Lecture 2: Corpus Linguistics

Today

- Corpus Linguistics
  - Historical background
  - Theoretical models
  - Some corpus example
  - Using corpora

References

- Essential references
  - Jurafsky, D. & Martin, J. H. (2000)
    Speech and Language Processing. Prentice-Hall.
    http://www.cs.colorado.edu/~martin/slp.html
    (ch. 2, 3 and 4.. not directly related to this class, but useful for the next two lectures)

- Extended references
  - Manning & Schütze (1999)
    Foundations of statistical natural language processing. MIT press.
  - Lazzari, Bianchi, Cadei, Chesi e Maffei (2010)
    Informatica umanistica. McGraw-Hill (capitolo 4)
    https://www.academia.edu/1836987/Informatica_umanistica
  - Lenci, Montemagni e Pirrelli (2016)
    Testo e computer. Carocci (II Edition)

Someone said...

- But it must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.
  Noam Chomsky (1969:57)

- Anytime a linguist leaves the group the recognition rate goes up.
  Fred Jelinek (IBM Speech Group Project Manager) (1988)
Trends

- Google trends

Hot topics

- Twitter word clouds

Big Data & Machine Learning

- Machine Learning approaches to massive linguist corpora ("big data")

Corpora: what they are, the structure they have

- Finite collections, which are homogeneous and representative with respect to a certain linguistic domain; collected in a systematic and controlled way, corpora should represent the real (qualitative and quantitative) phenomena that are meant to be studied

  - Unstructured (only textual linguistic information)
    ex. text files with non significant formatting (columns, justification...)
  - Structured (precise linguistic data annotation)
    ex. Morphosyntactic tagging (treebank)
  - Semi-structured (convention for conveying extra-linguistic information)
    ex. Html pages, formatted text (titles, paragraphs, turns in conversation ecc.)
Historical background

- Corpus Linguistics (Bloomfield, Harris)
- Advent of computers
  - Enormous storage capability for linguistic data archive
  - Simple and efficient query systems
  - Formal models of language
- Index Thomisticus ([http://www.corpusthomisticum.org/it/index.age](http://www.corpusthomisticum.org/it/index.age))
  - Padre Busa, Gallarate, Centro per l'automazione dell'Analisi Linguistica (1950)
  - Complete collection of Tommaso d'Aquino's writings
  - 10 Millions of tokens (words)
  - Machine readable dictionary
  - Concordances

Lecture 2 - Corpus Linguistics Computational Linguistics - C. Chesi

From punched cards (’50s)

(64 B)

To micro SD cards (2018)

(64 GB = 15,625,000 punched cards...
about 780 boxes containing 20,000
punched cards!)

Corpora: why we need them

- Linguistic documentation: ecological linguistic data sources
- Creation of dictionaries and grammars
- Language models based on frequencies and distributions
- Linguistic benchmark (for NLP tools)

Lecture 2 - Corpus Linguistics

Corpora: classification

- Genericity
  - specialist (or vertical) vs. general (horizontal)
- Modality
  - written vs. spoken vs. mixed
- Time
  - synchronous vs. diachronic
- Language
  - mono vs. multilingual
- Integrity
  - full texts vs. partial texts
- Coding
  - level of annotation
Corpora: other properties

- **Extension**
  «there is no data like more data» (Manning & Schütze 1999)
  ... but focusing only on dimension does not always pay you back (Leech 1991:10)

- **Representatively**
  Web corpus...
  (Google battles... noise...)

- **Closed corpora, monitoring corpora**

Example of (un-)annotated corpus: Brown Corpus

- **Brown corpus** (Francis and Kucera, 1964)
  - 1 Million tokens, representative of written English (500 texts, 1961)
  - 15 categories:
    A. press: reportage (44 texts)
    B. press: editorials (27 texts)
    C. press: periodicals (17 texts)
    D. religion (17 texts)
    E. Skills and hobbies (36 texts)
    F. Popular lore (48 texts)
  -...

Example of (un-)annotated corpus: Brown Corpus

- **Brown corpus** (Francis and Kucera, 1964)
  - Example:
    A01 0010 The Fulton County Grand Jury said Friday an investigation
    A01 0020 of Atlanta's recent primary election produced "no evidence" that
    A01 0030 any irregularities took place. The jury further said in term-end
    A01 0040 presentments that the City Executive Committee, which had over-all
    A01 0050 charge of the election, "deserves the praise and thanks of the
    A01 0060 City of Atlanta" for the manner in which the election was conducted.

