Adaptation of a Rule-Based Translator to Río de la Plata Spanish

Ernesto López
Instituto de Computación
Universidad de la República
Uruguay
ernesto.nicolas.lopez@gmail.com

Luis Chiruzzo
Instituto de Computación
Universidad de la República
Uruguay
luischir@fing.edu.uy

Dina Wonsever
Instituto de Computación
Universidad de la República
Uruguay
wonsever@fing.edu.uy

Abstract

Pronominal and verbal voseo is a well-established variant in spoken language, and also very common in some written contexts - web sites, literary works, screenplays or subtitles - in Rio de la Plata Spanish. An implementation of Río de la Plata Spanish (including voseo) was made in the open source collaborative system Apertium, whose design is suited for the development of new translation pairs. This work includes: development of a translation pair for Río de la Plata Spanish-English (back and forth), based on the Spanish-English pairs previously included in Apertium; creation of a bilingual corpus based on subtitles of movies; evaluation on this corpus of the developed Apertium variant by comparing it to the original Apertium version and to a statistical translator in the state of the art.

1 Introduction

In this multilingual world, easily accessible through Internet, machine translation is becoming increasingly important. While the problem as a whole remains yet to be solved, there are several systems which provide an interesting service, by automatically producing a translated version of a text. In these days Google (Google, 2013) provides translation services - at least from and into English - for 51 different languages. While, in general, translations provided by Google are not completely accurate, users will have a reasonable comprehension of the content of the source text. A language like Spanish, that is spoken by about 420 million people (Instituto Cervantes, 2012) and is the official language in 21 countries, covering a vast geographical region, has different regional variations, some of which are firmly well-established. Appropriate coverage and fluid texts, adjusted to the situation and the language registry of an utterance, are not possible unless machine translation systems contemplate the consolidated and accepted variants used.

Pronominal and verbal *voseo* is a well-established variant in spoken language, and also very common in some written contexts - web sites, literary works, screenplays or subtitles - in Rio de la Plata Spanish.[1] To include this variant in a statistical machine translation system requires the availability of a large corpus for the language pair involved. An implementation of Río de la Plata Spanish (including *voseo*) was made in the open source collaborative system Apertium (Forcada et al., 2011), whose design is suited for the development of new translation pairs.

This work includes: development of a translation pair Río de la Plata Spanish-English (back and forth), based on the Spanish-English pairs previously included in Apertium; creation of a bilingual corpus based on subtitles of movies; evaluation on this corpus of the developed Apertium variant by comparing it to the original Apertium version and to a statistical translator in the state of the art. There is also an improvement of the translation system, through the addition of a repertoire of proper nouns of Uruguayan geographical regions.

The following section briefly introduces machine translation systems and their current performance. Section 3 describes the use of *voseo* in Río de la Plata while sections 4 and 5 describe the system development, the creation of the corpus, the evaluation and its results. Conclusions are in section 6. The developed translation pairs and the evaluation corpus are both available.

[1] This region includes an important part of Argentina and almost the entire Uruguayan territory.

Proceedings of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants, pages 15–22,
Hissar, Bulgaria, 13 September 2013.

2 Background

Machine translation (MT) is a development area within Natural Language Processing, which relates to the use of automatic tools to translate texts from one natural language into another. The different approaches used to solve this problem are separated into two main groups: Rule-based Machine Translation (RBMT) and Statistical Machine Translation (SMT).

2.1 Statistical Machine Translation

The current state of the art in MT is provided by Statistical Machine Translation systems. The initial interest into these approaches was drawn by the work of Brown et al. (1993), which recommends developing a *translation model* between language pairs and a *language model* for the target language. The system finds the best sentence in the target language, maximizing both accuracy (translation model) and fluency (language model).

Today the best performances are provided by phrase-based systems (PBMT) (Koehn et al., 2003). These systems consider the alignment of complete phrases in their translation model, and incorporate a *phrase reordering model*.

SMT systems strongly depend on the existence of a large volume of linguistic resources. Particularly, they depend on a target language corpus and a parallel corpus in source and target languages. This information is not available in an important number of language pairs.

2.2 Rule-Based Machine Translation

The second group of MT methods are the Rule-Based Machine Translation methods. These methods apply manually crafted rules to translate the source language text into the target language.

