Compilation of a Mexican Spanish text corpora

S. N. GALICIA-HARO, A. F. GELBUKH & I. A. BOLSHAKOV
Computation Research Center
National Polytechnic Institute
Av. Juan de Dios Batiz S/N
07738 MEXICO, D. F.
{sofia, gelbukh, igor}@cic.ipn.mx

Abstract: -Collections of texts with syntactic annotation are nowadays useful resources. They are employed for diverse tasks in theoretical research and natural language applications. The most important collections are dedicated to English. But huge efforts have being realized to develop the corresponding to other languages. In this work we present the initial steps for the compilation of a Mexican Spanish text corpora with syntactic annotation.

Key-Words: - text collection, annotated corpus, corpus compilation

1 Introduction[1]

Obtaining usage information of language from texts corpora has been a common practice in lexicography [5]. Also in natural language processing the use of very large texts corpora is a common practice for phenomena research in unrestricted materials. This practice facilitates and accelerate the tasks that are being developed to realize computer language understanding.

Researches recognize the potential of use of very large texts corpora for problem solving in the lexical, syntactic and semantic levels of analysis. However, the useful texts corpora required in natural language processing need annotations, i.e., lexical, syntactic and semantic marks. Manual work is the most employed method for text corpora annotation.

The main usage of this kind of corpora has been for training methods to resolve distinct natural language processing tasks. This usage has a wide range, grammatical categories assignment for unknown words [13], prepositional phrase attachment [11], sense word disambiguation [10], etc. The main purpose of our work is the unrestricted Mexican texts parsing.

In this work we present some observations and results about text collection compilation, then we describe the development for our syntactic annotated corpus.

[1] Work done under partial support of CONACyT, SNI, and CGEPI-IPN, Mexico.

2 Characteristics of annotated corpus

The text corpus compilation implies different problem solutions of the self texts, i.e. text analysis. The input source should be analyzed upon diverse criterions: texts with required information acquisition, corpus coverage over linguistic phenomena required, corpus confidence, etc.

Ideally it is desirable to obtain a large and representative sample of general language. The reason for a large sample is that it could be expected a larger quantity of words as longer is the corpus. This quantity of words will imply a bigger dictionary language coverage and it mainly implies greater evidence of the diverse linguistic phenomena required. To be representative supposes several cultural language levels, several themes and genres. However these qualities not implies each other, instead in some cases they are contrary. One contraposition that must be considered is that between quality and quantity. A big corpus does not guarantee to posses the expected quality.

The corpus should be balanced among those qualities. However, it seems impossible to balance appropriately a corpus, not without a huge effort. Besides the sampling methods are unfortunately expensive, for example those for quality selection. So we must to assume the obvious problems related to work with unbalance data, because the construction of a balanced corpus requires much time and a huge cost.

Because of the impossibility of having a corpus with all desired qualities, we limited the corpus qualities to required information and size that are relevant for our goals. One of the goals for the

Sofía N. Galicia-Haro, Alexander Gelbukh, Igor A. Bolshakov. *Compilation of a Mexican Spanish text corpora.* Advances in Information Science and Soft Computing, WSEAS Press, 2002, p.70-75.

Mexican texts corpus compilation is the syntactic analysis of unrestricted texts, similar to newspapers texts. One of the main problems in the syntactic analysis is the correct attachment of noun and prepositional phrases. Therefore it is important that the corpus contain extensive use of prepositional phrases and predicates.

About information required in a corpus, for example, [2] notes the different use of prepositional phrases according the text genre. [12] found significant differences of subcategorization frequencies in different corpus. Discourse and semantic influence were identified as the sources for those differences. The former is caused by the language form changes used in different types of discourse. The semantic influence is based on semantic context of discourse. Then a corpus with different genres should be quite adequate.

About big size of corpus, the current corpora have a range from millions of words to hundred of millions, depending on its type, i.e. plane text or diverse type of annotations. [1] argue that a corpus must be big enough to avoid sparse data and reflect natural use of language in order to obtain a good probabilities approximation. They use the one million of word Wall Street Journal. Other authors, at the contrary, don't use the whole corpus for their research but a subcorpus with specific characteristics [14], [11]. Therefore we consider the use of several millions word corpus.

The main importance of the corpus considering our purpose is the possibility to obtain the arguments of verbs, adjectives and nouns. [12] explain that as the quantity of surrounding context increase (from one sentence to a connected discourse) the necessity of explicitly express all verb arguments decrease. This phenomena also appears in very long sentences. So in our work even not long sentences are useful contrary to the sentences required for syntactic analysis testing.

