A Galician Textual Corpus for Morphosyntactic Tagging with Application to Text-to-Speech Synthesis

Lorena Seijo Pereiro*, Ana Martínez Ínsua*, Francisco Méndez Pazó†, Francisco Campillo Díaz†, Eduardo Rodríguez Banga†

* Centro Ramón Piñeiro para a Investigación en Humanidades. Xunta de Galicia.
Estrada Santiago - Noia, Km. 3 - A Barcia 15896 - Santiago de Compostela - Galicia (Spain)
{lseijo, iaminsua}@usc.es

† Dpto. Teoría de la Señal y Comunicaciones. Universidad de Vigo.
Campus Universitario - Lagoas-Marcosende s/n 36200 -Vigo (Spain)
{fmendez, campillo, erbanga}@gts.tsc.uvigo.es

Abstract

This paper will present the morphosintactic tagger and the corpus of contemporary written Galician which are being employed in the development of the Galician version of our tex-to-speech synthesizer. Their quality and accuracy make them useful for speech technology applications and turn them into possible references for further investigation and research projects about Galician language. In essence, the tagger assigns automatically the morphosyntactic categories and other additional labels to the words in the corpus by resorting to a combination of both a reduced (although highly reliable) set of rules, and a stochastic language model that employs class n-grams whose probabilities are trained using the corpus itself. A bootstrapping technique is employed for tagging the texts contained in the corpus: a small amount of text is initially tagged automatically making use of a reduced set of linguistic rules and then, gathering together the results obtained at this stage of the process (after the manual revision of the tagging), an initial statistical model is built. The tagging process may be said to consist essentialy of a number of consecutive automatic-tagging stages that enclose: the use of the latest version of the statistical model, the manual revision, and the subsequent updating of the stochastic model with the correctly tagged text.

1. Introduction

Statistical methods have proven to be a well-founded technique for the problem of automated morphosyntactic analysis (Cutting et al., 1992). These methods help to resolve ambiguity on the basis of the most likely interpretation, but need a large labelled corpus for accurately training the probabilities of the morphosyntactic model. The acquisition of such a corpus may become an extremely difficult task, especially in the cases of minority languages, such as Galician. In this paper we present the creation of a Galician corpus for these purposes, as well as a (part–of–speech) POS–tagger in the context of text–to–speech systems (TTS).

Prosodic modelling is a fundamental stage in any text–to–speech system. The quality of the resulting synthetic speech is highly dependent on a proper estimation of the segmental durations and the intonation contour. As it is well–known, the distribution of the accents plays a central role in prosodic modelling. Like in many other systems, the prosodic modelling in our TTS (Campillo and Banga, 2002) is based on accent groups (defined as a sequence of non–stressed words ending in a stressed word). Given that stressed words are closely related to the function they perform in the sentence, it is easy to understand the significance of a precise morphosyntactic tagger in the performance of a speech synthesizer. Moreover, the syntactic relationships, related to the morphological analysis, contain much information about the fundamental frequency contour, and the suitable places for pause insertion.

As will be detailed in the following section, the corpus was designed with the speech synthesis necessities in mind, and was therefore extracted from a contemporary newspaper with multiple authors and different topics. Nevertheless, it is our intention to enlarge its size so as to reach 1 million words and include other types of texts such as novels or essays, making it useful for other purposes. Briefly, the structure of this paper will be as follows. First, we will describe the characteristics of the corpus, as well as the different stages of tagging and debugging that were carried out so as to minimize errors. Once we have described the corpus itself, we will describe the chosen tagset, based on the EAGLES recommendations (Leech and Wilson, 1996), and adapted to the Galician morphosyntactic peculiarities. After that, some attention will be paid to the tagging process. Finally, we will conclude the paper with an evaluation of the developed POS–tagger, showing the suitability of both corpus and tagger for the original purpose they were designed.

2. The Corpus

As hinted above, the textual corpus is inspired in other similar corpora widely recognised as reliable for other languages (e.g. P*enn Treebank* or *Brown Corpus* for English, *NEGRA* for German, or *LexEsp* for Spanish). It is made up of about 400.000 words drawn from journalistic texts of the last 6 years. The texts included within the compilation were not restricted in terms of their topics, styles or authors. Given its contemporary nature, the corpus may be said to be completely up–to–date and follow the linguistic norms prevailing at the moment when this paper was written.

