RiCL Research in Corpus Linguistics

POS-tagging a bilingual parallel corpus: methods and challenges

Irene Doval
University of Santiago de Compostela / Spain

Abstract – This paper reviews the author's experiences of tokenizing and POS tagging a bilingual parallel corpus, the PaGeS Corpus, consisting mostly of German and Spanish fictional texts. This is part of an ongoing process of annotating the corpus for part-of-speech information. This study discusses the specific problems encountered so far. On the one hand, tagging performance degrades significantly when applied to fictional data and, on the other, pre-existing annotation schemes are all language specific. To further improve accuracy during post-editing, the author has developed a common tagset and identified major error patterns.

Keywords – multilingual resources, parallel corpus, corpus annotation, POS tagging, tagset corpus building

1. INTRODUCTION

In the last decades, parallel and comparable corpora have played an increasingly important role and parallel corpus linguistics has emerged as a distinct field of research within corpus linguistics (Borin 2002: 1). As the name implies, the term 'parallel corpus' has now been established to refer to collections of bitexts,[1] consisting of original texts in one language, together with their translations into another language or of translations from original texts written in a third language. At the University of Santiago de Compostela a parallel corpus German/Spanish (PaGeS) is currently being developed.[2] The goal of this project is to create an open multifunctional tool for a variety of purposes ranging from general research in cross-linguistics and translation studies, to more practical applications, such as teaching and learning foreign languages or translation.

In order to achieve this objective, the corpus has to meet certain requirements regarding quality of texts and translations, size and representativeness and accuracy of the alignment. It also has to provide a high qualitative linguistic annotation of the language data. Specifically, the morphosyntactic annotation (part-of-speech tagging or POS tagging for short), to which this study is devoted, "is an important and widely-used preprocessing step in natural language processing applications, and it is almost indispensable for the exploitation of corpus data" (Giesbrecht and Evert 2009). This paper describes the process of providing the Spanish and the German data of the corpus PaGeS with part of speech tags, which includes the selection of the tagger software, the assessment of the results in both languages and the development of a unified tagset.

Although POS tagging is a common task in the creation of a corpus, when dealing with multilingual data, as in this case, this task involves additional challenges. On the one hand, Spanish and German languages each have specific grammatical features and terminology. Accordingly, each language uses annotation schemes that differ greatly between themselves in scope and detail. But to search efficiently for lexical categories through large amounts of data in both, Spanish and German, the user needs a

[1] The term 'bitext' was originally coined by Harris (1988) to refer to documents along with their translations into other languages. However, this term is now commonly used in a broader sense to refer to a wider range of parallel resources, not only original documents and their direct translations (see Tiedemann 2011: 1).

[2] This project is carried out by the research team SpatiAlEs, led by Prof. Irene Doval. For more detailed information on the PaGeS Corpus, see below, Doval (2016) and Doval et al. (forthcoming).

tagging scheme with a unified format and conventions for representing the tags. On the other hand, the composition of corpus PaGeS, consisting primarily of fiction, leads to a higher error rate of the taggers, commonly trained on very different data regarding genre and domain.[3]

The remainder of this paper is organized as follows. Section 2 describes the composition and current size of the PaGeS Corpus, as well as the steps completed in its construction. Section 3 presents the different challenges encountered in the tokenization of the German and Spanish texts. Section 4 focuses on the tagging process. It first describes the basic features of the tagger used for the annotation, TreeTagger, and evaluates its accuracy running it on a subset of the Spanish and German data. I consider some issues to explain the difference between the reported tagging accuracies and the results in PaGeS and I give an overview of the main error patterns, which can be subsequently manually corrected. Then, I discuss the differences across the standard tagsets used for Spanish and German and propose a harmonized tagset for the corpus PaGeS (see Appendix). Finally, section 5 offers a summary and some perspectives.

2. THE PAGES CORPUS: DESIGN AND DATA PROCESSING

As mentioned above, the PaGeS corpus (www.corpuspages.eu), is a bilingual parallel corpus consisting of German and Spanish original and translated texts as well as a small percentage of German and Spanish translations from a third language. The German and Spanish data have been linked together sentence by sentence. They form a growing collection of fiction (roughly 90% of novels and short stories) and nonfiction (essays and popular science texts). Most of the selected books are represented not by the full texts but by samples, allowing a better cross-section of texts.

As shown in Table 1, at the current stage (July 2017), PaGeS contains 19,017,837 words and 655,321 bisegments, i.e. pairs of aligned text chunks (sentences or smaller segments).

Language	Works	Types	Tokens
German Original	54	140,750	4,253,900
German Translation<Spanish	38	126,225	3,564,688
German Translation<3rd language	12	49,333	1,577,794
Spanish Original	38	99,710	3,584,908
Spanish Translation<German	54	89,109	4,507,832
Spanish Translation<3rd language	12	61,925	1,528,715
Total	208		19,017,837

Table 1: Size of the PaGeS corpus in terms of works, types and tokens, sorted by languages

After being digitized, the texts undergo a manual process to prepare them for the alignment. This consists of reducing the noise and achieving as much parallelism as possible between the source and target texts with a view to obtaining the best results in the alignment. This implies the removal of non-corresponding texts, bad characters and pictures and proofreading.

