# A corpus reader for Spanish books in Project Gutenberg

## 1. Corpus readers in NLTK

Corpus readers are NLTK objects that are used to access corpora (collection of texts). The *nltk.corpus* module already includes some reader instances, which grant access to specific corpora already included in NLTK (e.g. *nltk.corpus.brown*, to access the Brown corpus). Corpus readers define standard methods to access a corpus as a single string ( `.raw()` ), as a list of words ( `.words()` ), as a list of sentences ( `.sents()` ), or as a list of paragraphs ( `.pars()` ).

- Accessing the Brown corpus using the `.words()` method

```
>>> import nltk
>>> nltk.corpus.brown.words()
[u'The', u'Fulton', u'County', u'Grand', u'Jury', ...]
>>>
```

When a new corpus is created, i.e. a corpus not packaged with NLTK already, it is convenient to associate it with a new corpus reader. NLTK's *corpus* module provides some standard corpus reader constructors (e.g. *PlaintextCorpusReader()*). To create an instance of a corpus reader, the constructor takes a root arguemnt and one or more fileids.

In our case, we are interested in a corpus of late 19th. and early 20th. century Spanish language books, which are part of Project Gutenberg. The collected corpus contains 45 books, from both Peninsular and Latin American authors. It contains a mixture of fiction and non-fiction works, including novels, dramatic plays, and poetry.

Project Gutenberg books include a standard preface and back-matter, in English, with information about copyright and other editorial stuff. I have removed all that information from the textx in the corpus.

It is very easy to associate the corpus with one of the standard reader constructors from NLTK, to create a reader instance:

- Import a corpus reader constructor from nltk.corpus

```
>>> from nltk.corpus import PlaintextCorpusReader as PTR
>>>
```

- Identify the root directory for the corpus and pass it as an argument to the constructor to create an instance of a corpus reader

```
>>> root = './SpGut_cp/body'
>>> myreader = PTR(root, '.*\.txt')
>>>
```

- Get some of the initial characters from the corpus.

```
>>> print(myreader.raw()[0:100])
      DON PEDRO DE LARA
      DOÑA MATILDE, su hija
      DON EDUARDO DE CONTRERAS
      BRUNO, criado de DON PEDR
>>>
```

One immediately apparent issue that needs to be addressed is the encoding of Spanish special characters, like accented vowels, the "ñ", and diacritics not found in English (like the upside-down exclamation and interrogation signs). In computer systems, each character is assigned a code. The characters used in English orthography are assigned codes 0 to 127. This encoding, known as ASCII, uses 7 bit codes.

When most computers were able to handle up to 8 bits, non-ascii characters were assigned codes 128 to 255. But different computers assigned those codes to different characters. Unicode solves this by assigning a uniform numerical value (known as a code point) to each possible character of any alphabet. An **encoding**, like UTF-8, turns a code point into a sequence of bytes.

in python, strings are encoded as ASCII. A unicode object is a string with the "u" prefix. A non-ascii character (i.e. a character whose code-point is above 127) can be denoted by the escape sequence \x, followed by two hexadecimal digits.

The files in the Gutenberg corpus are encoded as UTF-8 unicode texts. Any corpus reader that can handle unicode will threfore be able to read the special characters from the files.

```
>>> print(myreader.sents()[10023])
[u'\xbf', u'A', u'qu\xe9', u'ese', u'rostro', u'abatido', u'y', u'melanc\xf3lico', u'
?']
>>>
```

# 2. Limitations of standard corpus readers in NLTK

A standard corpus reader seems to do a good job of splitting and tokenizing these texts. But there are areas where other particular aspects of Spanish, and the editorial conventions of Project Gutenberg, require a custom approach. First, there is the matter of numerals and abbreviations. In Spanish, numerals use "." for thousands and "," for decimals. The standard corpus reader breaks numerals at the period. *7.500* is tokenized as three different tokens.

