Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 29-36 recibido 10-11-14 revisado 27-01-15 aceptado 12-02-15

Choosing a Spanish Part-of-Speech tagger for a lexically sensitive task*

Selección de un etiquetador morfosintáctico primando la precisión en las categorías léxicas

Carla Parra Escartín
University of Bergen
Bergen, Norway
carla.parra@uib.no

Héctor Martínez Alonso
University of Copenhagen
Copenhagen, Dennmark
alonso@hum.ku.dk

Resumen: En este artículo se comparan cuatro etiquetadores morfosintácticos para el español. La evaluación se ha realizado sin entrenamiento ni adaptación previa de los etiquetadores. Para poder realizar la comparación, los etiquetarios se han convertido al etiquetario universal (Petrov, Das, and McDonald, 2012). También se han comparado los etiquetadores en cuanto a la información que facilitan y cómo tratan características intrínsecas del idioma español como los clíticos verbales y las contracciones.
Palabras clave: Etiquetadores morfosintácticos, evaluación de herramientas, lingüística de corpus

Abstract: In this article, four Part-of-Speech (PoS) taggers for Spanish are compared. The evaluation has been carried out without prior training or tuning of the PoS taggers. To allow for a comparison across PoS taggers, their tagsets have been mapped to the universal PoS tagset (Petrov, Das, and McDonald, 2012). The PoS taggers have also been compared as regards the information they provide and how they treat special features of the Spanish language such as verbal clitics and portmanteaux.
Keywords: Part-of-Speech taggers, tool evaluation, corpus linguistics

1 Introduction

Part-of-Speech (PoS) taggers are among the most commonly used tools for the annotation of language resources. They are often a key preprocessing step in many Natural Language Processing (NLP) systems. When using a PoS tagger in a workflow, it is important to know the impact of its error rate on any modules that use its output (Manning, 2011).

In this paper, we compare four different PoS taggers for Spanish. Our goal is to build an NLP system for knowledge extraction from technical text, and this requires very good performance on lemmatisation and right PoS assignment. Inflectional information is not relevant for our purposes. Instead of choosing a PoS system based solely on its reported performance, we benchmark the output of the 4 candidate systems against a series of metrics that profile their behaviour when predicting lexical and overall PoS tags, as well as the final quality of the resulting lemmatisation.

As the taggers had different tagsets, and we were only interested in retrieving the coarse PoS tag (*buys*_verb vs. *buys*_verb_3ps), we have mapped the tagsets to the universal PoS tagset proposed by Petrov, Das, and McDonald (2012).

The remainder of this paper is organised as follows: Section 2 discusses available PoS taggers for Spanish and describes them briefly. In Section 3 the different challenges encountered when comparing the four PoS taggers are discussed. Section 4 discusses the differences across the tagsets used by each PoS tagger and Section 5 shows the evaluation of each PoS tagger compared with the other three. Section 6 offers a summary.

2 Part-of-Speech taggers available for Spanish

For Spanish, several PoS tagger initiatives have been reported (Moreno and Goñi, 1995; Márquez, Padró, and Rodríguez, 2000; Car-

* The authors thank the anonymous reviewers for their valuable comments and the developers of the taggers we have covered in this article for making them available.

ISSN 1135-5948

reras et al., 2004; Padró and Stanilovsky, 2012, i.a.). The reported accuracies for these PoS taggers are reported to be 96–98%.

However, not all of these tools were available for our tests. One of the existing PoS taggers, the GRAMPAL PoS tagger (Moreno and Goñi, 1995) was not included in this comparison because it is not downloadable and it does not seem accessible via a web service[1].

To our knowledge, four PoS taggers for Spanish are the TreeTagger (*TT*) (Schmid, 1994), the IULA TreeTagger (*IULA*) (Martínez, Vivaldi, and Villegas, 2010), the FreeLing PoS tagger (*FL*) (Padró and Stanilovsky, 2012), and the IXA PoS tagger (*IXA*) (Agerri, Bermudez, and Rigau, 2014). Two of them (IULA and FL) are also available as web services developed during the PANACEA project. The fourth PoS tagger was recently released within the IXA pipes. The study reported in this paper compares these four PoS taggers.

2.1 Default TreeTagger TT

TT provides 22 already trained models for 18 languages and new models can be created by retraining the tagger with new data.

