Subject Expression on Social Media Platforms: A Reddit Corpus Analysis – A Brief Overview in English

Dana Schwarz

1 Introduction

As part of a linguistic seminar, I created a corpus using Python and the NLP library SpaCy to analyze texts from Reddit. The primary goal was to investigate how various factors influence subject expression in Spanish and whether social media has an impact on this.

2 Short summary of the theory

2.1 Null subject languages

Unlike English or German, Spanish is considered a "null subject" or "pro-drop" language, meaning that the expression of pronominal subjects is optional. For example, the sentence "I eat an apple" can be expressed in Spanish as either *"Yo como una manzana"* (with the pronoun) or *"Como una manzana"* (without the pronoun).

2.2 Social media and communication

Social media has transformed the way we communicate online, which has given rise to a new linguistic dynamic, where written language adopts a more colloquial form, similar to orality. Spelling changes have been noted, such as abbreviations or simplifications of words, as well as effects on the lexicon and morphosyntax.

3 Creation of the corpus

3.1 Preparation of the data

The corpus was built using texts from the Reddit communities r/argentina, r/Espana, and r/mexico, which represent some of the largest Spanish-speaking groups on the platform. For each country, five posts without images or videos were randomly selected, ensuring that the discussions focused entirely on written content. The posts were categorized into Politics, Economy, Health, Love/Relationships, Education/Work, Experiences, Humor, and Media. Additionally, between 2 and 5 comments were chosen for each post. As a result, 24 to 26 short texts were gathered per country for analysis using computational tools.

3.2 Tools used to write the programs

To process the data, I saved the selected texts as text files and analyzed them with programs written in Python and SpaCy. I employed SpaCy's largest Spanish language model, *es_dep_news_trf*, as it produced more accurate results than the smaller models,

and speed was not a priority. With the help of the pandas library, I exported the results into Excel files for further examination.
The main focus was to extract the verbs from the texts and check whether they had associated subjects. Additionally, I aimed to examine other linguistic features, such as, for example, verb tense and whether the subject was a pronoun or a noun.

3.3 Functions of the programs

I developed two programs for this analysis.
The first program detected and counted all verbs in the text, listing them by frequency. It provided the number of verb occurrences, nominative subjects, and null-subjects (cases where the subject was omitted). The program also gave an overall count of verbs and null-subjects, offering a general overview of subject omission.
The second program also identified verbs and their subjects but offered more detailed information. It analyzed each verb's lemma, type (whether a main or auxiliary verb), number, person, tense, mood, and subject expression. If a subject was present, it determined whether it was a pronoun or a noun. The program then listed these features alongside the original inflected verb and the sentence in which it appeared.

4 Conclusion

Based on my data, I found that pronominal subject expression tends to be reduced on social media platforms. When pronominal subjects were expressed, it was primarily in cases where the author wanted to emphasize their opinion, highlighting their subjective perspective.
However, the results also contained errors, largely due to typos or other user mistakes. Additional challenges arose in the classification of verb tenses, which stemmed from both the ambiguity of Spanish verb suffixes and the limitations of the Spanish language model in detecting tenses accurately. Overall, SpaCy seemed highly optimized for English, but its performance in Spanish was less robust.