Manual for participants

From Tweet-norm

The annotation process can be described with the following steps:

PRE-PROCESSING

1) Tweets have been processed using *Freeling* (http://nlp.lsi.upc.edu/freeling/).

```
  analyze -f es.cfg --flush --ftok es-twit-tok.dat --usr --fmap es-twit-map.dat  --outf morfo --noprob --noloc
```

where the *es-twit*.dat* files can be found on the SVN: http://devel.cpl.upc.edu/freeling/svn/trunk/src/main/twitter/

2) Words with no analysis are marked as OOV.

3) Afterward, OOV words are annotated, distributed and evaluated. Therefore, real-word-errors are not being taken into account for the time being.

4) In the annotation process, OOV words are identified as one of the following: VARIATION, CORRECT or NoES (no Spanish). In case of a VARIATION, we assigned the corresponding normalized alternative.

5) VARIATION, CORRECT and NoES are coded as 0, 1 and 2, respectively.

ANNOTATION RULES

CASE 1: Word included in RAE

Rule 1: the word is annotated as correct with no alterations, even in the event that the word is not correctly used within the context.

CASE 2: Word not included in RAE, which can be categorized as proper noun

Rule 2.1: In case of an acronym [*composed of all the needed letters according to the context*], written with some lower case letters, both the original form and the upper-cased form will be considered correct with no alterations.

```
 (e.g., CoNLL and CONLL)
```

Rule 2.2: In case of an acronym [*composed of all the needed letters according to the context*], completely upper-cased, will be considered correct with no alterations.

```
 (e.g., IBM and I.B.M.)
```

Rule 2.3: If it is not an acronym, is correctly spelled and capitalized, it will be considered as correct, regardless of it being a diminutive, nickname or alternative name.

```
(e.g., Tony, Anita, Yoyas)
```

Rule 2.4: If it is incorrectly spelled or does not agree with any of 2.1 to 2.3, it will be marked as VARIATION and its correct alternative will be provided.

```
(e.g., sanchez -> Sánchez, tamagochi -> Tamagotchi, abc -> ABC, a.B.c. -> A.B.C.,
[*CONL -> CONLL*])
```

CASE 3: Word not included in RAE which is not a proper noun

Rule 3.1: If it is a correctly spelled neologism or foreign word, it will be considered as correct with no alterations.

```
(e.g., mouriñistas, retuitear, retweetear)
```

Rule 3.3: If it is a correctly spelled diminutive or superlative, it will be considered as correct with no alterations.

```
(e.g., supergrande)
```

Rule 3.4: If it is incorrectly spelled (e.g., repeated, removed, altered letters), it will be considered as a VARIATION and its correct alternative will be provided.

```
(e.g., horrooorr -> horror, hacia-> hacía)
```

Rule 3.5: If it is an abbrevation, it will be considered as a VARIATION and its correct alternative will be provided.

```
(e.g., admin -> administración, sr -> señor)
```

Rule 3.6: If it is an onomatopoeia (usually with repeated letters) altered from the spelling accepted by RAE, it will be considered as a VARIATION. If it is not included in RAE, it will be considered as correct.

```
(e.g., aaaahhh -> ah, jajajajas -> ja, iiiii -> uy+ay)
```

Rule 3.7: If it contains more than a word put together, it will be considered as a VARIATION, and the corresponding set of words will be provided.

Rule 3.8: If the word is not in Spanish, it will be annotated as NoES

```
(e.g., parking)
```

Rule 3.9: A smiley will be annotated as NoES

Rule 3.10: A sequence of words in a language other than Spanish will be annotated as NoEsBeg ... NoEsEnd

- This page was last modified 10:51, 31 May 2013.