We used the already trained model for Spanish which is available on its website. This model was trained on the Spanish CRATER corpus and uses the Spanish lexicon of the CALLHOME corpus of the LDC.

Prior to tagging the text, the tool tokenises it. The tokeniser does not have specific rules for Spanish.

In the output of this tagger, every line contains one wordform with its PoS tag and its lemma (i.e. citation form). TT PoS tags do not contain inflectional information (e.g. tense, number, gender, etc.). This tagset is the most similar to the universal PoS tagset.

When the tagger fails to assign a lemma to a specific wordform, the value for the lemma is *<unknown>*. Nevertheless, a PoS tag is assigned to an unknown word. Examples 1-3 show words with unknown lemmas and their assigned PoS tags.

(1) NÖ NC <unknown>

(2) WFG NP <unknown>

(3) plurifamiliares ADJ <unknown>

TT concatenates Multiword Expressions (MWEs). Their wordforms are listed with whitespaces as they occur in the text, while their lemmas are joined by means of tildes. Examples 4 and 5 show MWEs tagged by TT.

(4) de conformidad con PREP de∼conformidad∼con

(5) junto con PREP junto∼con

2.2 IULA TreeTagger (IULA)

The IULA is an instance of the TT trained on the IULA technical corpus (IULACT). Additionally, each file undergoes a preprocessing step prior to tagging. This preprocessing step is described in detail in Martínez, Vivaldi, and Villegas (2010). It comprises the following tasks:

1. Sentence-boundary detection;
2. General structure-marking;
3. Non-analyzable element recognition;
4. Phrase and loanword recognition;
5. Date recognition;
6. Number recognition;
7. Named Entity (NE) recognition.

These tasks were introduced in what Martínez, Vivaldi, and Villegas (2010) call a *text handling* module. This module was developed in order to solve potential sources of errors prior to tagging with the aim of improving the overall quality of the PoS tagging process. The whole toolset is available through a web service where one can upload the corpus to be tagged and download the tagged corpus upon completion of the task.

Unlike the TT instance discussed in Subsection 2.1, the IULA PoS tagset provides inflectional information for the relevant PoS. The tagset is partially based on the EAGLES tagset for Spanish, and includes more fine-grained information.

When the tagger fails to assign a lemma to a specific wordform, instead of assigning to it the value *<unknown>* as TT does, IULA assigns wordforms as lemmas. Example 6 shows the tagging of the unknown word *plurifamiliares*. *Plurifamiliares* is in plural, and its lemma should be the singular wordform *plurifamiliar* but instead, the plural wordform is used.

(6) plurifamiliares JQ–6P plurifamiliares

Special elements such as MWEs are treated in a different way. The previous MWE examples 4 and 5 do not appear concatenated, but are tagged as separate words. MWEs such as dates or names are lemmatised with underscores. Examples 7-9 show some

[1] There is an available online demo limited to 5000 words.

tagged MWEs. Furthermore, as a result of the preprocessing step, the IULA adds additional xml-style tags to such elements.

(7) 18 de diciembre del 2001 T 18_de_diciembre_del_2001

(8) Baja Austria N4666 Baja_Austria

(9) Promoción MH-NEU N4666 Promoción_MH-NEU

2.3 FreeLing (FL)

FreeLing is an open source suite of language analysers. It offers a wide variety of services for several languages. Among these analysers are tokenisers, sentence splitters, morphological analysers, PoS taggers, etc. The PoS tagger has two different flavours, a hybrid tagger called *relax*, which combines statistical and manually crafted grammatical rules, and a model based on the Hidden Markov Model (HMM) similar to the TnT proposed by Brants (2000). In both cases, the tagger was trained with the *LexEsp* corpus (Sebastián, Martí, and Carreiras, 2000). Again, the web service offered by the PANACEA project was used. Since in the web service only the HMM model was deployed, this is the tagging model we have used in this paper.

FL displays first the wordform, followed by the lemma and then the PoS tag. It uses the EAGLES tagset for Spanish, which, similarly to the IULA tagset, also includes more fine-grained inflectional information.

Whenever FL analyses a sequence as a MWE, it displays both the wordform and the lemma joined with underscores. Examples 10 and 11 show some tagged MWEs.

