Computational Linguistics
A.Y. 2018/19 – C. Chesi

Lecture 2

CORPUS LINGUISTICS

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**
  - Kennedy, G., Leech, G., & Short, M. (1998)
    An introduction to corpus linguistics. London: Longman.
  - 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)

Today

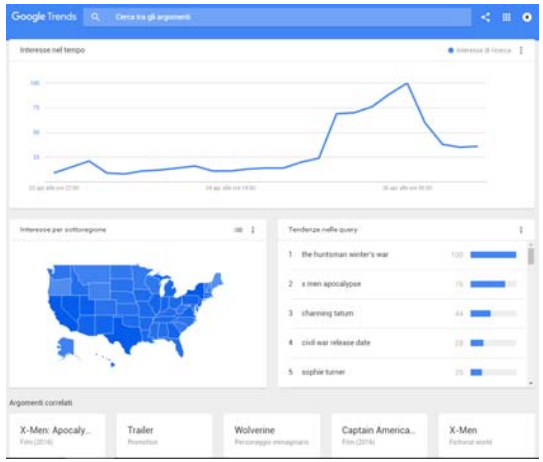
- **Corpus Linguistics**
  - Historical background
  - Theoretical models
  - Some corpus example
  - Using corpora

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

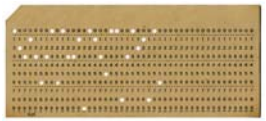
- **Finite** collections, which are **homogeneous** and **representative** with respect to a certain linguistic domain; collected in a **systematic** and **controlled** way, corpora should **represent the real** (qualitative and quantitative) **phenomena** that are meant to be studied
  - **Unstructured**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

Historical background

- 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)

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.
  - **Reference**: M. Baroni, S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston, M. Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian.Proceedings of LREC 2004.

Example of (**un-**)**annotated** corpus: Italian – COLFIS

- **Corpus e Lessico di Frequenza dell'Italiano Scritto (COLFIS)**
  - **Consistency**: 3.798.275 tokens
  - **Typology**: written corpus, texts taken from newspapers and magazines 1992-1994 ('La Repubblica', 'La Stampa', 'Il Corriere della Sera'), books:
    newspapers 1.836.119
    magazines 1.306.653
    books 655.503
    (the sampling has been carefully studied, using ISTAT data: representative lectures of Italian people; this is a nice **balancing** methodology)
  - **Reference**: Bertinetto P. M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., Thornton A. Maria. (2005) *Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)*.

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

- **Lessico di frequenza dell'italiano parlato, LIP** (http://badip.uni-graz.at/it/)
  - **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…) …
  - **Reference**:
    De Mauro T. , F. Mancini, M., Vedovelli, M. Voghera (1993)
    Lessico di frequenza dell'italiano parlato, Milano, Etaslibri.

Before using a corpus…

- **Text Normalization**
  Il sig. P. Pallino rappresentato e difeso dall'avv. Mario Rossi, notifica, ex-art.150 C.P.C., agli eredi e/od aventi causa di Gianni Bianchi, nato a Castelnuovo V.C. (PI) il 1° aprile 1908 e deceduto in Sassari (SA) l'11 aprile 2008, presso il Tribunale di Blablabla Sez. Distaccata, l'atto di sostituzione della locuzione 'Figura in Catasto alla partita terreni 953, foglio XI, mappale 335, are 1,59' con la seguente locuzione: 'figura in Catasto alla partita terreni 953, fogli X-XI, mappale 325, are 00,96'.

  sig. > signore (o signora?)
- **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/)

COMPLEMENTO DI → STATO IN LUOGO	Lo trovi **in** stazione
COMPLEMENTO DI → MOTO A LUOGO	Torniamo **in** Italia
COMPLEMENTO DI → MOTO PER LUOGO	Passò **in** corridoio come un fulmine
COMPLEMENTO DI → TEMPO DETERMINATO	**Nel** mese di aprile si seminano i pomodori
COMPLEMENTO DI → TEMPO CONTINUATO	Scriverò il nuovo libro **in due mesi**
COMPLEMENTO → PREDICATIVO DELL'OGGETTO	Gli ho dato in dono **un** cellulare
COMPLEMENTO DI → MATERIA	Tubi **in** titanio
COMPLEMENTO DI → LIMITAZIONE	Paolo è bravo **in** italiano
COMPLEMENTO DI → MEZZO O STRUMENTO	Ho viaggiato **in** treno
COMPLEMENTO DI → MODO O MANIERA	Bisogna fare **in** fretta
COMPLEMENTO DI → MISURA	Siamo **in** venti
COMPLEMENTO DI → PREZZO O STIMA	Ti tengo **in** grande considerazione
COMPLEMENTO DI → CAUSA	Esulto **nel** ricordo della vittoria
COMPLEMENTO DI → FINE O SCOPO	Mandarono l'autoambulanza **in** soccorso dei feriti
COMPLEMENTI DI → VANTAGGIO E SVANTAGGIO	L'ho fatto **nel** tuo interesse L'ha fatto **in** spregio di te

