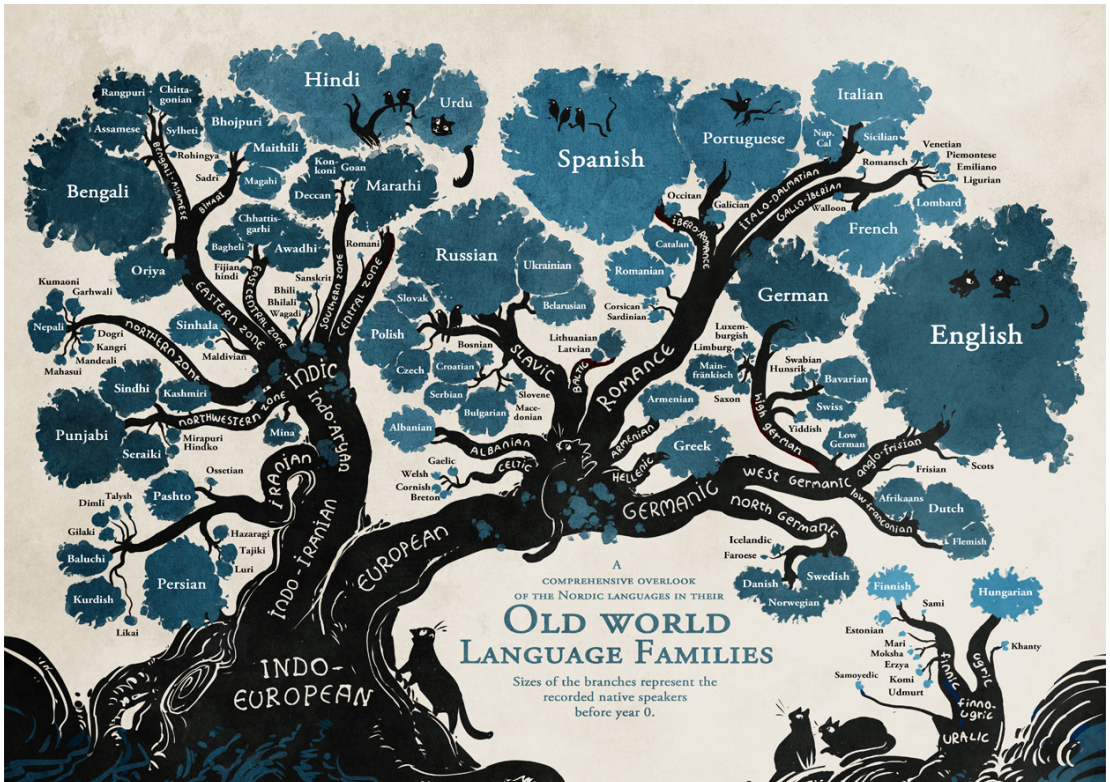
Risorse linguistiche e semantica computazionale