(10) Baja_Austria baja_austria NP00G00

(11) Promoción_MH-NEU promoción_mh-neu NP00V00

Another peculiarity is that all lemmas are lowercased regardless whether they correspond to a proper noun, an abbreviation or something else. This can be observed in Example 10, where the lemma for the proper noun *Baja Austria* ([EN]: Lower Austria) is lowercased to *baja_austria*.

Finally, dates are also treated differently in FL. Their wordform is the date itself joined by underscores, and their "lemma" is the same date converted to a numerical format, where month names are converted to their corresponding number. Examples 12 and 13 show this.

(12) 18_de_diciembre_del_2001 [??:18/12/2001:??.??:??] W

(13) 15_de_octubre_del_2002 [??:15/10/2002:??.??:??] W

2.4 IXA pipes (IXA)

The IXA pipes are "ready to use NLP tools" (Agerri, Bermudez, and Rigau, 2014) developed by the IXA NLP Group. Among the available tools there are a tokeniser and a PoS tagger for Spanish. The PoS tagger requires not only that the text is previously tokenised, but also, that it is in the NLP Annotation Format (NAF) (Fokkens et al., 2014). These requirements are met by using the provided tokeniser and piping it to the PoS tagger.

IXA has been trained and evaluated using the Ancora corpus (Taulé, Martí, and Recasens, 2008). Two PoS tagging models are available: one based on the Perceptron algorithm (Collins, 2002), and another one based on Maximum Entropy (Ratnaparkhi, 1999). We use the default model for Spanish, which is the Perceptron.

Its output format is xml-based and thereby differs from the taggers previously discussed in this paper. The resulting document is tagged in NAF and has a header specifying the tools that have been used, when they were used, how long the process has taken and the version of the tool used. Next, the tokenised text appears.

For each tokenised wordform, the tool provides its NAF required attributes for word *id* and the sentence where it appears (*sent*), as well as the optional attributes *para*, *offset*, and *length*, which correspondingly refer to the paragraph the word belongs to, the offset in number of characters and its length in number of characters.

Where the tokenised text ends, a new section starts with PoS and lemma information about each word as identified by its *id*. For each term, the following attributes are provided:

1. *id*: The term id. This is the only required attribute, all the other attributes are optional.
2. *type*: Whether the term belongs to an open PoS (e.g. nouns), or a closed one (e.g. prepositions).
3. *lemma*: The wordform lemma.
4. *pos*: The Part of Speech of the wordform.
5. *morphofeat*: the PoS tag assigned to the form, containing inflectional information. IXA uses the same tagset as FL:

the EAGLES tagset for Spanish.

MWEs are signalled by the sub-element *...*. When the PoS tagger identifies a MWE, the sub-element ** will have several *<target>* subelements. *<target>* subelements refer to the wordform ids assigned in the text part of the document. In our test, IXA failed to identify any MWE.

3 Challenges

Comparing and evaluating four different PoS taggers is not a straightforward task. Differences in their tagsets, output formats and tokenisation processes have to be addressed. Prior to the PoS tagging process, the text has to be tokenised. As pointed out by Dridan and Oepen (2012), tokenisation is often regarded as a solved problem in NLP. However, the conventions used in the tokenisation task have a direct impact in the systems which subsequently use the tokenised text. In recent years, several authors have highlighted the importance of tokenisation and researched new tokenisation strategies (Dridan and Oepen, 2012; Fares, Oepen, and Zhang, 2013; Orosz, Novák, and Prószéky, 2013, i.a.).

In our study, each PoS tagger had its own tokeniser either as an integrated component (TT, IULA and FL), or available as a separate tool to be piped to the PoS tagger (IXA). Each tagger subsequently tagged this text, following its internal tokenisation.

At this stage, there are also two particular features of the Spanish language that may be handled differently by the tokenisers and/or the PoS taggers:

- The portmanteau (contracted) wordforms *al* and *del* (cf. 3.1);
- Verbal clitics, which are attached to verbs without hyphenation (cf. 3.2).

Finally, an additional challenge is the way in which each tagger detects and tags MWEs, such as named entities, proper names, dates, and complex prepositions.

