COMPUTATIONAL LINGUISTICS FINAL PROJECT REPORT

TEDDY HEIDMANN
CORNELL UNIVERSITY
ATH55@CORNELL.EDU

Abstract

The project was to write a parser that recognizes the correct order of adjectives for both English and Spanish. For this, we can construct two grammars (one for each language), and write parsing rules for each grammar, as well as a variety of sentences to test with. Spanish has additional challenges such as gender agreement, the use of a coordinating conjunction when multiple adjectives are used, and a slightly different adjective order. Combining rules both languages, we can see the importance behind this phenomenon, and some considerations when scaling to larger grammars or multiple languages.

1. Problem

All of the Computational Linguistic assignments thus far have been dealing with English grammars that include nouns, verbs, and determiners. This is extremely limiting, as there are many more parts of speech in the English language than these three basic forms. This project extends a parser to include another of these parts of speech: adjectives. Specifically, it looks at the inclusion of adjectives in grammar. The final parser checks for the correct order of adjectives according to mnemonic OSASCOMP (opinion, size, age, shape, color, origin, material, purpose). As an example, in English we would say `the nice little old red brick house` rather than `the red old brick little nice house`. This second sentence does not make sense, and you might interpret brick as a different noun, rather than an adjective for house.

For the Spanish language, almost the same mnemonic is used, but the adjectives follow the noun. For example, instead of saying `the little red book`, you would say `el libro pequeño y rojo`, meaning "the book that is small and red". However, the rules are a little less restrictive in Spanish, and the mnemonic is only the suggested order, not the required one. As an example, `el libro rojo y pequeño` would also be a valid sentence, though most native speakers would still opt for the default order. For the purpose of this assignment, we will make the generalized assumption that every sentence must use the default order, since that is what is grammatically preferred.

2. Origin of OSASCOMP

While there is no exact origin for the OSASCOMP ordering of adjectives, a number of sources reference a passage from a book entitled: "The elements of eloquence: How to turn the perfect English phrase". In this, the author states "*…adjectives in English absolutely have to be in this order: opinion-size-age-shape-colour-origin-material-purpose Noun. So you can have a lovely little old rectangular green French silver whittling knife. But if you mess with that word order in the slightest you'll sound like a maniac. It's an odd thing that every English speaker uses that list, but almost none of us could write it out.*" [2].

Another site references a very similar mnemonic, but in a slightly different order. It

Date: Spring 2017.

uses the letters GSSSACPM instead of OSASCOMP [5], which has three (3) main differences. First, it breaks down 'Opinion' into general and specific opinions such as 'terrific' and 'slimy'. It also switches the order of age and shape, meaning the sentence `the round old window` would be valid, rather than `the old round window`.

While the article attributes these first two differences to linguistic choice, it does reference the third difference: lack of the word 'purpose' (in the new mnemonic, 'P' stands for provenance, i.e. origin). Seeing as the purpose adjective would not fall into any other category of adjectives, one assumption is that this adjective becomes part of the noun itself. Using the example mentioned above, the GSSSACPM mnemonic would classify 'reference' as part of the object, rather than an adjectival modifier. We could still add an adjective phrase to describe this new object in a similar fashion, but 'reference' would not be considered part of the AP.

Seeing as the OSASCOMP mnemonic covers all of the same categories as GSSSACPM, adds the purpose adjective, and is referenced much more often in articles [1][3][4], this project focuses solely on writing a parser for the OSASCOMP structure. In order to change from the selected mnemonic to the GSSSACPM version, only a slight modification is required to the code of the parser. The purpose adjective would become an optional part of the adjective phrase, and opinion would be split into two new categories: general and specific. Again, since these changes have been referenced as individual choices between linguists, this project focuses on writing a valid parser for OSASCOMP.

3. English Grammar

Starting with the strictest language, we can create a set of rules and a parser for English in order to recognize the OSASCOMP order of adjectives. The complete grammar for English is shown in Figure 1, which was used to generate the trees in Figure 2 and Figure 3. Rather than creating one rule for adjective phrases that list all categories of adjectives as optional additions to the rule, the decision was made to leave everything as binary, making it much clearer that the adjectives have to appear in the given order. The binary rules also simplify the logic of the code, making it much easier and cleaner to read.

