Saturnalia: A Latin-Catalan Parallel Corpus for Statistical MT

J. González-Rubio, J. Civera, A. Juan, F. Casacuberta

ITI/DSIC, Universidad Politécnica de Valencia
jegonzalez@iti.upv.es,{jcivera,ajuan,fcn}@dsic.upv.es

Abstract

Currently, a great effort is being carried out in the digitalisation of large historical document collections for preservation purposes. The documents in these collections are usually written in ancient languages, such as Latin or Greek, which limits the access of the general public to their content due to the language barrier. Therefore, digital libraries aim not only at storing raw images of digitalised documents, but also to annotate them with their corresponding text transcriptions and translations into modern languages. Unfortunately, ancient languages have at their disposal scarce electronic resources to be exploited by natural language processing techniques. This paper describes the compilation process of a novel Latin-Catalan parallel corpus as a new task for statistical machine translation (SMT). Preliminary experimental results are also reported using a state-of-the-art phrase-based SMT system. The results presented in this work reveal the complexity of the task and its challenging, but interesting nature for future development.

1. Introduction

Nowadays large historical document collections residing in libraries, museums and archives are being digitalised for preservation purposes and to make them available worldwide through large on-line digital libraries. The main objective, however, is not to simply provide access to raw images of digitised documents, but to annotate them with their real informative content and, in particular, with text transcriptions and, when convenient, text translations too.

Documents in historical collections are written in archaic forms of current official languages, as well as, in dead languages such as Latin or Greek. This fact limits the access of the general public to this information, which is not being fully exploited due to the language barrier. Unfortunately, there exist scarce electronic resources for these ancient languages suitable to be used in natural language processing (NLP), and more precisely in SMT.

Parallel texts for NLP purposes involving ancient languages, such as Latin, have been previously compiled (Resnik and others, 1999). However, to the best of our knowledge, they have never been published as a SMT task. This work presents the harvesting process of a new Latin-Catalan parallel corpus. This corpus was employed as a SMT task and initial experimental results obtained with a state-of-the-art phrase-based SMT system are reported.

The rest of the paper is structured as follows. Next section describes the compilation of the corpus, from text extraction and paragraph alignment to paragraph splitting and sentence alignment. Section 3. is devoted to the description of the experiments and results. Finally, Section 4. discusses the conclusions and future work ahead.

2. Corpus Collection

This section describes how the raw data originally in PDF format was transformed into a sentence-aligned Latin-Catalan parallel corpus. This involves five steps that are common in the compilation of almost any parallel corpus:

1. Text extraction: Obtaining the raw data.
2. Paragraph alignment: Extracting and mapping parallel segments of text.
3. Preprocessing: Normalisation and tokenisation of text in order to reduce corpus complexity.
4. Paragraph splitting: Dividing paragraphs into sentences.
5. Sentence alignment: Mapping sentences from one language to sentences in the other language.

The objective behind this process is to simplify the task of word alignment that lies at the core of state-of-the-art phrase-based SMT systems. From a probabilistic viewpoint, a word in one language can be a potential translation of any word in other language. Aligning at the sentence level allows us to reduce the number of possible words to which a word in one language can be aligned (translated) in the other language. This simplification improves the quality of word-alignment statistical models, and therefore the translation quality of phrase-based SMT systems.

2.1. Text Extraction

In this section, we briefly describe the extraction of raw text from files in PDF format, and the layout of the raw text.

The original book is the *Saturnalia* by *Ambrosius Theodosius Macrobius*[1], a Roman grammarian and Neoplatonist philosopher born in the fourth century. The Latin-Catalan version of the book[2] is part of the *Bernat Metge* collection, published by the *Institut Cambó*[3].

The Institut Cambó kindly provided us with files in PDF format of this book. Each file contains the original text in Latin and its corresponding translation in Catalan in a two-column format. As usual in Greco-Roman books, each paragraph is tagged with a sequential number between brackets, known as versicle number. As the reader could guess, versicle numbers are vital in paragraph alignment. This will be explained in the next section.

In order to extract the text from the PDF files, we initially converted these PDF files into files in HTML format using the standard *pdftohtml* tool found in Linux distributions. Finally, we adequately parsed the HTML files to extract the raw text.