It will be explained in the following section that the texts of the corpus were tagged using a bootstrapping

technique, which allowed us to build an initial statistical model as a subsequent help for the whole process.

As explained in the introduction, the corpus is intended to be enlarged and reach 1 million words by including samples of other types of non–journalistic texts (e.g. novels, essays, academic writing, etc.) in the near future. We aim to achieve a wide sample of real Galician language containing as many different styles and genres as possible. The enlargement of the types of texts of the corpus will provide us not only with morphological information but also with syntactic and prosodic data about Galician, from which the text–to–speech synthesis process will undoubtedly benefit. Similarly, it is our intention to provide the potential users of the corpus with the chronological information about the texts it contains as, for the time being, such data are not available.

3. The Tagset

All the texts contained in the corpus were morphosyntactically tagged. The set of labels employed have been normalized according to the recommendations for the morphosyntactic annotation of corpora that the EAGLES (*Expert Advisory Group on Language Engineering Standards*) group proposes in its 1996 recommendations document.

Generally speaking, the typology of categories proposed by the EAGLES group was maintained when elaborating ours, and both categorisations are considerably similar. Only slight modifications were made with respect to the EAGLES standard so as to adapt it to the language under investigation (i.e. Galician) and its defining features.

As Table 1 below shows, twelve main categories are distinguished in our typology: verb, substantive, pronoun, adjective, determiner, article, adverb, preposition, conjunction, interjection, residual and punctuation. Notice that, while the EAGLES group considered both pronouns and determiners as one single category, we separate them into two different ones. Besides that, while the EAGLES group granted numerals the status of independent category, we classify them as types of determiners and pronouns (depending on their function within the discourse), placing them at the same level as demonstratives and possessives, among others. The functional character of our corpus may be regarded as the underlying reason for having made these distinctions.

Another source of difference between the typology proposed by the EAGLES group and our classification is the category "Unique-unclassified". The EAGLES group gathers under this label a number of markers and particles that might not be taken to belong to any other category. By contrast, we did not include this category in our proposal, as Galician language does not really have any components of this kind (e.g. English negation mark "not", to mention only one). Finally, it is worth noting that, given that Galician lacks postpositions, the category "Adposition" (sub–divided by the EAGLES group into "Preposition" and "Postposition") was simplified in our model and the category "Preposition" was considered instead.

The total number of basic categories considered by the EAGLES group is 13 while those distinguished in our proposal amount to 12. The slight differences between one and the other typologies are not to imply, however, radical distinctions between them. Quite on the contrary, the similarities between them are obvious and our proposal not only is based on the EAGLES group's one but also aims to follow it closely and be compatible with it.

EAGLES GROUP	**OUR TAGGER**
Substantives	Substantives
Verbs	Verbs
Adjectives	Adjectives
Pronouns and determiners	Pronouns
Articles	Determiners
Adverbs	Articles
Adpositions	Adverbs
Conjunctions	Prepositions
Numerals	Conjunctions
Interjections	Interjections
Unique/unassigned	Residual
Residual	Punctuation marks
Punctuations marks	

Table 1: Main categories recognised by the taggers

It may be interesting now to go into further details and specify how these categories are treated in our proposal. Table 2 below displays the case of determiners.

Categorial tag	D
Category	Determiner
Type	Demonstrative; Contracted demonstrative; Indefinite; Contracted indefinite; Possessive; Distributive possessive Exclamative; Interrogative
Gender	Masculine; Feminine Ambiguous
Number	Singular; Plural
Person/number[1]	1st singular; 2nd singular; 3rd singular; 1st plural; 2nd plural; 3rd plural; none
Contracted forms	Demonstrative+indefinite Indefinite+definite article None

Table 2: Determiners

The whole tagset we propose, with its 760 tags, is available in our website:

[1]This is specified in the case of possesives

http://www.gts.tsc.uvigo.es/cotovia/doc/tags.htm

On certain occasions, working with Galician language, a highly inflected one, compelled us to search for labels with higher degrees of accuracy than those employed for tagging other languages. Such is, for instance, the case of Galician "compound" prepositions (*locucións preposicionais*), whose final component may be contracted with articles, demonstratives or indefinites and, therefore, specify gender and number. Likewise, even if no reference is made to them by the EAGLES group, we included in our typology the category *compound adverbs*, given their relevant presence in Galician language. Consequently, the tags for prepositions and adverbs were necessarily enlarged and, as Tables 3 and 4 below show, they came to specify such traces.

