SPOST: a Spanish Part-of-Speech Tagger

David Farwell, Stephen Helmreich, Mark Casper

Computing Research Laboratory
Box 30001/3CRL
New Mexico State University
Las Cruces, NM 88003

Tel: 505 646 5108
Fax: 505 646 6218

Abstract

This paper presents the structure, function and performance of SPOST, a broad coverage, rule-based part-of-speech tagger developed as a component of the Panglyzer Spanish Analysis Component of the Pangloss Knowledge-based Machine Translation System. It also describes preparations for the general access of SPOST by the research community in the fall of 1995.

Key Words: Spanish Part-of-Speech Tagging, NLP Development Tools, Broad Coverage Grammars, Machine Translation

The goal of this paper is to describe the structure and function of SPOST, a Spanish part-of-speech tagger, which has been developed at the Computing Research Laboratory (CRL) at New Mexico State University. In addition, we wish to announce its pending availability for general use by the research community via an e-mail List Server that is being set up at the CRL. After a brief summary of the context in which SPOST was developed, we describe the architecture and processing of the component, present suggestive results of its current performance, explain the set up for using it via e-mail and conclude with a description of our current activities and future plans.

While part-of-speech (POS) taggers have a relatively long history (e.g., Klein & Simmons, 1963; Greene & Rubin, 1971; Leech, Garside & Atwell, 1983), recent work has shown that, at least for English, this task can be performed with a high rate of accuracy. This recent work (e.g., DeRose, 1988; Brill, 1992; Schütze & Singer, 1994) relies heavily on large amounts of hand-tagged texts to provide training data for tagger development. This has hampered the development of POS taggers based on similar methods for many other languages such as Spanish. SPOST, however, relies heavily on a large lexicon derived from an on-line version of *Collins Spanish-English English-Spanish Dictionary* (1992) and a rule-based tagging algorithm to produce its results. As with Brill (1992), SPOST exploit a multipass algorithm which first assigns POS tags on the basis of

...tionary look-up and then applies fix-up rules to rewrite predictable errors during the ...tial pass. While SPOST is not as accurate as the best taggers of English, its perfor-...ance is solid and sufficiently accurate for use as a tool in producing large amounts of ...and-tagged text or for various applications where some inaccuracy is tolerable or can be ...ompensated for in later processing.

Background

SPOST has been developed as a component of the Panglyzer Spanish analysis system (Farwell, et al, 1994), which in turn, is part of the knowledge-based machine translation engine of the Pangloss human-assisted machine translation system (Frederk- ing, et al, 1994). The Pangloss system is the result of a joint research and development project of the CRL, the Center for Machine Translation at Carnegie-Mellon University and the Information Sciences Institute of the University of Southern California.

The function of the Panglyzer is to provide for each clause in an input text a set of possible meaning representations ranked on the basis of likelihood along with a set of possible relationships to any other clause in the input with which it is associated. These are then passed to an augmentor which either interactively or automatically selects a rep- resentation from among the set of the possibilities and fills in various sorts of missing information to produce an interpretation that is compatible with the context and that can act as a basis for generation.

The method for developing the system has been bottom up, focusing on provid- ing layer after layer of increasingly abstract analysis in a multi-pass process. Each level of analysis is based on a focussed type of knowledge and, to the extent possible, exploits pre-existing software or proven techniques. The architecture of the Panglyzer is a by- product of this bottom up methodology. But it is also compatible with the goal of provid- ing a "fail-soft" analysis, that is, one that is as deep as possible in the event the system fails to provide a full analysis, as well as with a strategy of repairing predictable types of errors of earlier components in later components when the relevant information is avail- able.

Panglyzer, then, consists of seven components that sequentially process the data. The first component, the Preprocessor, converts an input ASCII text into a file of PRO- LOG data structures corresponding to the sentences of the input text. SPOST is the sec- ond component and it provides, sentence-by-sentence, a part-of-speech tag for each item in the input. The third component is the Phrase Recognizer which groups contiguous ele- ments of the part-of-speech tagged input into phrase length chunks, inserting brackets around the chunks and assigning phrasal categories. The fourth component is the Proper- Noun Classifier which, operating over the entire text, assigns a semantic category (per- son, place, organization and so on) to each phrasal element of the input identified as a proper noun. The fifth component, the Phrase Analyzer, provides, sentence by sentence, a set of possible semantic representations for each phrasal element of the input. The sixth component is a Clause Recognizer which groups contiguous sets of phrase analyses in the input into clause length chunks, inserting labeled brackets indicating the clause type (finite, relative, infinitival, participial, etc.) around each chunk. The seventh component,

the Clause Analyzer, assigns, clause by clause, sentence by sentence, syntactic dependency relations (such as head, subject-of, object-of, circumstantial-modifier-of, and so on) to the various constituent phrases of each clause.

