11-712: NLP Lab Report

[Kartik Goyal –NAS]

April 26, 2013 [due date –NAS]

Abstract

[one paragraph here summarizing what the paper is about –NAS]

[brief introduction –NAS]

1 Basic Information about [Spanish –NAS]

Spanish is a Romance(Ibero romance group) language that originated in Castile region of Spain. Apart from the other Romance languages, it enriched it's vocabulary from Basque and Arabic. Spanish is closely related to other Iberian Languages like Italian and Portuguese with which, it has 82% and 89% lexical similarity respectively. Spanish is written in Latin script. The interrogative and exclamatory clauses are introduces with inverted question and exclamation marks.

It is the second most spoken language by number of native speakers and is one of the 6 official languages of the united nations. It is spoken by around 329 million people all over the world. Spanish is the primary language of 20 countries worldwide. Apart from Spain, it is official language of many Latin American regions(Argentina, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama, Uruguay, Venezuela, Peru,Puerto Rico). It is also, the official language of Equatorial Guinea in Africa.

There are some grammatical and lexical differences between the Spanish spoken in regions of Spain and the Spanish spoken in Latin America. The main grammatical variations between dialects of Spanish involve differing uses of second person pronouns. For eg. In most of Spain, 'ustedes' and 'vosotros' are used according to the degree of formality, but only 'ustedes' is used in Latin America. There are some important vocabulary differences too between the dialects.

Spanish has an S-V-O grammar and is a heavily inflected language. It has a two gender noun system and the inflections of nouns, adjectives and determiners are caused by the number and gender. There are about 50 conjugated forms per verb which are caused by tense, number, person, T-V distinctions(formal) for second person, mood, aspect and voice. Spanish has some irregular verbs with respect to the conjugation rules like 'sentir', 'requerir' etc. The modifying constituents tend to be placed after their head words. Also, usually the adjectives are placed after nouns.

Spanish is a morphologically rich language and this tool aims to carry out a resonably accurate morphological analysis of the language.

2 Past Work on the Morphology of [Your Language –NAS]

Tzoukerman and Liberman discus about a transducer which is essentially is a finite directed graph with edges labeleed by relations between surface and lexical forms.

Santiago Rodriguez and Jesus Carretero describe the formal approach behind their spanish morphological analyzer, COES.

Atserias et al. discus about the morphosyntactic analyzer which can be executed in GATE environment.

Carmona et al. discus about MACO+, a tool for morphological analysis of spanish.

3 Available Resources

The reference grammar being used for this project has two sources:

1) 'A Comprehensive Spanish Grammar'- Jacques de Bruyne 2) 'A Reference Grammar of Spanish'- R.E. Batchelor, Miguel Angel San Jose

The corpora had to be the textual data from Mexico. It has been built from the Judicial weekly and its Gazette provided on the official site of 'The Supreme court of Justice of the nation' from mexico :'http://www.scjn.gob.mx/Paginas/Inicio.aspx'. The file used for building the corpus was the 'September 2011' issue of the gazette.

These judicial proceeding contain clean data and is dominated by the formal official language. But, this data also contains quoted statements by authorities and people involved in the proceedings. Hence, some personal words are also expected in the corpus. All the numbers and words containing single letters like 'a' were removed as they are not interesting from the morphological point of view.

A total of 23,000 types were found in the corpus. Corpus A and Corpus B have 1000 types each. The test corpus has 10,000 words.

This corpus is expected to be a fairly clean and diverse because apart from being an official record, it contains people's personal statements and opinions.

4 Survey of Phenomena in Spanish

Spanish is a morphologically rich language. The verbs, nouns, pronouns and adjectives demonstrate inflections.