2.1 Corpus annotation

There are several levels for corpus annotation: lexical, syntactic, semantic, etc. Another levels of annotation could exist in each of those indicated levels.

Lemma and part of speech assignments are considered in lexical annotation. There could be diverse detailed grade of annotation. For example, the Penn Tree-bank [8] uses 36 marks for part of speech and 12 for punctuation and other symbols. While the Brown Corpus [5] uses 87 simple annotations and it permits compose annotations.

The sentence structure in the syntactic level is generally showed grouping words by parenthesis, and additionally labeling those groups. Because of a complete structure requires more learning time of the scheme by annotators and more time for sentence annotation there are different grades of sentence hierarchic structure realization. For example, in the Penn Tree-bank development the distinction of sentence arguments and adjuncts was ignored in a first stage. But the argument annotation is crucial for the semantic interpretation of verbs.

In the semantic level, it has been considered the signification annotation and a type or concept, for example in the development of the Italian Syntactic-Semantic Treebank [3].

The first stage in the compilation of a Spanish corpus with syntactic annotation only comprise the lexical and syntactic levels, future work will include the semantic annotation.

3 Corpus compilation

Given the defined size of millions words for the corpus and the objective of unrestricted text analysis, we consider text extracting from Internet as the quickest way to collect Spanish texts.

We selected four Mexican newspapers that are daily published in the Web with a considerable part of their complete publication. Their Web organization permitted us an automatic extraction for monthly and yearly periods. The texts correspond to diverse sections: economy, politics, culture, sport, etc. from 1998 to 2000. All texts are in HTML format.

The size of the original texts was 1540 MB from which we obtain 1092 MB in plain text with some annotations by the following steps:

1) HTML labels deleting.

2) Article structure assignment.

   It was automatically labeled with marks of title, subtitle, text body, paragraph, sentence.

3) Wrong or correct word assignment.

   We obtained all the different words of the texts. We annotated automatically the "correct" words, using the orthographic tool of a word processor and a Spanish dictionary. The correct words were those recognized by such resources.

   A manual non exhaustive work let us identify words used in Mexico that does not appear in

Type		Person	Gender		Number		Case	Possessor
Value	Key		Value	Key	Value	Key		
Demonstrative	D	1	Feminine	F	singular	S	0	0
Possessive	P	2	Masculine	M	Plural	P		
Interrogative	T	3	Common	C	Invariable	N		
Exclamatory	E							
Undefined	I							

Figure 1. Determinant characteristics.

DRAE[2], neither in María Moliner dictionary, like *ámpula*, but it appears in DEUM[3]. Also other correct words with Indian origin (náhuatl, maya, otomí, etc.) were identified. Some heuristics were used to detect diminutives and other word variations.

A future work will include some kind of error correction. From 747,970 total different words, 60% are marked as wrong words. Typographic error, spelling check errors and words in capital letters without accents were considered as wrong words.

4) Linking of composed prepositions.

There are many composed prepositions in Spanish besides simple prepositions. Words group as *al cabo de*, *con respecto a*, requires a manipulation as a set. According the preposition list of [8] the composed prepositions were automatically linked in the corpus of newspapers texts (*al_cabo_de*, *con_respecto_a*).

5) Proper names annotation.

One postponed task is the proper name identification, simple and composed proper names. This task usually has been manually realized. We expect to reduce the manual work employing some heuristics.

3.1 Lexical level annotation

There are Spanish corpus with lexical level annotations, for example the LEXESP[4] corpus. As LEXESP corpus has been used for research in our work group we decided to use the same 275 different labels. This quantity of labels is so big mainly because of Spanish gender, person and number agreement.

The LEXESP corpus has the PAROLE [4] categories. We present the POS classification in PAROLE, where we only detail the complete key for determinants, showing the considered features.

1. Adjective (A). Example: *frágiles* <AQ0CP00>
2. Adverb (R). Example: *no* <RG000>
3. Article (T). Example: *la* <TDFS0>
4. Determinant (D), see figure 1. Example: *tal* <DD0CS00>
5. Noun (N). Example: *señora* <NCFS000>
6. Verb (V). Example: *acabó* <VMIS3S0>
7. Pronoun (P). Example: *ella* <PP3FS000>
8. Conjunctions (C). Example: *y* <CC00>
9. Numerals (M). Example: *cinco* <MCCP00>
10. Prepositions (SPS00). Example: *a* <SPS00>
11. Numbers (Z). Example: *5000* <Z>
12. Interjections (I). Example: *oh* <I>
13. Abbreviations (Y). Example: *etc.* <Y>
14. Punctuation (F). All punctuation signs (.,:;-¡!'¿?"%). Example: "." <Fp>
15. Residuals (X). The words that does not fit in the previous categories. Example: *sine* <X> (a Latin word).

