NLP

Course of DSSC Master degree - University of Trieste

Matilde Trevisani

DEAMS

2020/05/25 (updated: 2020-06-12)

Introduction to corpora and dimensions

Basic concepts

1. constitution of the corpus
2. counting words
3. vocabulary
4. statistical terminology

Constitution of the corpus

A text is an object that has a beginning, a development, an end and has a communicative purpose.

A corpus is not simply a "collection of texts". A corpus must have characteristics of **breadth**, **coherence**, **homogeneity** and, in some cases, **exhaustiveness** that make it suitable for research purposes.

The corpus is a set of texts that responds to the needs of a specific **research question**.

You must first formulate a clear research question. Without forgetting that the "good" research questions must be **comparative** (which refers to the problem of identifying sub-corpora).

Note: many problems can be addressed with statistical tools but the path of definition, construction and analysis of data is often not evident.

Homogeneity and variation

Starting from the research question and from the strategy adopted for the creation of the corpus, it is immediately necessary to distinguish the criteria adopted for the evaluation of the quality of the corpus from those which are, instead, variations that you want to observe because they represent the object of the research study.

Some criteria:

- dimensions
- textual genre
- language, theme, style, the (Muller & Brunet, 1988)

Five dimensions of variation for the language (Berruto, 1987):

1. **diachronic** (variation due to chronological differences),
2. **diatopic** (variation due to differences in geographical location),
3. **diaphasic** (variation due to differences in the communicative situation, typical opposition: formal vs. informal),
4. **diastratic** (variation due to differences in the social groups of reference),
5. **diamesic** (variation linked to the medium of transmission, typical opposition: oral vs. written text).

About the diachronic variation

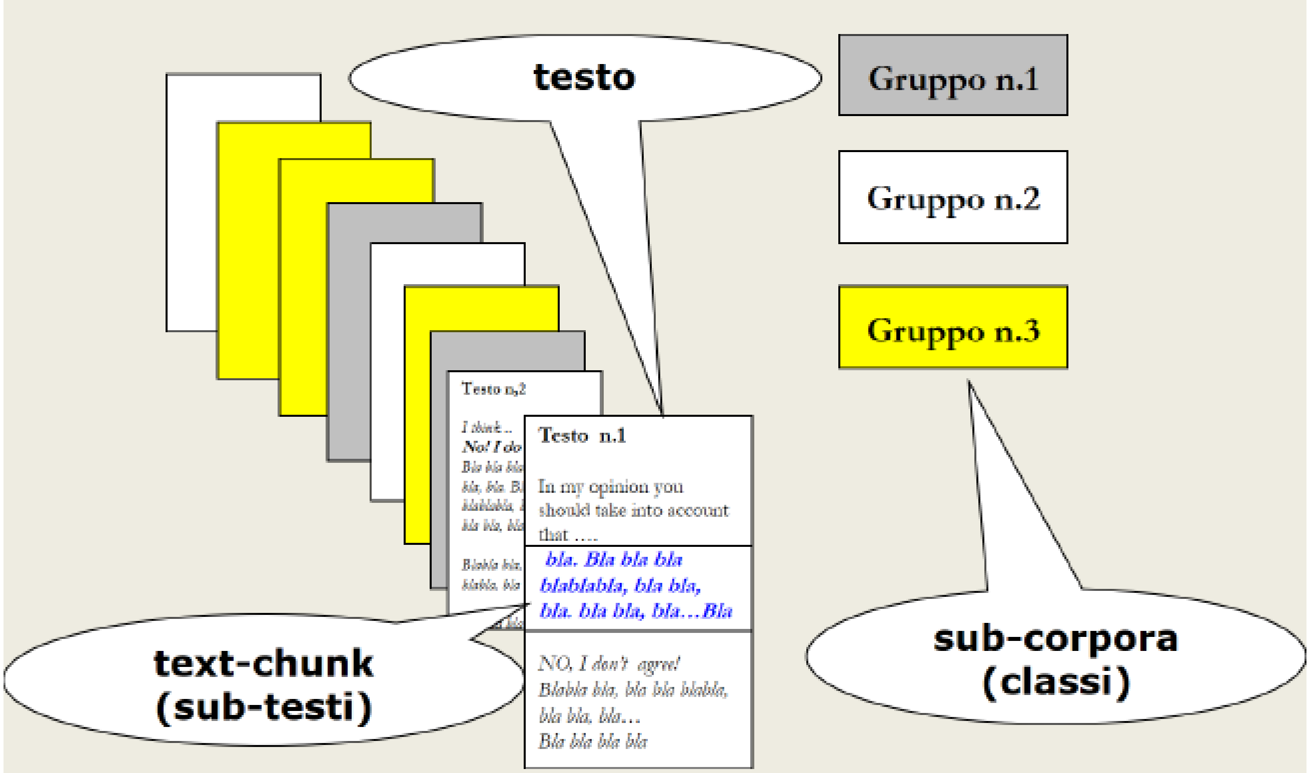
The language needs long periods of time (centuries) to show significant changes.

