stage finite-state based chunking for structural transfer. XML-based standard formats are used to encode the linguistic data, which are then compiled into the high-speed formats used by the engine. Further details are given in Armentano-Oller et al. (2006), and on the project website.

2.3 The pipeline

A typical translator built with Apertium consists of 8 modules which communicate between each other using standard Unix pipes.[7] The modules comprise of the following:

- A **deformatter** which encapsulates any formatting (e.g. HTML or XML tags etc.) information in the input stream.
- A **morphological analyser** which for each surface form in the stream returns a sequence of possible analyses.
- A **part-of-speech tagger** which out of the possible analyses for a given word returns the most probable analysis.
- A **lexical transfer** module which for each unambiguous source language lexical form returns a target language lexical form. Currently, no lexical selection is attempted: a single target equivalent is provided for each source lexical form. Multi-word units are added to dictionaries to partly compensate for this limitation.
- A **structural transfer** module which performs syntactic and morphological operations to convert the source language intermediate representation into the target language intermediate representation. Common operations include insertion, deletion and substitution of lexical units, agreement between lexical units for e.g. gender, number and case, etc. The structural transfer module calls the lexical transfer module.
- A **morphological generator** which for each target language lexical form returns a surface (inflected) form.
- A **postgenerator** which performs orthographic operations, for example elision (such as *de+el=del* in Spanish).
- A **reformatter** which de-encapsulates any formatting, leaving it untouched.

In translators built with versions 3.0 and higher of the platform (English–Catalan, Welsh–English, etc.), the structural transfer process is split into three parts. These are:

- The first stage (**chunker**) performs lexical transfer and local syntactic operations and segments the sequence of lexical units into *chunks*. A chunk is defined as a fixed-length sequence of part-of-speech tags that corresponds to some syntactic feature, for example a chunk

[6] `http://www.fsf.org/licensing/licenses/gpl.html`

[7] The modules use text-based formats to communicate, which eases diagnosis, the insertion of new modules, etc.

might encompass all or part of a noun phrase.

- The second stage (**interchunk**) performs more global operations on and between chunks.
- The third stage (**postchunk**) performs another round of local operations on each chunk and outputs the stream in the format accepted by the morphological generator.

The three-stage organization does not give the transfer any more computational power with respect to a single-stage transfer (chunker only), since it still works with finite-state patterns, but it does allow longer patterns to be treated in a manner which is more comfortable for the linguist or programmer writing the rules. In case that more than one rule may be applied at a given position of the text, both the *chunker* and the *interchunk* modules select the longest matching rule.[8]

2.4 Morphological analysis

The data used to create the monolingual Basque dictionary were obtained from the free data of the Matxin system, although some substantial changes were made in the way this system morphologically analyses Basque.

Basque is an agglutinative language, so that postpositions, articles and other affixes are attached to a main word. Matxin analyses the individual words as single lexical forms, treating case markers and definiteness as *declension* of the lexical form. Apertium on the other hand, considers these units as individual lexical forms (articles, postpositions), so that, when analysing a word, it decomposes it into its constituents and gives a different lemma to each one of the morphemes. So, for example, the word *etxean* ('in the house') is analysed by Matxin as "common noun, singular, case inessive", whereas the same word is analysed by Apertium as:

```
etxe+n a+det.art.sg an+post
```

where the output is three lexical forms: a noun, a singular article, and a postposition. These three lexical forms will be treated as three independent words by the subsequent modules (tagger, structural transfer module, etc.)

The absolutive case is not marked in the Apertium system. As with the other cases, the ergative case is also treated as a postposition and given a lemma (*k*):

```
etxe : etxe+n
etxeek : etxe+n a+det.art.pl k+post
```

Basque has another group of words which function like postpositions (and which are also translated into Spanish as prepositions) but are not attached to a word, that is, are written as independent lexical forms. They can come after a noun phrase or after certain postpositional phrases (noun phrase with a postposition). They are labelled in the system as "separate postpositions", with the tag *spost*, compared with the other attached postpositions, which have the label *post*.

Some examples of these are *arabera*, *gabe*, *buruz*, *ustez*, *kontra* (viz. 'according to', 'without', 'about', 'in the opinion of', 'against'). Their paradigms include the genitive postposition *-ko* which can be attached to them (so that other lexical forms which may come after *ko* may also be attached). See figure 3 for some examples.

