Project 1: Conversational Spanish Corpus

COMP-551: Applied Machine Learning

Jonathan Adalin
ID: 260636904
Email: jonathan.adalin@mail.mcgill.ca

Ryan Burgett
ID: 260615904
Email: ambrose.burgett@mail.mcgill.ca

Alice Scott
ID: 260631443
Email: alice.scott@mail.mcgill.ca

I. CORPUS URL

http://cs.mcgill.ca/~ascott41/adalin-burgett-scott_spa.xml

II. INTRODUCTION

The corpus contains 17054 dialogs. The sources are all formatted such that there is a main post or video, and a comment section in response. Consequently, any interaction between comments and their replies is considered a dialog. While Spanish is fairly standardized, there are certain regional variations in spelling. In an effort to include these variations, we included a variety of geographically related forums ranging from Latin America to Spain. When there was no indication of where the data originated, it was presumed to come from multiple regions.

III. DATASET DESCRIPTION BY SOURCE

A. Apple Forums

4303 dialogs were collected from the Spanish Regional Apple Forums using the BeautifulSoup API for Python. These dialogs were collected between the range of the below URLs:

- https://communities.apple.com/es/thread/160030016
- https://communities.apple.com/es/thread/160025017

Apple Forums is a support community where users can ask and answer questions and engage in discussions related to Apple products. This community website was chosen as a data source because it has thousands of threads made publicly available and because of the regional seperation it guaranteed a low probability of non-Spanish dialog appearing in the dataset. Consequently, the data collected from this source contains techonology-focused vocabulary and are of a question-answer format. However, each dialog is a great representation of the formulation of questions and answers in Spanish making it a useful addition to our dataset. Each utterance ranges from a couple sentences to a couple paragraphs in both the question and the answer turns.

The user who posted the question to the forum would rarely respond to the user who provided the answer. Rather they would rate the given answers and pins the best response. Due to this system it made sense to extract just the initial question and best response from each thread. Thus, each dialog only contains only two uids and is exactly two turns long.

B. YouTube

An additional 4726 dialogs were collected from YouTube. More specifically, 9 videos were selected and their comments extracted using an online scraper found on ytcomments.klostermann.ca. The topics of these videos ranged from international news to documentaries and their links can be found in the readme found in the source code section.

In order to find videos that had comments exclusively in Spanish, our search process included changing the country within the user profile to Spain. For this reason, one could assume the dialogues pulled from this source have geographical roots leaning towards that of Spain. However, as previously stated, it is impossible to guarantee this presumption without individually checking each user profile.

The dialogues from this source are of the form comment and reply, which means conversations are limited to two turns. This call-response format was chosen because of the difficulty of retrieving properly ordered recursive replies in YouTube's current comment system. Although this dialogue size limitation is not idea for a rich corpus, in the next section we describe a reference where longer discussion chains are easily retrievable.

C. Reddit

Another 8025 dialogues were collected from Reddit using PRAW (Python Reddit API Wrapper). The top 1000 posts were pulled from the following subreddits:

- /r/mexico
- /r/es
- /r/espanol
- /r/argentina
- /r/redditores
- r/programacion

Each top-level comment was treated as start of a new dialog in order to minimize the inclusion of the same data twice, since Reddit allows for nested comments. The API stores comments in a tree-like structure and therefore if the comment was not a leaf and its child did not have the same uid associated with it, the comment was included in the corpus since it consisted of at least a two-person dialog. Any reply further nested in the structure was considered to be part of the same dialog as its root. Usernames were assigned a uid and stored in a dictionary. The dialogs collected from Reddit is much more irregular than the previous two sources and widely vary in terms of number of turns. The number of uids per dialog varies between 2 and 35. Lengths of turns also vary from a single word to multiple paragraphs.

One complication encountered with data collected from Reddit was the presence of other languages, mainly English, in the dialogues. This was particularly prevalent in Latin American subreddits such as /r/Mexico where many conversations contain utterances in both English and Spanish. We discuss our decision to include these non-Spanish utterances in the corpus in the following section.

IV. KEY CHARACTERISTICS

The data from our sources contains a large quantity of improper Spanish words due to the frequent use of internet slang and other common internet conventions.

TABLE I. INTERNET SLANG OCCURANCES BY WORD

Word (case-insensitive)	Number of Occurrences
lol	410
lmao	24
haha	247
:)	451
;)	133
wtf	24
eli5	9
jaja	569
omg	24

Punctuation is treated informally as a way to express emotion online (for example, the use of '?!' for disbelief, '!!!!!' for emphasis, or more than three periods in an ellipsis for added suspension). In a similar manner, certain utterances have every character capitalized to make a point come across.

As stated in the previous section, a common difficulty encountered during data collection was the presence of utterances in languages other than Spanish. While the data from YouTube and Apple Forums is fairly well sanitized, the data from Reddit still contains multiple instances of non-Spanish languages. We decided to keep these dialogues in the corpus because they present an interesting aspect of the effect of the internet on language: interactions between people across the globe are now vastly easier than before, and online translation services make it possible to hold a conversation with someone speaking a completely different language.

