

# Corpus linguistics

HAP/LAP. Corpus Linguistics.

# Outline



- 1 Introduction
- 2 Corpus linguistics
- 3 Building a corpus
- 4 Uses of corpora
- 5 Corpus types
- 6 Corpora and annotation
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond

# Introduction: What is a corpus?



- Corpus: collection of words

# Introduction: What is a corpus?



- Corpus: collection of words
  - produced on a **natural communicative context**
  - it must be **representative** of a language or of a genre of that language
  - a hell of a lot of words **stored on a computer**
  - whose purpose is to **analyze language**
  - it allows to **search rapidly and reliably** through that collection of words
  - it acts as a **standard reference** about what is typical in the language
  - often **annotated** with additional linguistic information
  - so that you can start to measure it up against the research questions you have

# Introduction: What is a corpus?



- Corpus: collection of data about
  - textual language (the construction of written corpora has never been easier)
  - speech/spoken language: orthographically or phonemically transcribed
- Different points of view:
  - Linguistics: use corpus to validate linguistic hypotheses
    - for morphologists, set of word derivations
    - for syntax, set of phrases
    - ...
  - NLP: use corpus to learn linguistic models
    - speech recognition (ambiguity in speech)
    - machine translation (frequencies, collocations...)
    - if the corpus is annotated for almost any task
- Purpose:
  - language learning
  - linguistic research
  - ...

# Introduction: Main characteristics



1. Representative of language or genre
  - a corpus is different from a random collection of text
2. Large body of text (bounded in size)
  - it depends on the purpose
  - it might be 30,000 or 40,000 words or it could be billions of words
3. Machine readable (quick search)
4. Standard reference
  - e.g. British National Corpus, which had the aim of being broadly representative of British English
5. Additional linguistic information (annotated)

# 1. Representative



- If we want to study a particular aspect of language
  - analyze all occurrences (excessive!)
  - create a representative *sample* (a subset)
    - e.g. subcorpus of CREA
  - But sample has to be *representative*
- Some definitions

“A corpus is thought to be representative of the language variety it is supposed to represent **if the findings based on its contents can be generalized to the said language variety**” (Leech 1991)

“Representativeness refers to the extent to which **a sample** includes the full range of **variability in a population**” (Biber 1993)
- Depends on the purpose of the corpus
  - For a corpus which is representative of general English, a corpus representative of newspapers is not enough
  - For a corpus representative of newspapers, a corpus representative of The Times is not enough

- Language is infinite but a corpus has to be finite in size
  - we sample and proportionally include a wide range of text types to ensure maximum balance and representativeness
- Create an inventory of the aspects to be analyzed
  - purpose of the corpus, your research questions
- Goal: create a **balanced** corpus
  - cover a wide range of the aspects which are supposed to be representative of the language (variety) under consideration
  - don't oversample one genre wrt. the others
  - examples:
    - British National Corpus (BNC)
    - Corpus of Contemporary American English (COCA): monitor
    - XX. Mendeko Euskararen Corpusa
    - Lancaster-Oslo/Bergen (LOB) corpus: modern British English in the early 1960s



- Decide the **population**:
  - Written Spanish text published in Cuba, Speech text (dialect), etc.
- Decide **sampling unit**:
  - book, chapter, news article, etc.
- Decide the **sampling frame**
  - List of sampling units
- Two types of sampling:
  - random: sample units randomly from sampling frame
    - correlates with frequency in population, rare features not represented
  - stratified sampling:
    - divide population in groups and sample from each group at random

- Limitations:
  - not easy to create the inventory
  - license issues (once the choice is made)
- Alternative: opportunistic corpus
  - not adhere to a rigorous sampling frame
  - the data that it was possible to gather for a specific task
- Web as a Corpus ([Kilgarriff and Greffensette, 2003](#))
  - Gather all the data you need/can from the web
  - Not balanced, maybe repeated, but size is huge

## 2. Large body of text (bounded in size)



- Corpus should be big in size containing large amount of data
  - spoken or written form
- For example:
  - BNC: 100 M words
  - Gigaword (LDC Catalog): 1750 M words
  - CREA (Corpus de Referencia del Español Actual): 130 M words
  - CORDE (Corpus Diacrónico del Español): 136 M words
  - CORPES XXI (Corpus Español del Siglo XXI): 275 M words
  - XX.MECE: 4 M words
  - Corpus of Contemporary Basque (ETC): 270 M words
- Usually corpora are “closed”
  - exception: monitor corpora

- Not bounded text collection, updated regularly
  - very frequently (daily, monthly)
- Layers of sedimentary rocks: endless layering of sediment layers
- Lexicographic work
  - new words and terms, new senses, etc.
- For example:
  - Bank of English: more than 500 M of words
  - Corpus of Contemporary American English (COCA): balanced
  - CORDE (Corpus Diacrónico del Español): 136 M words
  - Lexikoaren Behatokia: 60 M words
- Pros:
  - not static (diachronic study)
  - the field of study keeps expanding
- Cons:
  - very expensive
  - not very reliable quantitatively speaking

### 3. Machine readable



- Most corpora are readable by machines nowadays: tokens, text structure...
- But look at some old book in Google Books
- Pros:
  - fast usage/query
  - machine tractable
  - tools to handle corpora
    - reveal rare cases, difficult to come up with
- Cons:
  - sometimes OCR must be made

## 4. Standard reference



- Researchers can work with the exact same text
- Reference point to which new work will be compared against
- Using the same data the difference lies in the method
- Ex: What is the best POS tagger?
  - POS annotated corpora:
    - English: Wall Street Journal corpus
    - Spanish: CREA corpus
    - Basque: EPEC corpus

## 5. Additional linguistic information



- If we want to study a particular aspect of language
  - usually hand annotated
  - different annotation tags:
    - Part of Speech
    - Named Entities
    - Word senses
    - ...

# Outline



- 1 Introduction
- 2 Corpus linguistics**
- 3 Building a corpus
- 4 Uses of corpora
- 5 Corpus types
- 6 Corpora and annotation
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond



- Corpus linguistics is a methodology for approaching the study of language
- Study of linguistic phenomena through corpora
  - you have a theory of how language works
  - go to real examples (the corpus) and test
- Science creates hypotheses (models) that are tested against an empirical reality
- It is based on empirical (“real”) data
- In linguistics? **Linguistic models**, but how to test them?

- Methodology:
  - use computers to develop language models
  - use the model to make predictions
  - test the prediction against “real” language
- Example: given  $N$  words, which is the next word? Which probability does it have?

$$P(\text{America} \mid \text{United States of}) = p_1$$

$$P(\text{I'm Erica} \mid \text{United States of}) = p_2$$

$$P(a \mid b) = \frac{\text{freq}(b, a)}{\text{freq}(b)}$$

- A statistical language model is a probability distribution over sequences of words
- Voice recognition, machine translation...