The texts are then aligned using LF-Aligner,[4] the resultant output being a TSV-file, a simple text format that stores the bisegments in a tabular structure. In the alignment process, two tasks are combined: a prior segmentation of the texts and the subsequent linking of those segments with the corresponding ones in the other language to form bisegments, i.e., aligned pairs of segments (Tiedemann 2011: 7). The segmentation is done monolingually and the alignment is based on this step. To guarantee quality, the corpus has been verified manually at different levels, including preprocessing, sentence splitting and sentence alignment.

Besides the language data, PaGeS contains three types of non-textual information added to enrich the corpus and to facilitate the linguistic exploration of the material: metadata, mark-up and linguistic annotation (McEnery and Hardie 2012: 29). The metadata include information about author, title, year of the first publication and, when applicable, the edition used. The textual mark-up corresponds to the internal structure of the texts, that is the division in parts, chapters and pages.

As for the linguistic annotation, the data are being currently lemmatized and POS-tagged in order to enhance the usefulness of the corpus. This is particularly important with regard to the aforementioned

[3] In this paper the terms 'genre' and 'domain' are used interchangeably, in keeping with the practice of the British National Corpus (see Lee 2001), which does not distinguish any domains within fiction ('written imaginative').

[4] LF Aligner is a freeware, open source, multiplatform alignment tool created by Farkas András (http://sourceforge.net/projects/aligner).

primary purpose, the contrastive analysis of the spatial relations in German and Spanish and as a first step for further annotation layers. This paper is devoted to this process, the issues found and the solutions proposed.

For this purpose, once the bisegments have been aligned and identified, the TSV-file has to be split back into two monolingual files, given that the subsequent task of tokenizing and aligning can only be performed on monolingual data. This workflow is summarized in Figure 1.

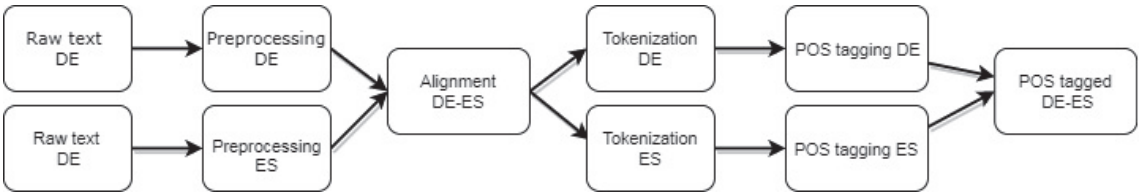

Figure 1: The workflow of the PaGeS Corpus construction

Before focusing on the POS tagging process, we will briefly explain the tokenization process that needs to take place beforehand.

3. TOKENIZATION

As shown in Figure 1, before tagging the data have to be previously tokenized. At first sight, tokenization, the division of the sentences into tokens, seems to be a trivial task, which it would be if tokenization consisted only of identifying strings delimited on both sides by spaces or punctuation.

However, tokenization is a much more challenging task. First, tokenization is not fully independent of the previous sentence splitting. This sentence segmentation is not straightforward either due to the ambiguity of some punctuation marks, particularly the dot character. The following German example illustrates different uses of the dot character as a marker of abbreviated forms, a marker of an ordinal number and full stop to mark the end of a sentence.

(1) Dr. Müller feiert heute seine 30. Geburtstag.

The example shows that it is necessary to disambiguate end-of-sentence punctuation from part-of-word punctuation as in abbreviations or ordinal numbers. For this, a finite list of common abbreviations for German and Spanish is integrated in the tokenizer. This list is obviously not exhaustive, since new or occasional abbreviations are created continuously.

Moreover, for languages with alphabetic writing systems, like German or Spanish, the orthographic word – delimited by a white space preceding and following – does not necessarily correspond to the morphosyntactic word, that is, the word token needed for further linguistic analysis. Therefore, as Leech (1997: 22) points out, these units should be both linguistically significant and methodologically useful.

These issues of tokenization are language-specific depending on the orthographic conventions of a particular language. In the case of German and Spanish some of the issues we had to face are listed below. The list is far from exhaustive and is intended to illustrate the difficulties that tokenization occasionally runs into.

The first case are multiwords, that is, when more than one orthographic word corresponds to one morphosyntactic word. Below are some examples:

- German and Spanish complex prepositions (*a pesar de*, *en relación a*, *in bezuf auf*)
- Spanish complex subordinating conjunctions (*para que*, *a pesar de que*, *so dass*)
- German verbs with separable particles (*anfangen*, *losrennen*)
- multiword named entities (*La Coruña*, *New York*)
- dates (11. *Januar* 2011), phone numbers and other numerical sequences that can contain a space inside.

Conversely, there is the case of mergers, that is, one orthographic word which corresponds to more than one morphosyntactic word. These cases involve:

- Spanish enclitic forms, that is, unstressed pronoun forms, which are orthographically attached to the end of some infinitive, participle, imperative and subjunctive exhortative verbal forms, such as *dímelo*, *vayámonos* or *dárselas*.

- Portmanteau word forms in German and Spanish: Contractions of prepositions and definite articles in German (*im*, *ins*, *zur*, *aufs*, *am*…) and Spanish (*al, del*) or the German postclitic *'s* (*geht's*, *gibt's*).