Although this consideration is obvious in a linguistic context, it is important to underline that the diachronic variations observable in the short term (decades) almost always concern the lexical level (typical examples are neologisms and foresters), which represents the most superficial part of the language.

A change at the syntactic level would represent a profound change in the language.

Representation of the corpus

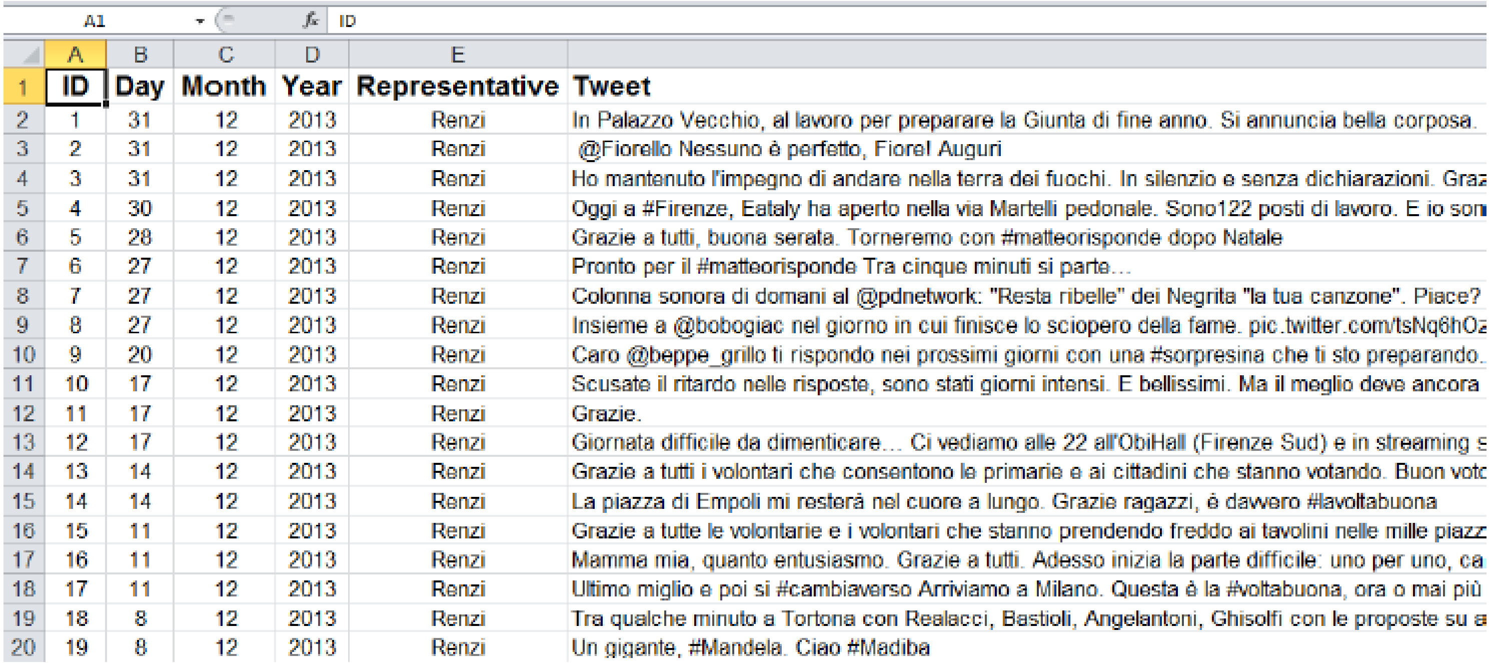
Corpus – texts – subcorpora – fragments


testo
Gruppo n.1
Gruppo n.2
Gruppo n.3
Testo n,2
Testo n.1
In my opinion you should take into account that ....
bla. Bla bla bla blablabla, bla bla, bla. bla bla, bla…Bla
NO, I don't agree! Blabla bla, bla bla blabla, bla bla, bla… Bla bla bla bla
text-chunk (sub-testi)
sub-corpora (classi)