Expression	**Tag**
A carón de (beside)	P#L (i.e. compound preposition of place)
A carón dela (beside her)	P#LNTFS3T (i.e. compound preposition of place contracted with a stressed personal pronoun of third person, feminine, singular and non–oblique).

Table 3: Compound preposition

Expression	**Tag**
De vagar (slowly)	B#M (i.e. compound adverb of mood

Table 4: Compound adverb

Other labels that required a considerable amount of enlargement were those employed for tagging the verbal system. As opposed to what happens in other languages such as English, Galician verbs are highly inflected and have different endings marking tense, person and number. Even infinitives may be inflected for person and number, as evinced by Table 5 below.

4. The Tagging Process

As regards the tagging process, we may say that, in essence, the tagger assigns automatically the morphosyntactic categories and other additional labels to the words in the corpus by resorting to a combination of both a reduced (although highly reliable) set of rules, and a stochastic language model that employs class n–grams whose probabilities are trained using the corpus itself.

Thus, the tagging process begins by assigning each word all its possible morphological categories. In order to do that, information is gathered from a number of dictionaries and tables elaborated by the linguists of the research team. Such dictionaries and tables first establish divisions between words belonging to closed classes (i.e. prepositions, adverbs, conjunctions, etc.), and subsequently, between substantives and adjectives (with their corresponding gender and number), verbs and periphrases.

Label	**Verbal form**	**Morphological information**
V1	*Comer* (to eat)	Infinitive
VDPS1	*(Eu) Como* (I eat)	Verb, indicative present tense, 1st singular person
VDPS2C	*(Vostede) Come* (you eat)	Verb, indicative present tense, 2nd singular person polite form
VCP1	*Comermos* (we - to eat)	Verb, inflected infinitive, 1st plural person

Table 5: Verbal tags

Once the initial tagging process was carried out, in those cases where it is possible to assign more than one category to one single word, it corresponds to the statistical model to select the most probable category. Expectedly enough, the process is not completely error–free, but the accuracy of the system is very high: generally speaking, it raises to nearly 97%, reaching 98% in the cases of gender and number. Consequently, the manual revision of the tags assigned is still necessary so as to avoid possible problems and mistakes in cases where the system encounters ambiguity.

A bootstrapping technique was employed for tagging the texts contained in the corpus. This means that, initially, a small amount of text is automatically tagged making use of a reduced set of linguistic rules. Gathering together the results obtained at this stage of the process, and after the manual revision of the tagging, an initial statistical model is built. The manual revision and tagging processes are carried out by a couple of linguists, who employ various tools so as to avoid possible mistakes such as spelling errors, for instance. Therefore, the process may be said to consist essentialy of a number of consecutive automatic–tagging stages that enclose: the use of the latest version of the statistical model, the manual revision, and the subsequent updating of the stochastic model with the correctly tagged text. It is important to mark that, no doubt, the manual revision gains precision and reliability when the source text has fewer errors.

It is equally important to clarify now the above mentioned question of prepositional and adverbial phrases. They are included (as sub–types) within the categories *preposition* and *adverb*, respectively, as their main component is one of these, and their meaning is to a large extent equivalent to the single prepositions, adverbs and conjunctions. As explained, given the peculiar nature of Galician language and the possibility that the final

component of these phrases is contracted with articles, demonstratives and indefinites, the tags assigned to these words were necessarily enlarged to specify number and genre. Thus, in these cases, our tagging process gains accuracy with respect to those of other languages, and our tags become further descriptive with respect to the standard ones proposed by the EAGLES group, for instance.

5. Evaluation of the tagger

As said above, our tagger employs a hybrid method that combines linguistic rules (which initially reduce the ambiguity in the morphological categories) and a statistical or machine–learning method (which decides the most probable sequence of morphosyntactic categories in the sentence). We use a pentagram–based language model for the contextual probabilities and the concept of ambiguity class (Tzoukermann et al., 1999) for the lexical probabilities.