3.1 Portmanteaux in Spanish

To a certain extent, the portmanteaux *al* and *del* are not difficult to tackle. They are the result of contracting the Spanish prepositions *a* and *de* with the determined masculine singular article *el*. Thus, *a* + *el* results in *al* and *de* + *el* results in *del*. Additionally, *al* can also be used to introduce subordinated infinite clauses in Spanish (eg. *al pasar*, [EN]: when/while passing). Each PoS tagger however, handles this phenomenon differently.

(a) **TT**: TT has a special tag for each of these wordforms: *PAL* for *al* and *PDEL* for *del*. TT treats the subordinated conjunction reading using a third tag: *CSUBI*. Thus, TT does not split these wordforms and handles them by using specific tags.

(b) **IULA**: The IULA assigns to these wordforms a double tag *P_AMS*, thus providing information from each of the components but joining the tags with an underscore. The lemmas are assigned correspondingly, retrieving the preposition and the article as separate items but joining these lemmas with an underscore: *a_el* and *de_el*.

(c) **FL**: FL retrieves the preposition and the article undoing the contraction completely. Thus, *al* and *del* become *a el* and *de el* and each word is analysed and tagged separately.

(d) **IXA**: IXA uses a strategy similar to that of the TT and uses one special tag *SPCMS* for contracted prepositions available in EAGLES for both *al* and *del*.

Each PoS tagger takes a different stance on this phenomenon. Two taggers tag the contracted prepositions with special tags (TT and IXA). The other two treat them as separate words and retrieve the underlying non-contracted wordforms (IULA and FL), although they represent them differently.

3.2 Verbal clitics

Verbal clitics are morphemes with the syntactic characteristics of a word which depend phonologically on another word. In Spanish, clitics are attached directly to the verb without hyphenation and fulfil a pronominal role. Moreover, it is possible to have two clitic pronouns, one referring to the direct object and the other to the indirect object. As a result of this agglutinative process, some wordforms will require an acute accent to comply with the Spanish orthographic rules. For instance, the wordform *enviármelo* ([EN]: 'send it to me') is composed of the verb *enviar* ([EN]: 'send') to which the clitics *me* ([EN]: 'me') and *lo* ([EN]: 'it') are attached.

Clitic handling is a challenge for PoS taggers. As Martínez, Vivaldi, and Villegas (2010) point out, "the appearance of an acute accent in some wordforms makes a brute-

force stemming difficult". They account for 32 wordforms with pronominal clitics for an infinite verb like *dar* ([EN]: 'give') and explain that as per now the verbal wordforms with clitics are kept in the lexicon of their application to determine if they belong to the language or not. In the four PoS taggers compared in the present paper, different strategies are used.

(a) **TT**: TT assigns special tags to verbs in which clitics are incorporated. It has three different tags available, depending on the verb wordform to which the clitics are attached: *VCLIinf* (infinitive), *VCLIger* (gerund), and *VCLIfin* (finite wordform). For instance, *utilizarla* ([EN]: use it) is assigned the tag *VCLIinf*.

(b) **IULA**: The IULA uses a different strategy. It separates the verb from the clitic, and analyses them as two different words. However, both the wordform and the lemma have additional information attached to them by means of underscores.

(c) **FL**: Like the IULA FL separates the verb from the clitic and analyses them separately. However, no additional marking is used to indicate that the verb and the clitic are one word.

(d) **IXA**: IXA ignores clitics and assigns to verbs with clitics the same tag that the verb would have without the clitic. Thus, *enviármelo* is assigned the tag *VMN0000*, which corresponds to the infinitival form of a main verb.

In conclusion, verbs with clitics are handled in different ways by the different taggers under investigation. While IXA seems to ignore the clitics, all the other taggers handle them differently. The TT has its own tag for this kind of phenomena, FL splits the verb and the clitic, and the IULA uses a mixed approach by splitting the verb but adding additional information to the wordform and the lemma.

4 *Tagset comparison*

The use of different tagsets, together with the other challenges previously discussed in section 3 makes the comparison of PoS taggers a challenging task. Of the four PoS taggers which we have investigated, only FL and IXA share the same tagset, while the others use different tagsets. Each tagger does not only provide a different granularity of categorial distinctions and features, but also treats intrinsic linguistic phenomena differently.

PoS Tagger	Tagset	Tags	Granularity
TT	TT ES	77	low
IULA TT	IULA tagset	435	high
FL	EAGLES	577	high
IXA	EAGLES	577	high

Table 1: PoS taggers tagset comparison.