Usually, translations produced by these methods are more mechanic than and not as fluent as those produced by SMT. However, users who have a fairly good command of both languages do not require large parallel corpora to elaborate translation rules (Forcada et al., 2011).

2.3 Hybrid Machine Translation

In recent years, new approaches have attempted to combine the best qualities of the two traditional groups of translation systems (Thurmair, 2009). Statistical Post-Edition (SPE) edits manually the output of a RBMT system to produce a higher-quality translation. Then, a corpus is created using the RBMT output and the edited translation, and a SMT is trained with this corpus [Simard 2007].

By using a parallel corpus, Molchanov (2012) extracts a bilingual dictionary and complements it with SPE between the RBMT output and the parallel corpus destiny. Dugast et al. (2008) trains a SMT with the correspondence of the source text and the RBMT translation, instead of using a parallel corpus or a manually corrected output.

There is a different hybrid approach which uses phrases translated by the Apertium system (RBMT) to enrich the translation model tables of the Moses system (SMT) (Sanchez-Cartagena et al., 2011).

2.4 Apertium

Apertium is a RBMT system developed by the *Transducens group* from the *Universitat d'Alacant*. Originally, it was a translation system for related-language pairs (particularly for languages spoken in Spain) (Corbi-Bellot et al., 2005), but later on modules were added to translate more distant language pairs, such as English and Spanish (Forcada et al., 2011).

It is an open-source machine translation platform and it includes a set of tools to develop new language pairs. For this reason, an important number of collaborators have contributed with new linguistic resources and there are currently 36 pairs (Apertium, 2013) of languages officially accepted to be translated by Apertium.

Apertium has proved to be very useful to develop translation systems between related languages (Wiechetek et al., 2010) and languages with few linguistic resources (Martinez et al., 2012). In other development areas Apertium has been integrated to other finite-state tools such as the Helsinki Finite-State Toolkit (Washington et al., 2012).

As mentioned above, besides being used as a standalone RBMT system, there have been some experiments regarding the use of Apertium jointly with statistical systems (Sanchez-Cartagena et al., 2011).

3 Río de la Plata Spanish

There are variants in all languages spoken in the world. These are differences in vocabulary, verb conjugation, pronunciation, and in some cases, even syntactic differences.

Language variants are caused by historical, cultural and geographical factors. There are many sociological studies which try to explain the reason for these variants. For example, Chilean Spanish, which has a fairly unusual pronunciation, is a combination of the language spoken by Mapuche natives and the Quechua language from the south. This Spanish variant is found even in Argentinean provinces bordering with Chile. It has a lot in common with Rio de la Plata Spanish. Even in Uruguay, with a small population – just over 3 million people – there are multiple variants of Spanish. In Uruguayan cities separated by street borders from Brazil, there is a particular combination of Spanish and Portuguese. There is another example in southern Brazil, where Portuguese language uses the personal pronoun 'tú' (you singular).

The Spanish spoken in Río de la Plata is no exception, with many differences with Spanish from Spain. This variant occurs mainly in coast cities along Río de la Plata and Río Uruguay, upstream to the mouth of Río Negro. But it is also found in Uruguayan remote inland, albeit with variants, with a stronger Portuguese influence. Likewise, fusion of Spanish variants is seen in northern Argentina provinces and in southern Paraguay (Elizaincín, 2009).

There are various differences between Río de la Plata Spanish and the other Spanish variants. Some of these differences are only phonetic, such as *yeísmo* [2], and others are related to verb conjugation and pronoun uses, such as *voseo*.

Voseo - albeit not exclusively from Río de la Plata - is one of the most distinctive particularities of this Spanish variant, and it itself has some variants. In the definition of RAE[3] *voseo* is the use of the pronominal *vos* (You, singular) to address the interlocutor (RAE, 2011). There are two separate types of *voseo*:

The **reverential *voseo***, is the ceremonial usage of *vos* pronoun to address the second person, both plural and singular, and it is rarely used today. It is found in old Spanish texts, ceremonial writings or those which recreate Spanish language from the past.