- Mainly used in:
  - Comparative linguistics (comparison of phonology, morphology, syntax and the lexicon of two or more languages)
  - Syntax and semantics (grammar induction based on corpus)
  - Spelling conventions (frequency distribution of letters and sequences)
  - Foreign language (word frequencies, vocabulary lists)
  - ...

# Corpus based (computational) linguistics



- Test the prediction against “real” language
- But, what is “real” language?
  - approximation: obtain a sample of real use
  - corpus → (big) collection of text
- Linguistics has turned from an introspective view to an empiricist view
- Corpora is a representative sample of language, not the whole view of a language

[Fillmore, 1992]

An **armchair linguist** [...] sits in a deep soft comfortable armchair, with his eyes closed and his hands clasped behind his head. Once in a while he opens his eyes, sits up abruptly shouting, "Wow, what a neat fact!", grabs his pencil, and writes something down. Then he paces around for a few hours in the excitement of having come still closer to knowing what language is really like.

An **corpus linguist** [...] has all of the primary facts that he needs, in the form of a corpus of approximately one zillion running words, and he sees his job as that of deriving secondary facts from his primary facts. At the moment he is busy determining the relative frequencies of the eleven parts of speech as the first word of a sentence versus as the second word of a sentence.

# Rationalism vs empiricism



- This is a simplification, but is representative of an open debate.
- There is a tension in AL: rationalism vs empiricism

[Leech, 1992]

*Corpus is a more powerful methodology from the point of view of the scientific method.*

[Fillmore, 1992]

*I don't think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore... [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn't imagine finding out any other way. My conclusion is that the two types of linguists need one another.*

It seems that both introspection and corpora analysis maybe complementary

# Generative linguistics (Chomsky)



- Both sentences below have never occurred in English:
  1. colorless green ideas sleep furiously
  2. furiously sleep ideas green colorless
- However, 1) is grammatical whereas 2) is not.
- Any English speaker can tell the difference (*competence*).

## Chomsky's arguments against the use of corpora

- Natural language is not finite; thus, corpus (which is finite) can not be a good model for human language
- Introspection allows to detect ungrammatical and ambiguous structures
- For example: how can a theory of syntax develop from the observation of utterances?
  - it is only a partly account for the true model of language!



# Rationalism vs empiricism



Example [Morrill, 2000]

- Would you say that? Is it correct?  
The dog that chased the cat that saw the rat that ate the cheese barked
- And this?  
The cheese that the rat that the cat that the dog chased saw ate stank
- Yes, I could say that - but I never would

# Rationalism vs empiricism



Are those sentences possible?

Amonak ekarri du liburua.

Amonak du ekarri liburua.

Amonak ekarri liburua du.

I have a meeting on Monday.

I have a meeting in Monday.

I have a meeting at Monday.

La abuela trajo la comida.

La abuela trajo comida.

La abuela trajo a la comida.

# Rationalism vs empiricism



- Rationalism: expert introspection, rules:

S → NP VP  
NP → Det N  
NP → Det N PP  
PP → Prep NP  
...

- Empiricism: corpus will tell:

*P*(I have a meeting on Monday)

*P*(I have a meeting in Monday)

# Rationalism vs empiricism



## Let's see

Go to [BNC](#) corpus and compare "*on Monday*" with "*in Monday*"

## Let's see

Go to a Spanish corpus and compare  $P(\textit{trajo comida})$  with  $P(\textit{trajo la comida})$

In [CREA](#)

In [Corpus Español](#)

## Let's see

Go to a Spanish corpus and compare  $P(\textit{pienso que})$  with  $P(\textit{pienso de que})$

# Rationalism vs empiricism



- the exact example is not in BNC (*I have a meeting on/in Monday*)
- but, we can decompose the probability

$$P = P(\text{I have}) \times P(\text{have a}) \times P(\text{a meeting}) \times P(\text{meeting on}) \times P(\text{on Monday})$$

$$P(w_1, \dots, w_N) = \prod_1^N P(w_{i+1} \mid w_i)$$

# Rationalism vs empiricism



- By querying corpora, we can come up with valuable information!

Let's see

Go to the BNC corpus and find what preposition follows “collapse”

Let's see

Go to the [CREA corpus](#) and find how many more times it happens (*noun + adj*) than (*adj + noun*)

Let's see

Go to the [CREA corpus](#) and find NOUN derivatives by searching words starting with “conduc\*”

## Problems with rule-based methods:

- Rules never end
  - new cases always arise
  - an informal, incomplete grammar for English runs over 1,700 pages
- Can not properly deal with ambiguity
- No robust: if a case is not foreseen, it will not return nothing
- Huge manual work building the rules
  - problems to maintain coherence among working teams
- Introspection lacks systematicity (mistakes also occur during introspection)

Why is empiricism successful?

- Big success in speech processing
- More and more textual corpora
- Engineering point of view:
  - robust
  - rapid development of tools
- Same method for different domains
- Commercial applications: MT, IR, IE, etc.



- A bag of related techniques: corpus-based, data-intensive, machine learning
- The methods are “fed” with corpora:
  - large set of text
  - many times the information we want to learn is manually annotated
  - corpora use tags to describe the information: POS of words, senses, syntax trees, semantic representation...
- Main problems: sparseness (few data)
  - even very big books contain less than fifty per cent of dictionary words
  - words follow “long tail” frequency patterns (Zipf's law)
  - cases not present in corpus have therefore zero probability in models (smoothing)

# Can we learn rules from examples?



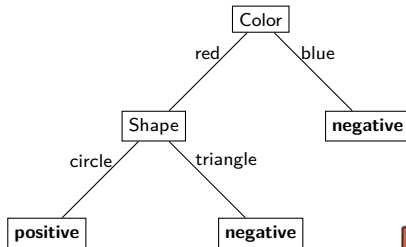
- Hand-made rules

$(Color = red) \wedge (Shape = circle)$   
 $\Rightarrow$  **positive**  
*otherwise*  $\Rightarrow$  **negative**

- Extracted from data

Example	Size	Color	Shape	Class
1	small	red	circle	positive
2	big	red	circle	positive
3	small	red	triangle	negative
4	big	blue	circle	negative

- Decision tree



# Outline



- 1 Introduction
- 2 Corpus linguistics
- 3 Building a corpus**
- 4 Uses of corpora
- 5 Corpus types
- 6 Corpora and annotation
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond

# Building a corpus



- Two main types of corpora
  - general corpus of language (big, balanced)
  - specialized corpus of some domain or genre (small)
- Almost anyone can build a specialized corpus:
  - retrieve documents (for instance, from the web)
  - use computers to extract information
- Build your own corpus to study some particular aspect of language
- Maybe annotate using annotation tools
- But more control over what goes into the corpus.
- More familiar with the content.