Although SPOST was developed as a component of the Panglyzer system, it has been used as a stand-alone POS tagger in other applications. Among these, it has been used as part of a tool within the Temple Translator's Workstation (Frederking, et al, 1994) which extracts proper names from texts for use in the construction of an onomasticon for the Mikrokosmos System (Onyshkevych & Nirenburg, 1994). It is also currently being used by Dublin University tagging a 200 MByte corpus (approximately 32 million words) for use in Tipster 4.

SPOST Architecture

SPOST, Panglyzer's Spanish part-of-speech tagger, automatically assigns parts of speech to the tokens of the input text stream along with a citation form and certain other morphological information. It does this by assigning an initial part of speech using a lexicon of single-word entries and a Spanish morphological analyzer and then applying context-sensitive fix-up rules which use a lexicon of predominately multi-word entries to rewrite specific POS tags. The first of the lexicons was derived automatically from an on-line version of *Collins Spanish-English Dictionary* and the fix-up rules were developed by examining results against a hand-tagged corpus of about 140 Spanish language news articles. This compares very favorably with other POS taggers that require much larger hand-tagged corpora. In addition, SPOST has the attractive feature of being modular in design, as shown in Figure 1. Since on-line dictionaries and morphologies are readily available for many languages, researchers can quickly build their own taggers using existing software. Only the fix-up rules need to be added and these can be either hand coded or learned automatically from hand-tagged text.
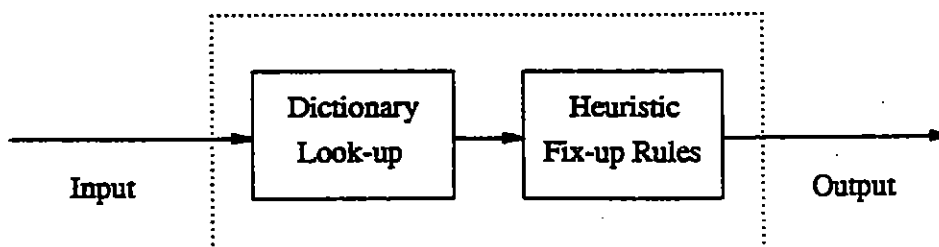

Figure 1. POS Tagger

The dictionary look-up component of the tagger takes as input a sentence-length sequence of tokens. For each token or sequence of tokens, it first generates a corresponding citation form and, for those forms requiring morphological analysis, associated morpho-syntactic information (e.g., tense, mood, number, gender, etc.). The morphological analyzer supports all verbs found in *Collins* and their inflected forms, as well as most inflectional morphology for nouns, adjectives, pronouns and articles. The citation form is then matched against the *Collins*-derived lexical database of approximately 65,000 items.

For both morphological analysis and look-up, simple heuristics based on the order of rule application are used to help the program select the most likely analysis in case of ambiguity. Where there is a match, the procedure returns a structure containing the token, the citation form and the part of speech annotated where appropriate by the results of morphological analysis. Where there is no match, the procedure returns the word, an empty list for the citation form, and the part-of-speech tag "unknown."

Altogether SPOST assigns any of 14 categories to an input token. In addition to "unknown," SPOST recognizes 11 traditional parts of speech including:

> verb,
> auxiliary,
> adjective,
> preposition,
> noun,
> proper noun,
> pronoun,
> article,
> adverb,
> conjunction,
> complementizer,

as well as "number" and "punctuation." The traditional parts of speech correspond to categories recognized by *Collins*. *Collins* also recognizes several further categories such as "interjection," "exclamation," "abbreviation," and so on which have been omitted from SPOST's tag set. In addition, most of these categories are annotated either by morphological information (verb, auxiliary, adjective, noun, proper noun, pronoun, article) or subcategory information (conjunction, complementizer, punctuation) or both (pronoun, article).