Second, there is the matter of abbreviations. Unless a list of abbreviations is provided, the reader will break them at the period as well. So *Vd.* is tokenized as *Vd* and *.*.

Third, punctuation marks need to be handled properly by the tokenizer. In Project Gutenberg, the style require a double dash "--" to indicate dialogs. If this mark is preceded or followed by a punctuation mark (period, interrogation sign, etc.), the tokenizer interprets that as a single token. The same happens with other expressions involving punctuation, like the ellipsis "...".

Moreover, the faulty tokenizaiton of punctuation marks results in an incorrect splitting of the text into sentences. For instance, the tokenizer will not insert a break between the period and the dash in ".--", and therefore will miss a sentence break there.

Finally, PG uses underscores to indicate emphasis or italics in the original text (following Markdown conventions). A standard reader will not separate the underscore from the word, thus rendering "Flor", "_Flor_" "_Flor", and Flor_" as four different types.

- Standard reader breaks numerals at the period

```
>>> print(myreader.sents()[4103])
[u'La', u'suscrici\xf3n', u',', u'contando', u'con', u'lo', u'gastado', u'en', u'las'
, u'municiones', u',', u'ha', u'producido', u',', u'por', u'nuestra', u'parte', u'7',
 u'.', u'500', u'pesos', u'fuertes', u'.']
>>>
```

- Reader mishandles punctuation and other marks, like "--" or "...", and misses sentence boundaries.

```
>>> print(myreader.sents()[10027])
[u'--', u'No', u',', u'se\xf1orita', u'...--', u'murmur\xe9', u'sonriendo', u'.--', u
'A', u'las', u'veces', u'se', u'me', u'va', u'el', u'pensamiento', u'hacia', u'Villav
erde', u',', u'en', u'busca', u'de', u'los', u'que', u'me', u'aman', u'....']
>>>
```

- Abbreviations and punctuation marks are not handled correctly by the tokenizer

```
>>> print(myreader.sents()[25500])
[u'--\xbf', u'Qu\xe9', u'hac\xeda', u'su', u'padre', u'de', u'Vd', u'.?']
```

- Reader mishandles underscores indicating emphasis.

```
>>> print(myreader.sents()[29300])
[u'La', u'compa\xf1\xeda', u'de', u'Garc\xeda', u'Delgado', u'cantaba', u'el', u'himn
o', u'nacional', u'y', u'representaba', u'la', u'_Flor', u'de', u'un', u'd\xeda_', u'
,', u'de', u'Camprod\xf3n', u'.']
>>>
```

# 3. A customized corpus reader for Spanish books

To overcome the limitations of standard corpus readers, I have created my own reader for the Spanish books in Project Gutenberg. The `SpanishCorpusReader()` constructor object is a sub-class of `PlaintextCorpusReader()`. The heart of the new corpus reader is a block reader that reads one paragraph at a time. This takes advantage of the fact that paragraphs in Project Gutenberg books are separated by blank lines. The StreamBackedCorpusView built on this block reader will load one paragraph at a time into memory from the underlying files to operate on it. The block reader is paired up with a word tokenizer that uses regular expression matching.

Tokenization is performed by regular expression matching.

- Numerals are sequences of characters that match the regular expression `\$*[\d\.]+,\d+|\$*[\d\.]*\d`, that is, a sequence that has a digit at the end, preceded by other digits, possibly separated from the last one by a comma or by periods, and maybe preceded by the dollar sign.
- There are also abbreviations, which are defined so that they match a sequence of characters including a final (or sometimes a medial) period. Examples: "Vd." (for *usted*), "L.E.". These also include the ellipsis sign "...".
- Punctuation includes the double dash "--" used in Project Gutenberg as an m-dash, and common Spanish marks like the opening question and exclamation marks. Underscores are used in PG to indicate emphasis, so they need to be treated as punctuation. Quotation marks before or after "whitespace" are considered a single mark with whitespace (necessary to aid in correct sentence splitting later on).
- Finally, any sequence of alphanumeric characters not matched by the previous regular expressions will be identified as a token. This would be the majority of the words in the text. An alphanumeric character is defined as the complement of the set of punctuation marks: `[—✚§/#´<>+’=:$;,“”»«";¿?¡!_\s.()\[\]{}*^-]`