# Building a corpus



- DIY approach
- Only a computer is needed
- Many tools (most of them free) for corpus building
  - Text capturing (BootCat)
  - Linguistic annotation (ixa-pipes, spacy, stanford nlp)
  - Text analysis (Antconc, ...)
  - Statistical analysis (R, SPSS, ...)
- It is always good to know how to program.

# Five stages of corpus building



Kennedy (1998) identifies five stages:

1. Design
2. Planning a storage system
3. Obtaining permissions
4. Text capture
5. Markup

# Five stages: 1. Design



- Need a solid design:
  - what is the corpus usage.
  - what research questions.
  - what other corpora to compare with.
- Examples:
  - speech or writing corpus?
  - different time periods and compare among them?
  - ...
- One important question: how **big** will the corpus have to be?
- No single answer.
- In general:
  - focus on rare linguistic feature: lot of data.
  - focus on relatively common linguistic feature: much smaller corpus.

# Five stages: 1. Design



- Size of individual files in the corpus.
- Usually, try to maintain a *balanced* corpus.
- Idea: take *samples* of data
  - BNC, Brown: roughly equally similar sized samples.
- Samples from different parts of text:
  - beginning: "In this essay", "I'm going to", ...
  - end: "in conclusion", "to conclude", ...
- Often useful to carry out a pilot study first.



# Five stages: 1. Design. Data to be collected



- As always, depends on your research questions
- Compare British English and American English
  - collect spoken and / or written data produced by native speakers of the two regional varieties of English
- How Chinese speakers acquire English as a second language,
  - collect the English data produced by Chinese learners to create a learner corpus
- How the language has evolved over centuries
  - collect samples produced in different historical periods to build a historical or diachronic corpus

## Five stages: 2. Storage system



- Deciding what the file system will look like.
  - All documents in a big file.
  - Each text snippet on a different file.
- Usually, store each document in a different file.
  - but again, problem of size.
- Decide the directory structure:
  - one directory per year.
  - one directory per genre.
  - ...

## Five stages: 3. Permissions



- Obtaining permissions from the owner of the texts.
- Texts regulated under copyright laws
  - Newspapers
  - Books, essays.
  - ...
- Vexed issue. Long process.
  - Collecting students texts: aged under 18.
- Rule of thumb:
  - use data for own purposes: no so strict (send letters explaining that the text will be used for some research project).
  - distribute the data: much more complex.

## Five stages: 4. Text capture



- If source is not electronic: convert them to electronic form.
  - By hand (typing on a text editor)
    - takes very long time.
    - error prone.
  - OCR
    - many errors
    - have to be manually corrected
- If spoken corpus: need transcriptions.
- Option: get text from the internet
  - *crawl* the web!
  - Web as a Corpus
  - Many tools o do this.
- Problems:
  - Less control over the quality
  - Copyright issues.
  - Boilerplate in HTML pages (but there are programs, i.e., *beautiful soup*).

## Five stages: 4. Text capture. Useful links



- Oxford Text Archive
  - Oldest text archive - thousands of texts (and many well-known corpora) in more than 25 different languages
- Project Gutenberg
  - Free electronic books
- Digital collections of university libraries e.g.
  - <http://www.digitalcurationsservices.org/digital-stewardship-services/etext-projects/>
  - <http://onlinebooks.library.upenn.edu/>

## Five stages: 5. Markup



- We will talk about markup later on.

# Outline



- 1 Introduction
- 2 Corpus linguistics
- 3 Building a corpus
- 4 Uses of corpora**
- 5 Corpus types
- 6 Corpora and annotation
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond

# Who reads a corpus?



- A corpus is usually too large for anyone to read, e.g. the BNC is very large:
  - It took 4 years to build
  - Contains over 100 million words
  - It comprises 4,124 texts
  - There are circa six and a half million sentences
  - Each word is annotated with a part of speech code
  - It occupies 1.5 gigabytes of disk space
  - Reading the whole corpus aloud at a rate of 150 words a minute, eight hours a day, 365 days a year, would take nearly 4 years
- A computer can scan in a few seconds more text than you can read in your whole life.



- Observe words in context: **concordances** (KWIC: Key Words In Context)
- Analysis of lexicon, terminology, grammar...
- Relationships among words: **collocations**, **colligations**, **semantic preferences**...
- Statistics: frequencies, mutual information...
- Examples: extraction of keywords...

# Concordances



- Comprehensive index of the words in the corpus
- A set of concordance lines
- Usually in a KWIC format - Key Word in Context
  - the word of interest appears in a central position with all lines vertically aligned around the node
- Reveal patterns of usage

# Words in context: concordances (KWIC)



AntConc 3.5.7 (Windows) 2018

File Global Settings Tool Preferences Help

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Corpus Files  
brown\_A.txt  
brown\_B.txt  
brown\_C.txt

Concordance Hits 58

Hit KWIC File

1 for two proposed junior high schools at a Monday night workshop session A bond issue v brown\_A.txt  
2 t approved will follow shortly Before adjournment Monday afternoon the Senate is expected to ap brown\_A.txt  
3 be an orchestra playing nightly except Sunday and Monday for the summer season Mrs. J. Edward brown\_A.txt  
4 rary assistant district attorney it was announced Monday by Charles E. Raymond District Attorne brown\_A.txt  
5 Achievement program begins at St. Thomas Aquinas Monday Subsequent assemblies will be held at brown\_A.txt  
6 one country The Portland school board was asked Monday to take a positive stand towards devel brown\_A.txt  
7 the local police station may be taken at Monday's special North Providence Town Coun brown\_A.txt  
8 well denied all motions made by defense attorneys Monday in Portland's insurance fraud trial Deni brown\_A.txt  
9 i crushing the Reds in a humiliating 13-5 barrage Monday in the loosely played finale With Micke brown\_A.txt  
10 Ave. who died Thursday in Portland will be Monday 1 p.m. at the Riverview Abbey Mr. Brett brown\_A.txt  
11 since Russia's detonation of a super bomb Monday was 4 on Tuesday and 7 last Saturday i brown\_A.txt  
12 affairs told members of the World Affairs Council Monday night Martin who has been in office in brown\_A.txt  
13 out of Texas Tech's sweat-suits drill Monday at Lubbock was tackle Richard Stafford brown\_A.txt  
14 Lawrence J. Sullivan according to a deed filed Monday at City Hall F. Morris Cochran universit brown\_A.txt  
15 The game players saw the Air Force film Monday ran for 30 minutes then went in while brown\_A.txt  
16 dwindled Volume was 1.23 million shares down from Monday's 1.58 million Gains of 2-3/4 were post brown\_A.txt

Total No.  
3  
Files Processed

Search Term Words Case Regex  
Monday  
Advanced  
Start Stop Sort Show Every Nth Row 1  
Kwic Sort  
Level 1 1L Level 2 0 Level 3 0  
Clone Results

# Uses of corpora



Analysis of lexicon, terminology

## Lexicon

In CREA (non annotated) corpus.