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV "Consolider Ingenio 2010" program (CSD2007-00018), the iTransDoc (TIN2006-15694-CO2-01) and iTrans2 (TIN2009-14511) projects and the FPU scholarship AP2006-00691. Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project.

[1] http://en.wikipedia.org/wiki/Ambrosius_Theodosius_Macrobius
[2] http://books.google.com/books?id=BQhlmI60N_EC
[3] http://www.bernatmetge.com

2.2. Paragraph alignment

In our case, the task of paragraph alignment was greatly simplified by the use of versicle numbers. Paragraphs that were translation of each other received the same versicle number. Thus, we just had to output those pairs of Latin-Catalan paragraphs sharing the same versicle number.

2.3. Normalisation and Tokenisation

Once the corpus was aligned at the paragraph level, we converted it into lowercase, and separate punctuation marks from words by a tokenisation process. Lowercasing and tokenisation are conventional preprocessing steps in harvesting parallel corpora. As a result, the vocabulary size is significantly reduced and word spelling differences are eliminated. In our case, lowercase and tokenisation scripts are those of the ACL 2009 Workshop on SMT (Callison-Burch and others, 2009).

2.4. Paragraph splitting and sentence alignment

Paragraph splitting is an ill-posed problem regarding the segmentation of paragraphs into smaller sense units. Paragraph splitting is then followed by a process of aligning sentences from one language to their corresponding translation sentence in other language.

To solve these problems we have followed two approaches. Initially, we employed automatic statistical techniques to perform paragraph splitting and sentence alignment. However, the poor translation results obtained with this approach led us to consider the manual segmentation and alignment by an expert. In this section, we provide details of both approaches and discuss the problems faced with the automatic approach.

2.4.1. Automatic splitting and sentence alignment

Automatic splitting is usually carried out on the basis of punctuation marks, such as the period, the semicolon, the colon and the comma. A subset of these four punctuation marks defines what we refer to as *anchor words*. However, we are fully aware that an anchor word may introduce ambiguity in the segmentation process. This is the case of the period that appears in abbreviations.

Regarding sentence alignment, there is considerable previous work. In (Gale and Church, 1993), the authors presented an algorithm that aligns sentences of similar length and merges sentences if necessary. One extension of this algorithm consists in incorporating word correspondences within sentence pairs (Melamed, 1999; Varga and others, 2005). In our case, we employed the RecAlign algorithm (Nevado and others, 2004). RecAlign is a greedy algorithm based on a statistical dictionary, that recursively computes sentence alignments on the basis of a predefined set of anchor words. The statistical dictionary was trained using a extended version of the well-known IBM Model 1 (González-Rubio and others, 2008). This model estimates a statistical dictionary from a predefined segmentation of paragraphs in both languages. The segmentation is provided by anchor words.

However, the sentence-aligned corpus obtained as a result of this automatic process presents a major drawback. As shown in Example 1, sentence pairs tend to be excessively long to adequately train a SMT system.

Paragraph splitting and sentence alignment is particularly difficult between Latin and Catalan due to substantial linguistic differences. On the one hand, Latin is a synthetic, fusional language that uses suffixes attached to fixed stems to express gender, number, and case in adjectives, nouns, and pronouns. This linguistic process is known as declension. However, declension in Catalan is replaced with the use of prepositions. This fact is reflected in Catalan sentences being much longer than their corresponding Latin translations. On the other hand, word order is not preserved in both languages. For example, the verb is located at the end of the sentence in Latin, while in Catalan, the verb usually appears in the middle of the sentence. Word reordering keep us from defining monotone word alignments, adding complexity to the task of splitting paragraphs and aligning sentences.

Latin: multas uariasque res in hac uita nobis , eustathi fili , natura conciliauit ; sed nulla nos magis quam eorum qui e nobis essent procreati caritate deuinxit , eamque nostram in his educandis atque erudiendis curam esse uoluit , ut parentes neque , si id quod cuperent ex sententia cederet , tantum ulla alia ex re uoluptatis , neque , si contra eueniret , tantum maeroris capere possint .