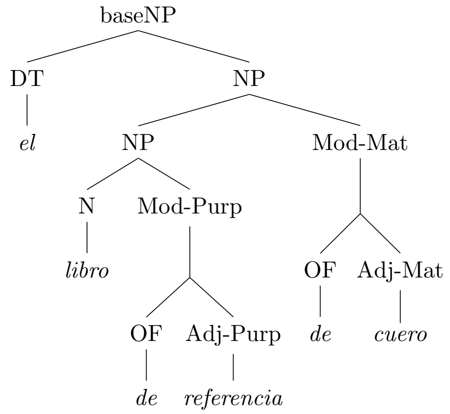
As seen in the resultant trees, we can still decipher the hierarchy of the sentence, and they are not overly cluttered by optional adjective categories that do not appear. In Figure 3, the structure is very linear, and it is clear that all of the adjectives fall under the 'AP' heading with no special modification or change in location. Switching to Spanish, we will see that this is not quite the same.

4. Spanish Grammar

While the Spanish language uses the same categories of adjectives as English does, there are a few main differences. As mentioned previously, the adjectives come after the noun phrase. In addition to this, Spanish also has the added complexity of gender agreement, and using coordinating conjunctions such as 'and' to combine multiple adjectives together. The parser handles these correctly, and ensures that all determiners, adjectives, and nouns must agree in gender. If multiple adjectives are used, the word 'y' must come before the final descriptor, resulting in valid parses.

The complete grammar for Spanish is shown in Figure 4, which was used to generate the trees in Figure 5 and Figure 6. In this grammar, masculine and feminine words were combined using '@', which should be replaced with 'o' for masculine and 'a' for feminine. This is also discussed further in subsection 5.1. Also, for simplicity (and because the OCaml makes dealing with special characters difficult) all accents have been removed.

4.1. **Placement of material and purpose adjectives.** In Spanish, while the same categories of adjectives as English are used, the material and purpose adjectives behave slightly differently. They appear as part of the noun phrase rather than adjective phrase. For example, `the leather reference book` translates to `el libro de referencia de cuero`, which is literally `the book of reference of leather`'. We notice that the purpose adjective is closest to the noun, followed by the material adjective, and can visualize this as in the diagram below.



In this example, the word ‘reference’ describes the book, and ‘leather’ describes the ‘book of reference’. This demonstrates a much deeper semantic description than the English translation, ‘leather reference book’, as this tells us that the book is first and foremost a book intended for reference. While it could be argued that English has this same semantic structure, it is not explicitly written in the sentence. Spanish makes it much easier to see the overall structure and not be confused if ‘leather’ is one object and ‘reference book’ is another.

4.2. **Gender agreement and multiple adjectives.** In the final parser, additional boolean parameters were used to dictate whether the determiner was masculine or not, and if an adjective came before the current adjective. For example, the definition for ‘aCO’, the adjective phrase for the part of the adjective phrase that comes after the shape adjective (with just color and origin remaining) is defined as:

```
aCO m a words =
    let color = if m then
        adjCol_m
    else
        adjCol_f in

    ((aO m a)
    |.  ( if a then
            ((y &. color)
                >. binary "AP")
        else
            color
        )
    |. ((color &. (aO m true))
        >. binary "AP")
    ) words
```

Notice the two additional parameters, `m` and `a`. The first is whether the determiner is masculine or not. Looking at this in the code, we reference two arrays, one for masculine and one for feminine color adjectives. The word must match one of the words in that array, so if the gender does not agree, the sentence is not valid.

We then use the second parameter to determine whether we need to add a coordinating conjunction before the terminal adjective. If there are more adjectives in the list, we can pass the same parameter along to the next part of the adjective phrase (ex: `aO m a`), and postpone the decision of whether this is a valid sentence or not. We can perform a similar action if the word matches the current category, but if it is the last adjective, then we use `a` to decide if the word ‘y’ should be added as a conjunction.

The rest of the definitions are very similar to the above one, with the only differences being the material and purpose categories, which are included with the noun phrase rule. Due to these slight differences, the same parsing algorithm as English was used, and the majority of the work was spent ironing out the changes with Spanish gender, modifiers, and plurality.