Which is used most: *manitas* or *manitos*?

Where is mostly *manitos* used?

## Lexicon

In CREA (annotated) corpus.

Nouns ending in 'o' preceded by article 'la'.

## Terminology

In LB corpus.

Is the term *programazio(-)hizkuntza* used?

Is the term *hizkuntza-teknologia* written with a hyphen?

Analysis of grammar

## Grammar

In CREA corpus.

Look for article + noun examples.

## Grammar

In LB corpus.

Which is the preferred auxiliary verb form for *eutsi*, *dio\** or *du\**?

## Relationships among words

- **Collocations:**  
Relationship between a lexical item and other lexical items (*very collocates with good*)
- **Colligations:**  
Relationship between a lexical item and a grammatical category (*very collocates with ADJ*)
- **Semantic prosody:**  
Meaning arising from collocation. Usually it stretches over more than one word.
- **Semantic preferences:**  
Set of words grouped semantically (by meaning) that refers to a specific subject

## Collocations

Which nouns collocates with *vencer* in [Corpus Español](#)?

Which are the objects of *vencer* (qué cosas son las que vencemos)

Relationships among words at the [BNC](#)

## Collocations

Which adjective collocates with *diamond*?

## Colligations

*he* colligates with verbs, *Mrs.* colligates with proper nouns and determiners  
colligate with nouns

# Uses of corpora: Semantic prosody



Collocational meaning arising from the **semantic relations between node and collocates**

## cause

Analyze collocations of *cause* in [BNC](#)

## causar

Does it happen also in Spanish with the word *causar* in [Corpus Español](#)?

## provide, create

Analyze co-occurrences of *provide, create*

## consequence

Analyze collocations of lemma *consequence* in [BNC](#)



# Uses of corpora: Semantic preference



Collocational meaning arising from the semantic **relations among collocates of a word**

glass of

Which is the semantic class of the collocates of *glass of* in [BNC](#)?

vaso de

Does it happen also in Spanish with the collocates of *vaso de* in [Corpus Español](#)?

large

Which is the semantic class of the collocates of *large* in [BNC](#)?

Polarity analysis

## Polarity

What polarity have the words *boy* and *girl* in BNC?

Compare both words and their adjectives

- Statistics at the LB corpus (<http://lexikoarenbehatokia.euskaltzaindia.eus/cgi-bin/kontsulta.py>)
  - Frequency of the inflected forms of *nabari*
  - Frequency of nouns preceding *nabari*
- Statistics at CREA corpus (<http://web.frl.es/CREA/view/inicioExterno.view>)
  - Statistics of collocations (*arbusto*)
  - Mutual information, log-likelihood simple...
- Statistics at ZT corpus (<http://www.ztcorpusea.eus/cgi-bin/kontsulta.py>)
  - hand corrected vs the whole corpus
  - *energia* (lema) *iturri* (lema)

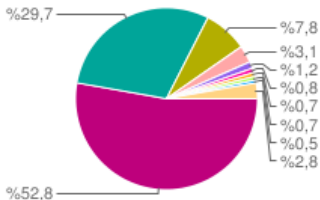
# Frequency of the inflected forms of *nabari*



## Forma

Forma	Kop
 nabari	1590
 nabaria	896
 nabariak	235
 nabariagoa	92
 nabariena	36
 nabariagoak	23
 nabariko	21
 nabariago	20
 nabarian	16
 Beste guztiak	83
Guztira	3012

## Guztien testuinguruak batera



# Outline



- 1 Introduction
- 2 Corpus linguistics
- 3 Building a corpus
- 4 Uses of corpora
- 5 Corpus types**
- 6 Corpora and annotation
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond

- Medium: speech vs written corpus
- Genre and subject
- Language: monolingual vs multilingual
- Other aspects: opportunistic vs balanced, learners', historical or diachronic, monitor vs static corpora...

## Medium: speech corpus



- Database of speech audio files (and sometimes orthographic transcriptions: spoken)
- Time-alignment: transcriptions linked back to the recording
- Contains formal and informal language usage
- Representative of some community, register or style  
e.g. COLT: the Bergen Corpus of London Teenage Language  
COLA: Corpus Oral de Lenguaje Adolescente

## Example



Basque:

*Ni saldu et, en jeneral asko itxekua, itxeko arraya, txipiroyak, berdelak, antxuak, dontxellak, krabak... oyek itxekuak, baño esanian, juate nitzan kayea, orduan emen saltzean besela esta, kayian an Donostiyan, bakisu nun dan, an bapore txikiyak an sartzen tzen, an arrayartu ta bagoitan sartu ta Antxoa saltzea, ta Antxo espasan, Sampedrotikanen onea, ta ordun kaletik barrena ijuka onea, ta espalimaseon, egualdi txarra ta bapore txikiyak espasian eteatzen, Trintxerpea, arrasteroko arraya billa, (...)*

English:

*This guy haxxxd me yo. Traded 6 rebb on 10 key buyout wit 2 buds and 1 bills. He has kickass unusual gibus tho, check it out yo cuz its sweetzzzzzz -rep for hax, +rep for gibus = 0*



- Usually use voice recordings
- Goal: obtain phonetic information to develop speech recognition and synthesis systems
  - **speech recognition:** automatic speech recognition (ASR) or speech to text (STT)
  - **speech synthesis:** the artificial production of human speech. Text-to-speech (TTS) system
- Symbolic representation by means of special phonetic vocabulary
- Sometimes the orthographic representation is also used
  - it is rarely a reliable source of evidence for research into variation in pronunciation

- Example: TIMIT Acoustic-Phonetic Continuous Speech Corpus
  - recordings of 630 speakers of 8 major dialects of American English
  - each reading 10 phonetically rich sentences
  - time-aligned orthographic, phonetic and word transcriptions
  - 16-bit, 16kHz speech waveform file for each utterance
- Sample: <https://catalog.ldc.upenn.edu/LDC93S1>
  - transcript: *She had your dark suit in greasy wash water all year*
  - phonemes sh, ix, hv, eh, dcl, jh, ih, dcl, ...  
phoncode: <https://labur.eus/0JniN>
  - *she, had, ...*
  - audio

- Need for a comprehensive, standardized speech corpus is threefold
  - acquire acoustic-phonetic knowledge for phonetic recognition
  - provide speech for training recognizers
  - provide a common test base for the evaluation of recognizers

# Study of speech corpus



- Converted to normal orthography
- Useful for analysis of discourse and conversations
- Compare textual corpora results with results derived from speech corpora
- Most useful when recording is aligned with transcriptions
- Usually size is smaller than textual corpora: more expensive to build