Further, the list of tokens is assembled into sentences. Thus, instead of splitting a paragraph into sentences by looking for punctuation marks followed by whitespace, for instance, I first tokenize the paragraph (which is bound to begin and end at a sentence boundary), using formulas to find the sentence boundaries between tokens. In this way, I avoid the pitfalls of breaking a sentence at the wrong place when two punctuation marks follow each other, or when a punctuation mark is followed by a quotation mark, for instance.

These functions are used to build three different block readers. `par_word_reader()` reads a paragraph and returns a list of tokens, while `par_sent_reader()` returns a list of sentences, each tokenized into words. `par_para_reader` takes that list of sentences and makes it the only element of a new list, corresponding to a paragraph. Each of these block readers is passed on to a stream-backed corpus viewer to define methods for viewing the corpus as a list of words ( `.words()` ), as a list of sentences ( `.sents()` ), or as a list of paragraphs ( `.paras()` ), respectively. The `.raw()` method for viewing the corpus as a string of characters is inherited from the parent class.

The reader is saved as a python module `SpanishReader.py` . The module can be loaded and the constructor used to create an instance of a Spanish corpus. We can compare the tokenization and sentence splitting methods of the Spanish corpus reader to the standard PlaintextCorpusReader. Many of the issues we identified have now been resolved.

- Import the module and use `SpanishPlaintextCorpusReader()` as a constructor. Use the same root directory as before.

```
>>> import SpanishReader
>>> myspreader = SpanishReader.SpanishPlaintextCorpusReader(root, '.*\.txt')
```

- Print some characters and words

```
>>> print(myspreader.raw()[0:100])
    DON PEDRO DE LARA
    DOÑA MATILDE, su hija
    DON EDUARDO DE CONTRERAS
    BRUNO, criado de DON PEDR
>>> print(myspreader.words()[90:100])
[u'DO\xd1A', u'MATILDE', u'Y', u'BRUNO', u'DO\xd1A', u'MATILDE', u'.', u'\xa1', u'Bru
no', u'!']
>>>
```

- Spanish reader does not break numerals at the period

```
>>> print(myspreader.sents()[4288])
[u'La', u'suscrici\xf3n', u',', u'contando', u'con', u'lo', u'gastado', u'en', u'las'
, u'municiones', u',', u'ha', u'producido', u',', u'por', u'nuestra', u'parte', u'7.5
00', u'pesos', u'fuertes', u'.']
>>>
```

- Spanish reader finds the right sentence and punctuation breaks

```
>>> print(myspreader.sents()[10708:10710])
[[u'--', u'No', u',', u'se\xf1orita', u'...', u'--', u'murmur\xe9', u'sonriendo', u'.
'], [u'--', u'A', u'las', u'veces', u'se', u'me', u'va', u'el', u'pensamiento', u'hac
ia', u'Villaverde', u',', u'en', u'busca', u'de', u'los', u'que', u'me', u'aman', u'.
..', u'.']]
>>>
```

- Spanish reader keeps periods with abbreviations

```
>>> print(myspreader.sents()[26293])
[u'--', u'\xbf', u'Qu\xe9', u'hac\xeda', u'su', u'padre', u'de', u'Vd.', u'?']
>>>
```

- Spanish reader separates underscores from words

```
>>> print(myspreader.sents()[30042])
[u'La', u'compa\xf1\xeda', u'de', u'Garc\xeda', u'Delgado', u'cantaba', u'el', u'himn
o', u'nacional', u'y', u'representaba', u'la', u'_', u'Flor', u'de', u'un', u'd\xeda'
, u'_', u',', u'de', u'Camprod\xf3n', u'.']
>>>
```