Some collection methods

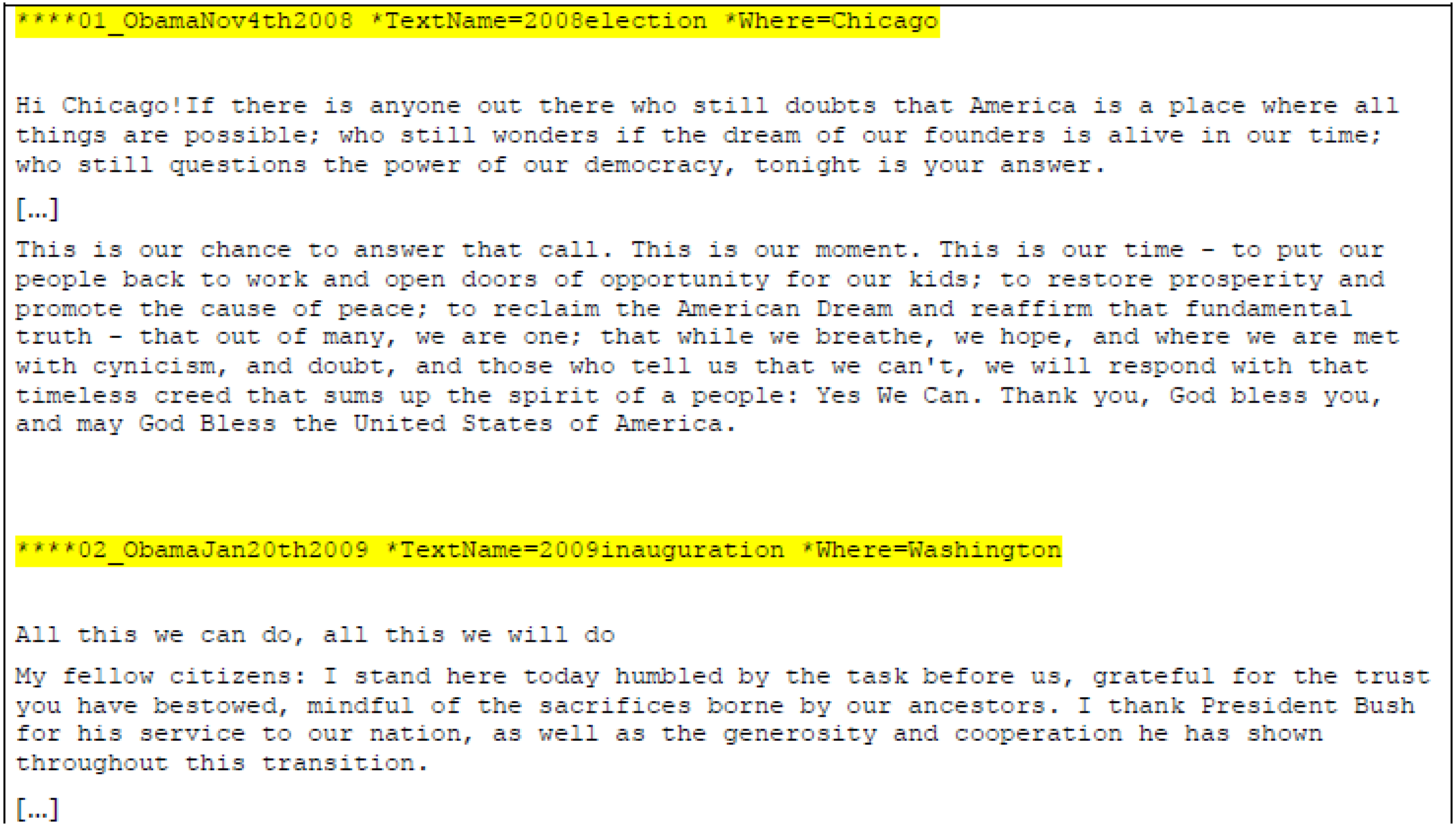
1. When you have a large number of texts of limited dimensions, for example interventions in focus groups, short answers to open questions, posts, messages (SMS and Whatsapp), advertising messages …
    - Excel sheet (a text in a cell)
2. When you have a limited number of medium-large sized texts, eg. speeches, letters, documents, topics, scientific articles …
    - Word document (one text below the other in the same file)
3. When you have a very large number of texts
    - folder that collects all the files (in .txt format)
4. The "tidy corpus" perspective (new approach)