Using an (un-)annotated corpus

- **KeyWord in Context (KWIC)**

Contesto sinistro	keyword	Contesto destro
esattezza: contare oggetti, ordinarli	**in**	figure geometriche, risolvere problemi
ambiguo, anch' egli si sentiva annegare	**in**	questa morbida penombra, non riusciva p cui
io. stava male: erano quelli i momenti	**in**	cui si sentiva venir meno; alle volte s
iasi cosa avesse davanti. o a metterle	**in**	fila, a ordinarle in quadrati o piramid
nti. o a metterle in fila, a ordinarle	**in**	quadrati o piramidi. l' applicarsi a q

Using an (un-)annotated corpus

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

Rank	Lemma	Gramm. Category	Frequency
1	lo	art/pron	48101,08
2	essere	v	43777,54
3	e	cong	41043,77
4	il	art	35677,16
5	uno	agg/art/pron	29119,51
6	di	prep	26673,87
7	a	prep	22277,41
8	che	agg/cong/pron	20081,16
9	avere	v	18371,11
10	io	pron	17333,47

- **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)

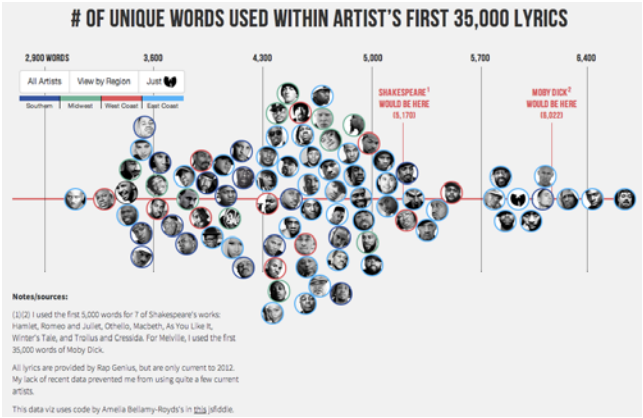
Using an (un-)annotated corpus

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

(F. = frequency, D. = Dispersion)

Lemma	Forma	Cat. gramm.	F. ass. totale	F. ass. quotidiani	F. ass. periodici	F. ass. libri	D. totale	D. quotidiani	D. periodici	D. libri	F. rel. totale	F. rel. quotidiani	F. rel. periodici	F. rel. libri	Rango
CASA		Sost.	2954	1214	1127	613	0.9572	0.9248	0.9778	0.9147	746.09	308.93	423.22	432.28	92
	casa	Sost.	2583	1063	970	550	0.9488	0.9135	0.9699	0.8989	646.98	267.67	361.81	382.00	133
	case	Sost.	338	144	138	56	0.9160	0.9091	0.8922	0.8669	81.88	36.09	47.93	37.68	972
	casa	Sost.	151	104	38	9	0.6890	0.6914	0.7489	0.4242	27.98	20.39	11.55	3.47	2692
	casetta	Sost.	23	6	10	7	0.5011	0.4673	0.4129	0.3116	3.17	0.90	1.90	1.96	13898
	ca'	Sost.	11	4	7	0	0.4098	0.0000	0.4738	0.0000	1.27	0.11	1.48	0.00	24018
	casette	Sost.	7	0	7	0	0.3592	0.0000	0.5097	0.0000	0.74	0.00	1.61	0.00	31955
	case	Sost.	3	3	0	0	0.0000	0.0000	0.0000	0.0000	0.05	0.10	0.00	0.00	84318
	casas	Sost.	1	0	1	0	0.0000	0.0000	0.0000	0.0000	0.03	0.00	0.11	0.00	88160
	ca'	Sost.	1	1	0	0	0.0000	0.0000	0.0000	0.0000	0.02	0.03	0.00	0.00	98248
	case	Sost.	1	0	1	0	0.0000	0.0000	0.0000	0.0000	0.01	0.00	0.04	0.00	129648
	ciasa	Sost.	1	0	1	0	0.0000	0.0000	0.0000	0.0000	0.01	0.00	0.02	0.00	129648
	ca'	Sost.	2	0	0	2	0.0000	0.0000	0.0000	0.0000	0.01	0.00	0.00	0.22	129648
	kasa	Sost.	1	0	1	0	0.0000	0.0000	0.0000	0.0000	0.00	0.00	0.01	0.00	170426