As for inflection, verbs are analysed with tags for tense, mood and tags for the values of *nor?* (`NR`), *nori?* (`NI`) and *nork?* (`NK`) (direct object/intransitive subject, indirect object and transitive subject respectively)[9]. For example for the inflected verbs *dituzte* 'they have (them)' and *nien* 'I had (it to them)',

```
dituzte : ukan+vbsint.pri.NR_HK.NK_HK
```

Here the lemma is *ukan* 'to have'. It is a synthetic verb, meaning it inflects for conjugation. It is conjugated in the present indicative with a third person plural (`HK`) direct object (*nor?*) and a third person plural subject (*nork?*).

```
nien : ukan+vbsint.pii.NR_HU.NI_HK.NK_NI
```

The lemma is as above, but the tense is changed to imperfect and the direct object is third person singular, the indirect object

[8] In a *left-to-right*, *longest-match* (LRLM) strategy similar to the one used by lexical scanners like `lex` (Lesk, 1975).

[9] The Basque words *nor?*, *nori?*, and *nork?* are respectively the nominative, dative and ergative forms of the interrogative pronoun *nor?* 'who?' and are used in Basque grammars to describe sentence and verb structure.

```
(1) iturri    ofizialen                                 arabera
    iturri+n  ofizial+adj.izo a+det.art.pl en+post      arabera+spost
    source    official-the-of                           according to
    “according to the official sources”

(2) historiari                         buruzko             liburua
    historia+n a+det.art.sg i+post     buruz+spost ko+post liburu+n a+det.art.sg
    history-the-to                     about-of            book-the
    “the book about history”
```

Figure 3: Analyses of two noun phrases with attached and separate postpositions (possible ambiguous analysis have been discarded for clarity)

(*nori?*) is third person plural and the subject is first person singular (NI).

Derivational affixes (that is, affixes that are attached to a word to form a new lexical form with different part of speech and different meaning) like *-pe* ‘under’, *-garri* (-able), *-dun* ‘having’, have been left out of paradigms, although they are part of them in the Matxin system. We found that including them caused overgeneration of lexical forms and increased dramatically the size of the dictionaries. The generation of the corresponding translation was also not straightforward, for example *txapeldun* is ‘champion’, but this can be decomposed as *txapel+dun* ‘hat (beret) having’, which results in an inadequate translation. Therefore, words that are the result of a derivational process must be entered in the dictionary separately as a word with a specific lemma.

The morphological analyser delivers all possible analysis of a word. Ambiguities are dealt with by the POS tagger, which uses a statistical bigram model to choose the most probable analysis. Some ambiguities in the Basque language can not be adequately resolved based on bigrams (e.g. the analysis of the morpheme *-ak* as article plural or as article singular + ergative case). We are planning to add a constraint grammar to a future version of the translator to improve this performance.

2.5 Overview of transfer

In the first module of transfer (**chunker**), the *chunks* are created, which broadly correspond to phrases, and the lexical transfer module is called to translate each word into the target language. A listing of the most important chunks can be found in table 1.

Type	Description
SN	Noun phrase
SV	Verb phrase
SPR	Prepositional phrase
SPGEN	Genitive prepositional phrase
SVsub	Subordinated verb phrase
Orel	Relative clause
SA	Adjective phrase
SADV	Adverbial phrase
PREP	Preposition

Table 1: The principle chunk types

The chunks are created using rules which detect fixed patterns of parts of speech. For example, the sequence *noun-adjective-determiner* forms a noun phrase (SN) chunk, and the sequence *noun-adjective-determiner-postposition* forms a prepositional phrase (SPR) chunk. Grammatical operations are performed inside the chunks, such as word reordering according to target language order, gender and number agreement, verb inflection. For example, the Basque phrase *datuen arabera* (“data-the-of according to”, ‘according to the data’) is segmented and managed by two rules, one for *noun-determiner-postposition* and one for *separate postposition*; the output after the first module of transfer would be:

```
[SPGEN de+pr el+det.art.m.pl dato+n.m.pl]
[PREP según+pr]
```

As can be seen, two chunks are created (SPGEN and PREP), and gender and number agreement is performed between determiner (*el*, ‘the’) and noun (*dato*, ‘datum’) of the first chunk, masculine plural for both (m.pl). Words are translated into Spanish.

In this module, chunks can also be output in a different order from the input words, and a chunk can be created which combines two non-consecutive words. The linguist has but to write a single rule which matches the pattern of words that needs to be addressed. This is needed for example for verbs, which often present a negative adverb or other words between the main verb and the auxiliary. So, for example, there is a rule which detects the pattern *gerund + negative adverb + auxiliary verb + causal conjunction* and outputs three chunks: `CONJ` (causal conjunction), `SADV` (negative adverb) and `SV` (the inflected verb). According to this, the input phrase *ezagutzen ez direlako* "because they don't know (lit. "knowing no [they] are because") is output as follows:

```
[CONJ porque+conj]
[SADV no+adv]
[SV conocer+vblex.pri.p3.pl]
```

("because no [they] know", 'because they don't know'), where the (source language) main verb and auxiliary verb are united into a single chunk with an inflected verb.