The subject of this work is the **South American dialectal *voseo***. It is the Spanish use of the plural second-person pronominal and (modified) verbal forms, to address a single interlocutor. It is common in different variants of Spanish in Latin America, and, unlike reverential *voseo*, it implies closeness and informality since it is not usually seen - at least in its pronominal form - in very formal situations, where *ustedeo* is commonly used (Kapovic, 2007). The conjugation pattern in this variant is also different to peninsular Spanish.

Pronominal *voseo*

Pronominal *voseo* is the use of *vos* as singular second-person pronoun, instead of *tú* or *ti*. *Vos* is used as:

- Subject: *Puede que **vos** tengás razón* (**You** might be right)
- Vocative: *¿Por qué la tenés contra Alvaro Arzú, **vos**?* (**You**, what is it that you have against Alvaro Arzú?)
- Preposition term: *Cada vez que sale con **vos**, se enferma* (Every time he goes out with **you**, he feels unwell)
- Comparison term: *Es por lo menos tan actor como **vos*** (He is so good an actor as **you**)

According to RAE, for pronouns used with pronominal verbs and in objects with no preposition (atonic pronoun), and for possessive pronouns, it is combined with *tuteo* form, e.g.: *Vos **te** lavaste las manos* (You washed your hands), *No cerrés **tus** ojos* (Don't close your eyes).

Verbal *voseo*

Verbal *voseo* is more complex than pronominal *voseo*. RAE defines "verbal *voseo* is the use of the original verb suffixes of the plural second-person, more or less modified, in the conjugating forms of the singular second-person: *tú vivís, vos comés, vos comís* (you live, you eat, you eat)". Verbs vary differently in their form and tenses in each region. Complexity of verbal *voseo* lies on the fact that its use varies considerably in each region, some of which do not accept it as correct language. The subject of this work is the Río de la Plata variant. In fact, *voseo* is acknowledged as correct language only in Argentina, Uruguay and Paraguay (Kapovic, 2007). The Argentine Academy of Letters did not accept *voseo* as correct language – and only in some of its modalities - until 1982.

Voseo, as mentioned above, implies closeness and informality, and this is strongly related with its origin. Originally, it was rejected by purists and considered vulgar and demeaning by grammarians of the time. The use of *vos* was firmly rejected, particularly by the upper-class society.

[2] ***Yeísmo*** consists of a phonological variant, where consonants /Ǐ/ and /y/ are merged into a single sound /y/. It is a phonological process which merges two phonemes originally different (González, 2011).

[3] Royal Spanish Academy

Present tense verbal voseo

It may be found in indicative present tense forms combined with the plural diphthongs (*habláis (*You talk*)*); in some cases the ***s*** at the end of the verb is silent, particularly in Andean regions. In Río de la Plata, diphthongs consist of a single open vowel (*sabés (*You know*)*), although there are documents in which the vowel is closed (*sabís (*You know*)*). For first conjugation verbs, where infinitive forms end in *-ar*, verbs do not end in *–ís* with *vos* in this present tense form (RAE, 2011).

In present subjunctive structures *voseo* is seen in plural diphthongs as well (*habléis (…*you to talk), in some regions the *s* at the end of the verb is silent. In Río de la Plata, diphthongs consist of a single open vowel (*subás (…*you to climb*)*), although there are documents in which the vowel is closed (*hablís (…*you to talk*)*). Here, the *–ís* suffix only appears in first conjugation verbs.

Verbal voseo in imperative tenses

Voseo in imperative tenses is the variation of the plural second-person with omission of the *d* at the end of the verb. For example: *tomá (tomad (*take*))*, *poné (poned (*put*))*. These forms do not follow irregularities of the singular second-person characteristic of *tuteo*, therefore, *di (tell), sal (leave), ven (come), ten* (take) become *decí, salí, vení, tené* in verbal *voseo*.

These verb forms have accent marks since they are words stressed on the last syllable with a vowel at the end. When there is a pronoun attached to the verb, as a suffix, these forms follow general accentuation rules. For example: "*Compenetrate en Beethoven, imagináteIo. Imaginate su melena"* (RAE, 2011). ("Think about Beethoven, picture him. Picture his long hair" (RAE, 2011).