VERBS:

Handling the verb inflection is extremely important because they demonstrate extremely rich inflectional phenomena. The verbs undergo inflection according to following categories:

Tense: Past, Present , Future

Number: Singular, Plural

Person: First, Second, Third

Mood: Indicative, Subjunctive, Imperative

Aspect: Perfective aspect, Imperfective aspect

Voice: Active, Passive

Regular verbs:

The regular verbs can be classified into 3 categories based on their endings:

-ar ended- Most frequent

-er ended- Fairly frequent

-ir ended- Least Frequent

All the regular verbs are conjugated in a standard manner. All of the following phenomena inflect the verbs according to 3 Person X 2 Number ways:

PRESENT INDICATIVE- The second person plural ending has a written accent.

IMPERFECT INDICATIVE- 'er' and 'ir' ended verbs are inflected in the same way.

PRETERITE- 'er' and 'ir' ended verbs are inflected in the same way.

FUTURE INDICATIVE- 'ar', 'er and 'ir' are inflected in the same way. Second person plural ending has a written accent.

CONDITIONAL- 'ar', 'er and 'ir' are inflected in the same way.

PRESENT SUBJUNCTIVE-'er' and 'ir' ended verbs are inflected in the same way. The second person plural ending has a written accent.

IMPERFECT SUBJENCTIVE- It has two forms and both the forms are popular for all the number and person combinations. 'er' and 'ir' ended verbs are inflected in the same way.

FUTURE SUBJUNCTIVE- It is extremely rarely used.

THE IMPERATIVE- The imperative has distinct forms only in the second person singular and plural.

The rest of the phenomena which can be classified as Compound tenses of the indicative are formed from the auxiliary verb 'haber'(to have)+past participle. These phenomena are:

PERFECT

PLUPERFECT

FUTURE PERFECT

CONDITIONAL PERFECT

PERFECT SUBJUNCTIVE

PLUPERFECT SUBJUNCTIVE

PERFECT INFINITIVE

These phenomena are actually not needed to be analysed from the Morpholofical point of view as only the auxiliary verb 'Haber' which has an 'er' ending inflects in these cases and the rules to cover inflection of 'er' ended verbs will be implemented in the analyzer.

Irregular verbs have slightly different inflection rules than those of the regular verbs. These irregular verbs are some of the most common verbs in Spanish. These verbs are 'Haber', 'Tener', 'Ser' and 'Estar'.

Apart from the verbs above, some verbs change the stem's spelling. In a number of words, spelling of stems ending in 'c','g','gu','qu' and 'z' has to change before ending beginning with 'e' in order for the pronunciation of the stem to be preserved

NOUNS

The nouns inflect according to 2 genders X 3 Person cases. Nouns ending in -o are masculine, with the only notable exception of the word mano ("hand"); -a is typically feminine, with notable exceptions.

A small set of words of Greek origin and ending in -ma, "-pa", or "-ta" are masculine. Many nouns have same spellings for both female and male forms, thereby, making a deterministic conclusion about the gender pf a noun difficult.

Plurals of nouns in Spanish are formed in the following way:

When the noun ends in an unstressed vowel, '-s' is added.

When the noun ends in a consonant, a stressed vowel '-es' is added.

Nouns ending in 'z', change the 'z' to 'c' before 'es' in plural.

ADJECTIVES

Adjectives too inflect according to 2 genders X 3 persons. The adjectives ending in 'o' are male and their female forms end in 'a'.

Adjectives ending in -ete,-ote are male and their female forms have the 'e' replaced by 'a'.

Adjectives ending in '-an', '-in' and '-on' are male and their female forms have a appended at the end.

The adjectives inflect with number in exactly the same way as nouns inflect with number.

The adjectives also inflect according to the degree of comparison.'-isimo' at the end and 're-','rete-', and 'requte-' at the beginning are used for expressing a superlative quality.

ARTICLES

Both Definite and Indefinite articles inflect with 2 gender X 3 Person. They have a fixed form without any variations.

However 'de' and 'el' are combined to 'del' and 'a' and 'el' are combined to 'al'.

PRONOUNS

They are seldom used but they too have fixed inflections with 2 gender X 3 Person

Finally some adverbs are generated from adjectives by aadding a suffix to them. This is a very common phenomenon.

5 Initial Design

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work