Trivia: Matt Daniels hip-hop corpus



Using an (un-)annotated corpus

- **Balancing psycholinguistic experiments**

  a. il **poliziotto** che il **maestro** ha riconosciuto...
  *the **policeman** that the **teacher** recognized*

  b. Il **poliziotto** che lo **spazzacamino** ha riconosciuto...
  *the **policeman** that the **chimneysweep** recognized*

Lemma	Cat. gramm.	F. ass. totale	F. rel. totale	Rango	Len
POLIZIOTTO	Sost.	250	43.93	1723	10
MAESTRO	Sost.	276	61.33	1293	7
PAZZACAMINO	Sost.	1	0.02	47102	12

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

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:
    (S (PP (IN **Of**) (NP (NN **course**))) (, ,) (S (S (NP (DT **some**) (PP (IN **of**) (NP (PRP$ **my**) (NN **color**) (NNS **values**)))) (AUX (VBP **do**)) (NEG (RB **not**)) (VP (VB **match)** (NP (NP (DT **the**) (JJ **old**) (NN **Master**)) (POS **'s**)))) (CC **and**) (S (NP (DT **the**) (NN **perspective**)) (VP (VBZ **is**) (ADJP (JJ **faulty**)))) (CC **but**) (S (NP (PRP **I**)) (VP (VBP **believe**) (S (NP (PRP **it**)) (AUX (TO **to**)) (VP (VB **be**) (NP (DT **a**) (JJ **fair**) (NN **copy**)))))))))

Example of an **annotated** corpus: Penn Treebank

- **Penn Treebank** (Marcus & al. , 1989-1992)
  - Formatted example:

```
(S
   (PP
         (IN Of)
         (NP
               (NN course)
         )
   )
   (, ,)
   (S
         (S
               (NP
                     (DT some)
                     (PP
                           (IN of)
                           (NP (PRP$ my)
                           (NN color) …
```


Example of an **annotated** corpus: Penn Treebank

- PENN Tag Set (Marcus & al. , 1989-1992)

POS Tag	Description	Example	POS Tag	Description	Example
CC	coordinating conjunction	and	PRP$	possessive pronoun	my, his
CD	cardinal number	1, third	RB	adverb	however, usually, naturally, here, good
DT	determiner	the			
EX	existential there	*there* is	RBR	adverb, comparative	better
FW	foreign word	d'hoevre	RBS	adverb, superlative	best
IN	preposition/subordinating conjunction	in, of, like	RP	particle	give *up*
			TO	to	*to* go, *to* him
JJ	adjective	green	UH	interjection	uhhuhhuhh
JJR	adjective, comparative	greener	VB	verb, base form	take
JJS	adjective, superlative	greenest	VBD	verb, past tense	took
LS	list marker	1)	VBG	verb, gerund/present participle	taking
MD	modal	could, will			
NN	noun, singular or mass	table	VBN	verb, past participle	taken
NNS	noun plural	tables	VBP	verb, sing. present, non-3d	take
NNP	proper noun, singular	John	VBZ	verb, 3rd person sing. present	takes
NNPS	proper noun, plural	Vikings	WDT	wh-determiner	which
PDT	predeterminer	*both* the boys	WP	wh-pronoun	who, what
POS	possessive ending	friend*'s*	WP$	possessive wh-pronoun	whose
PRP	personal pronoun	I, he, it	WRB	wh-abverb	where, when

Other **annotated** corpora: Tag Sets

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

Tag	Description	Example
A	aggettivo	bello
B	avverbio	velocemente
C	congiunzione	e, o
D	determinante	questo, quello
E	preposizione	di, a, da
F	punteggiatura	. , ! ?
I	Interiezione	beh
N	numerale	uno, due
P	pronome	suo, io
R	articolo	il, lo
S	nome	cane
T	predeterminante	tutti, ogni
V	verbo	corre
X	classe residuale	SpA

Other **annotated** corpora: Tag Sets

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

categoria	descrizione	esempi	contesto d'uso
A	aggettivo	bello, buono, bravo	una bella passeggiata, una persona brava
AP	aggettivo possessivo	mio, tuo, nostro	a mio parere, il tuo libro
B	avverbio	bene, fortemente, malissimo, domani	arrivo domani sto bene
BN	avverbio negativo	non	non sto bene
CC	congiunzione coordinativa	e, o, ma	i libri e i quaderni, vengo ma non rimango
CS	congiunzione subordinativa	mentre, quando	quando ho finito vengo, mentre parlava rideva
DD	determinante dimostrativo	questo, codesto, quello	questo denaro, quella famiglia
DE	determinante esclamativo	che, quale, quanto	che disastro! quale catastrofe!
DI	determinante indefinito	alcuno, certo, tale, parecchio, qualsiasi	alcune telefonate, parecchi giornali, qualsiasi persona
DQ	determinante interrogativo	cui, quale	i cui libri, quale intervista
DR	determinante relativo	che, quale, quanto	Che cosa, quanta strada, quale formazione
E	preposizione	di, a, da, in, su, attraverso, verso, prima_di	a casa, prima_di giorno verso sera
...			