Table 1 shows the differences across the tagsets used by each of the PoS taggers. A full comparison of the output of four different tools with such big differences regarding the number of tags and the information encoded in such tags could be a very tedious and inaccurate task. We were only interested in retrieving the coarse PoS, the best approach to map the tagsets seemed to be to map each tagset to the universal PoS tagset. Petrov, Das, and McDonald (2012) distinguish twelve PoS tags: "NOUN (nouns), VERB (verbs), ADJ (adjectives), ADV (adverbs), PRON (pronouns), DET (determiners and articles), ADP (prepositions and postpositions), NUM (numerals), CONJ (conjunctions), PRT (particles), '.' (punctuation marks) and X (a catch-all for other categories such as abbreviations or foreign words".

We have developed a mapping from each Spanish tagset to the universal PoS tagset[2]. When projecting to the universal PoS tags we loose the inflectional information. Furthermore, past participles have been mapped to the PoS *VERB*, regardless of whether they function as part of a periphrastic verb tense or function as modifiers or predicates in the same way as adjectives. The adjective-participle ambiguity is addressed differently in the annotation guidelines of all training data, and the behaviour of the tagger in this aspect is a consequence of the data it is trained on. This simple approach was chosen in order to avoid manual revision. A similar approach was taken in other cases, such as the verbs with attached clitics and the portmanteaux when these had not been previously preprocessed and split. In these cases, the categories *VERB* and *ADP*, respectively, were used as defaults.

[2]The mappings are available at the Universal-Part-of-Speech-Tags repository.

5 POS tagger performance

Bearing in mind all the discrepancies described in sections 3 and 4, a completely fair comparison of all the PoS taggers against a unique Gold Standard (GS) is not feasible. A measure like accuracy requires the input data to be linearly comparable—i.e. be tokenized using the same convention—and the output data to rely on the same number of labels—i.e. the datasets should be the same.

The ideal downstream evaluation method would be parsing, but it is too sensitive to tokenisation changes to be applicable when comparing taggers that tokenize differently, besides using different tagsets.

Nevertheless, the four different systems rely on the same sentence boundaries, and are thus sentence-aligned. Given the aforementioned limitations, we use a series of metrics that aim to assess the PoS tagging and lemmatisation performance of the systems by measuring matches—i.e. set intersections—at the sentence level.

In addition, we use metrics that assess the similarity between tag distributions (KL divergence) in order to assess whether the bias of each tagger is closer to the desired performance determined by a reference GS, and metrics that evaluate how many of the predicted lemma+PoS appear Spanish wordnet. The wordnet check serves as a downstream evaluation of the joint predicted values and is one of the instrumental parts of the knowledge-extraction system mentioned in Section 1.

Table 2 summarises the main features of each tagger. Only TT fails to provide inflectional information. Portmanteaux, verbal clitics and reflexive verbs are treated in different ways. While IULA and FL split them, TT and IXA do not. Finally, the tagsets differ also greatly in terms of their overall tag number and the number of non-lexical tags. IULA offers the greatest absolute number and greatest proportion of tags dedicated to non-lexical elements.

The discrepancies between the different tagsets can be addressed by mapping them to the universal PoS tagset. However, then we are losing some of the inflectional analyses produced by some of the taggers. Furthermore, a comparison against one Gold Standard might be biased against one of the various choices of handling MWEs and other special features.

	TT	IULA	FL	IXA
Morphosyntactic info	-	✓	✓	✓
Splits portmanteaux	-	✓	✓	-
Splits verbal clitics	-	✓	✓	-
Joins dates	-	✓	✓	-
Reflexive verb lemmatisation	-	✓	✓	-
Tagset size	77	435	577	
Number of non-lexical tags	42	241	173	

Table 2: PoS taggers features.

In order to allow a comparison, we created a GS following a procedure that attempts to minimise the starting bias in favour of a certain system. We chose two short texts from the freely available technical corpus TRIS (Parra, 2012). We processed this material with FL, because this tagger tokenizes most aggressively (cf. 2.1-2.4). MWEs detected by FL where manually split and tagged, aiming at facilitating the evaluation of this GS against the outputs of the other PoS taggers. Then, we converted the FL tagset to the universal PoS tagset (Petrov, Das, and McDonald, 2012) and manually corrected the tagged text. Each tagger was then compared against this GS.