Pronominal and verbal *voseo* may be combined with *tuteo*. These are the modalities of *voseo*:

- Verbal and pronominal use of *vos*: Very frequently used in Río de la Plata. The subject, *vos*, is combined with verbal *voseo* forms, e.g.: "*Vos no podés entregarles los papeles antes de setenta y dos horas" (*You cannot give him the documents for the next three days*)*
- Exclusively verbal *voseo*: The subject of the verbal forms in this case is exclusively *tú*. It is commonly used in Uruguay, particularly in fairly informal situations.
- Exclusively pronominal *voseo*: *Vos* is the subject of singular second-person verbs, e.g.: "*Vos tienes la culpa para hacerte tratar mal*" (You are the one to blame for his abusive behaviour). This is rare in Río de la Plata.

4 Río de la Plata Apertium

Apertium decomposes the translation process into modules, executed in sequence. Figure 1 describes the pipeline of Apertium modules. It may be divided into the following steps:

- De-formatter: Separates the text in the input file from the format information.
- Morphological analyzer: In this module the text is segmented into lexical units. The units are supplied with morphological information. This step requires finite-state transducers (FST) technology.

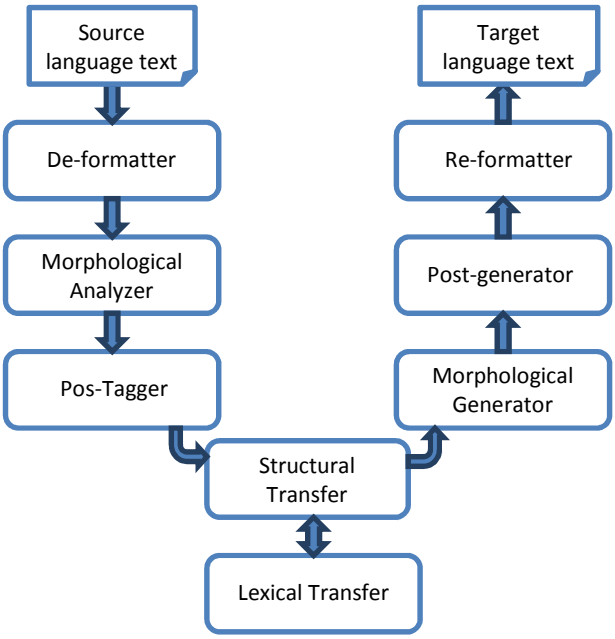

Figure 1 - Apertium modules

- POS-Tagger: The part-of-speech tagger chooses one of these analyses for the lexical unit.
- Lexical transfer: Establishes correspondence of the lexical units in the target language with the lexical units from the source text.
- Structural transfer: There is shallow parsing or chunking of text and a set of rules are applied, established specifically for each language pair, to transform the source language structure into the structure of the target language. Therefore, Apertium is classified as a shallow transfer system. (Forcada et al., 2010)
- Morphological generator: The morphological generator inflects target-language lexical units to produce the surface forms.

- Post-generator: Applies target-language orthographic rules.
- Re-formatter: Restores format information encapsulated by the de-formatter, to produce a translated file format similar to the source file format.

Voseo is a discourse phenomenon, occurring basically at morphological level. In Apertium, this process is carried out by the morphological analyzer. There is a transducer generated from an XML file, including the necessary rules (Forcada et al., 2011). The input of the module is a set of lexical units which are separately processed, inflections are analysed, and based on inflections, the attributes of the unit, such as lexical category, number or gender (for verbs) are tagged. The XML file information consists simply of rules which assign a set of attributes to a particular morphology.

In the particular case of verbs and verb tenses, verbs with similar morphological inflection – even if their lemma is different – belong to the same group. For example, in Spanish the verbs *cantar* (to sing) and *abandonar* (to abandon) have similar morphological inflection. Therefore, inflection paradigms are independent from lemmas.

	Cantaría	**Abandonaría**
Lemma	Cant	Abandon
Inflection	aría	aría
Attributes	Verb, Cond., Indic., Sing., First Person	Verb, Cond., Indic., Sing., First Person

Table 1 - Verbal paradigms in Apertium

It is clear that there are multiple analyses for each lexical unit. The selection of the corresponding analysis occurs in the next module. A new verb may be added by simply identifying its lemma and selecting an inflection paradigm. Therefore, since verbal *voseo* modifies verb inflections, this variation may be included by simply adding the new inflections to the paradigms already defined for traditional Spanish. So all the verbal paradigms defined in the Apertium dictionary were extracted, and the inflections studied and their corresponding attributes for imperative tenses and indicative present tenses were added. There were 170 inflection paradigms modified.