We present the POS for the word *bajo* which could be verbal form, preposition, adverb, noun or adjective:

bajar<VMIP1S0> bajo<SPS00> bajo<RG000> bajo<NCMS000> bajo<AQ0MS00>.

The "common" value in gender is employed for both feminine and masculine, for example *alegre* (glad). The "invariable" value in number is used for both singular and plural, for example: *se* (pronoun).

The POS annotation was realized with the MACO program developed by the Natural language processing group of the Artificial intelligence section of the Software Department in the Polytechnic University of Catalonia in collaboration with the Computational Linguistic Laboratory of the Barcelona University.

[2] Spanish language dictionary of the Real Spanish Academy. Espasa, Calpe, 21 ed. 1995

[3] Usual Spanish in Mexico Dictionary. Ed. Colegio de México. México, 1996.

[4] The LEXESP corpus was kindly provided by H. Rodríguez from the Polytechnic University of Catalonia, Barcelona, Spain.

3.2 Syntactic level annotation

The syntactic annotation is actually being developed. We propose this annotation based on dependency grammars. Dependencies are established between pairs of words, where one is principal or government and the other is subordinated or dependent of the first one. The root of the tree is the only word that is not subordinated to other one.

The syntactic annotation of English corpora as Penn Tree-bank and Brown Corpus is based on constituent grammars, mainly because of its stricter word order. The knowledge described in constituent grammars is the classification and segmentation of sentences based on the POS of the sentence words. This grammatical knowledge is directly codified in rewriting rules, i.e. in context free grammars.

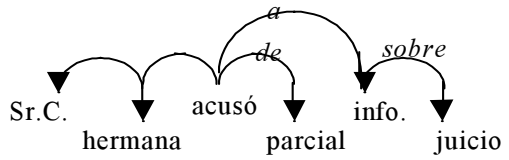
However, the dependency grammars are considered more adequate for languages with relaxed word order constrains. Besides the syntactic description alone of complements does not permit to establish computer rules defining the specific words with which they combine. We considered the syntactic annotation in two levels: dependency structure and syntactic relations.

Nowadays an automatic syntactic annotation is not possible because all the syntactic analysis problems are not yet resolved. The combinations of all the complements in a sentence introduce certain complexity in the parsing task.

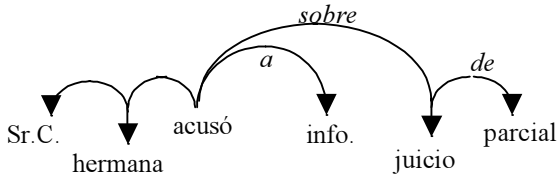
For example, in the phrase *Acusó la hermana del señor Carrillo a la información televisiva sobre el juicio de parcial y progolpista*5 (Mr. Carrillo's sister accuse the trial television information of being partial and pro coup d'état), there are four prepositional phrases (*del señor Carrillo*, etc.) introduced by three prepositions: *a, de, sobre*, and one noun phrase (*la hermana,* the sister). The possible combinations are no random but the complements could be linked in different combinations. However all variants are possible for a computer, for example:

- *Acusó la hermana del señor Carrillo.*
- *Acusó a la información televisiva.*
- *Acusó sobre el juicio.*
- *Acusó de parcial y progolpista.*
- *sobre el juicio de parcial y progolpista.*
- *etc.*

Employing some of the combinations the correct syntactic analysis is obtained:



But many other wrong syntactic structures6 are also obtained, among them we show the following structure:



This example shows how the number of syntactic structure variants could be reduced specifying the verb arguments.

In the work presented in [8] the authors use an English parser that only gives one structure variant in the output or groups of isolated constituents otherwise. From that kind of output the annotators corrected the sentence structure. They required well trained annotators with high percentages of efficiency.

We propose the development of a tool for syntactic annotation that permits the selection of the complete structure for any sentence, considering the following characteristics:

1. Use of a constituent grammar to describe the syntactic structure of each sentence.

It is one of the simplest models for syntactic ambiguity resolution but one more easy for applying and for compiling the necessary resources.

We use a Spanish context free grammar for this tool, it was developed in the Natural Language and Text Processing Laboratory of the Computational Research Center of National Polytechnic Institute.

The compiled rules cover the most common employed syntactic structures. The following improvements were introduced into the grammar:

- *Agreement restriction (gender, number, etc.) to avoid overgeneration.*
- *Inclusion of the government element, marked with sign @.*
- *Syntactic relations, for example an adverb has a modifier relation respect a government verb.*
- *Punctuation elements.*

5 This phrase belongs to LEXESP corpus.

6 These structures are simplified dependency structures since they do not have syntactic relations.

- *Semantic annotations for time in noun phrases.*
- *Weights to classify the rules employed in the analysis.*

The first improvement is obligatory for a language with inflections as Spanish. The rest of improvements are not common in this type of grammars.