Other **annotated** corpora: Tag Sets

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

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), g (gerundive), f (infinite), p (participle)
tense	p (present), i (imperfect), s (past), f (future)

Principal category	Category with features	Example
A (aggettivo)	Ams (agg. masc. sing.)	tossico, doppio, italiano ...
	Amp (agg. masc. plur.)	chimici, tossici, giudiziari ...
	Afs (agg. fem. sing.)	moderna, splendida, clamorosa ...
	Afp (agg. masc. plur.)	numerose, belle, antiche ...
	Ans (agg. genere non spec. sing.)	speciale, londinese, lunghista ...
	Anp (agg. genere non spec. plur.)	trasparenti, mondiali, pesanti, naturali ...
	Ann (agg. genere e numero non spec.)	top_secret, ex, pari ...

Other **annotated** corpora: POS Tagging

- *PoS tagging* example

token	PoS Tag (TANL tagset)
A	E
ben	B
pensarci	Vfc
,	FF
l'	RDns
intervista	Sfs
dell'	EAns
on.	SA
Formica	SP
è	VAip3s
stata	VApsfs
accolta	Vpsfs
in	E
genere	Sms
con	E
disinteresse	Sms
.	FS

XML annotation

- Inclusion indicates **constituents**:
  - Parentheses [[A] [B C]]
  - HTML 123 Mario Rossi
  - 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)
  - (**Chi**ld **L**anguage **D**ata **E**xchange **S**ystem) is an archive of spontaneous speech transcription between children and adults (each transcription is about 20-60 minutes long).
  - http://childes.psy.cmu.edu
    more than 130 corpora, 1500 published articles…

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)

Words
@ special form markers
xxx unintelligible speech, not treated as a word
xx unintelligible speech, treated as a word
yyy unintelligible speech transcribed on %pho line, not treated as a word
yy unintelligible speech transcribed on %pho line,treated as a word
www untranscribed material
0 actions without speech
& phonological fragment
[?] best guess
text(text)text noncompletion of a word
0word omitted word
0*word ungrammatical omission
00word (grammatical) ellipsis

Basic Utterance Terminators
. period
? question
! exclamation

Tone Unit Marking
-? rising final contour
-! final exclamation contour
-. falling final contour
-'. rise-fall final contour
-,. fall-rise final contour
-, level nonfinal contour
-_ falling nonfinal contour
- low level contour
-' rising nonfinal contour
, syntactic juncture
„ tag question
pause between words

-: previous word lengthened

Prosody Within Words
/ stress
// accented nucleus
/// contrastive stress
: lengthened syllable

Dependent Tiers
%act: actions
%add: addressee
%alt: alternative transcription
%cod: general purpose coding
%eng: English translation
%err: error coding
%exp: explanation
%fac: facial actions
…

Using **semi-structured** corpora

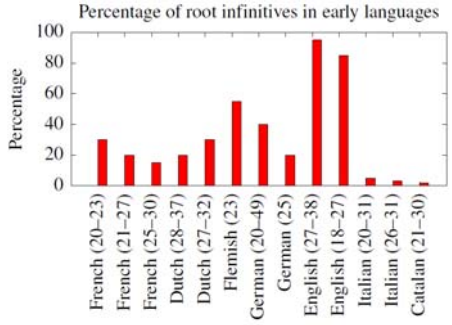
- Example of linguistic questions:
  «are children sensitive to the finiteness of the verb?»

  in French we can use **negation** («je **ne** **mange** **pas**» vs. «**ne** **pas** **manger**»)
  in Italian **clitics** distribution
  ("*lo* mangio" Vs. *"mangio *lo*"; "mangiar-*lo*" Vs. *"*lo* mangiare")
  (Guasti 1993-94):

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

Using **semi-structured** corpora

- **Root Infinitives** (Haegeman 1995, Bromberg & Wexler 1995, Crisma 1992 …)



Using **semi-structured** corpora

- **Root Infinitives** (Haegeman 1995, Bromberg & Wexler 1995, Crisma 1992 …)

	Finite	Non-finite	
• Declaratives	3768	721	(about 20%)
• *Wh*-questions	80	2	(about 2%)

This supports the **truncation** thesis (Rizzi 1993-94)

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…)