For pronominal *voseo*, the lexical unit *vos* was added to the dictionary. It was assigned with the attribute of tonic pronoun.

To improve the identification of named entities, 120 locations of Uruguay were added to the Apertium dictionary. They were extracted from the Geonames database (Geonames, 2011).

5 Evaluation and metrics

It is extremely complex to evaluate a translation system, mainly because there is usually more than one correct translation. Translations may vary in the word order, and even use different words. Yet translations will have many things in common and this is what metrics tries to measure to evaluate machine translation systems (Papineni et al., 2002).

A reference translation is always used to evaluate the MT system and sentences are the basic evaluation units every time. One of the most acknowledged metrics is BLEU, which weighs adequacy and fluency of sentences. This requires considering not only the number of lexical units in common between the translation to be evaluated and the reference translation, but also the length of common n-grams. BLEU also penalizes translation lengths which do not match the reference translation (Papineni et al., 2002).

NIST metrics - also used to evaluate translation systems - was also taken into account in this work. NIST is based on BLEU. The only difference is that NIST gives a higher score to less common n-grams, which actually provide more information to the content of the sentence.

5.1 Evaluation corpus

The corpus to evaluate the adaptation of Apertium should have the following particularities: It should be a bilingual Spanish – English corpus, for Río de la Plata Spanish, contemplating that *voseo* is more frequent in dialogues and conversations.

There are some texts that contain dialogues and conversations, which naturally have translations: movie subtitles. There are many movies subtitled in several languages. These subtitles may be read as transcriptions of the same text, in different languages, so subtitles are bilingual texts. Although it fails to be a perfectly aligned corpus, a very valuable asset of subtitles is the time window where they must be shown on screen. This provides more information, which is very useful to align two subtitles from the same movie.

It is very simple to find subtitles in the web. However, the only subtitles of interest for this work were those which included texts in Río de la Plata Spanish. IMDb highest-ranked movies from Argentina and Uruguay were used based on the premise that it is more likely to find the corresponding subtitles in both languages. Whenever possible, the original transcription extracted from the movies' official version was used. Otherwise, the subtitles used were those created by Internet users, based on the same highest-ranked premise. A corpus with about 100000 words was elaborated by using subtitles from 26 movies (Table 2).

Name	Year
The Pope's Toilet	2007
Son of the Bride	2001
Valentín	2002
Waiting for the Hearse	1985
Merry Christmas	2000
Official Story	1985
Night of the pencils	1986
The Die is Cast	2005
Rain	2008
Nine Queens	2000
Chinese Take-Away	2011
Avellaneda's moon	2004
A Matter of Principles	2009
Made Up Memories	2008
Martin (Hache)	1997
Camila	1984
Tierra del Fuego	2000
Seawards Journey	2003
Whisky	2004
25 Watts	2001
A place in the world	1992
Burnt Money	2000
Autumn sun	1996
Chronicle of an Escape	2006
Anita	2009
On Probation	2005

Table 2 - Movies used for the evaluation corpus

Sentences were aligned in all subtitles based on (Tyers and Pienaar, 2008; Tiedemann, 2007; Gale and Church, 1991; Brown et al., 1991). Sentences from subtitle pairs were aligned with relative precision, using the start/end time of the lines in the screen. In general, the parallelization algorithm groups together those sentences that appear in the same time frame in the subtitle pair. Then sentences are aligned based on their length, given similar sentence lengths in both languages. Accuracy was about 80% for random samples.

5.2 Evaluation and results

NIST and BLEU metrics were used. Adaptations made were compared with Apertium in its traditional version and with Google translator. Evaluation scripts (NIST, 2011) used were those developed in the 2008 edition of the NIST (National Institute of Standards and Technology) Open Machine Translation Evaluation.

In Spanish to English translations, all results provided by Apertium adapted to Río de la Plata Spanish were more correct translations than those obtained with Apertium's traditional version. As expected, Google translator provides significantly better results. Table 3 shows the average results for these metrics, in the Spanish into English direction.