Catalan: la natura , eustati , fill meu , ens ha donat moltes i diverses coses en aquesta vida , però no ens ha dispensat cap lligam tan fort com l ' afecte envers aquells qui han estat procreats per nosaltres , i ha volgut que posem la nostra cura en llur criança i llur educació , fins al punt que els pares no poden atènyer cap plaer més immens si s ' esdevé satisfactòriament allò que anhelen i , en canvi , els sobrevé la més gran de les tristeses si succeeix tot el contrari .

Example 1: Sentence pair generated by the automatic sentence splitting and alignment process.

As mentioned above, experimental results obtained on this version of the parallel Latin-Catalan corpus were very poor. Therefore, we had to revert to human expertise to successfully address the task of paragraph splitting and sentence alignment.

2.4.2. Manual segmentation and alignment

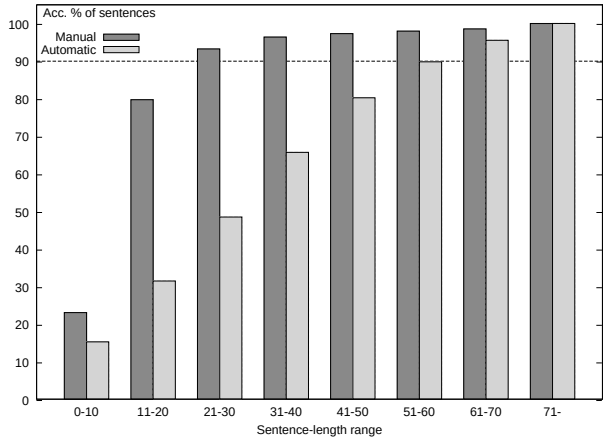
The manual segmentation was carried out by a human expert that was instructed to divide paragraphs into minimum sense units that could be aligned monotonically. An additional constraint was to keep sentence lengths under 40 words in both languages, whenever possible. Moreover, the expert was asked to annotate text fragments in Greek appearing in the Latin text, as well as their corresponding translations into Catalan. A manually split and aligned sentence is shown in Example 2.

Latin: multas uariasque res in hac uita nobis , eustathi fili , natura conciliauit ;

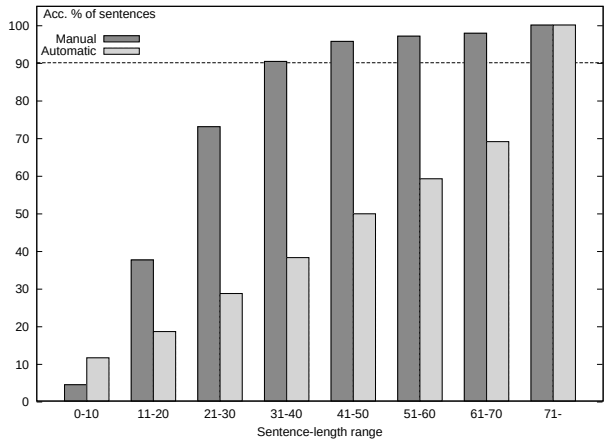
Catalan: la natura , eustati , fill meu , ens ha donat moltes i diverses coses en aquesta vida ,

Example 2: Sentence pair generated by a human expert.

Figure 1 shows, for Latin and Catalan, the accumulated percentage of number of sentences up to a certain sentence-length range. Light grey bars represent sentence lengths obtained by applying the automatic approach, while dark grey bars show the manual approach. As observed in Figure 1, the manual segmentation provided a parallel corpus with shorter sentences than the automatic segmentation. In-



(a) Latin



(b) Catalan

Figure 1: Accumulated histogram for the percentage of number of sentences (y axis) in (a) Latin and (b) Catalan up to a given sentence-length range (x axis), after paragraph splitting and sentence alignment. Light grey bars represent sentence lengths obtained by applying the automatic approach, while dark grey bars show the manual approach.

deed, 90% of the sentences given by the manual segmentation are below 30-40 words, while lengths of 70 or more words are needed to cover the 90% of the sentences for automatic segmentation. For this reason, hereafter the experimental results reported in this paper were computed on the manually split and aligned version of Saturnalia.