As an example, suppose the input to the dictionary look-up component is:

> ['El',grupo,'Roche',',',a,través,de,su,compañía,en,'España',',',adquirió,el, laboratorio,farmacéutico,'Doctor','Andreu',',',se,informó,hoy,aquí,'.'].

The initial dictionary look-up process will produce an output of the form:

> ['El','el',article(masculine,definite,singular)].
> ['grupo','grupo',noun(masculine,singular)].
> ['Roche','[]',unknown].
> [',',',',punc(comma)].
> ['a','a',preposition].
> ['través','través',noun(masculine,singular)].
> ['de','de',preposition].
> ['su','su',adjective(Gender,singular)].
> ['compañía','compañía',noun(feminine,singular)].
> ['en','en',preposition].
> ['España','[]',unknown].

```
[',',',',punc(comma)].
['adquirió','adquirir',verb(pret_ind,sg_3)].
['el','el',article(masculine,definite,singular)].
['laboratorio','laboratorio',noun(masculine,singular)].
['farmacéutico','farmacéutico',adjective(masculine,singular)].
['Doctor','doctor',noun(masculine,singular)].
['Andreu','[]',unknown].
[',',',',punc(comma)].
['se','se',pronoun(Gender,reflexive,singular)].
['informó','informar',verb(pret_ind,sg_3)].
['hoy','hoy',adverb].
['aquí','aquí',adverb].
['.','.',punc(period)].
```

In each line of this output, the first element is the token that appeared in the input, the second is the citation form used for accessing the part-of-speech information in the lexicon, and the third element is the part of speech, including, where relevant, the associated morphological information.

The fix-up rule component takes the above as input and uses local morphosyntactic cues to repair common mistakes. The fix-up rules themselves match predetermined sequences of tokens, citation forms, parts of speech or morphological information against the input. If a match is found, it triggers the revision of some particular part-of-speech tag in that sequence. For example, the following rule is used to identify certain cases of nouns mistagged as verbs. It looks at the input for a sequence *su*, tagged as an adjective, followed by a verb and, if it finds one, rewrites the verb as a noun.

```
[['su','su',adjective(Gender,singular)],
[Word,Citation_form,verb(Morphology)]]
=>
[['su','su',adjective(Gender,singular)],
[Word/Citation_form_1/noun(Gender_1,singular)]]
;
get_tag(Word,Citation_form_1,noun(Gender_1,singular)).
```

Note that the fix-up rules can call arbitrary Prolog predicates, here **get_tag**, to support the fix-up process. In this case, the procedure looks the token up in the lexicon as a singular noun so as to confirm the fix-up and retrieve the appropriate morphological information.

While this rule has no application to the example, three other fix-up rules do apply. The first is a rule that looks for sequences of tokens that match against the a custom built lexical database of 1000 items primarily containing fixed multi-word expressions. Here, it applies to rewrite the sequence *a, través, de* as:

```
[ ...
['a través de','a través de',preposition].
... ]
```

The second rule rewrites certain tokens tagged "unknown" as proper nouns. It applies to rewrite the tags for *Roche*, *España* and *Andreu* as:

```
[ ...
['Roche','Roche',proper_noun(unknown)].
...
['España','España',proper_noun(unknown)].
...
['Andreu','Andreu',proper_noun(unknown)].
... ]
```

The third rule rewrites capitalized words in certain contexts as proper nouns. It applies to rewrite the tag for *Doctor* as:

```
[ ...
['Doctor','doctor',proper_noun(noun(masculine,singular))].
... ]
```

As a result of applying these fix-up rules, SPOST's final output is:

```
['El','el',article(masculine,definite,singular)].
['grupo','grupo',noun(masculine,singular)].
['Roche','Roche',proper_noun(unknown)].
[',',',',punc(comma)].
['a través de','a través de',preposition].
['su','su',adjective(Gender,singular)].
['compañía','compañía',noun(feminine,singular)].
['en','en',preposition].
['España','España',proper_noun(unknown)].
[',',',',punc(comma)].
['adquirió','adquirir',verb(pret_ind,sg_3)].
['el','el',article(masculine,definite,singular)].
['laboratorio','laboratorio',noun(masculine,singular)].
['farmacéutico','farmacéutico',adjective(masculine,singular)].
['Doctor','doctor',proper_noun(noun(masculine,singular))].
['Andreu','Andreu',proper_noun(unknown)].
[',',',',punc(comma)].
['se','se',pronoun(Gender,reflexive,singular)].
['informó','informar',verb(pret_ind,sg_3)].
['hoy','hoy',adverb].
['aquí','aquí',adverb].
['.','.',punc(period)].
```