In the corpus, tokenization was performed by the tokenizer integrated into the POS tagger (see below). In general, we used the default assumption that an orthographic word is the most appropriate unit for grammatical tagging. Exceptions to this were listed in a separate file for each language. We decided not to use pre-existing files for this purpose, given that multiwords are not consistently gathered in any of them. This produced many inconsistencies in the tagging (e.g. the compound preposition *al lado de* may receive a single tag, while *a través de* may receive three tags, because it is missing in the file). Furthermore, the grammatical status of these units is not clear: reference books on Spanish grammar do not agree as to whether these strings are complex prepositions or freely constructed sequences. Similarly, multiword German adverbs like *nach wie vor* or *ab und zu* were tagged separately.

Regarding the orthographically merged forms, they are not split during tokenization, but the tagset handles them using specific tags in the cases of Spanish clitics and German and Spanish contractions of preposition and article.

When all the aforementioned restrictions are in place, tokenization immediately becomes a challenging and crucial task, since tokenizer errors propagate not only to POS tagging but also to all subsequent components in the pipeline. Moreover, tokenization is only preceded by sentence splitting and takes place without any previous grammatical analysis of the context, which would be helpful for achieving a better performance.

4. PART-OF-SPEECH TAGGING

POS tagging is the process of assigning a part-of-speech label to each token in an input text based on both its form and its context. Tagging therefore consists of two steps: (a) tag assignment: each token is assigned the corresponding set of possible tags for a particular tagset; and (b) tag disambiguation: using different procedures, the number of tags per token is reduced to the correct one for the specific context. This is not a trivial task since word forms are ambiguous and tagging without a complete syntactic analysis is often difficult even for humans (Jurafsky and Martin 2017: 11).

Table 2 shows the ambiguity rates in the PaGeS corpus using the 50 word tags of the Stuttgart-Tübingen tagset (STTS) for German and the 52 tags of the Spanish tagset for Spanish (see below). Most word types (87-88%) are unambiguous; that is, they have only a single tag. The ambiguous words, although they account for only about 12% of vocabulary, are some of the most commonly used words in Spanish (*que*, *como*, *la*, …) and German (*der*, *die*, *das*, *und*), hence about half of word tokens in running text are ambiguous. Note that there are no major differences between the two languages.

	Types		Tokens	
	German	Spanish	German	Spanish
Unambiguous (1 tag)	88%	87%	50.5%	48.1%
Ambiguous (1 tag)	12%	13%	49.5%	51.9%

Table 2: The amount of tag ambiguity for word types and tokens in the PaGeS Corpus from the TreeTagger tagging

4.1. The tagger

POS taggers can be grouped into two main types: rule-based and stochastic. Rule-based taggers start by assigning all possible tags to words using a dictionary. Then they apply hand-written rules to selectively remove tags. Whereas most of the early taggers were predominantly rule-based, stochastic taggers are more widely used these days. They are mostly HMM based,[5] i.e. they choose the tag sequence that is most probable given the observation sequence of *n* words. Figure 2 shows a schematic of HMM-based tagging for a sample German sentence, showing the set of possible tags for each word and the correct tag sequence as the highlighted path.

[5] HMM is the abbreviation for 'Hidden Markov Model'. An HMM-part-of-speech tagger is a probabilistic sequence model: given a sequence of words, it computes the most probable POS tag sequence. HMM-taggers are fast and have successfully been applied to a wide range of languages and training corpora (Schmidt and Laws 2008: 777).

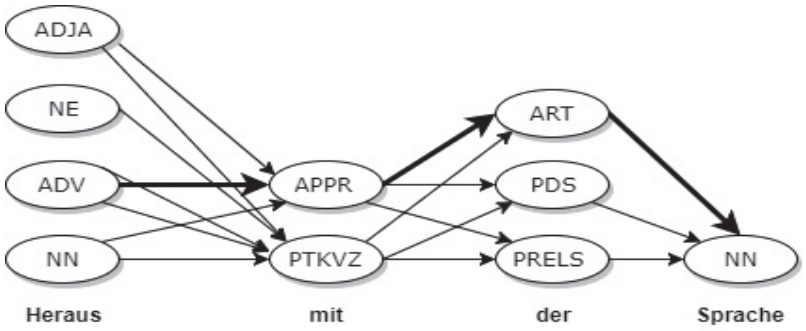

Figure 1: Schematic of HMM-based tagging

In the PaGeS Corpus, we have chosen to use TreeTagger[6] (TT), a popular free application developed by Schmidt (Schmidt 1994, 1995) at the Institute for Computational Linguistics at the University of Stuttgart. TT is a language-independent HMM-tagger, which uses decision trees to disambiguate word forms. It needs a training corpus that has been hand-tagged with POS. From this tagged training corpus, a full form lexicon is created. Thus, TT consists of two programs: on the one hand, the training program, which creates a parameter file from the full form lexicon and the hand-tagged corpus; on the other, the tagger program, which reads the parameter file and annotates the text with part-of-speech and lemma information. As we intended to use the same tagger for the German and Spanish texts, TT seemed to be the best option, as it is adaptable to different languages on the basis of a lexicon and a manually-tagged training corpus. TT provides trained models for German and Spanish, among many other languages. The German parameter file available on the TT website was trained on the TIGER corpus[7] and the Spanish parameter file was trained on the CRATER corpus.[8]

In the input file, each line must contain only one token (words or punctuation characters). The tokens may contain blanks, thus allowing multiword tokens. The output file also has a one-word-per-line format. It contains three columns separated by tabs, with tokens, tags and lemmas, as illustrated in Table 3.