A set of basic statistics of Saturnalia is reported in Table 1. A priori, the first problem that we can observe is the small size of the corpus with only seven thousands sentences, compared to the millions of sentence pairs available in large parallel corpus. Furthermore, the high ratio between singletons[4] and vocabulary size denotes the complexity of the task. For instance, this ratio goes up to 62% in Latin, that is, 62% of the words only occurs once in the whole corpus. Another figure supporting the difficulty of the task is the perplexity for Latin computed as the average of a 10-fold cross validation setup.

[4]Words with a single occurrence in the whole corpus

	Latin	Catalan
# sentences	7172	
# running Kwords	118	183
vocabulary (Kwords)	23	17
singletons (Kwords)	14	10
singletons ratio (%)	62	56
Average length	16	25
Perplexity (5-gram)	368	103

Table 1: Basic statistics for Saturnalia.

3. Experiments and results

Once the Saturnalia corpus was preprocessed and sentence aligned, a series of experiments were performed in order to test the capabilities of a state-of-the-art phrase-based SMT system when translating from Latin to Catalan.

To deploy our Latin-Catalan phrase-based SMT system, we used the publicly available *Moses toolkit* (Koehn and others, 2007). Moses allows us to train a state-of-the-art phrase-based SMT system (Koehn et al., 2003) with little effort, as well as, to smoothly integrate linguistic information from diverse sources (Koehn and Hoang, 2007). This feature is specially important in the case of working with a small corpus, where morphological, syntactic, or semantic sources of information boost the performance of SMT systems (Nießen and others, 2000).

3.1. Experimental Setup

Given the small size of the Saturnalia corpus, we design a 10-fold cross validation experiment to obtain more robust results. This means that the corpus was split into 10 partitions of equal size. Eight partitions were devoted to train the phrase-based SMT system, while the other two partitions were used as development and test sets, respectively. We repeated this process 10 times shifting training, development and test sets across the 10 partitions.

As mentioned above, the training set was used to generate a phrase-based SMT system. Current SMT systems are a log-linear combination of models whose weights need to be optimise on a development set according to a translation quality criterion. In our case, the Moses toolkit provides MERT (Och, 2003), a piece of software to optimise the weights of a log-linear model according to the *BiLingual Evaluation Understudy* (BLEU) (Papineni and others, 2002) score. BLEU and *Translation Edit Rate* (TER) (Snover and others, 2006) metrics are computed on the test set. BLEU score is an accuracy measure of the degree of n-gram[5] overlapping between the system and the reference translation. TER is an error metric that measures the number of edit operations to convert the system translation into the reference translation. Finally, BLEU and TER scores reported are the average values calculated over the test set of the 10 folds.

3.2. Incorporating linguistic information

To collect and extract linguistic information from Latin and Catalan texts we used two freely available linguistic tools, *Words*[6] for Latin and *FreeLing* (Atserias and others, 2006) for Catalan. Words and Freeling work with individ-

[5]A sequence of n consecutive words in a sentence.

[6]http://users.erols.com/whitaker/words.htm

ual words providing lemma and suffix, along with morpho-syntactic information.
Thus, we define four Latin-Catalan SMT systems depending on the linguistic information integrated into the system:

- *Baseline*: Conventional SMT system trained on the original corpus, adding no linguistic information at all.
- *Baseline + Catalan Morpho*: We incorporate morpho-syntactic (POS-tagging) information in the baseline SMT system to improve number and gender agreement in Catalan. This is achieved by first generating POS-tags from translated words, and then using a POS-tag language model smoothly integrated in a log-linear fashion into the baseline system.
- *Baseline + Morpho*: Latin and Catalan morpho-syntactic information is provided not only to translate at the word level (word translation), but also at the POS-tag level (POS-tag translation). Independently from the POS-tag translation, translated Catalan words generate POS-tags that have to be compatible with those obtained by the POS-tag translation. This process adds an additional constraint with respect to the previous scenario, and it is thought to produce better results since compatible POS-tags must be generated by two alternative paths.
- *Lemma + Suffix*: In this scenario Latin words are split into lemma and suffix according to the tool Words, and employed to train a phrase-based SMT system. The aim behind this scenario is to partially revert the declension process in Latin, so that suffixes in Latin can be translated into prepositions in Catalan. This splitting process reduces the Latin vocabulary by half, the number of singletons by 20% and the perplexity of Latin almost by third.