# Examples of speech corpora



Name	Size	Lang.
Ahotsak	2.8 M	eu
CREA oral	9 M	es
Spoken BNC	10.3 M	en
Common Voice	–	multilingual

- Spoken BNC <http://bnc.phon.ox.ac.uk/data/>:
  - 10% of BNC
  - “BNC spoken demographic”: informal conversations demographically sampled (4.2 M)
  - “BNC spoken context governed”: speech recorded at specific locations for specific events (6.1 M)
- CREA Oral <http://www.rae.es/recursos/banco-de-datos/crea-oral>
  - 1,600 documents: 9 M forms from transcriptions of spoken language
  - 1 M forms aligned text-sound
  - the texts have synchronization marks
  - the segment of the sound file to which the transcription belongs

# Examples of speech corpora



Name	Size	Lang.
Ahotsak	2.8 M	eu
CREA oral	9 M	es
Spoken BNC	10.3 M	en
Common Voice	–	multilingual

- Common Voice <https://commonvoice.mozilla.org/en/datasets>
  - Multilingual
  - Thousands of speakers / recordings hours
  - Anybody can contribute by recording/validating sentences
    - Goal: at least 10k hours validated per language
  - Free (CC0 license)

# Medium: textual corpus



- Collection of text
- Tells us about tendencies and what is normal or typical in real-life language use
- Reveals instances of very rare or exceptional cases: hard to come with from looking at single texts or from introspection
- Basic language resources for creating applications
- Can be easy to gather: for instance, by web crawling

# English corpora: examples



British National Corpus	100 M	British
COCA	385 M	American
TIME Magazine corpus	100 M	(from 1923 - )
OED Corpus historical English	57 M	1200 - XX century
COBUILD	450 M	
GigaWord	1,700 M	
Google Ngrams	1,0 B	(13 M unigrams)
ukWaC...		

- “BNC Informative writing”: world affairs; leisure; arts; commerce and finance; belief and thought; social, applied, natural, and pure sciences (70.9 M)
- “BNC Imaginative writing”: fiction (16.4 M)
- Google Ngrams: useful for statistical language modeling



# Example: BNC



Mode	Category	N
Written 87M words	"Informative writing" World affairs Leisure Arts Commerce and finance Belief and thought Social science Applied science Natural and pure science	70.9 million
	"Imaginative writing" Fiction	16.4 million
Spoken 10M words	"Spoken demographic": informal conversations demographically sampled.	4.2 million
	"Spoken context governed": speech recorded at specific locations for specific events.	6.1 million

- Clarin project <http://clarino.uib.no/korpuskel/corpus-list?session-id=245791836103237>
- Top Ten LDC Corpora <https://catalog.ldc.upenn.edu/topten>
- Corpus Listing for Babel <http://verbs.colorado.edu/corpora/>
- Common Crawl <https://commoncrawl.org/>
  - Huge corpora in many languages (English contains over 100 billion pages)
  - Collected over 8 years of crawling
- Oscar (Open Super-large Crawled ALMAAnaCH) <https://oscar-corpus.com/>
  - language classification and pre-processing of common crawl
  - 166 languages
  - shuffled due to licence issues

# Datasets derived from corpora



- Datasets used to build deep language models
- Derived from crawled corpora
  - cleaned and de-duplicated
- **cc100** dataset
  - Used to train XLM-R model
  - Using common-crawl corpus
  - Cleaned and de-duplicated
- **c4** dataset (Colossal Clean Crawled Corpus)
  - Used to build T5/mT5 models
  - Used common crawl corpus (~ 7 TB)
  - Clean and de-duplicate (~ 335 CPU days)
  - Also multilingual version (**mc4**)

# Basque corpora: examples



EPEC	0.3 M	
XXMECE	4.6 M	
Egungo Testuen Corputa	270 M	2001-2015
Ereduzko Prosa Gaur	25 M	
ZT Corputa	8.5 M	
Lexikoaren Behatokia	60 M	2017
Web as Corpus	124 M	2013
Euscrawl	300M	2021

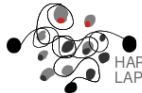
# XX. MENDEKO EUSKARAREN CORPUS ESTATISTIKOA



- Contains 4.658.036 tokens
- Source:
  - most of the Basque material published in the XX century, organized following specific criteria
  - proportional sample of these texts
  - in total there are 6,351 works (“pages”)

- Since 1999 the corpus is closed
- Mostly contains written text, but also some speech
- Organization criteria:
  - 1900-1936: from beginning of century to start of 1936 war
  - 1940-1968: after war texts until Basque standardization
  - 1969-1990: from Basque standardization to the creation of the HLEH dictionary
  - 1991-1999: new criteria for selecting texts

## XX. MECE

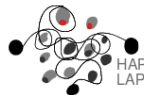


- Covers many dialects:
  - bizkaiera
  - gipuzkera
  - zuberera
  - lapurtera-nafarrera
  - batua (standard)
  - miscellaneous:
    - journals and newspapers
    - collaborative works by many authors

- Type of text:
  - articles from journals (*Euskera, Egan, Euzko Gogoa, Jakin...*)
  - administrative texts
  - learning books
  - literature
  - poems
  - drama
  - bertsoak
  - research works
  - books for kids
  - speech transcriptions
  - liturgy
  - newspapers



## XX. MECE. More characteristics.



- XML encoding (following TEI)
- Manually POS-tagged
- Lots of information for the words (derivations, multiwords, etc.)
- <http://www.euskaracorpora.net/>

# Spanish corpora: examples




- Real Academia Española <http://www.rae.es/>

CREA	130 M	All
CREA	3.6 M	Cuba
CORDE	136 M	Diachronic
CORPES XXI	275 M	30% Spain, 70% America
Corpus Español	2.000 M	Oportunistic

# Spanish corpora (and datasets): more links



- 
- A vertical decorative bar on the left side of the slide, composed of several overlapping oval shapes in shades of grey and black, with a prominent red oval in the middle.
- TIMM resources hub
  - SINAI research group
  - TASS workshop
  - NLP reseach Spanish group
  - DaSCI research institute

- Corpora comprising text from more than one language
- Types:
  - different texts on each language
  - parallel corpora: direct translation of text
  - comparable corpora:
    - same text on different languages (but not translations)
    - usually sentence aligned
    - need to be manually checked
- Example of parallel text:
  - *El libro está en la mesa.*
  - *The book is on the table.*
  - *Liburua mahaian dago.*
- Example of comparable corpora: Wikipedia

# Examples of multilingual corpora



Name	URL	Languages
Europarl	<a href="http://www.statmt.org/europarl/">http://www.statmt.org/europarl/</a>	21 languages
CRATER	<a href="https://labur.eus/ffzk8">https://labur.eus/ffzk8</a>	en; es; fr
Bible	—	many
OPUS	<a href="http://opus.nlpl.eu/">http://opus.nlpl.eu/</a>	> 30 languages
Hansard	<a href="https://labur.eus/XWi9z">https://labur.eus/XWi9z</a>	en; fr
HAC	<a href="https://www.ehu.eus/ehg/hac/">https://www.ehu.eus/ehg/hac/</a>	eu; es; en; fr
ClueWeb09	<a href="https://labur.eus/nCwLP">https://labur.eus/nCwLP</a>	10 langs. (1,000 M web pp.)