		Output file	
Input file	Word	tag	Lemma
cuando	cuando	CSUBX	cuando
se	se	SE	se
miraba	miraba	VLfin	mirar
a	a	PREP	a
la	la	ART	el
calle	calle	NC	calle
,	,	CM	,

Table 3: Sample of input and output of the TreeTagger with the CRATER tagset

Every tagger lexicon is incomplete, however large it may be. New proper names and acronyms are always being created and new nouns and verbs enter the language continuously. When TT does not find a word in the lexicon, the value for its lemma is unknown, but nevertheless it assigns a POS tag to the unknown word. To do this, TT first excludes the closed word classes (prepositions, determiners, conjunctions), all of which are listed in the full-form lexicon. To disambiguate the remaining open word classes, TT uses an automatically created suffix lexicon that assigns tag probabilities to words based on their endings, applying a similar pruning strategy as that used for decision trees (Schmid 1995: 4). Table 4 shows Spanish and German words with unknown lemmas and the POS tags assigned to them.

[6] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

[7] The TIGER Corpus, produced by the Institute for Natural Language Processing (University of Stuttgart), consists of about 900,000 tokens (50,000 sentences) of German newspaper text. (http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html).

[8] The Corpus Resources and Terminology Extraction project (CRATER), produced by the Department of Linguistics & Modern English Language, Lancaster University (http://catalog.elra.info/product_info.php?products_id=636), consists of texts of the agency for information and communication technologies (ITU) of the United Nations. The project CRATER 2 "extended the bilingual annotated English-French International Telecommunications Union corpus to include Spanish (…). The offer consists of a multi-lingual aligned corpus of 1,000,000 tokens per language for English, French and Spanish, with morphosyntactic annotations (human-edited)" (http://catalog.elra.info/product_info.php?products_id=84).

Word	Tag	Lemma
desvanecía	VLfin	unknown
Bastián	NP	unknown
wurmartigen	ADJA	unknown
dahinkrochen	VVFIN	unknown

Table 4: Presentation of unknown lemmas

Since TT is based on a training corpus, its accuracy level is influenced by the size and genre of this training set. When TT tags texts from genres other than those that were used for training the model, its performance decreases significantly.

The statistics in Table 5 show the results of running TT on a subcorpus of the PaGeS data.[9] The first column displays the number of tokens of the subcorpus used in the test. The second column shows the proportion of unknown words. The last two columns indicate TreeTagger's accuracy on unknown words and overall.

	Tokens	Unknown words	Accuracy unknown words	Accuracy overall
German	312,522	4%	86.1%	93.2%
Spanish	322,405	7.7%	81.2%	91.3%

Table 5: TreeTagger accuracy on PaGeS data (percentages)

For this test, we did not use any additional training corpus with our data but only the standard parameter files for both languages distributed with TT. Compared with the performance of ca. 97% reported by Schmidt (1995), the results above seem relatively low and the current state of the art in POS tagging.[10]

However, three issues should be taken into account when interpreting these findings. First, the proportion of unknown words is much higher than the values of about 2% reported by Schmid (1995). The percentage of words tagged as unknown rises to 7.7% in the Spanish texts of PaGeS and to 4% in the German texts. As expected, TreeTagger's performance for unknown words is much lower and our results are more similar to those reported by Volk and Schneider (1998). Second, most accuracy figures result from an evaluation of the tagger using the same text type with which it was developed. However, when training and test datasets stem from different domains, as in this case (training data belonged to the newswire or ICT domain and PaGeS data to narrative fiction), the accuracy drops significantly (see Manning 2011). Giesbrecht and Evert (2009) have also pointed this out: "Therefore, the reported tagging accuracies of 97%–98% have to be understood as optimistic estimates, representing an ideal case (…) when training and test data are very similar (usually from the same volume of the same newspaper)". The much lower performance we observed for TT in Spanish is, to a large extent, also due to the fact that the domain of the training corpus for Spanish, information and communication technologies (see footnote 8), is less representative and more distant from the fictional texts (many word forms in the corpus data do not occur in the training text) than the journalistic texts of the German training corpus and, consequently, the results are here much less satisfying.[11] The third factor that may account for the decrease in accuracy is that most of the tests do include punctuation marks in their statistics (Manning 2011). They thus achieve a much higher level of accuracy, since they are always unambiguous. In our test, punctuation marks, representing 19% of all tokens, were not taken into account.

In view of these modest results, it is crucial to increase significantly the performance of the tagger. Tagging accuracies below 90% allow neither for a proper direct use of the corpus nor for further linguistic analysis and erode all subsequent processes, such as word alignment. For improving the results of an HMM-tagger such as TT, there is no way around hand-tagging a representative amount of the corpus data. This task is currently being carried out by two independently working annotators and disagreements

[9] This subcorpus aims to be as representative of the main corpus as possible. For this reason, three works were selected (three German and three Spanish texts), each of them with a different original language: German, Spanish and English (see Table 1).

[10] See Schmid (2008: 541): "The state-of-the-art accuracy in POS tagging is between 95% and 98% depending on the language, the tagset, the size of the training corpus, the coverage of the lexicon, and the similarity between training and test data".