2. Transforming the constituent structure to a dependency structure.

Introducing the government element @ in the Spanish CFG rules permits by means of an algorithm to realize the transformation from a constituent structure to a dependency structure. One example is the following rule:

NOM(nmb,gnd,pers) → @:N(nmb,gnd,pers)
Adj(nmb,gnd)

For this simple case, the resulting structure is a government node for the noun with a dependent node for the adjective.

3. Classifying the syntactic structure variants for each sentence.

Introducing the statistics of words (verbs, adjectives and nouns) with their corresponding prepositions introducing their predicates [6] it is possible to classify the variants obtained by the generative grammar-based parser. The result is a group in the top of the most probable correct variants.

This classification permits the annotator to choose the correct variant from a reduced group.

4. Permitting to link substructures and adding syntactic relations.

In the cases that the annotator could not find the correct structure among the group in the top, he or she should select the root in the structure and from it the tool should narrow the group of variants showed.

In all the cases the tool should have to permit the modification of syntactic relations.

4 Conclusions

We exposed the utility of corpus with annotation on several levels, for diverse work in theoretical research and natural language applications.

We presented the development of a resource for the linguistic processing of Mexican Spanish texts: a collection of texts with syntactic annotation. We explained the process required to obtain the collection of texts and the part of speech annotation. We presented the proposed method for syntactic annotation.

The main advantage of our method to compile the annotated corpus is that most of the work was realized in an automatic form to reduce time and costs of compilation.

References

[1] Berthouzoz, C. and Merlo, P. *Statistical ambiguity resolution for principle-based parsing*. In Proceedings of the Recent Advances in Natural Language Processing. Pp. 179-186, 1997

[2] Biber, D. Using Register. Diversified *Corpora for general Language Studies*. Computational Linguistics 19 (2) pp. 219—241, 1993.

[3] Calzolari, N. Corazzari, O. & Zampolli, A. *Lexical-Semantic tagging of an Italian Corpus*. Second Conference on Intelligent text processing and Computational linguistics. CICLing-2001.

[4] Civit, M e I. Castellón. Gramesp: *Una gramática de corpus para el español*. Revista de AESLA, La Rioja, España, 1998.

[5] Francis, W. N. and Henry Kučera. *Frequency Análisis of English Usage: Lexicon and Grammar*. Houghton Mifflin. 1982

[6] Gelbukh, A., Bolshakov, I., Galicia-Haro, S.N. *Statistics of parsing errors can help syntactic disambiguation*. CIC-98 - Simposium Internacional de computación, Noviembre 11 - 13, Mexico D.F., pp. 405 – 515. 1998

[7] Lara, L. F. y Ham Chande, R. *Base estadística del Diccionario del español de México*. En Lara, L. F.; Ham Chande, R.; García Hidalgo, M. I. (eds.) Investigaciones lingüísticas en Lexicografía. El Colegio de México 1979.

[8] Marcus, M., Santorini, B. and Marcinkiewicz, M. *Building a large annotated corpus of English The Penn Treebank*. Computational Linguistics 19, 2, 1993.

[9] Nañez Fernández, E. *Diccionario de construcciones sintácticas del español. Preposiciones*. Ed. de la Universidad Autónoma de Madrid, España 1995.

[10] Pedersen, T. *An Ensemble Approach to Corpus-based Word Sense Disambiguation*. Conf. on Intelligent text processing and Computational linguistics. CICLing-2000

[11] Ratnaparkhi, A. *Statistical Models for Unsupervised Prepositional Phrase Attachment*. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics. Montreal, Canada, 1998 http://xxx.lanl.gov/ps/cmp-lg/9807011

[12] Roland. D. and D. Jurafsky. *How Verb Subcategorization Frequencies are Effected by Corpus Choice*. In Proceedings International Conference COLING-ACL'98. Quebec, Canada, pp. 1122-1128, 1998.

[13] Weischedel, Ralph; Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeoe Palmucci. *Coping with ambiguity and unknown words through probabilistic models.* Computational Linguistics, 19(2): 359-382, 1993.

[14] Yeh, Alexander S., M. B. Vilain. *Some Properties of Preposition and Subordinate Conjunction Attachments.* In Proceedings International Conference COLING-ACL'98. August 10-14 Quebec, Canada, pp. 1436-1442, 1998. http://xxx.lanl.gov/ps/cmp-lg/9808007