Example of (un-)annotated corpus: Italian – La Repubblica

- **Corpus «La Repubblica»**
  - Consistency: 380.000.000 tokens
  - Typology: written corpus based on Italian newspaper Repubblica (articles from 1985 to 2000)
    Various topics: culture, economy, education, news, society, science, sport...
    Semiautomatic POS annotation.
Example of (un-)annotated corpus: Italian – COLFIS

- Consistency: 3,798,275 tokens

Example of (un-)annotated corpus: Italian – LIP

- Consistency: 490,000 tokens
- Typology: spoken language; this is one of the most used corpus in psycholinguistics. Built in 1990-1992 by Tullio De Mauro and colleagues; used using Fondazione IBM Italia technology, the first spoken Italian frequency lexicon. 469 texts collected in 4 cities (Milano, Firenze, Roma e Napoli), 5 macro classes of productions: Type A: face to face conversation (e.g. home-based conversations, workplace conversations, school conversation…). Type B: bidirectional mediated conversation (telephone conversations…).

Before using a corpus...

- Tokenization: What’s a word/token? (spaces, punctuation, quotes, subscripts, numbers...)
- Lemmatization: bello for bello, belli, bella, belle...

Using an (un-)annotated corpus

- Ambiguities: the case of “in” preposition in Italian (http://www.treccani.it/)
Using an (un-)annotated corpus

- **KeyWord in Context (KWIC)**

- **Frequency lexicon** (from «Lessico Elementare», Zanichelli, 1994)

- **Type/Token Ratio (TTR)**

Using an (un-)annotated corpus

- **Frequency lexicon** (from «Lessico Elementare», Zanichelli, 1994)

- **Trivia: Matt Daniels hip-hop corpus**

### Frequency lexicon (from «Lessico Elementare», Zanichelli, 1994)

<table>
<thead>
<tr>
<th>Lemma Forma</th>
<th>Cat.</th>
<th>Rango</th>
</tr>
</thead>
<tbody>
<tr>
<td>casa Sost.</td>
<td>2954</td>
<td>1214</td>
</tr>
<tr>
<td>case Sost.</td>
<td>338</td>
<td>144</td>
</tr>
<tr>
<td>casa Sost.</td>
<td>151</td>
<td>104</td>
</tr>
<tr>
<td>ca' Sost.</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>casette Sost.</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>ca' Sost.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ciasa Sost.</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>kasa Sost.</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

### Type/Token Ratio (TTR)

- Richness of vocabulary, calculated by dividing forms (types) by occurrences (tokens).
- The value goes from 0 (low richness) to 1 (high form variety).

### Trivia: Matt Daniels hip-hop corpus

- # of unique words used within artist's first 35,000 lyrics.
Using an (un-)annotated corpus

- Balancing psycholinguistic experiments
  a. il poliziotto che il maestro ha riconosciuto...
  b. Il poliziotto che lo spazzacamino ha riconosciuto...

- N-grams & Language Models (LM)
  Next word probability: $P(w_n|w_0...w_{n-1})$
  Bayesian approximation: $P(w_n|w_0...w_{n-1}) \approx P(w_n|w_{n-1})$

<table>
<thead>
<tr>
<th>lemma</th>
<th>caption</th>
<th>ass. totale</th>
<th>col. totale</th>
<th>range</th>
<th>set</th>
</tr>
</thead>
<tbody>
<tr>
<td>POLIZIOTTO</td>
<td>Sost.</td>
<td>250</td>
<td>43.93</td>
<td>1721</td>
<td>10</td>
</tr>
<tr>
<td>MAESTRO</td>
<td>Sost.</td>
<td>276</td>
<td>61.33</td>
<td>1293</td>
<td>7</td>
</tr>
<tr>
<td>PAZZACAMINO</td>
<td>Sost.</td>
<td>1</td>
<td>0.02</td>
<td>47002</td>
<td>12</td>
</tr>
</tbody>
</table>

Example of an annotated corpus: Penn Treebank

- Penn Treebank (Marcus & al., 1989-1992)
  - 1 million of tokens (taken from Wall Street Journal 1989)
  - Plus small excerpt from ATIS-3 (Automatic Terminal Information Service)
  - “standard” Treebank II style tagging