-> 1. A large number of texts of limited size



ID	Day	Month	Year	Representative	Tweet
1	31	12	2013	Renzi	In Palazzo Vecchio, al lavoro per preparare la Giunta di fine anno. Si annuncia bella corposa.
2	31	12	2013	Renzi	@Fiorello Nessuno è perfetto, Fiore! Auguri
3	31	12	2013	Renzi	Ho mantenuto l'impegno di andare nella terra dei fuochi. In silenzio e senza dichiarazioni. Graz
4	30	12	2013	Renzi	Oggi a #Firenze, Eataly ha aperto nella via Martelli pedonale. Sono122 posti di lavoro. E io son
5	28	12	2013	Renzi	Grazie a tutti, buona serata. Torneremo con #matteorisponde dopo Natale
6	27	12	2013	Renzi	Pronto per il #matteorisponde Tra cinque minuti si parte…
7	27	12	2013	Renzi	Colonna sonora di domani al @pdnetwork: "Resta ribelle" dei Negrita "la tua canzone". Piace?
8	27	12	2013	Renzi	Insieme a @bobogiac nel giorno in cui finisce lo sciopero della fame. pic.twitter.com/tsNq6hOz
9	20	12	2013	Renzi	Caro @beppe_grillo ti rispondo nei prossimi giorni con una #sorpresina che ti sto preparando.
10	17	12	2013	Renzi	Scusate il ritardo nelle risposte, sono stati giorni intensi. E bellissimi. Ma il meglio deve ancora
11	17	12	2013	Renzi	Grazie.
12	17	12	2013	Renzi	Giornata difficile da dimenticare… Ci vediamo alle 22 all'ObiHall (Firenze Sud) e in streaming s
13	14	12	2013	Renzi	Grazie a tutti i volontari che consentono le primarie e ai cittadini che stanno votando. Buon votc
14	14	12	2013	Renzi	La piazza di Empoli mi resterà nel cuore a lungo. Grazie ragazzi, è dawero #lavoltabuona
15	11	12	2013	Renzi	Grazie a tutte le volontarie e i volontari che stanno prendendo freddo ai tavolini nelle mille piazz
16	11	12	2013	Renzi	Mamma mia, quanto entusiasmo. Grazie a tutti. Adesso inizia la parte difficile: uno per uno, ca
17	11	12	2013	Renzi	Ultimo miglio e poi si #cambiaverso Arriviamo a Milano. Questa è la #voltabuona, ora o mai più
18	8	12	2013	Renzi	Tra qualche minuto a Tortona con Realacci, Bastioli, Angelantoni, Ghisolfi con le proposte su a
19	8	12	2013	Renzi	Un gigante, #Mandela. Ciao #Madiba

-> 2. A limited number of medium-large texts



****01_ObamaNov4th2008 *TextName=2008election *Where=Chicago

Hi Chicago!If there is anyone out there who still doubts that America is a place where all things are possible; who still wonders if the dream of our founders is alive in our time; who still questions the power of our democracy, tonight is your answer.

[...]

This is our chance to answer that call. This is our moment. This is our time – to put our people back to work and open doors of opportunity for our kids; to restore prosperity and promote the cause of peace; to reclaim the American Dream and reaffirm that fundamental truth – that out of many, we are one; that while we breathe, we hope, and where we are met with cynicism, and doubt, and those who tell us that we can't, we will respond with that timeless creed that sums up the spirit of a people: Yes We Can. Thank you, God bless you, and may God Bless the United States of America.

