Spanish-English Code-switching Annotations for Twitter

1. WORD-LEVEL ANOTATIONS

Tokens that start with a @ character, urls, emoticons or any token that does not contain any letters such as punctuation marks and numbers (examples: ♥, !, -. , •• >, @____)), and the {symbol} tokens should all be labeled as ‘None of the above’.

If a number represents a word in the sentence it should be labeled as the language of that word instead of ‘None of the above’. An example is ‘I like 2 party.’, but not ‘Meet me in 2 hours.’

For tokens beginning with a # tag consider them as a single token and label them according to the regular word level guidelines.

1.1. Language

For each word in the Source, identify whether it is Spanish, English, Mixed, Other, Ambiguous, or NE (for named entities, which are proper names that represent names of people, places, organizations, locations, movie titles, and song titles). Below is an example showing the correct tags (labels) for each token in the source.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>English</td>
<td>Tuesdays</td>
<td>English</td>
</tr>
<tr>
<td>always</td>
<td>English</td>
<td>Around</td>
<td>English</td>
</tr>
<tr>
<td>tell</td>
<td>English</td>
<td>6</td>
<td>None of the above</td>
</tr>
<tr>
<td>him</td>
<td>English</td>
<td>pero</td>
<td>Spanish</td>
</tr>
<tr>
<td>to</td>
<td>English</td>
<td>it</td>
<td>English</td>
</tr>
<tr>
<td>sing</td>
<td>English</td>
<td>'s</td>
<td>English</td>
</tr>
<tr>
<td>to</td>
<td>English</td>
<td>not</td>
<td>English</td>
</tr>
<tr>
<td>me</td>
<td>English</td>
<td>worth</td>
<td>English</td>
</tr>
<tr>
<td>pero</td>
<td>Spanish</td>
<td>it</td>
<td>English</td>
</tr>
<tr>
<td>nunca</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quiere</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ambiguous words

Ambiguous words are words that, in context, could belong to either language. This can happen because words such as red, a, doctor, me, and can are valid words in both languages. However, every instance of such a word is not ambiguous – only those instances where there is not enough context to decide whether the word is being used as English or Spanish. The fragment on the left shows an example where a potentially ambiguous word, me, is not ambiguous because the context helps identify the language, while the example on the right shows a truly ambiguous word, NO, which could be in either English or Spanish. Note that typos and misspellings should be labeled with the corresponding language.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>English</td>
<td>Johnny</td>
<td>NE</td>
</tr>
<tr>
<td>always</td>
<td>English</td>
<td>Depp</td>
<td>NE</td>
</tr>
<tr>
<td>tell</td>
<td>English</td>
<td>para</td>
<td>Spanish</td>
</tr>
<tr>
<td>him</td>
<td>English</td>
<td>Dr.</td>
<td>NE</td>
</tr>
<tr>
<td>to</td>
<td>English</td>
<td>Strange</td>
<td>NE</td>
</tr>
<tr>
<td>sing</td>
<td>English</td>
<td>?..</td>
<td>None of the above</td>
</tr>
<tr>
<td>to</td>
<td>English</td>
<td>NO</td>
<td>Ambiguous</td>
</tr>
<tr>
<td>me</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pero</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nunca</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
<tr>
<td>quiere</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mixed words
Mixed words are words that are partially in one language and partially in another. This can occur when the first part of a word is in English and the second part is in Spanish, or vice versa. The mixed
category should only be used if the word clearly has a portion in one language and another portion in a
different language. It is not for words that could exist entirely in either language (see Ambiguous).

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ayy</td>
<td>Spanish</td>
</tr>
<tr>
<td>que</td>
<td>Spanish</td>
</tr>
<tr>
<td>pepe</td>
<td>NE</td>
</tr>
<tr>
<td>snapchateame</td>
<td>Mixed</td>
</tr>
<tr>
<td>el</td>
<td>Spanish</td>
</tr>
<tr>
<td>arreglo</td>
<td>Spanish</td>
</tr>
</tbody>
</table>

Named Entities (NE)

This is a difficult section. Please read carefully. NEs are proper names. Examples of NEs are names
that refer to people, places, organizations, locations, movie titles, and song titles. Named entities are
usually, but not always, capitalized, so capitalization can’t be the only criterion to distinguish them.