Example of an annotated corpus: Penn Treebank

- PENN Tag Set (Marcus & al., 1989-1992)
  - POS Tags: 
    - CC coordinating conjunction
    - CD cardinal number
    - DT determiner
    - EX existential there
    - FW foreign word
    - IN preposition/subordinating conjunction
    - JJ adjective
    - JJR adjective, comparative
    - JJS adjective, superlative
    - LS list marker
    - MD modal
    - NN noun, singular or mass
    - NNS noun plural
    - NNP proper noun, singular
    - NNPS proper noun, plural
    - PDT predeterminer
    - POS possessive ending
    - PRP personal pronoun
    - PRP$ possessive pronoun
    - RB adverb
    - RBR adverb, comparative
    - RBS adverb, superlative
    - RP particle
    - TO to
    - UH interjection
    - VB verb, base form
    - VBD verb, past tense
    - VBG verb, gerund/present participle
    - VBN verb, past participle
    - VBP verb, sing. present, non-3d
    - VBZ verb, 3rd person sing. present
    - WDT wh-determiner
    - WP wh-pronoun
    - WP$ possessive wh-pronoun
    - WRB wh-adverb
  - POS Tag Description Example: 
    - VBD: verb, past tense
Other annotated corpora: Tag Sets

**TANL** (Text Analytics and Natural Language, Attardi & Simi 2009)

<table>
<thead>
<tr>
<th>Tag Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>adjective</strong></td>
<td>bello, buono, brutto</td>
</tr>
<tr>
<td><strong>article</strong></td>
<td>questo, quello</td>
</tr>
<tr>
<td><strong>adverb</strong></td>
<td>velocemente</td>
</tr>
<tr>
<td><strong>conjunction</strong></td>
<td>e, o, e, o</td>
</tr>
<tr>
<td><strong>preposition</strong></td>
<td>di, a, da</td>
</tr>
<tr>
<td><strong>punctuation</strong></td>
<td>., , !, ?</td>
</tr>
<tr>
<td><strong>interjection</strong></td>
<td>beh</td>
</tr>
<tr>
<td><strong>numeral</strong></td>
<td>uno, due</td>
</tr>
<tr>
<td><strong>pronoun</strong></td>
<td>suo, io</td>
</tr>
<tr>
<td><strong>determiner</strong></td>
<td>il, lo</td>
</tr>
<tr>
<td><strong>noun</strong></td>
<td>cane</td>
</tr>
<tr>
<td><strong>predeterminer</strong></td>
<td>tutti, ogni</td>
</tr>
<tr>
<td><strong>verb</strong></td>
<td>come</td>
</tr>
<tr>
<td><strong>class residual</strong></td>
<td>SpA</td>
</tr>
</tbody>
</table>

Other annotated corpora: Tag Sets

**TANL** (Text Analytics and Natural Language, Attardi & Simi 2009)

<table>
<thead>
<tr>
<th>Category</th>
<th>Gender</th>
<th>Number</th>
<th>Person</th>
<th>Mode</th>
<th>Tense</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (adjective)</td>
<td>m (male), f (feminine), n (non-specific)</td>
<td>s (singular), p (plural), n (non-specific)</td>
<td>1 (first), 2 (second), 3 (third)</td>
<td>i (indicative), m (imperative), c (subjunctive), d (conditional)</td>
<td>p (past), i (imperfect), f (future)</td>
<td>bello, buono, brutto</td>
</tr>
</tbody>
</table>

Other annotated corpora: POS Tagging

**PoS tagging example**

<table>
<thead>
<tr>
<th>Italian</th>
<th>PoS Tag (TANL Tagset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ben</td>
<td>A</td>
</tr>
<tr>
<td>pensò</td>
<td>Vt</td>
</tr>
<tr>
<td>fui</td>
<td>R</td>
</tr>
<tr>
<td>intervistò</td>
<td>Vrs</td>
</tr>
<tr>
<td>delf</td>
<td>EArS</td>
</tr>
<tr>
<td>è</td>
<td>isolare</td>
</tr>
<tr>
<td>sono</td>
<td>3p</td>
</tr>
<tr>
<td>Vab</td>
<td>Vab</td>
</tr>
<tr>
<td>voleva</td>
<td>VrH</td>
</tr>
<tr>
<td>in</td>
<td>I</td>
</tr>
<tr>
<td>porto</td>
<td>Sis</td>
</tr>
<tr>
<td>con</td>
<td>S</td>
</tr>
<tr>
<td>interessato</td>
<td>Sis</td>
</tr>
<tr>
<td>ai</td>
<td>Is</td>
</tr>
</tbody>
</table>