Table 3 summarises the results of the automatic evaluation we carried out for all taggers against the GS. Three of the metrics have two variants; in an *overall* metric ($_o$), we compare the performance across all parts of speech, whereas in a *lexical* metric ($_l$), we only take into consideration the PoS that are tagged as ADJ, NOUN or VERB after projecting onto Petrov's tagset.

The metrics are the following:

- *Matching lemmas ($_o$/$_l$)*: Proportion of lemmas that match with the GS;
- *Matching PoS ($_o$/$_l$)*: Proportion of PoS tags that match. A match is defined as the minimum count of a given PoS tag matching that in the GS;
- *KL total ($_o$/$_l$)*: Kullback-Leibler divergence (Kullback and Leibler, 1951) between the PoS distribution of the system and that of the GS;
- *GS token ratio*: The relation between the amount of tokens and that of the GS. As explained earlier, we have chosen a GS with the most fragmentary tokenisation, so the ratio will always be equal or less to one. The higher the number, the most similar the tokenisation convention of the system is to the GS tokenisation.

- *WN GS matches*: Number of Spanish wordnet (Gonzalez-Agirre, Laparra, and Rigau, 2012) hits for all predicted lemma-PoS combinations in the GS;
- *WN sys matches*: Number of Spanish wordnet hits for all predicted lemma-PoS combinations matching those of the GS; and
- *WN intersection*: Number of Spanish wordnet hits that also appear in the GS.

	TT	IXA	IULA	FL
Matching lemmas$_o$	0.77	0.85	0.7	**0.89**
Matching lemmas$_l$	0.63	0.63	0.66	**0.78**
Matching PoS$_o$	0.86	0.87	0.85	**0.88**
Matching PoS$_l$	0.93	0.93	**0.94**	0.92
KL_o	**0.041**	0.053	0.042	0.063
KL_l	0.0029	0.0095	**0.0005**	0.001
GS token ratio	0.97	**0.99**	0.94	0.93
WN GS matches	402	402	402	402
WN sys matched	378	367	**386**	384
WN intersection	361	350	366	**373**

Table 3: PoS taggers evaluation.

As Table 3 illustrates, FL has the highest proportion of both overall and lexical matching lemmas (cf. *Matching lemmas o* and *Matching lemmas*$_l$ in Table 3). Its highest proportion of lexical matches is remarkable, as the lemmatisation of lexical words is more important (and error-prone) than that of closed grammatical classes such as articles and prepositions.

As previously stated, FL is also the system which uses a more fragmented tokenisation, and this makes it the most similar to the GS in that respect. However, it gets the lowest *GS token ratio* score. This might be because FL joins the lemmas of MWEs with underscores (cf. 2.3). IXA is the tagger which achieves a better score as regards the ratio of tokens in the GS (cf. *GS token ratio* in Table 3). This may be due to the fact that we split all MWEs in our GS. As mentioned earlier, IXA failed at identifying and tagging them in our test files.

KL divergence offers a different perspective on the different systems. The lower the KL divergence between PoS distribution is, the more similar to the GS is the expected prior for a certain PoS. Interestingly, FL has the highest overall KL divergence in spite of it having the highest performance on lemma retrieval, and the second lowest lexical KL divergence. This difference is due to FL having different conventions on the way that some function words are annotated and thus later converted to the universal PoS tags.

With regard to the overall results, FL has the highest PoS overlap (cf. *Matching PoS*$_o$ in Table 3) with the GS, followed by IXA by a close second. The better accuracy on lemma retrieval, paired with the lowest KL divergence on lexical PoS is also reflected in the highest wordnet hit count for FL (cf. *WN intersection* in Table 3). However, the IULA had a better performance when tagging lexical words (cf. *Matching PoS*$_l$ in Table 3). In fact, it manages to correctly tag lexical words, despite not necessarily achieving to lemmatise them correctly. This explains also, why it has the highest WordNet hit count of lemma-PoS combinations (cf. *WN sys matched* in Table 3).