	BLEU	NIST
Traditional Apertium	0.118183333	3.414683333
Río de la Plata Apertium	0.1246	3.553916667
Google Translator	0.226116667	4.810316667

Table 3 - Spanish into English translation results

The amount of *voseo* occurrences contained in the source text is difficult to establish, yet there is a 5.4% increase in the performance of Apertium in relation to the traditional version of the system. The modified system identifies the *voseo* verbs and its contractions, as well as all the uses of the *vos* pronoun.

In terms of recognition, the analysis of the morphological analyser output showed 13% and 14% improvement in the recognition of verbs and pronouns, respectively. Recognition improvement of named entities was 4.4%, reflecting that 44% more locations were identified.

While Google Translator provides better results, this is mainly due to the fact that generally translations are structurally and lexically more accurate. Many lexical units are not included in Apertium's dictionaries, which could explain its recognition problems, as shown in Table 4. In terms of *voseo*, Google Translator does not handle the *vos* pronoun properly: Google translator

translates '*Vos pensás en él*' as '**Vos you think on it*'.

Translation for:	Vos te lo merecés
Apertium	* Vos You it * merecés
R.P. Apertium	You deserve it

Table 4 - Translation before and after adaptation

In English to Spanish translations (Table 5), a fact to consider is that the Río de la Plata Apertium translator may operate in two modes to produce Spanish text: in the traditional mode (exclusively use of *tú*) or in the mode with exclusive use of *vos*. The traditional mode and the system without modifications provide identical results. Therefore, the work studied the operation of the system in the modality with exclusive use of *vos*.

	BLEU	NIST
Traditional Apertium	0.112733333	3.35635
Río de la Plata Apertium	0.111433333	3.374283333
Google Translator	0.21005	4.597533333

Table 5 - English into Spanish translation results

6 Conclusions

Apertium machine translations were improved by generating Río de la Plata Spanish – English pairs in the system. This is a free tool, and will be useful to translate colloquial language texts, such as web sites, blogs, literary works, screenplays or subtitles.

A Río de la Plata Spanish – English bilingual corpus was compilated from movie subtitles. This corpus was aligned and used for evaluation. There was clear improvement in relation to the previous version of Apertium. Apertium was compared with Google Translator at all times and in this context, Google Translator clearly surpasses Apertium. However, while Google Translator's performance is always better, there were some examples in which it failed to deal with the *voseo* particularity.

Translation was also improved by the addition of geographical entity names from the Geonames repository, filtered by their importance.

Overall, in translations from Río de la Plata Spanish into English, there is clear improvement, while not in the opposite direction, since *voseo* and traditional variants co-exist. So a more refined mechanism is required, to capture the speech registry in each statement and to select the corresponding mode. In future works, communicative situations and participants should be contemplated, as well as the symmetric and asymmetric interpersonal relations involved.

References

Apertium. 2013. Wiki – Apertium, Main Page. http://wiki.apertium.org/wiki/Main_Page (7th July, 2013)

Peter F. Brown, Jennifer Lai and Robert Mercer. 1991. *Aligning sentences in parallel corpora.* ACL.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation.* IBM T.J. Watson Research Center. Computational Linguistics - Special issue on using large corpora: II archive Volume 19 Issue 2, June 1993. Pages 263-311. MIT Press Cambridge, MA, USA.

Antonio M. Corbí-Bellot, Mikel L. Forcada, Sergio Ortiz-Rojas, Juan Antonio Pérez Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez-Martínez, Iñaki Alegria, Aingeru Mayor and Kepa Sarasola. 2005. *An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain.* Proceedings of the European Association for Machine Translation, 10th Annual Conference (Budapest, Hungary, 30-31.05.2005), p. 79-86.

Loïc Dugast, Jean Senellart and Philipp Koehn. 2008. *Can we relearn an RBMT system?* Proceedings of the Third Workshop on Statistical Machine Translation, pages 175–178, Columbus, Ohio, USA, June 2008.

Adolfo Elizaincín. 2009. *Geolingüística, sustrato y contacto lingüístico: español, portugués e italiano en uruguay.* ROSAE – Congresso em Homenagem a Rosa Virgínia Mattos e Silva.