---

**Other annotated corpora:**

**Tag Sets**

- **TANL** (Text Analytics and Natural Language, Attardi e Simi 2009)

<table>
<thead>
<tr>
<th>Tag Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>adjective</strong></td>
<td>bello, buono, brutto</td>
</tr>
<tr>
<td><strong>article</strong></td>
<td>questo, quello</td>
</tr>
<tr>
<td><strong>adverb</strong></td>
<td>velocemente</td>
</tr>
<tr>
<td><strong>conjunction</strong></td>
<td>e, o, e, o</td>
</tr>
<tr>
<td><strong>preposition</strong></td>
<td>di, a, da</td>
</tr>
<tr>
<td><strong>punctuation</strong></td>
<td>., , !, ?</td>
</tr>
<tr>
<td><strong>interjection</strong></td>
<td>beh</td>
</tr>
<tr>
<td><strong>numeral</strong></td>
<td>uno, due</td>
</tr>
<tr>
<td><strong>pronoun</strong></td>
<td>suo, io</td>
</tr>
<tr>
<td><strong>determiner</strong></td>
<td>il, lo</td>
</tr>
<tr>
<td><strong>noun</strong></td>
<td>cane</td>
</tr>
<tr>
<td><strong>predeterminer</strong></td>
<td>tutti, ogni</td>
</tr>
<tr>
<td><strong>verb</strong></td>
<td>come</td>
</tr>
<tr>
<td><strong>class residual</strong></td>
<td>SpA</td>
</tr>
</tbody>
</table>

**Features Values**

- **gender:** m (male), f (feminine), n (non specific)
- **number:** s (singular), p (plural), n (non specific)
- **person:** 1 (first), 2 (second), 3 (third)
- **mode:** i (indicative), m (imperative), c (subjunctive), d (conditional)
- **tense:** p (past), i (imperfect), f (future)

**Principal category**

- **N (nouns):**
  - Arts (agg. mas. sing.)
  - Amr (agg. mas. plur.)
  - Arn (agg. gen. sing.)
  - Arn (agg. gen. plur.)
  - Term (adj.)
- **V (verbs):**
  - Vbc (imp. coniug.)
  - Vbc (infinitive)
- **A (adjectives):**
  - Ade (agg. mas. sing.)
  - Adr (agg. mas. plur.)
  - Arn (agg. gen. sing.)
  - Arn (agg. gen. plur.)
- **P (prepositions):**
  - Diff (definite)
  - Int (indefinite)
- **O (other):**
  - Altro (other)

**Features in prevalent use**

- **gender:** m (male), f (feminine), n (non specific)
- **number:** s (singular), p (plural), n (non specific)
- **person:** 1 (first), 2 (second), 3 (third)
- **mode:** i (indicative), m (imperative), c (subjunctive), d (conditional)
- **tense:** p (past), i (imperfect), f (future)

**Principal category Category with features**

- **Ams (agg. mas. sing.):**
  - bello, buono, brutto
- **Amp (agg. mas. plur.):**
  - buono, brutto
- **Arn (agg. gen. sing.):**
  - bello, buono
- **Arn (agg. gen. plur.):**
  - buono, brutto

**Features in prevalent use**

- **gender:** m (male), f (feminine), n (non specific)
- **number:** s (singular), p (plural), n (non specific)
- **person:** 1 (first), 2 (second), 3 (third)
- **mode:** i (indicative), m (imperative), c (subjunctive), d (conditional)
- **tense:** p (past), i (imperfect), f (future)

**Principal category**

- **Ams (agg. mas. sing.):**
  - bello, buono, brutto
- **Amp (agg. mas. plur.):**
  - buono, brutto
- **Arn (agg. gen. sing.):**
  - bello, buono
- **Arn (agg. gen. plur.):**
  - buono, brutto

**Features in prevalent use**

- **gender:** m (male), f (feminine), n (non specific)
- **number:** s (singular), p (plural), n (non specific)
- **person:** 1 (first), 2 (second), 3 (third)
- **mode:** i (indicative), m (imperative), c (subjunctive), d (conditional)
- **tense:** p (past), i (imperfect), f (future)

**Principal category**

- **Ams (agg. mas. sing.):**
  - bello, buono, brutto
- **Amp (agg. mas. plur.):**
  - buono, brutto
- **Arn (agg. gen. sing.):**
  - bello, buono
- **Arn (agg. gen. plur.):**
  - buono, brutto