[11] These initial disappointing results of the TreeTagger for Spanish have led us to consider the use of other POS taggers, such as Freeling (http://nlp.lsi.upc.edu/freeling/node/1) or a combination of both of them, by using the output of one tagger to help the other. The final choice will depend on the results obtained once the TreeTagger has been trained on a sufficient amount of data from the PaGeS Corpus.

are reconciled in order to obtain a reliable gold standard. Given that this is a very time-consuming task, at the time of writing we cannot yet offer any preliminary results.

4.2. Harmonization of the tagsets

Traditional part-of-speech descriptions of German usually contain 10 tags, based on Adelung's (1781) "Zehn-Wortarten-Lehre": substantive, verb, adjective, article, conjunction, interjection, numeral, pronoun, preposition and adverb. Spanish traditional word classes are very similar (see Bello 1847).

There is to date no uniform tagset for the tagging of corpora. The number of tags in existing corpora varies depending on language, corpus, level of development and the criteria that are considered. However, all of the most common tagsets contain a number of tags that exceeds the number of traditional word classes. Looking at some of the currently best-known tagsets for English, the situation is as follows: the pioneering Brown Corpus distinguishes 87 simple tags, the Penn Treebank tagset uses 48 tags, the basic tagset of the British National Corpus (CLAWS C5) distinguishes 61 categories and its more detailed tagset (known as C7) 160.

For German, the Stuttgart-Tübingen Tagset (STTS) (Schiller et al. 1999), used by TT, is now the de-facto standard for the tagging of German texts and is used by the vast majority of German POS-tagged resources (Telljohann et al. 2013:1).[12] It distinguishes 54 tags[13] (51 for word classes and 3 for punctuation). On the other hand, the Spanish tagset used by TT[14] comprises 75 tags, of which 52 are word tags.

In what follows, I give a brief overview of both tagsets, describing their basic features and their main differences. Both of them are limited to a determination of the word class without further morphosyntactic (e.g. case, tense, number, etc.) or semantic information.[15]

The STTS-tagset contains 11 core word classes (noun, adjective, numeral, verb, article, pronoun, adverb, conjunction, adposition, particle and interjection). In addition, there are three tags for special cases: foreign words, first member of a compound noun (*Damen- und Herrensalon*) and non-words. The Spanish tagset comprises the same basic word classes and also includes a tag for foreign words and another for acronyms. Within the nouns, a distinction is made between common and proper nouns. While the Spanish tagset has only one tag for adjectives, the STTS divides them into two groups: attributive and adverbial or predicative adjectives. For verbs, three categories are defined: full, modal and auxiliary verbs. The Spanish tagset has specific tags for the auxiliary verbs *haber*, *estar* and *ser.* Within each verb class, the STTS has tags for finite (indicative and subjunctive) verb forms, imperatives, infinitives and participles. The Spanish tagset additionally has a tag for gerunds and the finite verb forms include the imperative forms as well. There are no tags to indicate tense. As for the pronouns, the German tagset is more fine-grained and makes a distinction between the adjectival (*mein Buch*) and the pronominal (*Das Buch ist meins*) uses of pronouns within the different classes of pronouns (personal, demonstrative, indefinite, relative, interrogative). The Spanish tagset subsumes both uses under the same tag. In the German tagset, within the category of prepositions (adpositions) there are tags for prepositions in the strict sense, postpositions (*mir* zuliebe) and circumpositions (*von* Anfang *an*). As for the category of particles, the German tagset divides them into: negation particles, answer particles, the particle *zu* with infinitives and verb particles, whereas the Spanish tagset includes only the negation marker *no* in this group.

As can be seen from this overview, TT applies different tagsets for German and Spanish, tailored to each respective language. Of course, a tagset needs to take account of the specific features of the language with which it is being used. Relevant examples of these features are, for instance, the lack of postpositions, circumpositions and verbal particles in Spanish, and the lack of gerunds or verbal clitics in German. However, these cross-linguistic differences are only partially responsible for the discrepancies between the tagsets. These discrepancies are rather due to specific guidelines for the annotation of each

---

[12] The STTS-tagset was jointly developed in the 90s by the Universities of Stuttgart and Tübingen and is used by the following German corpora, among others: the treeband NeGra TIGER, TüBa-D/S and TüBa-D/Z (Telljohann et al. 2013: 2ff.).

[13] https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/hagen/STTS_Tagset_Tiger (Annotation scheme of Tiger).

[14] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/spanish-tagset.txt. It is a simplified version of the CRATER tagset for Spanish (Sánchez León 1994).

[15] For languages with strong case syncretism like German (a word like *Menschen* can be nominative (pl.), genitive, dative or accusative), to avoid errors in disambiguation, it might be wiser to leave inflectional information to the components at the syntactic level (Seeker and Kuhn 2013: 24). Beyond this consideration, there is a tool developed by Schmid and Laws (2008), RFTagger, that annotates texts with fine-grained part-of-speech tags and can easily be integrated into the STTS-tagset.

corpus and to the specific grammatical tradition of each language. This is why, for instance, the German tagset distinguishes between the pronominal and the attributive uses of pronouns while the Spanish one does not. Moreover, the tags' names are language-dependent and the abbreviations are based on the respective German or Spanish category name.

