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.

Source	**Language**
i	English
always	English
tell	English
him	English
to	English
sing	English
to	English
me	English
pero	Spanish
nunca	Spanish
quiere	Spanish

Source	**Language**
Tuesdays	English
Around	English
6	None of the above
pero	Spanish
it	English
's	English
not	English
worth	English
it	English

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.

Source	**Language**
i	English
always	English
tell	English
him	English
to	English
sing	English
to	English
me	**English**
pero	Spanish
nunca	Spanish
quiere	Spanish

Source	**Language**
Johnny	NE
Depp	NE
para	Spanish
Dr.	NE
Strange	NE
?..	None of the above
NO	**Ambiguous**

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).

Source	**Language**
@Sof_1D17	None of the above
Ayy	Spanish
que	Spanish
pepe	NE
snapchateame	**Mixed**
el	Spanish
arreglo	Spanish

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.

Source	**Language**	**Source**	**Language**	**Source**	**Language**
Mejor	Spanish	and	English	@username	None of the above
Vente	Spanish	I	English	it	English
para	Spanish	told	English	‘s	English
el	Spanish	her	English	on	English
West	**NE**	to	English	**telemundo**	**NE**
Coast	**NE**	record	English	**el**	**NE**
and	English	**La**	**NE**	**señor**	**NE**
visit	English	**Reina**	**NE**	**de**	**NE**
me	English	**del**	**NE**	**los**	**NE**
lol	English	**Sur**	**NE**	**cielos**	**NE**

Abbreviations

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

Source	**Language**	**Source**	**Language**	**Source**	**Language**
Mr.	**English**	**lol**	**English**	jajaja	Spanish
Smith	NE	yeah	English	**ntc**	**Spanish**
was	Spanish	I	English	gracias	Spanish
quejandose	Spanish	hear	English	por	Spanish
como	Spanish	you	English	todo	Spanish
siempre	Spanish	wey	Spanish		

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.

Source	**Language**	**Source**	**Language**
eu	**Other**	**Zaaas**	**Other**
voto	**Other**	viejas	Spanish
por	**Other**	zorras	Spanish
um	**Other**		
mundo	**Other**		
onde	**Other**		