Named entities can be multiple words, including articles (see the examples). Examples of NEs and
their tags are shown below.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mejor</td>
<td>Spanish</td>
<td>and</td>
<td>English</td>
<td>@username</td>
<td>None of the above</td>
</tr>
<tr>
<td>Vente</td>
<td>Spanish</td>
<td>I</td>
<td>English</td>
<td>it</td>
<td>English</td>
</tr>
<tr>
<td>para</td>
<td>Spanish</td>
<td>told</td>
<td>English</td>
<td>‘s</td>
<td>English</td>
</tr>
<tr>
<td>el</td>
<td>Spanish</td>
<td>her</td>
<td>English</td>
<td>on</td>
<td>English</td>
</tr>
<tr>
<td>West</td>
<td>NE</td>
<td>to</td>
<td>English</td>
<td>telemundo</td>
<td>NE</td>
</tr>
<tr>
<td>Coast</td>
<td>NE</td>
<td>record</td>
<td>English</td>
<td>el</td>
<td>NE</td>
</tr>
<tr>
<td>and</td>
<td>English</td>
<td>La</td>
<td>NE</td>
<td>seño</td>
<td>NE</td>
</tr>
<tr>
<td>visit</td>
<td>English</td>
<td>Reina</td>
<td>NE</td>
<td>de</td>
<td>NE</td>
</tr>
<tr>
<td>me</td>
<td>English</td>
<td>del</td>
<td>NE</td>
<td>los</td>
<td>NE</td>
</tr>
<tr>
<td>lol</td>
<td>English</td>
<td>Sur</td>
<td>NE</td>
<td>cielos</td>
<td>NE</td>
</tr>
</tbody>
</table>

Abbreviations

Abbreviations should be labeled according to the full word(s) they represent. Some examples are
shown below.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr.</td>
<td>English</td>
<td>lol</td>
<td>English</td>
<td>jajaja</td>
<td>Spanish</td>
</tr>
<tr>
<td>Smith</td>
<td>NE</td>
<td>yeah</td>
<td>English</td>
<td>ntc</td>
<td>Spanish</td>
</tr>
<tr>
<td>was</td>
<td>Spanish</td>
<td>I</td>
<td>English</td>
<td>gracias</td>
<td>Spanish</td>
</tr>
<tr>
<td>quejandose</td>
<td>Spanish</td>
<td>hear</td>
<td>English</td>
<td>por</td>
<td>Spanish</td>
</tr>
<tr>
<td>como</td>
<td>Spanish</td>
<td>you</td>
<td>English</td>
<td>todo</td>
<td>Spanish</td>
</tr>
<tr>
<td>siempre</td>
<td>Spanish</td>
<td>wey</td>
<td>Spanish</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Other

Languages other than Spanish or English should be labeled as Other. This category includes gibberish
and unintelligible words. The example on the left shows some content that is not in English or Spanish
(it is in Portuguese). The example on the right is an example of gibberish.

<table>
<thead>
<tr>
<th>Source</th>
<th>Language</th>
<th>Source</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>eu</td>
<td>Other</td>
<td>Zaaas</td>
<td>Other</td>
</tr>
<tr>
<td>voto</td>
<td>Other</td>
<td>viejas</td>
<td>Spanish</td>
</tr>
<tr>
<td>por</td>
<td>Other</td>
<td>zorras</td>
<td>Spanish</td>
</tr>
<tr>
<td>um</td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mundo</td>
<td>Other</td>
<td></td>
<td></td>
</tr>
<tr>
<td>onde</td>
<td>Other</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>