5. Future Improvements

Given more time, this project could be improved in numerous ways. Restricting it to the order of adjectives, there are a couple possibilities for extension: (1) automated gender matching, and (2) ability to import a grammar

5.1. **Automated gender matching.** One improvement to this project could be automated gender matching. Currently, there are two arrays for each category of adjectives: one for masculine and one for feminine. This adds the extra step of having to decide which array to use, and some words such as ‘grande’ appear in both lists. In addition, most of the words that differ between the lists differ in only one letter: the ‘o’ or ‘a’ at the end of the word. There could be an additional function that takes all but the last letter of the word along with whether the determiner is masculine or feminine, and adds the correct ending. Computationally, this adds an extra step to the calculation of each word, but would be worthwhile and take less space in memory for the duplicate entries in the masculine/feminine arrays.

Another option would be to integrate with FOMA in order to avoid repeating words twice. In this case, the individual words would be looked up in FOMA to see if their adjective categories and endings match up with the current parsing rule and gender agreement.

5.2. **Ability to import a grammar.** Another improvement would be to have an external file that contains the grammar rules, and an additional file for words. The grammar file would be imported each time, and quickly adapted for multiple languages, so extensions to more than English and Spanish could be easily adapted to. The other file would contain a list of words and their corresponding parts of speech categories. The addition of this file would require a bottom-up parsing algorithm rather than the current top-down version, but would allow a much quicker dictionary lookup to figure out the types of each words for larger dictionaries. Another option would be to use FOMA to figure out the gender and parts of speech of each word, then use these facts to determine which rules to apply to the sentence fragment.

6. Personal Reflection

Throughout the course of this project, I was able to learn a lot about the structure and purpose of adjectives. It surprises me that there is no specific origin to the OSASCOMP mnemonic, and that tens of websites and articles I found do not cite where they originally retrieved the information from. This was incredibly frustrating from a research standpoint, and even more so that I was unable to find a reputable source for the Spanish language. The majority of sources that I found were discussion boards where users referenced the order in English. While all of them agreed on the same adjective categories, the rules that I settled on were from consulting with a number of Spanish-speaking friends, a native speaker, and two online translation systems (Google and Online-Translator.com).

References

[1] Ordering multiple adjectives — english grammar guide.

[2] Mark Forsyth. *The elements of eloquence: How to turn the perfect English phrase.* Icon Books Ltd, 2013.

[3] Jespersen Otto. Language, its nature, development and origin, 1922.

[4] BBC Trending. Why the green great dragon can't exist, Sep 2016.

[5] Katy Waldman. Why do we say “big red barn,” but “red big barn” sounds wrong?, Aug 2014.

List of Figures

S	→	NP VP
NP	→	DT *adjNP*
VP	→	V
adjNP	→	(AP) N
AP	→	A-OSASCOMP
A-OSASCOMP	→	(Adj-Op) (*A-SASCOMP*)
A-SASCOMP	→	(Adj-Size) (*A-ASCOMP*)
A-ASCOMP	→	(Adj-Age) (*A-SCOMP*)
A-SCOMP	→	(Adj-Shape) (*A-COMP*)
A-COMP	→	(Adj-Col) (*A-OMP*)
A-OMP	→	(Adj-Orig) (*A-MP*)
A-MP	→	(Adj-Mat) (*A-P*)
A-P	→	Adj-Purp
DT	→	[the, a, one, his, her]
N	→	[book, dictionary, tower]
V	→	[fell, burned, rang]
Adj-Op	→	[ugly, beautiful]
Adj-Size	→	[large, small, big]
Adj-Age	→	[old, young, ancient]
Adj-Shape	→	[round, square]
Adj-Col	→	[brown, red]
Adj-Orig	→	[english, cornellian]
Adj-Mat	→	[leather, stone]
Adj-Purp	→	[announcement, reference]

Figure 1. Grammar for the English language parser

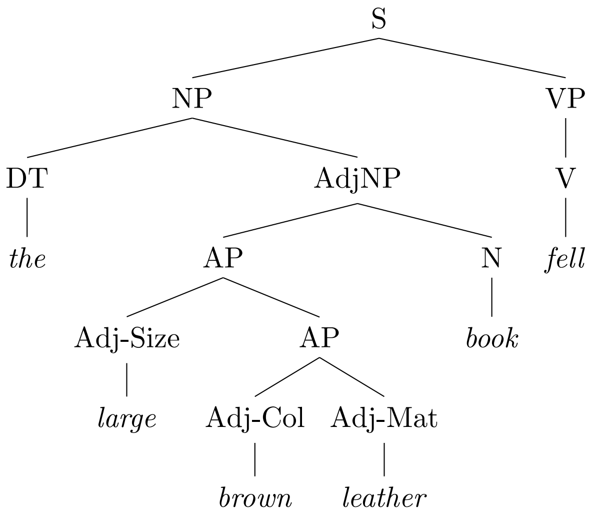


FIGURE 2. Tree for the English sentence ‘the large brown leather book fell’