Currently there are some 52 fix-up rules. Of these, 5 are for rewriting a multi-word sequence as a single item based on the multi-word lexicon. 9 of the rules are single category rewrite rules (such as the proper noun rules in the example). 18 are for rewriting a category in the context of a second category to its left or its right. 15 are for rewriting a

category in some sequence of three categories. 5 are for rewriting a category in some sequence of 4 categories.

Currently, SPOST has about 66,000 lines of code of which about 64,500 are related to the *Collins*-derived lexicon and custom-built multi-word lexicon. There are approximately 250 rules, including about 100 fix-up rules and 50 rules for morphological analysis. It is implemented in Quintus Prolog and runs in batch mode on a SUN workstation. The central algorithm was designed and implemented by Jim Hargrave, under the direction of Stephen Helmreich, with assistance in the development of the morphological analyzer by Margarita Gonzalez. A major extension of the fix-up component was designed and implemented by Hammam Yusuf, under the direction of David Farwell, with assistance by Mark Casper, Jeff Longwell and Maria Luz Arroyo.

System Performance

It should be pointed out at the outset that the performance results described in this section are over text drawn from a development corpus and as such are not necessarily reliable. Rather, they should be taken as suggestive. We are in the process of constructing an independent evaluation corpus and expect to present the results of a valid performance evaluation in September.

In the meantime, based on an internal performance evaluation in January, 1995, which compared the output from SPOST against a hand-tagged corpus of 6 Spanish language news articles containing 2561 tokens or 2499 words or phrases, we can report that SPOST accurately assigned 95.44% of the part-of-speech categories and 89.80% of the full tags which include morphological information. The texts in the test set were randomly selected from a development corpus of some 140 such articles. The overall figures appear to hold over that entire set. We carry out such evaluations every 4 months or so for the express purpose of providing data for diagnosing inadequacies in SPOST which we then repair.

The text-by-text breakdown of the results are presented in Table 1.

Text	Items	POS	Tag	C-Form
Total:	2499	95.44%	89.80%	96.12%
Text 1:	378	93.39%	88.62%	94.46%
Text 2:	424	96.23%	89.86%	96.46%
Text 3:	448	92.63%	87.28%	97.99%
Text 4:	447	96.21%	88.86%	95.75%
Text 5:	372	98.39%	94.09%	91.67%
Text 6:	430	96.05%	91.40%	96.52%

Table 1: Text-by-Text Breakdown of Results

...states, for each text, the number of items to be tagged, the per cent of items that ...signed a correct part of speech, the per cent of items that were assigned correct ...ological information, and the per cent of items that were assigned a correct citation

    The mean deviation for the assignment of part of speech is 1.93% and the stan-...deviation is 2.12%. SPOST performed best on Text 5 with an accuracy of 98.39% ...worst on Text 3 with an accuracy of 92.63%. In regard to the assignment of the full ...the mean deviation is 2.24% and the standard deviation is 2.45%. Here, SPOST per-...med best on Text 5 with an accuracy of 94.09% and worst on Text 3 with an accuracy ...7.28%. As for the correct assignment of the citation form, the overall performance ...r the 6 texts was 96.12%, best on Text 3 at 97.99% and worst on Text 5 at 91.67%.

    The category-by-category breakdown of these results are provided in Table 2.