Since our current research focuses on lexical words, the most important measures for our purposes are *Matching PoS*$_l$ and *Matching lemmas*$_l$. IULA may be our best choice when focusing on the assignment of right PoS tags. It performs better than the TT system which we have previously been using. FL, on the other hand, seems to be better for general lemmatisation tasks, regardless of the PoS tag.

6 Conclusions and future work

We have compared four PoS taggers for Spanish and evaluated their outputs against a common GS. For our purposes, we concluded that IULA would be the best choice, followed by FL. It is possible that a combined output of these two taggers may outperform each single tagger. Given the difficulties of combining different tokenisations in a voting scheme for PoS tagging, we leave this question for future work.

Evaluating on a technical text makes the task more difficult for most of the PoS taggers. The better performance of IULA may also be due it being the only tagger trained on a technical corpus. It is not impossible that for more general (i.e. less technical) texts the differences across the different taggers may be smaller.

We have also proposed a method to compare the output of different PoS taggers by mapping their respective tagsets to the universal PoS tagset and evaluating matches at the sentence level. As indicated earlier in Section 1, for our particular purposes no inflectional information was needed, and thus this method was enough. In case a more fine-grained tagging is needed, the universal PoS

tagset may need to be expanded.

This non-linear evaluation method is useful for tasks that depend on lemmatisation and coarse PoS tagging. Nevertheless, the method would have to be expanded for tasks that require inflectional information.

References

Agerri, R., J. Bermudez, and G. Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *LREC'14*. ELRA.

Brants, T. 2000. TnT: A Statistical Part-of-speech Tagger. In *ANLP'00*, pages 224–231, Sheattle, Washington.

Carreras, X., I. Chao, L. Padró, and M. Padró. 2004. Freeling: An Open-Source Suite of Language Analyzers. In *LREC'04*. ELRA.

Collins, M. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP'02*, volume 10, pages 1–8. ACL.

Dridan, R. and S. Oepen. 2012. Tokenization: Returning to a Long Solved Problem a Survey, Contrastive experiment, Recommendations, and Toolkit. In *ACL'12*, volume 2, pages 378–382. ACL.

Fares, M., S. Oepen, and Y. Zhang. 2013. Machine Learning for High-Quality Tokenization Replicating Variable Tokenization Schemes. In *Computational Linguistics and Intelligent Text Processing*, volume 7816. Springer Berlin Heidelberg, pages 231–244.

Fokkens, A., A. Soroa, Z. Beloki, N. Ockeloen, G. Rigau, W. R. van Hage, and P. Vossen. 2014. NAF and GAF: Linking Linguistic Annotations. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 9–17.

Gonzalez-Agirre, A., E. Laparra, and G. Rigau. 2012. Multilingual Central Repository version 3.0. In *LREC'12*, pages 2525–2529. ELRA.

Kullback, S. and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Manning, C. D. 2011. Part-of-speech Tagging from 97Linguistics? In *CICLing'11*, pages 171–189. Springer-Verlag.

Márquez, L., L. Padró, and H. Rodríguez. 2000. A machine learning approach to POS tagging. *Machine Learning*, (39):59–91.

Martínez, H., J. Vivaldi, and M. Villegas. 2010. Text handling as a web service for the IULA processing pipeline. In *LREC'10*, pages 22–29. ELRA.

Moreno, A. and J. M. Goñi. 1995. GRAMPAL: A Morphological Processor for Spanish implemented in Prolog. In *GULP-PRODE'95*.

Orosz, G., A. Novák, and G. Prószéky. 2013. Hybrid Text Segmentation for Hungarian Clinical Records. In *MICAI (1)*, volume 8265 of *Lecture Notes in Computer Science*, pages 306–317. Springer-Verlag.

Padró, L. and E. Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *LREC'12*. ELRA.

Parra, C. 2012. Design and compilation of a specialized Spanish-German parallel corpus. In *LREC'12*, pages 2199–2206. ELRA.

Petrov, S., D. Das, and R. McDonald. 2012. A Universal Part-of-Speech Tagset. In *LREC'12*. ELRA.

Ratnaparkhi, A. 1999. Learning to Parse Natural Language with Maximum Entropy Models. *Machine Learning*, 34(1–3):151–175.

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*, pages 44–49.

Sebastián, N., M. A. Martí, and M. F. Carreiras. 2000. *Léxico informatizado del español*. Edicions Universitat de Barcelona.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *LREC'08*. ELRA.