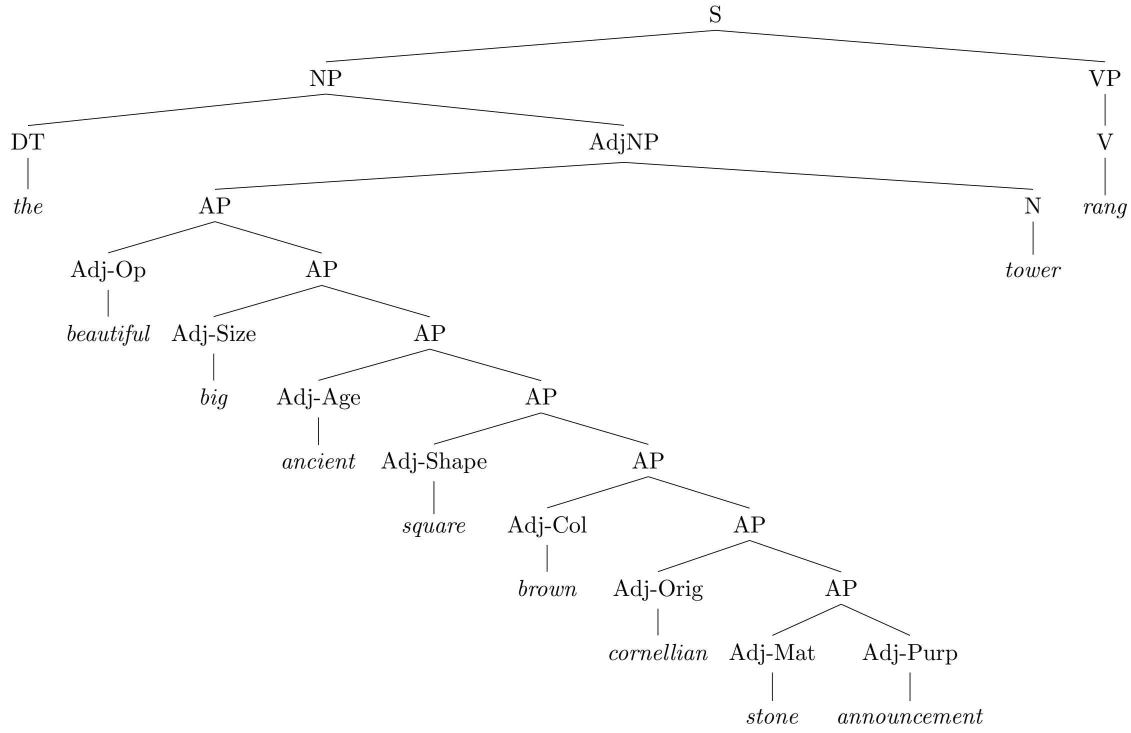


FIGURE 3. Tree for the English sentence 'the beautiful big ancient square brown cornellian stone announcement tower rang'

S	→	baseNP VP
baseNP	→	DT NP (*A-OSASCO*)
VP	→	V
NP	→	*nPurp* (DE Adj-Mat)
nPurp	→	N (DE Adj-Purp)
A-OSASCO	→	(Adj-Op) (*A-SASCO*)
A-SASCO	→	(Adj-Size) (*A-ASCO*)
A-ASCO	→	(Adj-Age) (*A-SCO*)
A-SCO	→	(Adj-Shape) (*A-CO*)
A-CO	→	(Adj-Col) (*A-O*)
A-O	→	Adj-Orig
DE	→	de
DT	→	[el, la, un, una]
N	→	[libro, plato, coche, biblioteca, torre, mesa]
V	→	[quemado, cayo, condujo, sono]
Adj-Op	→	[fe@, bonit@]
Adj-Size	→	[grande, pequen@, larg@]
Adj-Age	→	[viej@, joven, antigu@]
Adj-Shape	→	[redond@, cuadrad@]
Adj-Col	→	[marron, roj@, negr@]
Adj-Orig	→	[chin@, cornellian@]
Adj-Mat	→	[cuero, piedra]
Adj-Purp	→	[golf, referencia, anuncios]