To overcome these differences in the tagsets of both languages, it was necessary to create a common tagset. As Feldman and Hana (2010: 63) point out, a harmonized tagset makes a transfer of morphological information across languages much easier and allows for a quick and efficient comparison of the properties of both languages. Moreover, the user does not have to struggle with two language specific tagsets with differently named tags. According to Nivre's (2015: 5) principles, in our tagset a single tag is assigned to a common feature in both languages. A third principle we followed was to disregard, for practical reasons, tags for forms that can be easily retrieved with a string search. The Spanish tags PAL and PDEL were therefore omitted, since they tagged only the portmanteau article forms *al* and *del*. The same happens with the tag CQUE, which labels all *que* tokens, ignoring their syntactic function (relative, comparative or subordinating conjunction). Note that the query language of the PaGeS corpus supports a combined search for string and category.

For our purposes, we needed to map the more fine-grained tags of the specific tagsets onto an inventory of slightly more broadly defined tags. Therefore, according to the universal part-of-speech tagset (UPS) proposed by Petrov et al. (2012),[16] we developed a tagset for PaGeS composed of 43 word tags, grouped in 12 main part of speech categories: adjective (ADJ), noun (NOUN 2 tags), numeral (NUM), article (ART), determiner (DET 5 tags), pronoun (PRON 8 tags), verb (V 15 tags), adverb (ADV), adposition (ADP 3 tags), conjunction (CONJ 3 tags), particle (PART 3 tags) and interjection (INTJ). These categories are then subject to varying degrees of further subdivision. A common tagset necessarily involves a reduction in the number of tags, given that both tagsets have to distinguish the same word category. Consequently, we had to merge the two classes of adjectives of the German tagset, because they are not differentiated in the Spanish one. On the other hand, we subsumed the Spanish specific tags for the verbs *estar*, *ser* and *haber* under the category auxiliary verbs, similar to the German tagset. This simplification, ignoring subclasses, was also carried out if the criteria for distinguishing the tags were not applicable or if they were a major source of errors.

As for naming the tags, the UPS convention also served as a guideline for the PaGeS tagset, while we also tried to be as universal and mnemonic as possible in our choice of names. The tag of a word form consists of the name of the core category in uppercase letters followed, when necessary, by the name of the subcategory in lowercase: ADJ, DETpos, PRONint, ADPpos, etc. Table 6 in the Appendix provides a complete mapping between the PaGeS tagset and that of the STTS and the CRATER tagsets.

Considering that for some needs the search can require a more fine-grained tagset than that of PaGeS, the information with the finer distinctions within the specific tagsets has been kept in the query system so that the user can search for them as well.

4.3. Main error patterns

For the review of annotation with POS tags, it is interesting to estimate reliably where and what mistakes or ambiguities are to be expected. We analyzed the results in the test in order to find error patterns to help us to primarily focus on tags in which errors are more likely to occur in our manual post-processing work. The following is a list, by no means exhaustive, of the most common errors we have detected.

(i) Some issues result from inadequate tokenization. The tokenizer does not recognize the inverted question marks (¿) and exclamation marks (¡) used in Spanish to open interrogative and exclamatory clauses, respectively. Consequently, it does not split them from the following word, resulting in an unknown and often mistagged word: ¡*Maldita*/ADJ unknown, ¿*dónde*/ADJ unknown, ¿*No*/ADJ unknown.

(ii) New or unknown words are often a mix of common nouns, proper nouns and foreign words. Many unknown words are proper nouns that do not occur in the training data.

(iii) As stated in section 3, the tokenizer does not concatenate German multi-word adverbs like *nach wie vor* or *ab und zu*. Volk et al. (2016: 300) have already pointed out that TT is likely to mistake adverb usage for prepositions or verb prefix particles, since the adverb usage of these words is far less common. For example, the tags for *nach wie vor* are: nach/APPO, wie/KOKOM, vor/APPR, where the adverbs *nach* and *vor* are wrongly tagged as prepositions. Similarly, in the adverb *ab und zu*, tagged as ab/PTKVZ, und/KON, zu/ADV, the adverb *ab* is erroneously labelled as a verbal particle.

[16] The UPS tagset, developed for multilingual applications, is a revised and extended version of the Google Universal Part-of-Speech Tagset, which in turn is based on a generalization over tagsets in the Conference on Computational Natural Language Learning (CoNLL)-X shared task on multilingual dependency parsing. The UPS only represents the major word classes; in the first version there were 12 and in the new version there are 17.

(iv) In the category of verbs, the finite and infinite verb forms often have the same surface form in German, e.g. *bekommen*, *zerschlagen*, *erhalten* (finite, infinitive or participle). These cases are mostly disambiguated in context, but errors still occur. Participles are tagged as adjectives (*die geöffnet*/ADJD wurde), infinitives are tagged as participles (*das die Bilder zerfallen*/VVPP *lassen könnte*) and imperatives are tagged as finites (*nehmt*/VVFIN *die Axt*). The Spanish verbal forms show less homonymy. Major issues occur between the categories participle, adjective and noun; nouns are sometimes labelled incorrectly as participles: *le habían asignado como abogado*/VLadj.

(v) A further problem is the fact that German particle verbs are tokenized separately if they occur with a verb stem and particle split (e.g. in the sentence *du fängst spät an*) and as a single token if their forms are written together (e.g. *du kannst anfangen*).[17] So in the first case the lemma is *fangen* and in latter case *anfangen*.