Mikel L. Forcada, Boyan Ivanov Bonev, Sergio Ortiz Rojas, Juan Antonio Perez Ortiz, Gema Ramirez Sanchez, Felipe Sanchez Martinez, Carme Armentano-Oller, Marco A. Montava and Francis M. Tyers. 2010. *Documentation of the Open-Source Shallow-Transfer Machine Translation Platform Apertium.*

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez and Francis M. Tyers. 2011. *Apertium: a free/open-source platform for*

rule-based machine translation. Machine Translation: Volume 25, Issue 2 (2011), p. 127-144.

William A. Gale and Kenneth W. Church. 1991. *A program for aligning sentences in bilingual corpora.* ACL '91 Proceedings of the 29th annual meeting on Association for Computational Linguistics.

Geonames Web Site. *About Geonames.* http://www.geonames.org/about.html (23rd September, 2011).

Google Translate. 2013. http://translate.google.com/ (23rd July, 2013)

Rosario González Galicia. 2011. *Mi querida elle.* http://www.babab.com/no09/elle.htm (2nd August, 2011)

Instituto Cervantes. 2012. *Primer estudio conjunto del Instituto Cervantes y el British Council sobre el peso internacional del español y del inglés.* Instituto Cervantes. 21st June, 2012. http://www.cervantes.es/sobre_instituto_cervantes/prensa/2012/noticias/nota-londres-palabra-por-palabra.htm

Marko Kapovic. 2007. *Fórmulas de tratamiento en dialectos de español, fenómenos de voseo y ustedeo.* HIERONYMUS I, 2007.

Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. *Statistical Phrase-Based Translation.* HLT/NAACL 2003.

Juan Pablo Martínez Cortes, Jim O'Regan and Francis M. Tyers. 2012. *Free/Open Source Shallow-Transfer Based Machine Translation for Spanish and Aragonese.* LREC 2012.

Alexander Molchanov. 2012. *PROMT DeepHybrid system for WMT12 shared translation task.* Proceedings of the 7th Workshop on Statistical Machine Translation, pages 345–348, Montreal, Canada, June 7-8, 2012.

NIST. 2008. Mt08 scoring scripts. http://www.itl.nist.gov/iad/mig//tests/mt/2008/scoring.html, (23rd September, 2011)

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: A method for automatic evaluation of machine translation.* 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Real Academia Española. 2011. *Diccionario panhispánico de dudas, 1era edición 2da tirada.* http://buscon.rae.es/dpdI/SrvltGUIBusDPD?lema=voseo, (5th October, 2011)

Víctor M. Sanchez-Cartagena, Felipe Sánchez-Martínez and Juan Antonio Perez-Ortiz. 2011. *Integrating shallow-transfer rules into phrase-based statistical machine translation.* Proceedings of the 13th Machine Translation Summit : September 19-23, 2011, Xiamen, China, pp. 562-569

Michel Simard, Cyril Goutte and Pierre Isabelle. 2007. *Statistical Phrase-based Post-editing.* Proceedings of NAACL.

Jörg Tiedemann. 2007. *Improved sentence alignment for movie subtitles.* In Proceedings of RANLP 2007, Borovets, Bulgaria, pages 582–588, 2007.

Gregor Thurmair. 2009. *Comparing different architectures of hybrid MachineTranslation systems.* MT Summit XII: proceedings of the twelfth Machine Translation Summit, August 26-30, 2009, Ottawa, Ontario, Canada; pp.340-347.

Francis M. Tyers and Jacques A. Pienaar. 2008. *Extracting bilingual word pairs from wikipedia.* Proceedings of the SALTMIL Workshop at Language Resources and Evaluation Conference., LREC08:19–22.

Jonathan North Washington, Mirlan Ipasov and Francis M. Tyers. 2012. *A finite-state morphological transducer for Kyrgyz.* LREC 2012.

Linda Wiechetek, Francis M. Tyers and Thomas Omma. 2010. *Shooting at flies in the dark: Rule-based lexical selection for a minority language pair*. Proceeding IceTAL'10 Proceedings of the 7th international conference on Advances in natural language processing. Pages 418-429.