Figure 4. Grammar for the Spanish language parser

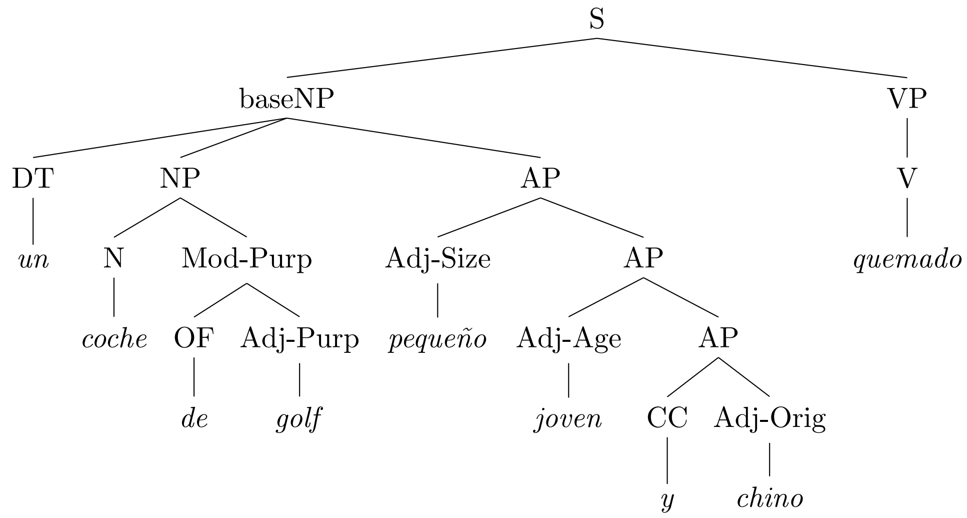


Figure 5. Tree for the Spanish sentence ‘un coche de golf pequeño joven y chino quemado’

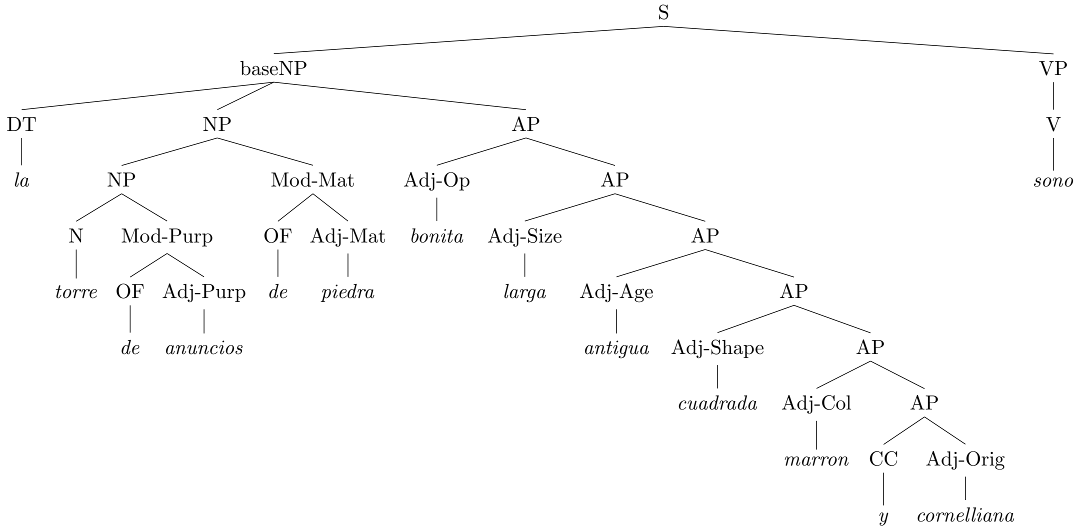


FIGURE 6. Tree for the Spanish sentence 'la torre de anuncios de piedra bonita larga antigua cuadrada marron y cornelliana sono'