Andrea Amelio Ravelli | Lorenzo Gregori

Università degli Studi di Firenze

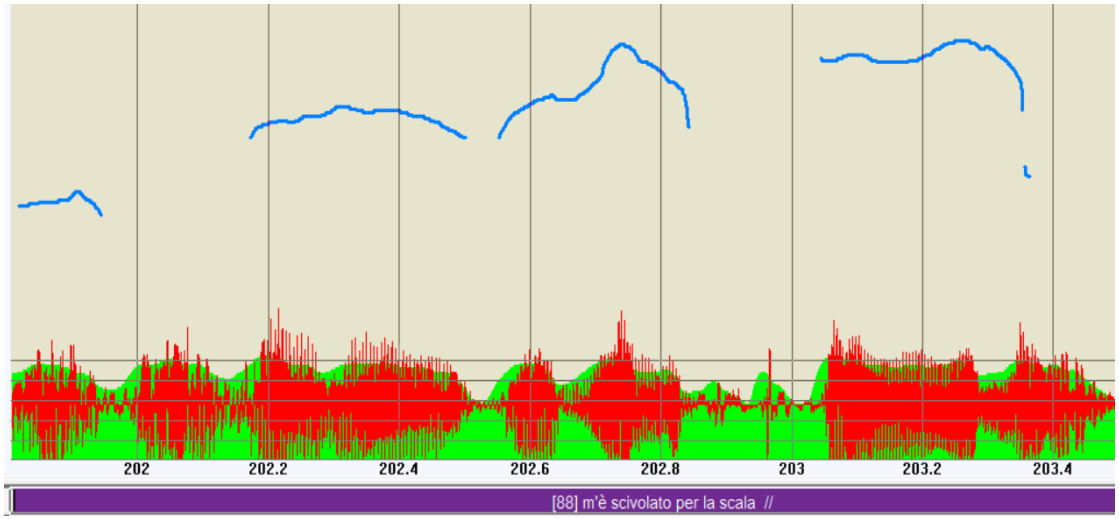
Linguistica

- Studio del linguaggio umano (=lingue naturali) e delle sue manifestazioni (=eventi comunicativi)
- Lingue naturali: 7.000 + (?)



Una lingua, varie manifestazioni

- Lingua scritta → [questa è una frase scritta]
- Lingua parlata →



- Forme intermedie:
  - parlato-scritto → chat, sms (mimesi del parlato nello scritto)
  - scritto-parlato → discorso dei presidenti di camera e senato

Processare la lingua

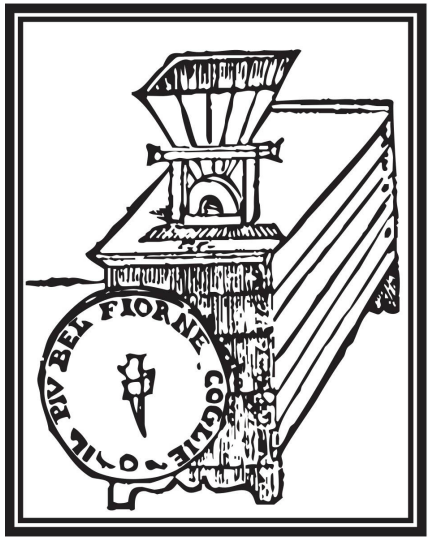
- 1612: 1ª ed. Vocabolario della Crusca
- 5 edizioni → spoglio di 1.684 opere letterarie
- Database di schede redatte a mano per ogni parola

VOCABOLARIO
DEGLI
ACCADEMICI
DELLA
CRVSCA,
CON TRE INDICI DELLE VOCI,
locuzioni, e prouerbi Latini, e Greci, posti per entro l'Opera.
CON PRIVILEGIO DEL SOMMO PONTEFICE,
Del Re Cattolico, della Serenissima Repubblica di Venezia, e degli
altri Principi, e Potentati d'Italia,
E FVOR D'ITALIA, DELLA MAESTA CESAREA,
Del Re Christianissimo, e del Serenis. Arciduca Alberto.

IN VENEZIA MDCXII.
Appresso Giouanni Alberti.