(vi) The tagger does not distinguish in Spanish between *que* as conjunction and *que* as relative pronoun, both of them are labeled as conjunction (CQUE). This is an obvious shortcoming of the tagger and this fundamental distinction must be clearly established through manual post-editing.

5. Concluding remarks

An accurate POS tagging is a very useful corpus annotation for many tasks and is the basis for a variety of higher level analyses. In this paper, we discussed the problems that arose while tagging a bilingual corpus of German and Spanish fictional texts. Considering that the corpus is primarily designed for linguistic research and language teaching, it was necessary to develop a common tagset for the annotation of both languages that brings out the similarities and differences between German and Spanish by maximizing parallelism in annotations. Our preliminary evaluation results show that the reported accuracy of TT, where punctuation marks are included, is only achieved under ideal conditions and is contingent upon high quality annotations and extensive training data of the same genre as the corpus. However, in the PaGeS corpus, where the domain—fictional texts—is distant from the domain of the training data, the performance of the tagger drops significantly to levels close to 90%, which is inadequate for most applications. Consequently, manual post-editing is necessary to obtain a POS-tagged PaGeS gold-standard corpus. To minimize this time-consuming and labor-intensive task, two approaches are proposed: producing a reduced tagset sufficient for use in many applications and focusing on common error patterns during manual post-processing work. One open question is how large the training corpus has to be in order to obtain a reliable estimate. Our aim is to achieve a practicable compromise between linguistic usefulness, feasibility and reliability of human tagging and the performance of automatic tagging software.

Acknowledgment

This work has been carried out within the research project PaGeS Corpus, funded by the Spanish Ministry of Economy and Competitiveness (FFI2013-42571-P) and the Galician Government (GI-1954).

References

Adelung, Johann C. 1781. *Auszug aus der deutschen Sprachlehre für Schulen*. Berlin: Voss.

Bello, Andrés. 1847. *Gramática: gramática de la lengua castellana destinada al uso de los americanos*. Santiago de Chile: Imprenta del Progreso.

Borin, Lars ed. 2002. *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University*, Sweden, 22-23 April, 1999. Amsterdam: Rodopi.

Doval, Irene. 2016. PaGeS: design and compilations of a bilingual parallel corpus German Spanish. Compilation of bilingual corpora for linguistic research. In Antonio Moreno Ortiz and Chantal Pérez-Hernández eds. *EPiC Series in Language and Linguistics Volume 1*, *CILC2016. 8th International Conference on Corpus Linguistics*, 88–96.

Doval, Irene, Santiago Fernández Lanza, Tomás Jiménez Juliá, Elsa Liste Lamas and Barbara Lübke. Forthcoming. Corpus PaGeS: a multifunctional resource for language learning, translation and cross-

[17] To tackle this problem, Volk et al. (2016: 298ff) have developed a re-attachment algorithm to compute the lemmas of verbs with separated prefixes, searching after POS tagging for a separated verb prefix (tagged as PTKVZ) and the most recent preceding finite full verb (VVFIN) or imperative verb (VVIMP) in the same sentence.

linguistic research. In Irene Doval and María Teresa Sánchez Nieto eds. *Parallel corpora for contrastive and translation studies: new resources and applications*. Amsterdam: John Benjamins.

Feldman, Anna and Jirka Hana. 2010. *A resource-light approach to morpho-syntactic tagging*. Amsterdam: Rodopi.

Giesbrecht, Eugenie and Stefan Evert. 2009. Part-of-speech tagging – a solved task? An evaluation of POS taggers for the Web as corpus. In Iñaki Alegria, Igor Leturia and Serge Sharoff eds. *Proceedings of the 5th Web as Corpus Workshop (WAC5). San Sebastian, Spain*. http://www.stefan-evert.de/PUB/GiesbrechtEvert2009_Tagging.pdf.

Harris, Brian. 1988. Bi-Text, a new concept in translation theory. *Language Monthly* 54: 8–10.

Jurafsky, Daniel and James H. Martin. 2017. *Speech and language processing. An introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Chapter 10: Part-of-Speech Tagging. https://web.stanford.edu/~jurafsky/slp3/ed3book. pdf (accessed 23 October 2017).

Lee, David Y.W. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5/3: 37–72.

Leech, Geoffrey. 1997. Grammatical tagging. In Roger Garside, Geoffrey Leech and Tony McEnery eds. *Corpus annotation. Linguistic information from computer text corpora*. London: Longman, 19–33.

Manning Christopher D. 2011. Part-of-Speech tagging from 97% to 100%: is it time for some linguistics? In Alexander Gelbukh ed. *Computational Linguistics and Intelligent Text Processing. CICLing 2011*. Berlin: Springer, 171–189.

McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press.

Nivre, Joakim. 2015. Towards a Universal Grammar for Natural Language Processing. In Alexander Gelbukh ed. *Computational Linguistics and Intelligent Text Processing. CICLing 2015*. Berlin: Springer, 3–16.

Petrov, Slav, Dipanjan Das and Ryan McDonald. 2012. A universal Part-of-Speech tagset. In *Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, 2089–2096.