	Items	% Total	POS	Tag	C-Form
Total:	2499	100.00%	95.44%	89.80%	96.12%
Verbs:	208	8.32%	92.31%	86.06%	81.25%
Auxiliaries:	22	0.88%	81.82%	77.27%	95.45%
Adjectives:	200	8.00%	92.00%	53.00%	97.00%
Prepositions:	452	18.09%	98.23%	98.23%	100.00%
Nouns:	500	20.01%	92.60%	89.20%	99.40%
Proper nouns:	236	9.44%	94.07%	90.25%	98.73%
Pronouns:	28	1.12%	75.00%	10.71%	100.00%
Articles:	286	11.44%	99.30%	99.30%	100.00%
Adverbs:	38	1.52%	78.95%	78.95%	100.00%
Conjunctions:	71	2.84%	95.77%	95.77%	100.00%
Complementizers:	38	1.52%	100.00%	100.00%	100.00%
Numbers:	155	6.20%	100.00%	100.00%	70.97%
Punctuations:	265	10.60%	100.00%	99.62%	100.00%

Table 2: Category-by-Category Breakdown of Results

The table states the category (e.g., verbs), the number of items that occurred in the hand-tagged test corpus (e.g., 208), the per cent of the total number of items that appeared in the test corpus (e.g., 8.32% of 2499), the per cent of items in the category that were assigned a correct part of speech (e.g., 92.31%), the per cent of items in the category that were assigned correct morphological information (e.g., 86.06%), and the per cent of items in the category that were assigned a correct citation form (e.g. 81.25%).

    From the table, it is clear that SPOST performs best on items that do not undergo inflectional morphology. The results on prepositions (98.23%), articles (99.30%), complementizers (100.00%), numbers (100.00%) and punctuation (100.00%), which represent nearly 48% of the items to be tagged, are certainly acceptable. It is also clear the SPOST performs worst on pronouns (75.00%) and adverbs (78.95%), which represent 2.64% of the items to be tagged. Curiously, these items do not undergo inflectional

morphology either, but rather have simply not been given serious attention. The major inadequacies, then, are with the tagging of verbs and auxiliaries, adjectives, nouns and proper nouns, which together represent about 47% of the items to be tagged. Since these are the categories which undergo inflectional morphology, are relatively prone to categorial ambiguity, and are open ended, the results may be attributed to inadequacies in the morphological analyzer, in the default heuristics for disambiguation, and limitations on the size of the lexicons.

The major categories of errors in part-of-speech assignment are presented in Table 3.

Category	Total	V	Adj	N	PN	Art	Conj	Other
Verbs:	15	-	3	7	4	-	-	1
Auxiliaries:	4	3	-	-	-	-	-	1
Adjectives:	15	3	-	6	3	-	-	3
Prepositions:	4	-	-	-	-	-	2	2
Nouns:	33	5	17	-	10	-	-	1
Proper nouns:	10	-	-	6	-	-	2	2
Pronouns:	7	-	-	-	-	7	-	-
Articles:	2	-	-	-	-	-	-	2
Adverbs:	7	-	2	-	-	-	5	-
Conjunctions:	3	-	-	-	-	-	-	3

Table 3: Mistaggings by Category

The table states for each category (e.g., verbs) the total number of incorrectly tagged items (e.g., 15) followed by a summary of how the items were mistagged (e.g., 3 as adjectives, 7 as nouns, 4 as proper nouns and 1 as either auxiliary, preposition, pronoun, adverb, complementizer, number or punctuation).

The table points out that the categories taking inflectional morphology, verbs, auxiliaries, adjectives, nouns and proper nouns, account for 77% of the total number of errors in part-of-speech assignment that SPOST committed (100). In addition, it shows that the majority of these errors are due to a misassignment to one of the other categories taking inflectional morphology (14 of 15 for verbs, 3 of 4 for auxiliaries, 12 of 15 for adjectives, 32 of 33 for nouns, 6 of 10 for proper nouns). As for the other categories, prepositions (4 errors), pronouns (7 errors), articles (2 errors), adverbs (7 errors) and conjunctions (3 errors), the results can mainly be attributed, on the one hand, to lack of attention (bugs in the code) and, on the other, to inadequate heuristics for disambiguation.

The were 134 errors related to morphological information altogether. Of these, 112 concerned the treatment of variable gender adjectives and nouns which have been mistakenly identified masculine or neuter, 10 concerned the identification of third, singular finite verbs as first, singular, 8 concerned incorrect subclassification parts of speech for proper nouns (e.g., "Doctor" in "Doctor Andreu" as "proper_noun(unknown))" rather than "proper_noun(noun(masculine,singular))"), and the remaining 4 related to verb morphology. None of the error types presents a significant problem. Only the errors

...verb morphology arise from ambiguity and, thus, require a default treatment that may ...certain cases lead to mistagging.