- Method for reading paragraphs

```
>>> print(myspreader.paras()[4007])
[[u'--', u'Bueno', u'...', u'vete', u',', u'y', u'\xa1', u'que', u'Dios', u'te', u'be
ndiga', u'!'], [u'Escribe', u'luego', u'que', u'puedas', u'.'], [u'Saludas', u'de', u
'nuestra', u'parte', u'al', u'se\xf1or', u'Fern\xe1ndez', u',', u'y', u'a', u'la', u'
se\xf1orita', u'.'], [u'Escribe', u'con', u'frecuencia', u'.'], [u'Acaso', u'tengas',
 u'que', u'tratar', u'con', u'los', u'mozos', u'...', u'.'], [u'Te', u'encargo', u'mu
cha', u'prudencia', u',', u'mucha', u'seriedad', u'...', u'.'], [u'Vamos', u',', u'da
me', u'otro', u'abrazo', u',', u'y', u'\xa1', u'que', u'Dios', u'te', u'lleve', u'con
', u'bien', u'!']]
>>>
```

# 4. Appendix

## 4.1. List of abbreviations:

Arqueol., &c., Mme., Lib., Lic., Antrop., núm., Dres., Descrip., —tom., rs., lám., Sec., Liv., Introd., Excmo., Caps., Amer., oct., Antigs., Ses., Moderns., Moralíz., Esp., Lam., act., Europ., Geog., CC., Eneid., Nat., M., Crón., Ntra., men., Láms., Orth., Gam., tam., Arg., Op., caps., Agust., fol., Sr., Tam., Janr., MS., Bol., Mr., S.A.S., Núms., Civiliz., Figs., DR., Orígs., Vocabuls., cits., L.E., Dicc., paj., Amér., Lám., ESQ., op., Argent., NE., Sres., Esp., Lam., Exmo., Espagn., pag., Conq., Cont., Sr., SR., SO., Ind., ded., cuads., Oct., Psch., Ed., Sta., Fot., sec., Part., JUV., Arqueolog., Sto., pp., Antig., vol.Cod., Srta., Col., lib., Congr., lin., Colec., Instit., Cong., Cient., Mlle., Rev., LLOR., nat., gr., ROB., Ge., Ord., lec., FR., Fr., ILMO., Colecc., Pág., Tuc., Prov., EXCMO., Págs., p.m., sc., capits., Pl., PP., lug., Sra., a.m., Antich., Gen., Apénd., Cap., Bs., pags., MSS., cap., Vds., nos., tom., Lug., Dr., págs., id., pág., verb., Or., sigtes., SEB., Hist., Vd., ci., vol., cit., etc., Cía., Id., Nos., Ibid., LLO., Ud., Fig., Geográf., Internat., Sant., ps., part., Luxemburg.,