## Let's see

Go to the OPUS corpus and select: *en-es / wikipedia / query / "Basque"*  
*alignment=es*

## Let's see

Go to the HAC corpus and select: *eu / emakume*

# Example of parallel corpora



Hansard French/English: the debates in the House and Senate of the 36th Canadian Parliament

- neither the sentence splitting nor the alignments are perfect
- you may want to filter these out before you do any statistical training

Total number of sentences

Name	en	fr	Aligned pairs
House Debats	1,925 K	1,894 K	1,070 K
Senate Debats	281 K	276 K	208 K

# Example of parallel corpora I



`<p id="p1@eu">`  
    `<s id="s1@eu" corresp="s1@sp">` Gipuzkoako Aldizkari Ofiziala 189. Zenbakia  
    Data 1999-10-01 14590 orria 7 UDAL ADMINISTRAZIOA LAZKAOKO UDALA Trafikoko  
    zehapen espedientea irekitzea.`</s>`  
    `<s id="s2@eu" corresp="s1@sp">` José Antonio González Sepúlveda. `</s>`  
`</p>`  
`<p id="p1@sp">`  
    `<s id="s1@sp" corresp="s1@eu s2@eu">` Boletín Oficial de Gipuzkoa Número 189  
    Fecha 01-10-1999 Página 14590 7 ADMINISTRACION MUNICIPAL AYUNTAMIENTO DE  
    LAZKAO Notificación de expte. sancionador por infracción en materia de  
    tráfico a José Antonio González Sepúlveda.`</s>`  
`</p>`

# Example of comparable corpora



To be similar, texts need to share some named entities  
Comparable texts need to be on the same topic

- Wikipedia: 'Computational linguistics' and 'Hizkuntzalaritza aplikatu'
- International Corpus of English (ICE); monolingual  
<http://ice-corpora.net/ice/>
  - around one million words in each of many varieties of English
  - assembled following the same model: common corpus design, and a common scheme for grammatical annotation.
  - prescribes genres and the target quantity of words
- MAtrixware REsearch Collection (MAREC)  
<https://www.kaggle.com/bigquery/marec>



# Outline



- 1 Introduction
- 2 Corpus linguistics
- 3 Building a corpus
- 4 Uses of corpora
- 5 Corpus types
- 6 Corpora and annotation**
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond

# Why annotate a corpus?



- Enables fast and efficient retrieval for humans to analyze
- Explicitly records linguistic analysis
- Provide a standard reference and stable base of linguistic analyses
  - successive studies can be compared and contrasted on a common basis
- Reusable and multifunctional
  - they can be annotated with a purpose and be reused with another

# Corpus annotation vs. mark-up



- Annotation requires interpretation, e.g.
  - part-of-speech
  - word sense (meaning)
  - etc.
- Mark-up provides objective information:
  - metadata: author, document created time, etc.
  - segmentation: paragraph and sentence boundaries, etc.

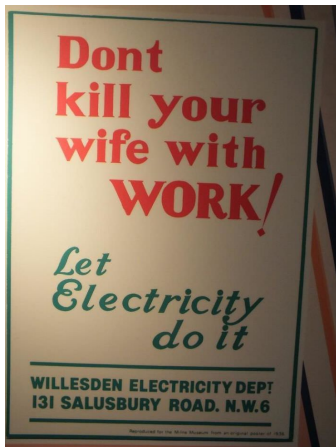
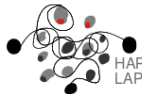
- Unannotated corpus: not markup at all
  - simple raw text
  - linguistic information is implicit:
    - No information of *present* as a noun
    - Computers can't use the linguistic information.
  - for instance, \*.txt, \*.pdf...
- Annotated corpus
  - text as well as linguistic and extra-linguistic information
  - explicitly annotated *hidden* information
  - *present* as a noun, adj or verb.
- Examples:
  - had V VBD
  - darama ADT A1 NOR\_HU NORK\_HU
  - tiene V PR 3 SG
  - La Habana NE.LOC

- **Ambiguity** is a pervasive problem in NLP, one of the most important  
Ambiguity resolution  $\Rightarrow$  Classification

# Ambiguity



# Ambiguity



- Let's see an example:

*He was shot in the hand as he chased the robbers in the back street*

(The Wall Street Corpus)

- Morphosyntactic ambiguity
- Lexical ambiguity
- ...



*He was shot<sub>NN</sub> in the hand<sub>NN</sub> as he chased<sub>JJ</sub> the robbers in  
the back street*

(The Wall Street Corpus)

JJ: Adjective

- **Morphosyntactic ambiguity**

Part of Speech (POS) tagging

*He was shot<sub>NN</sub> in the hand<sub>NN</sub> as he chased<sub>JJ</sub> the robbers in  
the back street<sub>VB</sub> hand<sub>VB</sub> chased<sub>VB</sub>*

(The Wall Street Corpus)

- **Lexical ambiguity**

Word Sense Disambiguation

*He was shot in the hand* body-part *as he chased the robbers in the*  
clock-part  
*back street*

(The Wall Street Corpus)

- **Lexical Ambiguity**

Word Sense Disambiguation

*He was shot in the hand **body-part** as he chased the robbers in  
clock-part  
the back street*

(The Wall Street Corpus)

- **Structural ambiguity (syntactic)**

*He was shot in the hand as he chased the robbers **in the back street***

(The Wall Street Corpus)

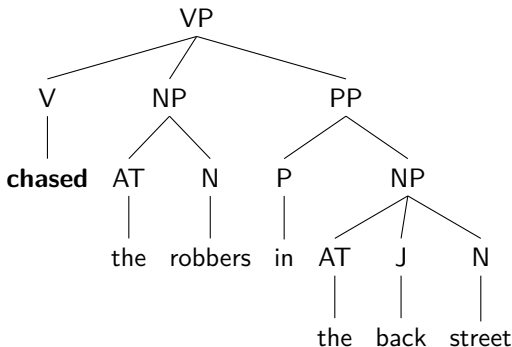
- What does **in the back street** modify: *robbers* or *chased*?