FIGURE 1. EXAMPLE OF MULTILINGUAL DIALOG

```
<s>
    <utt uid="1">
         My Iphone takes a long time to sync with iTunes, it
         stays in "PREPARING THE COPY OF ITEMS"Can you help me
         ?What i can do ?Thank you.
</utt>
<utt uid="2">
         Hola chechio!!Quizás la comunidad de soporte de Apple en
         inglés sería la más adecuada a tu idioma De todas
         formas, echa un vistazo a esta web de Apple: Cómo
         solucionar los problemas entre iTunes y el software de
         seguridad de terceros - Soporte técnico de Apple. Hay
         instrucciones para intentar solucionar problemas de
         sincronización con iTunes.Principalmente, te recomiendo
         que tengas actualizado tanto el ordenador como iTunes y
         que desactives el antivirus, si usas un PC, mientras
         trabajes con el iPhone conectado.Prueba y nos
         cuentas.Saludos!!
    </utt>
</s>
```

Other difficulties included the common use of hyperlinks in forums, as well as bot scripts in subreddits that overwrote comments, essentially rendering a conversation one-sided and therefore nonsensical.

These deviations from standard Spanish are what distinguish the corpus we collected from previously existing Spanish corpa. The Spanish corpa available mostly focus on the collection of a big lexicon, and incorporate resources from publications. The *Spanish Billion Words Corpus and Embeddings* includes translated documents coming from Wikipedia, United Nations documents, books, and news commentary [1]. Similarly, the *Corpus Del Español* and Spanish corpus by *Corpora From the Web (*assembled by Freie Universität Berlin) contain data collected from full documents and are preprocessed (*Corpus Del Español* is tokenized, lemmatized, and stemmed [2], and *Corpora From the Web* is processed into bags of sentences [3]). There's a particular focus among the current Spanish corpa on historical changes and regional dialect variations. In contrast to this, our dataset is a representation of colloquial Spanish and focuses on the ways people express themselves in the context of the Internet.

On a statistical level, our corpus contains utterances in which the word count is more likely to be lower (less than 20 words). Additionally, the most popular words and word pairings are predominantly prepositions.

FIGURE I. NUMBER OF WORDS PER UTTERANCE

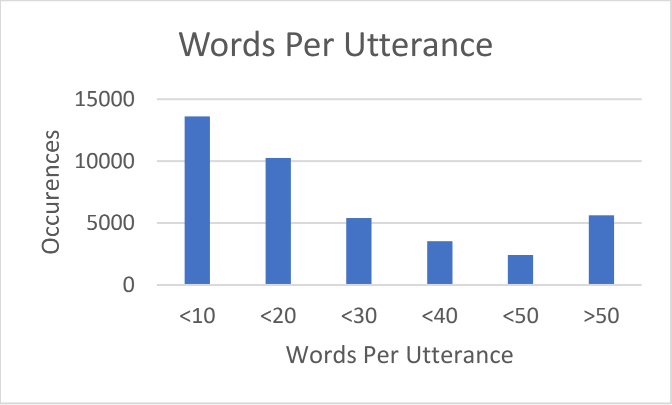


FIGURE II. MOST COMMON WORDS

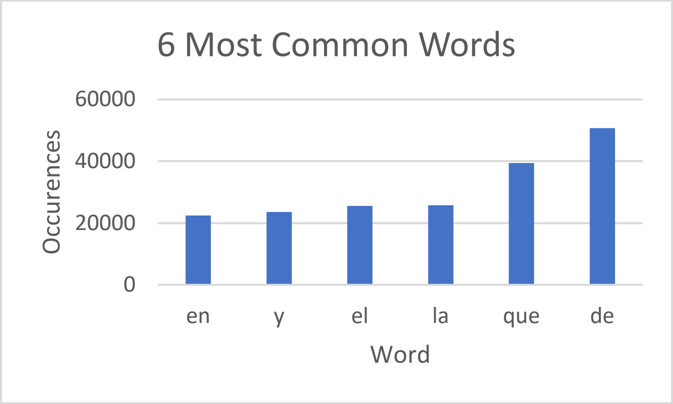
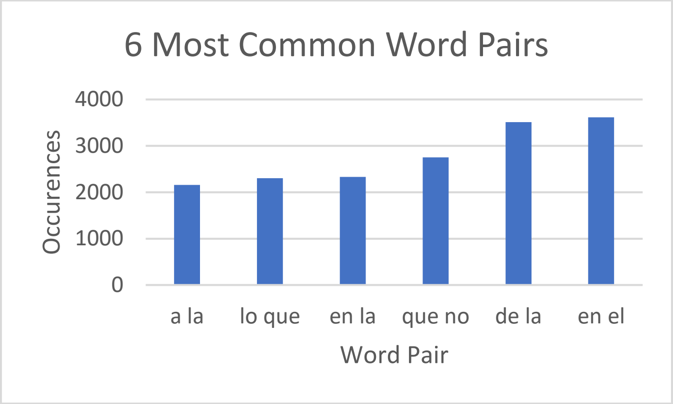


FIGURE III. MOST COMMON WORD PAIRS



V. SOURCE CODE

https://github.com/JonathanAdalin/comp551-project1

VI. STATEMENT OF CONTRIBUTIONS

A. Jonathan Adalin

Wrote a script for YouTube data collection. Additonally, created graphs and contributed towards writing the report.

B. Ryan Burgett

Wrote a script for Apple Forums data collection, as well as creating and contributing towards writing the report.

C. Alice Scott

Wrote a script for Reddit data collection and set up the webpage to host the corpus. Contributed towards writing the report.

We hereby state that all the work presented in this report is that of the authors.

VII. REFERENCES

[1] http://crscardellino.me/SBWCE/

[2] http://www.corpusdelespanol.org/size.asp

[3] http://corporafromtheweb.org/category/corpora/spanish