****02_ObamaJan20th2009 *TextName=2009inauguration *Where=Washington

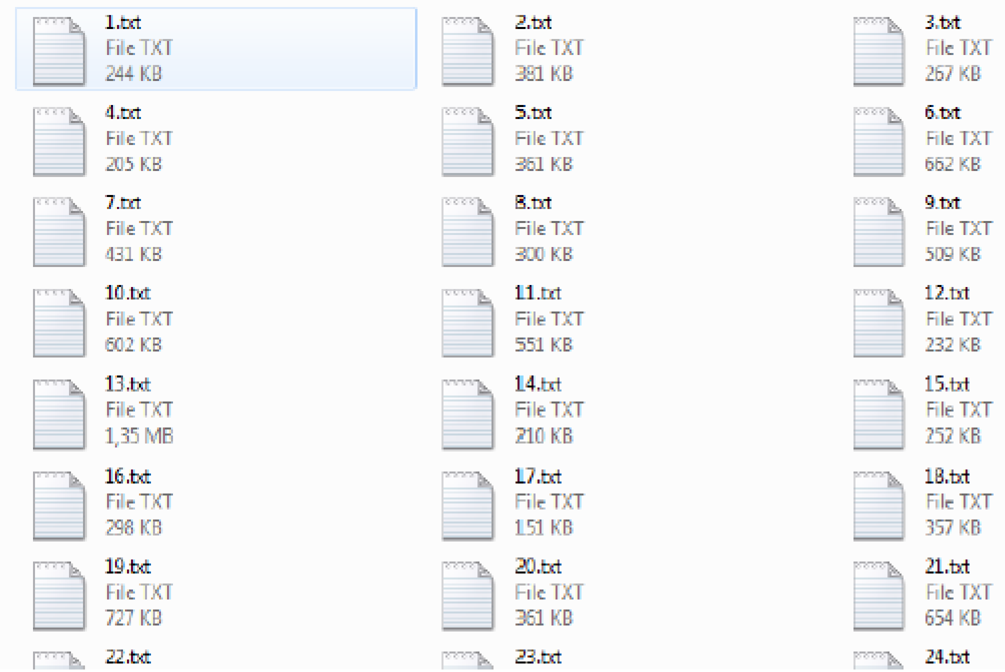
All this we can do, all this we will do

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors. I thank President Bush for his service to our nation, as well as the generosity and cooperation he has shown throughout this transition.

[...]

-> 3. A very large number of texts

cartella di file (di solito in formato .txt)