In the second module of the transfer (**interchunk**), rules are defined to perform operations between chunks, such as long-distance agreements, reorderings, change of tags, etc. Here the single words cannot be accessed nor changed, only chunk names and information are available. We'll take the above example *datuen arabera*, which was output by the first transfer module as `SPGEN + PREP`. In the second module, there is a rule which detects this sequence and outputs it in the inverse order:

```
[PREP según+pr]
[SPR-SN de+pr el+det.art.m.pl dato+n.m.pl]
```

Also, as it is a prepositional phrase followed by a preposition, the rule changes the name of the `SPGEN` chunk to `SPR-SN`, which means that the prepositional phrase should be changed to a noun phrase, that is, without the preposition (to avoid the translation "according to of the data"). This work will be done in the third module of transfer, where a rule detects all the chunks named `SPR-SN` and deletes the preposition (this operation is postponed as deletion of a word inside a chunk can not be performed in the second module). More examples of chunk reorderings can be found in figure 4.

The third transfer module (**postchunk**) does the final and local operations needed for certain chunks, and outputs the words without the chunk information. The previous example would be output:

```
según+pr el+det.def.m.pl dato+n.m.pl
```

without the preposition *de*. Other operations include management of articles, verb tense modifications and some lexical changes.

3 Current status

Version 0.3.1 of the system, as released on the 24th April 2009 contains approx. 6,300 monolingual dictionary entries in the Basque morphological dictionary along with 294 inflectional paradigms generating a total of 11,819,561 mappings from surface forms to lexical forms. There are 175 chunking rules, 54 rules for interchunk movement and agreement and 20 *postchunk* rules.

3.1 Coverage

Table 2 presents the figures for *naïve coverage* of the morphological analyser of the system over two available corpora, the Basque Wikipedia[10] and the online version of *Berria*,[11] a Basque daily newspaper. *Naïve coverage* is calculated as follows: if for a given token in the text, at least one possible analysis is returned, it is taken to be *covered* (even if other analyses are missing).

The ambiguity rate was also calculated based on the Wikipedia corpus. Excluding unknown words and numerals there were on average 1.37 analyses per surface form, ranging from 12 *frantziarrenak* to 1 *eta*.

Inspecting the top 1,000 unknown words in the Wikipedia corpus, many of them appeared to be non-Basque words, *the*, *a*, etc. and proper names *Karlos*, *John*, etc.

4 Preliminary evaluation

We have performed a preliminary evaluation of the current version of the Basque–Spanish machine translation system in an assimilation setting. In order to perform the evaluation we prepared a two-step procedure inspired in one used in the 2009 WMT workshop (Callison-Burch et al., 2009).

In the first step, an evaluation corpus made of 50 sentences of length less than 25 words drawn from recent editions of *Berria*

[10] `http://eu.wikipedia.org`; Database dump from the 5th April 2009.

[11] `http://www.berria.info`

```
(1) Gipuzkoa, Araba eta Bizkaiko           beste sei ikastetxerekiko          lana
    Guipuzkoa, Araba and Biscay-of         other six school-with-of           work
-   [SPGEN_1 Gipuzkoa, Araba eta Bizkaiko] [SPGEN_2 beste sei ikastetxerekiko] [SN lana]
-   [SN El trabajo] [SPGEN_2 con los otros seis colegios]
       [SPGEN_1 de Guipúzcoa, Álava y Vizcaya]
```

Apertium output: “El trabajo con los otros seis colegios de Guipúzcoa, Álava y Vizcaya”
(“The work with the other six schools of Guipuzkoa, Araba and Biscay”)

```
(2) Gobernu    demokratikoaren         aurkako         kolpearen          ostean
    Government democratic-the-of       against-of      coup-the-of        after
-   [SPGEN_1 gobernu demokratikoaren]  [PREPGEN aurkako] [SPGEN_2 kolpearen] [SADV ostean]
-   [SADV Después] [SPGEN_2 del golpe] [PREPGEN contra] [SPGEN_1 el gobierno democrático]
```