The complete set of categories (over 700 tags) we are presenting in this paper has been considered to be too ample and detailed, due to the problem of training data sparseness, for using it in the estimation of the probabilities of those language models. Therefore, a reduced tagset of 53 morphosyntactic labels was used for the training process of the statistical models. This reduction was done by suppressing superfluous (for the task of disambiguating morphosyntactic categories inside our text–to-speech system) information (i.e. subtype in conjunctions; person and tense in verbal forms; type in adverbs; etc.). Needless to say that the same reduced tagset is internally employed during the POS tagging an dissambiguation tasks in our TTS system. It may be said that this reduction process is, in most cases, perfectly reversible.

Aiming to evaluate the performance of the tagger, we took 50.000 words as our test material, we kept them apart from the rest of the corpus, and used the remaining words for training the statistical models for the tagger. At the present moment, the accuracy of our tagger reaches nearly 97% of correct tagged words, taking as reference the reduced tagset.

Further results are shown in Figure 1 below, where we may observe the evolution in the precision of the tagger versus the size of the training text during the bootstrapping procedure of creation the corpus. The figure shows a baseline value of 89.75% correctly tagged words, obtained as a result of applying exclusively the reduced set of linguistic rules of the tagger. It is worth noticing that results began to be reasonable from 50.000 words of training text. Multiplying this amount by 6 barely improved the accuracy in 0.6%. For that reason, we claim that, while a bigger corpus might be desired from a linguistic point of view, its current size is sufficient for its use as training material for our morphosyntactic tagger. Thus, with a corpus of this size, the results obtained are reliable enough for this stage of the synthesis process, which was one of our initial objectives.

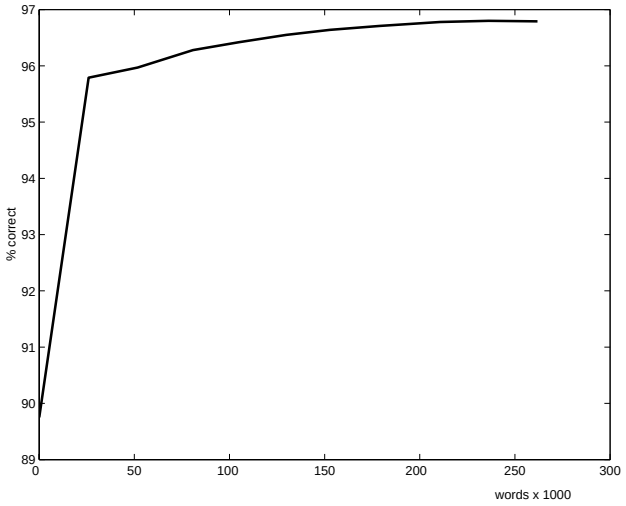

Figure 1: Tagging accuracy vs. size of training text

Summing up, suffice it to say here that part–of–speech tagging processes constitute vital stages in any TTS conversion mechanism. In this sense, the correct morphological analysis of a given text is essential for determining its adequate prosody. As explained above, the morphosyntactic tagger we presented in this paper was born as a tool for the TTS–synthesizer we are elaborating, and it has been integrated in the current version of such TTS–synthesizer (www.gts.tsc.uvigo.es/cotovia).

6. Acknowledgements

This work has been partially supported by the "Ministerio de Ciencia y Tecnología", FEDER funds and the "Xunta de Galicia" under the projects TIC2002–02208, PGIDT01PXI32205PN and PGIDT02PXI32201PR.

7. References

Campillo, F. and Banga, E. R. (2002). Combined prosody and unit selections for corpus-based text-to-speech systems. In *Proceedings of ICSLP*, volume 1, pages 141–144, Denver.

Cutting, D., Kupiec, J., Pederson, J., and Sibun, P. (1992). A Practical Part–of–speech Tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, ANLP*, pages 133–140, Trento, Italy.

Leech, G. and Wilson, A. (1996). Eagles: Recommendations for the Morphosyntactic Annotation of Corpora. . EAG–TCWG–MAC/R, http://www.ilc.cnr.it/EAGLES96/home.html.

Tzoukermann, E., Radev, D., and Gale, W. (1999). Tagging French without Lexical Probabilities - Combining Linguistic knowledge and statistical learning. In Armstrong, S. et al, editor, *Natural Language Processing using Very Large Corpora*, pages 43–66. Kluwer Academic Publishers, Dordrech.