1.txt
File TXT
244 KB

2.txt
File TXT
381 KB

3.txt
File TXT
267 KB

4.txt
File TXT
205 KB

5.txt
File TXT
361 KB

6.txt
File TXT
662 KB

7.txt
File TXT
431 KB

8.txt
File TXT
300 KB

9.txt
File TXT
509 KB

10.txt
File TXT
602 KB

11.txt
File TXT
551 KB

12.txt
File TXT
232 KB

13.txt
File TXT
1,35 MB

14.txt
File TXT
210 KB

15.txt
File TXT
252 KB

16.txt
File TXT
298 KB

17.txt
File TXT
151 KB

18.txt
File TXT
357 KB

19.txt
File TXT
727 KB

20.txt
File TXT
361 KB

21.txt
File TXT
654 KB

22.txt

23.txt

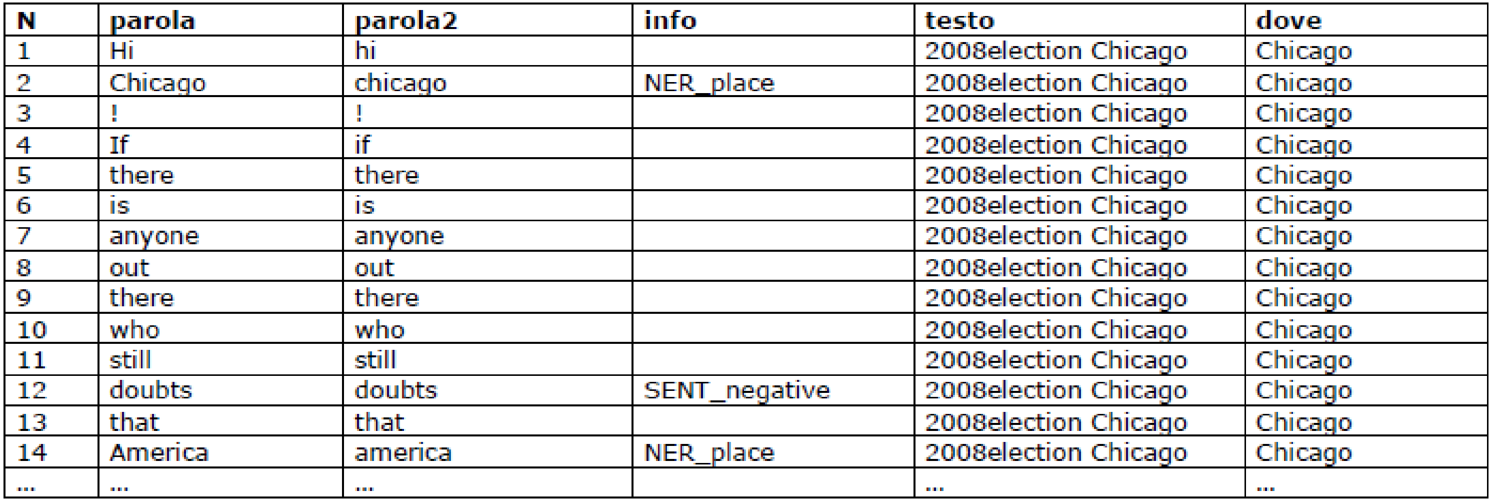
24.txt

[4.] "tidy corpus" perspective

The "tidy data" approach has a specific structure:

- each variable is a column
- each observation is a row
- each type of textual unit observed is a table

The "tidy text" format is a table with one word per row:



N	parola	parola2	info	testo	dove
1	Hi	hi		2008election Chicago	Chicago
2	Chicago	chicago	NER_place	2008election Chicago	Chicago
3	!	!		2008election Chicago	Chicago
4	If	if		2008election Chicago	Chicago
5	there	there		2008election Chicago	Chicago
6	is	is		2008election Chicago	Chicago
7	anyone	anyone		2008election Chicago	Chicago
8	out	out		2008election Chicago	Chicago
9	there	there		2008election Chicago	Chicago
10	who	who		2008election Chicago	Chicago
11	still	still		2008election Chicago	Chicago
12	doubts	doubts	SENT_negative	2008election Chicago	Chicago
13	that	that		2008election Chicago	Chicago
14	America	america	NER_place	2008election Chicago	Chicago
...	...	...		...	...

Size of corpora

The main strength of the statistical analysis of textual data is the opportunity to extract information from **large corpora**, which represent a source of complex unstructured data, overcoming the obstacles posed by the amount of text which, conversely, represents the main limitation for qualitative analyzes.

The availability of software tools makes it possible to extract information from large corpora and **results become more reliable as the size grows**.

These software tools are **not suitable for studying small corpora** and this proves to be a problem in many disciplines where obtaining large text sets is complicated.

Texual data

Ttextual data that can be analyzed are of a different type. In linguistics there is a distinction between:

- phonetics
- grammar (morphology and syntax)
- lexicon

In this part of the course we will focus on:

- **lexicon** (lexical-based analysis)
- **bag-of-words** approaches

Basics

- a **corpus** is a collection of texts;
- a **text** is made up of letters, spaces and other symbols (for example: punctuation, numbers, mathematical symbols);
- a **word** is a sequence of letters isolated by means of separators (spaces and punctuation marks);

In order to count the words contained in the corpus it is necessary to distinguish two (or more) concepts ...

How do we count words?

common sense is not so common

N = 6 word **tokens**

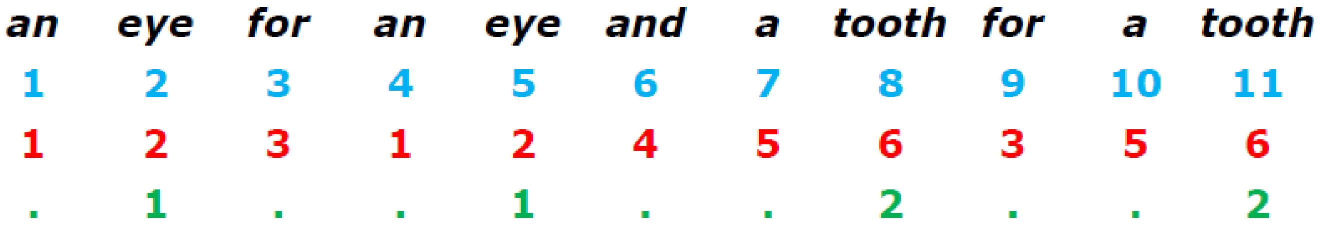
(occurrences / total words)

V = 5 word **types**

(different types / words)

How do we count words?

an eye for an eye and a tooth for a tooth



an	eye	for	an	eye	and	a	tooth	for	a	tooth
1	2	3	4	5	6	7	8	9	10	11
1	2	3	1	2	4	5	6	3	5	6
.	1	.	.	1	.	.	2	.	.	2

N = **11 word tokens**

V = **6 word types**

No. of **content words** = **2**

Sizes

A word token is an occurrence in the text of a word type.