Apertium output: “Después del golpe contra el gobierno democrático”
(“After the coup against the democratic government”)

```
(3) Iazko entzierroetan              baino        zazpi lagun  gutxiagok
    Yesterday-of runnings_of_bulls-in but         seven  people less-erg
-   [SPR Iazko entzierroetan]          [CONJ baino] [SN zazpi lagun gutxiagok]
-   [SN Siete personas menos] [CONJ que] [SPR en los encierros del año pasado]
```

Apertium output: “Siete personas menos que en los encierros del año pasado”
(“Seven people less than in last year’s runnings of bulls”)

```
(4) Estatu frantseseko        gobernu berria       osatzeko         eztabaida luzea
    State French-of           government new-the   build-to         discussion long-the
-   [SPGEN Estatu frantseseko] [SN_1 gobernu berria] [SVsub osatzeko] [SN_2 eztabaida luzea]
-   [SN_2 La discusión larga] [SVsub para componer] [SN_1 el gobierno nuevo]
       [SPGEN del estado francés]
```

Apertium output: “La discusión larga para componer el gobierno nuevo del estado francés”
(“The long discussion to form the new French government”)

Figure 4: Examples of reorderings of chunks performed in the second module of the transfer system

Corpus	Running tokens	Tokens found	Coverage (%)
Wikipedia	2,531,313	1,958,836	77.38
Berria	3,665,880	3,335,363	90.98

Table 2: Naïve coverage of the system

was translated into Spanish using the version of the Apertium Basque–Spanish translator from the 21st April 2009. The previous and subsequent sentences in the article were also translated as a minimal context to the problem sentence (see figure 5). The sentences were handed to a speaker of Spanish with no knowledge of Basque, who had to post-edit each sentence without looking at the original sentence until it was adequate Spanish.[12] To perform the task, they had to build a possible meaning for the sentence. A variety of situations may occur during this guessing process:

- When the system finds an unknown word, it leaves it untranslated, in Basque. Guessing what it means may be very difficult when the Spanish word is not a cognate, but the monolingual posteditor is asked to provide the best possible equivalent given the context, and without using a dictionary.
- The posteditor can even change a Basque word it it seems clear that the system has made an obvious mistake.

In the second step, the resulting sentences, together with the original Basque sentences and the context, were handed to a bilingual Basque–Spanish speaker, who had to judge

[12] The original instructions in Callison-Burch et al. (2009) said: *Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select “No corrections needed.” If you cannot understand the sentence well enough to correct it, select “Unable to correct.”* We have, however, asked the posteditor to work hard and produce a sentence anyway.

```
==  Por tanto, el de tres del Gobierno no tiene este año el presupuesto adelante
    dificultad para sacar ser.
53. Por medio de esa decisión en los últimos años la actitud que ha tenido
    completamente ha cambiado PSE.
==  De hecho, con PNV tenían el acuerdo de Gobierno allí abajo desde el que detuvo,
    la actitud rígida contra los proyectos del Gobierno han tenido.
```

Figure 5: Example evaluation text shown to Spanish speaking evaluator

whether the translation of the sentence produced by the posteditor was actually an adequate translation of the original Basque sentence (that is, whether the posteditor understood its meaning adequately).[13]

The bilingual speaker gave scores in the range 0%–100% (in steps of 10%) and wrote comments regarding the inadequate translations.

The histogram in figure 6 shows the distribution of scores. Instead of giving the number of sentences receiving each score, the number of sentences receiving a score equal or higher than a given score is plotted. Taking scores equal to or above 70% as *adequate*, 62%, that is, 31 out of 50 sentences, were judged to be *adequate*. The average score over all sentences was 69,4%.

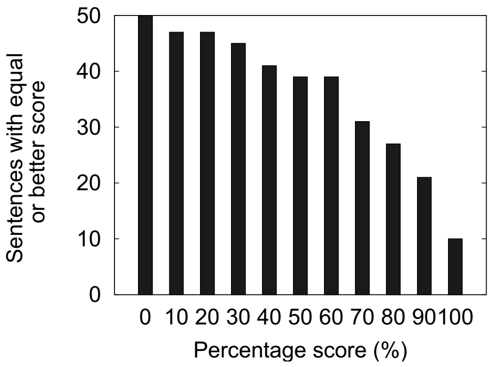

Figure 6: Number of sentences having a percentage score equal or higher than a certain score

Some of the comments made by the bilingual speaker follow:

- A couple of source sentences were strangely worded or had typographical errors.

- Many errors were due to the absence of frequent multiword units like *Audiencia Nacional* (*Auzitegi Nazionala*, *National Court*) or *quedar en suspenso* (*bertan behera gelditu* 'to be suspended')

- The posteditor had to guess untranslated Basque words, really worked hard, and got interesting results in some sentences using the context.