# Annotation and ambiguity



*He was shot in the hand as he (chased (the robbers)<sub>NP</sub> (in the back street)<sub>PP</sub>)*

(The Wall Street Corpus)

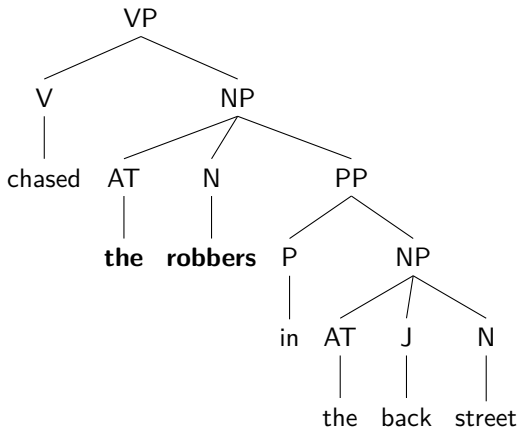


# Annotation and ambiguity



*He was shot in the hand as he (chased ((the robbers)<sub>NP</sub> (in the back street)<sub>PP</sub>)<sub>NP</sub>)*

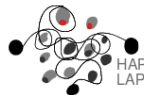
(The Wall Street Corpus)



- The whole NLP may be considered as an ambiguity problem
  - selecting a word (MT)
  - POS
  - syntax (PP attachment...)
  - semantics (polysemy, semantic role...)
  - pragmatics (anaphora resolution...)
- For this:
  - corpus representative of the problem at hand
  - linguists shifted from writing rules to corpus annotation



# See some examples



- Ad-hoc encoding:
  - `ikas_8.lem.esk`
  - `CONLL2009-ST-English-trial.txt`
- XML encoding:
  - `epec_ana.xml`
  - `obama.naf.xml`
- What encoding seems better to you?

# How to annotate corpora



- Automatic: by means of automatic tools
  - annotate large amount of data quickly and at a low cost
  - complete automatic annotation always has errors, difficult to identify
    - post-edition by humans
- Manual
  - Expensive and time consuming
  - Only feasible for small corpora (datasets)
  - There are many tools to help annotating corpora:
    - Brat annotation tool (free)
    - Doccano (free)
    - Prodigy
    - tagtog

# Leech's 7 maxims of annotation



1. Annotation has to be reversible
  - obtain original raw corpus from annotated corpus
2. Possible to extract the annotations by themselves from text
  - flip side of maxim 1
3. The annotation scheme should be based on **guidelines** which are available to the end user
4. It should be made clear **how and by whom the annotation was carried out**

# Leech's 7 maxims of annotation




5. The end user should be made aware that the corpus annotation is not infallible, but simply a potentially useful tool:
  - corpus annotation is an act of interpretation
6. Annotation schemes should be based as far as possible on widely agreed and theory-neutral principles
7. No annotation scheme has the a priori right to be considered as a standard:
  - **standards emerge through practical consensus**

There are three main strategies for annotating text:

- Ad-hoc markup
  - `ikas_8.lem.esk`
  - `CONLL2009-ST-English-trial.txt`
- Use *standard* markup language (XML, but also JSON)
  - in-place annotation: markup is mixed with text
    - TIME-ML: `AQUAINT_timeml.0261.xml`
  - stand-off annotation: markups organized in layers
    - AWA: Annotation Web Architecture (TEI) `amarauna.pdf`
    - NAF: `obama.naf.xml`

- Segmentation and structure of text:
  - title
  - paragraph
  - sentence
  - ...
- Problem: sometimes the structure is too complex
  - how to mark up automatically the file  
`EuropeSEnvironment-TheFourthAssessment.EN.pdf`
- Content vs appearance
- Possible solution: stand-off annotation

- 
- Segmentation (sentences, paragraphs...)
  - Tokens
  - Lexical units:
    - lemma
    - morphosyntax
    - senses
  - Named Entities:
    - date
    - place
    - person
    - quantity
    - senses (Wikipedia, DBpedia)

- Syntax:
  - chunks
  - components
  - dependencies
- Semantics:
  - semantic roles
  - events
- Pragmatics:
  - coreference (anaphora)
  - discourse markers
- Misc.:
  - temporal expressions
  - sentiment analysis
  - factuality

- Also POS tagging
- Convert word forms to root forms (lemmata)
  - *comes*  $\Rightarrow$  *come*
  - *went*  $\Rightarrow$  *go*
  - *darama*  $\Rightarrow$  *eraman*
  - *llevamos*  $\Rightarrow$  *llevar*
- We can analyze all derived forms of lexemes
- Also frequencies and distributions
- Some corpora are annotated at lemma level



# Lemmatization example: English



POS	Form	Lemma
JJ	German	german
NN	luxury	luxury
NN	car maker	car_maker
NNP	Audi	Audi
V	has	have
VBN	overtaken	overtake
NNP	Mercedes	Mercedes
TO	to	to
VB	grab	grab
DT	the	the
NN	number	number

# Lemmatization example: Basque



POS-SUBC	Form	Lemma
DET-DZH	Bi	bi
IZE-ARR	korrikalari	korrikalari
ADI-SIN	hil	hil
ADL	ziren	izan
IZE-LIB	Alemanian	Alemania
IZE-ARR	urtezahar	urtezahar
IZE-ARR	eguneko	egun
ADJ-IZO	San	San
IZE-IZB	Silvestre	Silvestre
IZE-ARR	proban	proba

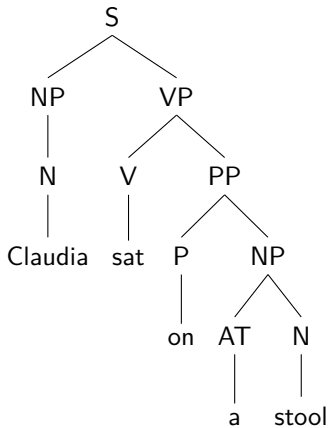
- English morphology is not complex
  - many times a simple stemmer will do the job
  - or a closed list (dictionary) of lemmas
- As a consequence, not many corpora annotated at morphosyntax level
- Basque is another story entirely
  - agglutinated morphology
  - many inflection cases
- Such languages need lemmatization for any analysis

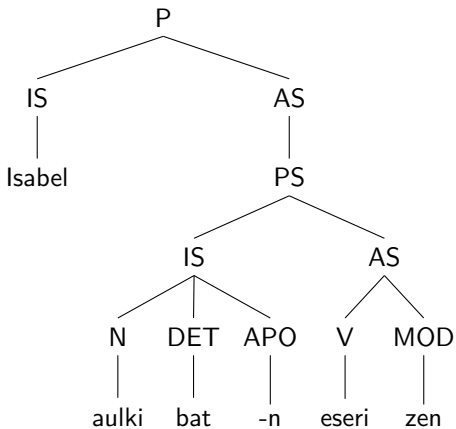
# Morphosyntax example: Basque



POS-SUBC-CASE	Form	Lemma
DET-DZH-NUMP	Bi	bi
IZE-ARR-ABS-MG	korrikalari	korrikalari
ADI-SIN-PART	hil	hil
ADL-B1-NK_HK	ziren	izan
IZE-LIB-INE	Alemanian	Alemania
IZE-ARR	urtezahar	urtezahar
IZE-ARR-GEL	eguneko	egun
IZE-ARR-INE	proban	proba