3.3. Results

Table 2 presents the translation quality assessment of the four Latin-Catalan SMT systems described in Section 3.2.. Results are reported in terms of BLEU score and TER.

	BLEU	TER
Baseline	10.9	80.9
Baseline + Catalan Morpho	10.8	82.3
Baseline + Morpho	10.9	81.4
Lemma + Suffix	11.6	80.8

Table 2: BLEU and TER evaluation of the four Latin-Catalan SMT systems described in Section 3.2.

Generally speaking, BLEU score and TER obtained when translating from Latin to Catalan are lower than those reported between European languages (Callison-Burch and others, 2007). As mentioned in Section 2., the complexity of this task poses a real challenge for a SMT system. First, the most notable drawback is the size of the corpus, a few thousands sentence pairs are not enough to adequately train a SMT system. This drawback is accentuated due to the complexity of the Latin language as proved by the figures presented in Table 1. Other peculiarities of this Latin-Catalan corpus are its literary style and the existence of text fragments in Greek.
Taking this into consideration, we observe in Table 2 that the incorporation of morpho-syntactic information does not help to improve the translation quality of the system. We believe that this unexpected behaviour is due to the large number of singletons and the poor quality of the word alignments that feed our phrase-based SMT systems.
However, if we simplify the task by splitting Latin words into lemma and suffix, we observe an improvement of 5% in BLEU score. This fact is explained by Latin declension, since splitting Latin words into lemma and suffix allow to separately align them to content and function words, respectively, in Catalan. These better alignments lead us, in this case, to better translation quality.

4. Conclusions and future work

This paper describes the acquisition and preprocessing of Saturnalia, a Latin-Catalan parallel corpus for SMT. Preliminary experimental results using a state-of-the-art phrase-based SMT system are also reported.
We are currently working on increasing the size of the Latin-Catalan parallel corpus. To this purpose, we have at our disposal a new book from the Bernat Metge collection, the *Epistolaria* by *Sidonio Apolinar*[7]. This book is being preprocessed to be appended to the Latin-Catalan corpus.
Finally, given the limited size of the corpus, we plan to exploit other linguistic resources such as bilingual dictionaries in order to enhance word alignments and ensure their correctness. We expect that superior alignments in this task will result in better phrase-based SMT systems.

5. References

J. Atserias et al. 2006. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. In *LREC*.
C. Callison-Burch et al. 2007. (Meta-) Evaluation of Machine Translation. In *WSMT*, pages 136–158.
C. Callison-Burch et al. 2009. Findings of the 2009 Workshop on SMT. In *WSMT*, pages 1–28.
W. Gale and K. Church. 1993. A program for aligning sentences in bilingual corpora. *Comp. Ling.*, 19(1):75–102.
J. González-Rubio et al. 2008. A novel alignment model inspired on IBM Model 1. In *EAMT*, pages 47–56.
P. Koehn and H. Hoang. 2007. Factored translation models. In *EMNLP-CoNLL*, pages 868–876.
P. Koehn et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180.
P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *NAACL*, pages 48–54.
I. Melamed. 1999. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25:107–130.
F. Nevado et al. 2004. Bilingual corpora segmentation using bilingual recursive alignments. In *JTH*.
S. Nießen et al. 2000. Improving smt quality with morpho-syntactic analysis. In *ACL*, pages 1081–1085.
F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167.
K. Papineni et al. 2002. BLEU: a method for automatic evaluation of MT. In *ACL*, pages 311–318.
P. Resnik et al. 1999. The bible as a parallel corpus: Annotating the "book of 2000 tongues".
M. Snover et al. 2006. A study of TER with targeted human annotation. In *AMTA*, pages 223–231.
D. Varga et al. 2005. Parallel corpora for medium density languages. In *RANLP*, pages 590–596.

[7] http://en.wikipedia.org/wiki/Sidonius_Apollinaris

Parallel corpora are central to translation studies and contrastive linguistics. Many of the parallel corpora are accessible through easy-to-use concordancers which considerably facilitates the study of interlinguistic phenomena. Such corpora are also a rich source of materials for language teaching. Furthermore, parallel corpora serve as training data for statistical machine translation systems. The parallel corpora are our largest resource family, as the CLARIN ERIC infrastructure provides access to 84 parallel corpora, the majority of which are available for download from national repositor