W3 Assignment – Corpus Linguistics 1

Computational Linguistics

John Gamboa

April 22, 2022

1 What is Corpus Linguistics?

Answer *TRUE* or *FALSE*

a) We defined Corpus Linguistics as

A methodology that analyses corpora to address linguistic questions

This definition is an attempt to summarize the discussion from Gries (2009)

`answer:` __________________________

b) Most researchers believe that Corpus Linguistics is not just a new methodoly, but a completely new research enterprise, or even a new philosophical approach to Linguistics

`answer:` __________________________

c) There is not much debate about what Corpus Linguistics is

`answer:` __________________________

2 Corpora

2.1 Additional Corpus Data

Choose the correct alternative:

a) Corpora can contain additional information, such as Metadata and Annotations.

b) Metadata may contain, among others, the source of the data, or copyright information related to the corpus files

c) Metadata may contain, among others, the constituent tree of the sentences in the corpus files

d) Annotations may contain, among others, the age of each of the participants in the conversations recorded in the corpus

e) Both the first and second alternatives are correct

f) Both the first and the third alternatives are correct

g) Both the first and the fourth alternatives are correct

h) The first, the second, and the third alternatives are correct

`answer:` ______________________________

2.2 Criteria for defining a corpus

As we saw, a corpus is a collection of texts. This collection needs to be:

a) ______________________________ , i.e., different parts of the linguistic variety are all present in the corpus

b) ______________________________ , i.e., these parts need to appear in the corpus according to the proportions they appear in real life

c) ______________________________ , i.e., "the texts were spoken or written for some authentic communicative purpose"

2.3 Types of Corpora

We saw 5 dimensions in which corpora can be classified. These are:

a) Depending on whether the corpus captures how the language varies over time or just keeps a snapshot of the language at a particular point of time.

   ______________________________ vs. ______________________________

b) Depending on whether the corpus focuses only in one language or has the same data in several languages

______________________________ vs. ______________________________

c) Depending on whether the corpus focus on the language as whole or only in a particular variety/dialect/register of the language.

______________________________ vs. ______________________________

d) Depending on whether or not the corpus has additional corpus data that represents a particular linguistic analysis

______________________________ vs. ______________________________

e) Depending on whether or not the corpus can be extended over time with new data

______________________________ vs. ______________________________

3 Answering Linguistic Questions

3.1 What do we normally calculate?

What is normally calculated in Corpus Linguistics are ______________________________.

Generally, we are interested in two types of them:

______________________________ and ______________________________,

i.e., how often certain linguistic elements appear in the data, and how often they appear along with other linguistic elements.

4 NLTK / Python programming

4.1 Accessing Corpus Data

In the class, we have seen that the NLTK has a subpackage called `gutenberg`, which you can use to get access to some books from the Project Gutenberg. Consider the following lines of code (In), and their output (Out):

```
In [1]: import nltk
In [2]: nltk.corpus.gutenberg.fileids()
Out[2]: ['austen-emma.txt',
         'austen-persuasion.txt',
         'austen-sense.txt',
         'bible-kjv.txt','blake-poems.txt',
         'bryant-stories.txt',
         'burgess-busterbrown.txt',
         'carroll-alice.txt',
         'chesterton-ball.txt',
         'chesterton-brown.txt',
         'chesterton-thursday.txt',
         'edgeworth-parents.txt',
         'melville-moby_dick.txt',
         'milton-paradise.txt',
         'shakespeare-caesar.txt',
         'shakespeare-hamlet.txt',
         'shakespeare-macbeth.txt',
         'whitman-leaves.txt']
```

Write the line of code you'd use to get all the words in the file `'chesterton-thursday.txt'`:

code: ____

Write the line of code you'd use to get all the sentences in the same file:

code: ____

Write the line of code you'd use to get a string containing the entire file data, unsegmented by any NLTK algorithm:

code: ____

4.2 List comprehensions

Assume the following lines of code have been run:

```
num_list = [0,1,2,3,4,5,6,7,8,9]
str_list = ['a', 'b', 'c', 'd', 'e', 'f']
lst_list = [[1,2], [3,4], [5,6,7], [8], [9]]
```

Decide whether the statements below are correct. (Ideally, try out these lines in Jupyter notebook, to see what they do)

- The line

  ```
  a = [2*i for i in num_list if i < 5]
  ```

  will produce the same effect as

  ```
  a = []
  for i in num_list:
      if (i < 5):
          a.append(2*i)
  ```

  `Is correct: ___________________________`

- The line

  ```
  a = [2*i for i in num_list]
  ```

  will produce the same effect as

  ```
  a = []
  for 2*i in num_list:
      a.append(i)
  ```

  `Is correct: ___________________________`

- The line

  ```
  a = [2*i for i in num_list]
  ```

  will produce the same effect as

  ```
  a = []
  for i in 2*num_list:
      a.append(i)
  ```

  `Is correct: ___________________________`

- The line

  ```
  a = [len(i) for i in lst_list]
  ```

  will produce the list

  ```
  [2,2,3,1,1]
  ```

  which is a list containing the length of each list inside `lst_list`.

  `Is correct: ___________________________`

- The line

  ```
  a = [i for i in num_list + str_list]
  ```

  will produce the same as

  ```
  a = num_list + str_list
  ```

  Is correct: __________________________

- The line

  ```
  a = [2*i for i in num_list]
  ```

  will produce the same effect as

  ```
  a = []
  for i in num_list:
      a.append(2*i)
  ```

  Is correct: __________________________