**Features in prevalent use**

- **gender:** m (male), f (feminine), n (non specific)
- **number:** s (singular), p (plural), n (non specific)
- **person:** 1 (first), 2 (second), 3 (third)
- **mode:** i (indicative), m (imperative), c (subjunctive), d (conditional)
- **tense:** p (past), i (imperfect), f (future)

**Principal category**

- **Ams (agg. mas. sing.):**
  - bello, buono, brutto
- **Amp (agg. mas. plur.):**
  - buono, brutto
- **Arn (agg. gen. sing.):**
  - bello, buono
- **Arn (agg. gen. plur.):**
  - buono, brutto

**Features in prevalent use**

- **gender:** m (male), f (feminine), n (non specific)
- **number:** s (singular), p (plural), n (non specific)
- **person:** 1 (first), 2 (second), 3 (third)
- **mode:** i (indicative), m (imperative), c (subjunctive), d (conditional)
- **tense:** p (past), i (imperfect), f (future)

**Principal category**

- **Ams (agg. mas. sing.):**
  - bello, buono, brutto
- **Amp (agg. mas. plur.):**
  - buono, brutto
- **Arn (agg. gen. sing.):**
  - bello, buono
- **Arn (agg. gen. plur.):**
  - buono, brutto
XML annotation

- Inclusion indicates constituents:
  - Parentheses: 
    ```
    [[A] [ B C ]]
    ```
  - HTML: 
    ```
    <p> <i>123</i> <b>Mario Rossi</b> </p>
    ```
  - XML: 
    ```
    <student> <id> 123 </id> 
    <name> Mario Rossi </name> 
    </student>
    ```

Using annotated corpora

- Grammar extraction
- Benchmark for POS Tagging & Parsing tools
- Linguistic studies: frequencies of forms and syntactic patterns (retrieved/counted using specific queries)

Using semi-structured corpora

- Childes (MacWhinney & Snow, 1985)
  - Child Language Data Exchange System is an archive of spontaneous speech transcription between children and adults (each transcription is about 20-60 minutes long).
  - [http://childes.psy.cmu.edu](http://childes.psy.cmu.edu) more than 130 corpora, 1500 published articles...

- CHAT coding sample
  ```
  @UTF8
  @Begin
  @Participants: CHI Cam Target_Child, DON Mother
  @Age of CHI: 3:4:9
  @Sex of CHI: female
  @Birth of CHI: 3-MAY-1988
  @Date: 12-SEP-1991
  *DON: quale volevi ?
  *CHI: io volevo questo .
  *DON: si ma cosa, che canzoni ci sono, sopra .
  *CHI: non lo so .
  *DON: come non lo sai ?
  (...) 
  @End
  ```

Using semi-structured corpora

- Childes (MacWhinney & Snow, 1985)
  - CHAT coding sample
  ```
  @UTF8
  @Begin
  @Participants: CHI Cam Target_Child, DON Mother
  @Age of CHI: 3:4:9
  @Sex of CHI: female
  @Birth of CHI: 3-MAY-1988
  @Date: 12-SEP-1991
  *DON: quale volevi ?
  *CHI: io volevo questo .
  *DON: si ma cosa, che canzoni ci sono, sopra .
  *CHI: non lo so .
  *DON: come non lo sai ?
  (...) 
  @End
  ```
Using semi-structured corpora

- Childes (MacWhinney & Snow, 1985)

Example of linguistic questions:
«are children sensitive to the finiteness of the verb?»

in French we can use negation («je ne mange pas» vs. «ne pas manger»)
in Italian clitics distribution ("lo mangio" Vs. "mangio lo"; "mangiare-lo" Vs. "lo mangiare")

● non puoi fam-mi questo (Diana 2 anni e 5 mesi)
● mi son fatta male


This supports the truncation thesis (Rizzi 1993-94)

Finite Non-finite
- Declaratives 3768 721 (about 20%)
- Wh-questions 80 2 (about 2%)
Today’s key concepts

- What’s a **Corpus** (finite collection of linguistic information)
- **Corpus typologies** (unannotated vs annotated)
- **Corpus examples** (Brown Corpus, PENN Treebank, Repubblica... CHILDES)
- What’s a corpus for (frequencies, grammar extraction, benchmark, linguistic questions...)