Sánchez León, Fernando. 1994. *Spanish tagset for the CRATER project*. https://arxiv.org/pdf/cmp-lg/9406023.pdf (accessed 12 October 2017).

Schiller, Anne, Simone Teufel and Christine Stöckert. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS*. Universitäten Stuttgart und Tübingen. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf (accessed 13 September 2017).

Schmid, Helmut. 1994. Probabilistic Part-of-Speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing, Manchester*. http://www.cis.uni-muenchen.de/~schmid/tools /TreeTagger/data/tree-tagger1.pdf (accessed 13 September 2017).

Schmid, Helmut. 1995. Improvements in Part-of-Speech tagging with an application to German. In *Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland*. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (accessed 15 October 2017).

Schmid, Helmut. 2008. Tokenizing and part-of-speech tagging. In Anke Lüdeling and Merja Kytö eds. *Corpus linguistics. An international handbook. Volume 1*. Berlin: Walter de Gruyter, 527–551.

Schmid, Helmut and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Manchester*, 777–784.

Seeker, Wolfgang and Jonas Kuhn. 2013. Morphological and syntactic case in statistical dependency parsing. *Computational Linguistics* 39/1: 23–55.

Telljohann, Heike, Yannick Versley, Kathrin Beck, Erhard Hinrichs and Thomas Zastrow. 2013. STTS als Part-of-Speech-Tagset in Tübinger Baumbanken. *Journal for Language Technology and Computational Linguistics* 28/1: 1–16.

Tiedemann, Jörg. 2011. *Bitext alignment*. Toronto: Morgan & Claypool.

Volk, Martin, Simon Clematide, Johannes Graen and Phillip Ströbel. 2016. Bi-particle adverbs, PoS-tagging and the recognition of German separable prefix verbs. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 296–305.

Volk, Martin and Gerold Schneider. 1998. Comparing a statistical and a rule-based tagger for German. https://arxiv.org/pdf/cs/9811016.pdf (accessed 10 October 2017).

Corresponding author
Department of English and German Philology
Facultade de Filoloxía
Campus Norte · Avda. de Castelao, s/n
15782 Santiago de Compostela
e-mail: i.doval@usc.es

received: November 2017
accepted: December 2017

Appendix: The tagset mapping

Word class	Tag TT-DE	Tag PaGeS	Tag TT-ES
Attributive, adverbial or predicative adjective	ADJA ADJD	ADJ	ADJ
Common noun	NN	NOUNcom	NP
Proper noun	NE	NOUNprop	NC NMEA NMON
Foreign word	FM	FW	PE
Cardinal number	CARD	NUM	CARD
Definite or indefinite article	ART	ART	ART
Demonstrative determiner	PDAT	DETdem	DM
Indefinite determiner	PIAT	DETindef	QU
Adding interrogative determiner	PWAT	DETint	INT
Possessive determiner	PPOSAT	DETpos	PPO
Relative determiner	PRELAT	DETrel	REL
Pronominal adverb	PAV	PRONadv	---
Demonstrative pronoun	PDS	PRONdem	DM
	PIDAT PIS	PRONindef	QU
Adverbial interrogative or relative pronoun	PWAV PWS	PRONint	INT
Interrogative pronoun			
Personal pronoun	PPER	PRONpers	PPX PPC
Possessive pronoun	PPOSS	PRONpos	PPO
Reflexive personal pronoun	PRF	PRONrefl	SE
Relative pronoun	PRELS	PRONrel	REL
Auxiliary finite verb	VAFIN VAIMP	VAUXfin	VEfin VSfin VHfin
Auxiliary verb: gerund	---	VAUXger	VEger VHger VSger
Auxiliary verb: infinitive	VAINF	VAUXinf	VEinf VHinf VSinf
Auxiliary verb: participle	VAPP	VAUXpart	Veadj Vsadj Vhadj
Verb with clitic: finite	---	VCLIfin	VCLIfin
Verb with clitic: gerund	---	VCLIger	VCLIger
Verb with clitic: infinitive	---	VCLIinf	VCLIinf
Lexical verb: finite	VVFIN VVIMP	VLEXfin	VLfin
Lexical verb: gerund	---	VLEXger	VLger
Lexical verb: infinitive	VVINF VVIZU	VLEXinf	VLinf
Lexical verb: participle	VVPP	VLEXpart	VLadj
Modal verb: finite	VMFIN	VMODfin	Vmfin
Modal verb: gerund	---	VMODger	VMger
Modal verb: infinitive	VMINF	VMODinf	VMinf
Modal verb: participle	VMPP	VMODpart	VMadj
Adverb	ADV	ADV	ADV
Preposition with article folded in	APPRART	ADPart	PAL PDEL
Postposition, right part of circumposition	APPO APZR	ADPpos	----
Preposition	APPR	ADPpre	PREP
Comparative conjunction	KOKOM	CONJcomp	CSUBX
Coordinating conjunction	KON	CONJcoor	CC CCAD CCNEG
Subordinating conjunction	KOUI KOUS	CONJsub	CSUBI CSUBF CQUE CSUBX
Particle with adverb or adjective	PTKA PTKANT	PART	ADV
Answer and negation particle	PTKNEG		NEG
Particle, part of separable verb, *zu*+infinitive	PTKVZ PTKZU		---
Interjection	ITJ	INTJ	ITJN

Table 6: Mapping between lexical descriptions and the tagsets