Examples:

- "the" is a word type that has numerous word tokens in any English text
- in any text there are word types that have only one occurrence, that is, they occur only once (*hapax legomena* or simply *hapax*)

The number (**N**) of word tokens is the **size of the corpus** (size in terms of occurrences or number of total words);

The number (**V**) of word types is the **size of the vocabulary** (size in terms of the number of different words);

The **frequency** of a word type is the number of corresponding word tokens in the corpus (number of repetitions, possibly expressed in the form of relative frequency or rate);

The list of word types with frequencies is the vocabulary (of frequency) of the corpus.

> When vocabulary is the basis of statistical analysis, the approach is **bag-of-words**.

A text

Give a man a fish and you feed him for a day.

Teach a man to fish and you feed him for a lifetime.

Is it ok?

A text

No, it isn't ...

(Typically, in pre-processing, upper/lower case must be handled)

give a man a fish and you feed him for a day.

teach a man to fish and you feed him for a lifetime.

- N = 24 word tokens
- V = 13 word types

For a software tool package "Give" is not the same as "give"

-> a pre-processing phase is required that performs a **normalization** of texts (for example, keep only the relevant capital letters, or transform everything into lowercase)

Vocabulary

word type	frequency
(entry)	(occorrences)
a	5
and	2
feed	2
fish	2
for	2
him	2
man	2
you	2
day	1
give	1
lifetime	1
teach	1
to	1
CORPUS	24

Using the language of statistics

The two different ways of counting words correspond to crucial concepts for understanding the logic of statistical analysis of textual data.

N = is the **number of observations** of the corpus (statistical units)

-> word tokens

V = is the **number of modes** (or number of categories of the "vocabulary" variable)

-> word types

The vocabulary is a **frequency table** (data summary).

The vocabulary is a **representation** of the corpus.

Sizes

Once in the manuals we read:

> a great corpus must exceed 100 thousand words
>
> a corpus of 500 thousand words is a good basis for building a frequency lexicon

Today

> corpora of millions of occurrences are quite common

(OK, we are in the era of BIG DATA, digital humanities, etc.)

However … Are we facing the problem from the right perspective?

Sizes

Some questions:

- **How many words** (word tokens) are enough to work with a quantitative approach?
- Is the number of words (word tokens) a good measure of corpus **size**?
- Is the number of different words (word types) a good measure of the (lexical) **richness** of the corpus?

Dimensions

To get a very general idea of dimensions there are empirical measures based on sizes and lexical richness that can be used to understand if a corpus is large enough to allow an effective use of quantitative methods.

- **TTR - Type-Token Ratio** = ratio of number of word types / number of word tokens (in%)
    - the TTR should be less than 20%
- **% of hapax** = number of hapax / number of word types (in%)

  hapax = word type that only one occurrence in the corpus

    - the hapax% should be less than 50%

Warning! These are empirical measures to be taken very carefully. They are very experienced on some languages (Italian, French), less on others (English).

The basic idea is that the use of statistical tools based on word frequency requires **redundancy**.

Bag-of-words and software

The fundamental difference with *computer-assisted qualitative data analysis software* (**CAQDAS**) is that in this case we work on the frequency of **elementary units** such as sequences of characters (what we commonly call "words" in common language) and not complex units such as portions of text.

They do not serve to label portions of text.

They start from a **lexical perspective**, not from the idea of "abstraction of concepts" and "labeling" of sentences / periods / paragraphs of complete meaning.

Examples:

- Alceste
- AntConc
- Iramuteq
- Hyperbase
- KH Coder
- Lexico
- R stylo
- Spad-T (today: Spad)
- Tableau
- Taltac
- Textable (Orange)
- T-Lab
- TXM
- Voyant-tools

- packages available in statistical environments such as R, SAS, Python.

Let's not confuse!

A large corpus does not fall under the definition of **big data**.

Big data must be called into question only if **V V V** (3V)

1. the **volume** of information is large (the dataset must exceed the processing capacity of a personal computer)
2. information is characterized by **velocity** (the dataset must be constantly updated or modified)
3. information is of **various** kinds (the data set contains data, text, audio, video, etc.)

... and after 3V we go on with Veracity, Value, Variability ... complexity ...

But we are working in the "big data era". If one day we will have

- a large corpus (Volume),
- who experiences changes and updates (Velocity)
- linked to other sources of information (Variety)

we will also need integrated software which is not currently available.