- They were sometimes misled by proper nouns with postpositions attached, and some of the translations were not correct due to this.

- Lexical selection misled the posteditor in some sentences (*elkarrizketa* →*diálogo* 'dialog' instead of →*entrevista* 'interview'

- In one sentence, the posteditor forgot to include part of a sentence. In another one, a word was left out. This kind of errors may be expected during postediting but have a negative impact on the final scores.

- The system once produced two translations for the same word and this misled the posteditor, who had no way to be aware of that.

- In two sentences, the bilingual speaker had trouble finding an alternate, fluent translation to improve a partially good one.

- Many errors made by the posteditor were due to bad handling of compound verbs by the system.

To get an estimate of the effort invested by post-editors, a parallel evaluation was performed. A bilingual speaker was asked to post-edit the output of the system on a separate set of 100 sentences, without context, but having access to the original sentence. Table 3 gives an idea of the postediting effort involved. The word error rate

[13] This is different from the procedure in Callison-Burch et al. (2009) which provided reference translations: *Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence.*

No. sentences	100
No. words (raw translation)	1312
No. words (post-edited translation)	1364
No. 1-word corrections	950
Word error rate (WER)	72.41%
Position-independent (PER)	39.86%

Table 3: Results of an additional evaluation run where a post-editor had access to the original sentence

and the position-independent word error rate were computed automatically from the raw and the postedited translation. The figures clearly indicate that the output of the system is far from being suitable for dissemination, although getting over 60% of correct word-for-word translations is a promising start.

A more complete evaluation is still pending.

5 Future work

There are many avenues open for future development. The coverage of the dictionaries can always be improved, as can the part-of-speech tagging and the organization and the coverage of transfer rules. Rules in the form of regular expressions may be added to dictionaries so that some entities such as place and person names may be recognized by the specific postpositions associated to them and handled accordingly. We have been looking at the possibility of using a constraint grammar for resolving long distance ambiguity.[14]

6 Concluding remarks

We have presented to our knowledge the first Basque to Spanish machine translation system. The system has taken advantage of existing data released under a free licence. Initial results are promising, and although the system is not suitable for producing text for dissemination, performance on the assimilation task has been found to be adequate, albeit with substantial room for improvement.

References

Aduriz, I., J. Arriola, X. Artola, A. Díaz de Ilarraza, K. Gojenola, and M. Maritxalar. 1997. Morphosyntactic disambiguation for Basque based on the constraint grammar formalism. In *Proceedings of Recent Advances in Natural Language Processing*, pages 282–288, Bulgaria.

Alegria, I., O. Ansa, X. Artola, N. Ezeiza, K. Gojenola, and R. Urizar. 2004. Representation and treatment of multiword expressions in Basque. In *ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 48–55.

Alegria, I., A. Díaz de Ilarraza, G. Labaka, M. Lersundi, A. Mayor, K. Sarasola, M. L. Forcada, S. Ortiz, and L. Padró. 2005. An Open Architecture for Transfer-based Machine Translation between Spanish and Basque. In *Proceedings of Machine Translation Summit X*, Phuket, Thailand.

Armentano-Oller, C., R. C. Carrasco, A. M. Corbí-Bellot, M. L. Forcada, M. Ginestí-Rosell, S. Ortiz-Rojas, J. A. Pérez-Ortiz, G. Ramírez-Sánchez, F. Sánchez-Martínez, and M. A. Scalco. 2006. Open-source Portuguese-Spanish machine translation. In *Proceedings of the VII Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada*, pages 50–59, Itatiaia-RJ, Brazil.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March. Association for Computational Linguistics.

Lesk, Michael. 1975. Lex – lexical analyzer generator. Technical Report Computer Science technical report #39, Bell Telephone Laboratories.

Matthews, Peter H. 1997. *The Concise Oxford Dictionary of Linguistics.* Oxford University Press, Oxford, UK.

Ortiz-Rojas, S., M. L. Forcada, and G. Ramírez-Sánchez. 2005. Construcción y minimización eficiente de transductores de letras a partir de diccionarios con paradigmas. *Procesamiento del Lenguaje Natural*, 35:51–57.

Tyers, F. M. and K. Donnelly. 2009. apertium-cy: A collaboratively-developed free RBMT system for Welsh to English. *Prague Bulletin of Mathematical Linguistics*, (91):57–66.

[14] A constraint grammar for Basque has already been written (Aduriz et al., 1997); unfortunately this is not publically available under a free licence and so the work would need to be redone.