- Syntactic structure of the sentences
- Syntactically analyzed corpora are often called *treebanks*
- Tree-like structure markup





- Usually represented using brackets:

(P (IS Isabel ) (AS (PS (IS (N aulki ) (DET bat ) (APO -n ) ) (AS (V eseri ) (MOD zen ) ) ) ) )

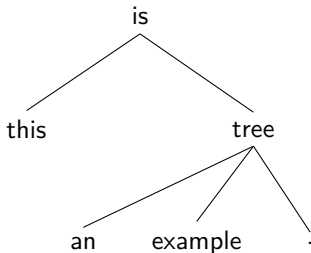
(S (NP (N Claudia )) (VP (V sat ) (PP (P on ) (NP (AT a ) (N stool ) ) ) ) ) )



# Dependency syntax



- Syntactic functions of sentence words
- Binary relation between words
  - CoNLL2009-ST-English-trial.txt
  - obama.naf.xml

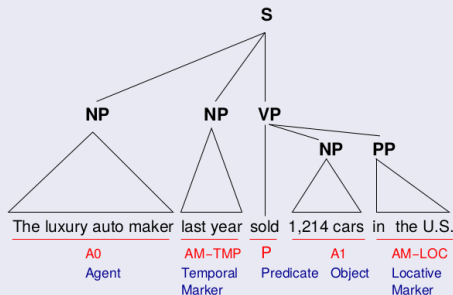


*This is an example tree.*

- **Semantic role:** semantic relations among text elements
- Detecting basic event structures such as:  
*who did what to whom, when and where*
- Task: Semantic Role Labeling (SRL)
- Two models: Framenet or PropBank
  - PropBank: syntax-based approach (focused on verbs)
  - Framenet: situation-based approach (based on semantic frames)

## SRL

The luxury auto maker last year sold 1,214 cars in the U.S.



1

- Framenet: mostly BNC
  - 800 semantic frames
  - 9,000 lexical units
  - 150,000 annotated sentences
  - <http://framenet.icsi.berkeley.edu>
- PropBank: all verbal predicates in WSJ (Penn Treebank)
  - languages other than English (Spanish, Catalan, Chinese...)
  - Basque semantic role: <http://ixa2.si.ehu.es/e-rolda/>
  - <http://verbs.colorado.edu/~mpalmer/projects/ace.html>
- Samples:
  - `wsj-*_prop(_skel)`
  - `CoNLL2009-ST-English-trial.txt`
  - `obama.naf.xml`

- **Word senses:** the different meanings of a word
- Some words have multiple meanings (e.g. *baso*)
- Task: Word Sense Desambiguation (WSD)
- Given a word in context and a fixed inventory of potential word senses:
- Decide which sense of the word this is
- <http://ixa2.si.ehu.es/mcr/>
- Samples:
  - `wsj-*_sense`
  - `obama.naf.xml`

- Textual expressions (mentions) that refer to the same real-world objects (entities)
- Gives cohesion to the text
- Pronouns refer to previous text elements
- Example:

*"I voted for Nader because he was most aligned with my values" she said*

- Samples:
  - obama.naf.xml (coreference)
  - egun.06-1-p2901.2000-06-01.mundua-coref\_level.xml

- Interfaces for manual annotation are needed
- Disagreements between annotators (metrics: *kappa coefficient*)
- When done automatically, how to know the confidence level?
  - obama.naf.xml
- Dealing with ambiguity
  - amarauna.pdf
- How to link annotation layers?
  - obama.naf.xml
  - amarauna.pdf

# Outline



- 1 Introduction
- 2 Corpus linguistics
- 3 Building a corpus
- 4 Uses of corpora
- 5 Corpus types
- 6 Corpora and annotation
  - Annotation and ambiguity
  - How to annotate corpora
  - Usual NLP marks
- 7 And beyond**



*You shall know a word by the company it keeps (Firth)*

- Word cooccurrences (collocations, keywords, etc)
- Hearst patterns:

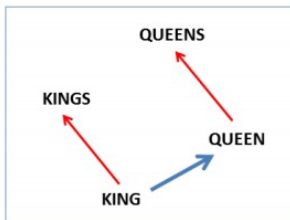
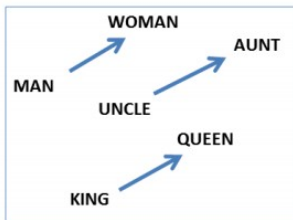
Pattern Name	Pattern structure	Example
HEARST 1	CONCEPT such as (INSTANCE)+ ((and   or) INSTANCE)?	Cities such as Barcelona or Madrid
HEARST 2	CONCEPT (,?) especially (INSTANCE)+ ((and   or) INSTANCE)?	Countries especially Spain and France
HEARST 3	CONCEPT (,?) including (INSTANCE)+ ((and   or) INSTANCE)?	Capitals including London and Paris
HEARST 4	INSTANCE (,)+ and other CONCEPT	Eiffel Tower and other monuments
HEARST 5	INSTANCE (,)+ or other CONCEPT	Coliseum or other historical places

- But much more!

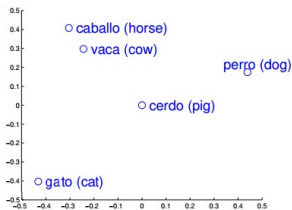
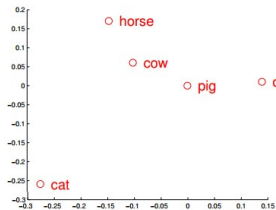
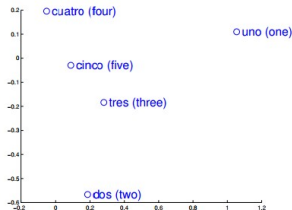
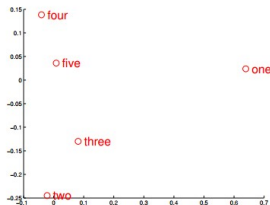
## Applications:


- Word embeddings
  - Continuous representations of words
- Word maps  
<http://vectors.nlp1.eu/explore/embeddings/en/>
- Analogy
- Bilinguality / Multilinguality

# Analogy



# Bilingual vectors



- 
- LMs for generating text
  - Bert's Masked Language Model
  - Another example, GPT2