## 4.2. Works in the Spanish corpus:

23957 Pereda, José María de, 1833-1906. Al primer vuelo (1896)
21906 Pérez Galdós, Benito, 1843-1920. Cádiz (1874)
26929 Picón, Jacinto Octavio, 1852-1923. Cuentos de mi tiempo (1895)
44120 Saurí i Marsal, Manuel, 1837-1924. La Caza de La Perdiz Con Escopeta, Al Vuelo y con Perro de Muestra (1877)
35882 Valera, Juan, 1824-1905. A vuela pluma: colección de artículos literarios y políticos (1897)
49013 Muñoz Seca, Pedro, 1881-1936. La venganza de Don Mendo (1918)
49280 Baroja, Pío, 1872-1956. Los Caminos del Mundo (1921)
47388 Espina, Concha, 1869-1955. Agua de Nieve (1919)
36573 Palacio Valdés, Armando, 1853-1938. La aldea perdida (1911)
42727 Palacio Valdés, Armando, 1853-1938. La alegría del capitán Ribot (1899)
14318 Pérez de Ayala, Ramón, 1880-1962. Belarmino y Apolonio (1921) 14311 Pérez Galdós, Benito, 1843-1920. Bailén (1873) 49149 Unamuno, Miguel de, 1864-1936. Amor y Pedagogía (1902) 44512 Unamuno, Miguel de, 1864-1936. Abel Sánchez: Una Historia de Pasión (1917) 50757 Zamacois, Eduardo, 1873-1971. La cita: novelas (1913) 41337 Castro, Cristóbal de, 1880?-1953. Catálogo Monumental de España; Provincia de Álava. Inventario general de los monumentos históricos y artísticos de al nación (1915)
26545 Martínez Ruiz, José, 1873-1967. Antonio Azorín: pequeño libro en que se habla de la vida de este peregrino señor (1913)
49437 Machado, Antonio, 1875-1939. Páginas escogidas (1917)
12368 Gorostiza, Manuel Eduardo de, 1789-1851. Contigo Pan y Cebolla (1833)
41575 Cané, Miguel, 1851-1905. Juvenilla (1901); Prosa ligera (1903)

16082 Delgado, Rafael, 1853-1914. Angelina (novela mexicana) (1893)
45945 Lastarria, José Victorino, 1817-1888. Antaño i Ogaño: Novelas i Cuentos de la Vida Hispano-Americana (1885)
18166 Martí, José, 1853-1895. Amistad funesta: Novela (1885)
28281 Villaverde, Cirilo, 1812-1894. Cecilia Valdés o la Loma del Ángel (1879)
31724 López, Lucio Vicente, 1848-1894. La gran aldea; costumbres bonaerenses (1908)
52469 Nieves, Juan B. ????-???? La Anexión de Puerto-Rico a los Estados Unidos de America (1898)
26771 Bunge, Carlos O. (Carlos Octavio), 1875-1918. Thespis (novelas cortas y cuentos) (1907)
29920 Larreta, Enrique, 1875-1961. La gloria de don Ramiro: una vida en tiempos de Felipe segundo (1911)
25054 Leumann, Carlos Alberto, 1886-1952. Adriana Zumarán (novela) (1921)
13507 Quiroga, Horacio, 1878-1937. Cuentos de Amor de Locura y de Muerte (1917)
26231 Vedia, Enrique de, 1847-1917. Transfusión (1908)
53798 Viana, Javier de, 1868-1926. Ranchos (Costumbres del Campo) (1920)
52050 Darío, Rubén, 1867-1916. Autobiografía (1918 [1912])
53927 Gutiérrez, Juan María, 1809-1878. Apuntes biograficos de escritores, oradores y hombres de estado de la Republica Argentina (1860)
34565 Palma, Angélica, 1883-1935. Crónicas de Marianela (1917)
54064 Quiroga, Adán, 1863-1904. La cruz en América (1901)
22899 Rodó, José Enrique, 1872-1917. Ariel (1900)
51458 Darío, Rubén, 1867-1916. Canto a la Argentina, Oda a Mitre y otros poemas (1918 [1906])
51711 Darío, Rubén, 1867-1916. El Canto Errante (1918 [1907])
50341 Darío, Rubén, 1867-1916. Cantos de Vida y Esperanza, Los Cisnes y otros poemas (1918 [1905])
10909 Hartzenbusch, Juan Eugenio, 1806-1880. Los Amantes de Teruel (1837)
29731 Alarcón, Pedro Antonio de, 1833-1891. El Capitán Veneno (1881)
14944 Blasco Ibáñez, Vicente, 1867-1928. La Barraca (1898)
16413 Blasco Ibáñez, Vicente, 1867-1928. Arroz y tartana (1894)
27295 Fernández y González, Manuel, 1821-1888. Amparo (Memorias de un loco) (1858)