As for the errors in citation form assignment, 45 of a total 79 errors were due to the assignment of a plural form of a noun rather than singular (e.g., "millones" as opposed to "millón"), 25 were due to the assignment of a past participle form of a verb rather than the infinitive (e.g., "dedicado" vs "dedicar"), 4 were due to the assignment of some other form (finite, present participle) of a verb rather than the infinitive, and the remaining 5 were due to assorted other problems. Again, none of the above types of errors presents a significant problem.

System Accessibility

The process of setting up SPOST for public access has already begun. We are setting up an e-mail List Server at "spost@crl.nmsu.edu" to which you will be able to mail in a text to be tagged. When it is received, it will be queued for tagging, tagged, and returned to the sender, possibly the following day. Texts should be submitted as ASCII files (Latin-1 ISO standard). Initially, all text formatting characters other than new lines should be removed prior to submission and some limitations on file size may have to be imposed. We also expect to distribute TagView and an automatic scoring program. TagView is the interactive interface developed at the CRL for viewing and correcting SPOST tagged texts. We are in the process of revising certain bugs in its tag revision procedure and of planning its distribution to sites interested in correcting SPOST's output. The effort related to the scoring program is essentially the same.

Current Activities and Future Directions

At this point, SPOST is a mature part-of-speech tagger. Nonetheless, while we are now in the process of making the component accessible to the research community at large, we also continue to work on improving its performance. Thus, our current efforts are directed at:

* annotating a test corpus and running a valid evaluation,

* improving SPOST's performance.

The process of annotating a test corpus has already begun. At the time of writing, we have assembled a 25 text corpus (approximately 400 words each or 10,000 words altogether) drawn from the EFE news service and other sources. Half the articles are in the domain of financial news reports while the other half are in other domains (general news, society, entertainment, sports). This Corpus will be tagged automatically by SPOST and then corrected by language experts using an interactive interface, TagView, which was developed for this precise purpose. When the annotation process is completed, the most recent version of SPOST will be used to tag the corpus again and the performance automatically scored against the hand-tagged version. Both the interface and the scoring algorithm are currently in use at the CRL.

In regard to improving SPOST's performance, we are currently focusing our efforts on:

- improving the tagging of verbs and auxiliaries to 100%,

- improving the tagging of prepositions, conjunctions and adverbs to 100%,

- improving the tagging of pronouns to 100%,

- debugging the procedures for assigning morphological information and citation forms,

- improving the tagging of nouns and proper nouns,

- improving the tagging of adjectives.

The improvements are mainly related to a revision and extension of the fix-up rules, although we are in the process of introducing an improved morphological analyzer at the time of writing. The reason for focusing on verbs and auxiliaries first and attempting to achieve such a high accuracy is that they play a central role in Panglyzer's clause level analysis. As for prepositions, adverbs, conjunctions and pronouns, our analysis of SPOST's current performance indicates that improvements are readily obtainable with a limited number of straightforward changes. Similarly, we expect a relatively small effort to lead to large gains in the assignment of morphological information and citation form. Improvements in the assignment of noun, proper noun and adjective categories, on the other hand, will be more difficult.

Apart from the current activities described above, we also have set for ourselves a number of further objectives for the future. These include:

- setting up the phrase analysis component and its interface for viewing and correcting output for public access,

- setting up the clause analysis component with its interface for viewing and correcting output for public access,

- experimenting with alternative techniques for developing fix-up rules (e.g, empirically deriving statistical rules),

- experimenting with the introduction of multiple tags where ambiguities exist,

- adapting the SPOST methodology to the development of taggers for other languages.

All of these projects are in an initial planning stage.

Conclusion

We have presented the architecture, processing and performance of SPOST, the Spanish part-of-speech tagger currently being used in the Panglyzer Spanish analysis component of the Pangloss knowledge-based machine translation system. SPOST is a two stage, rule-based tagger that uses a large lexicon derived from an on-line version of the *Collins Spanish-English Dictionary*. While only suggestive, a performance evaluation over a small corpus indicates that SPOST has a accuracy rate of about 95% for part-of-speech category assignment and 90% for additional morphological information. It is