Processare la lingua – ieri e oggi

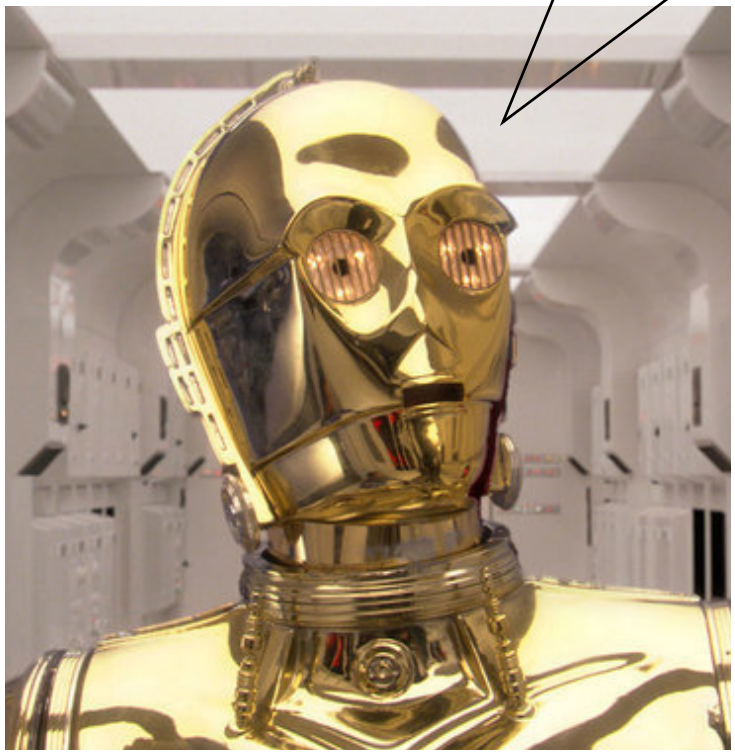
XVII° secolo



XX° secolo

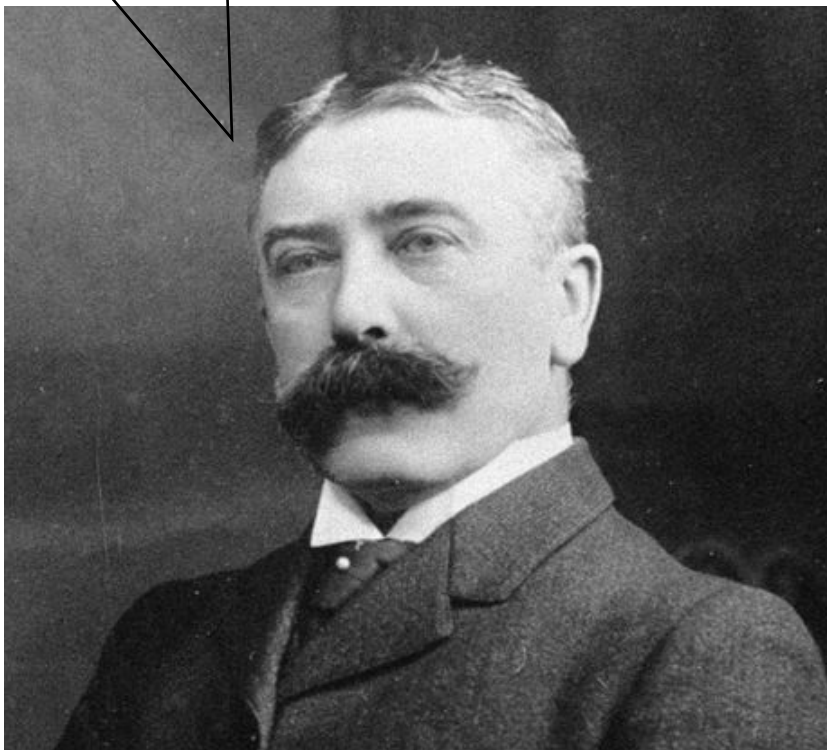
Processare la lingua – domani?

Salve, sono un droide protocollare!



C-3PO – droide protocollare

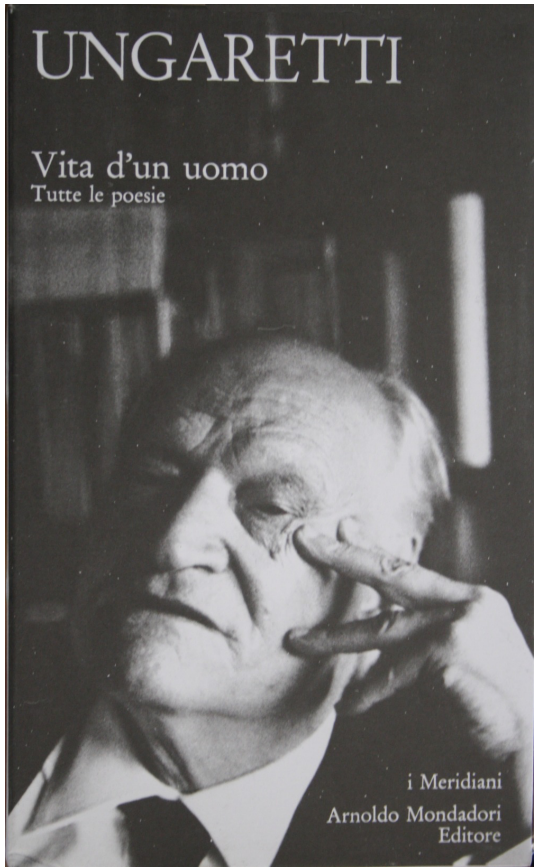
E io sono la Regina d'Inghilterra!



Ferdinand de Saussure - linguista

Che cos’è un corpus?

- Una raccolta di testi
  - Es. il corpus dell’opera ungarettiana



Che cos'è un corpus?

- Una raccolta di testi in formato elettronico
  - Es. il corpus dell'opera ungarettiana su un file .txt

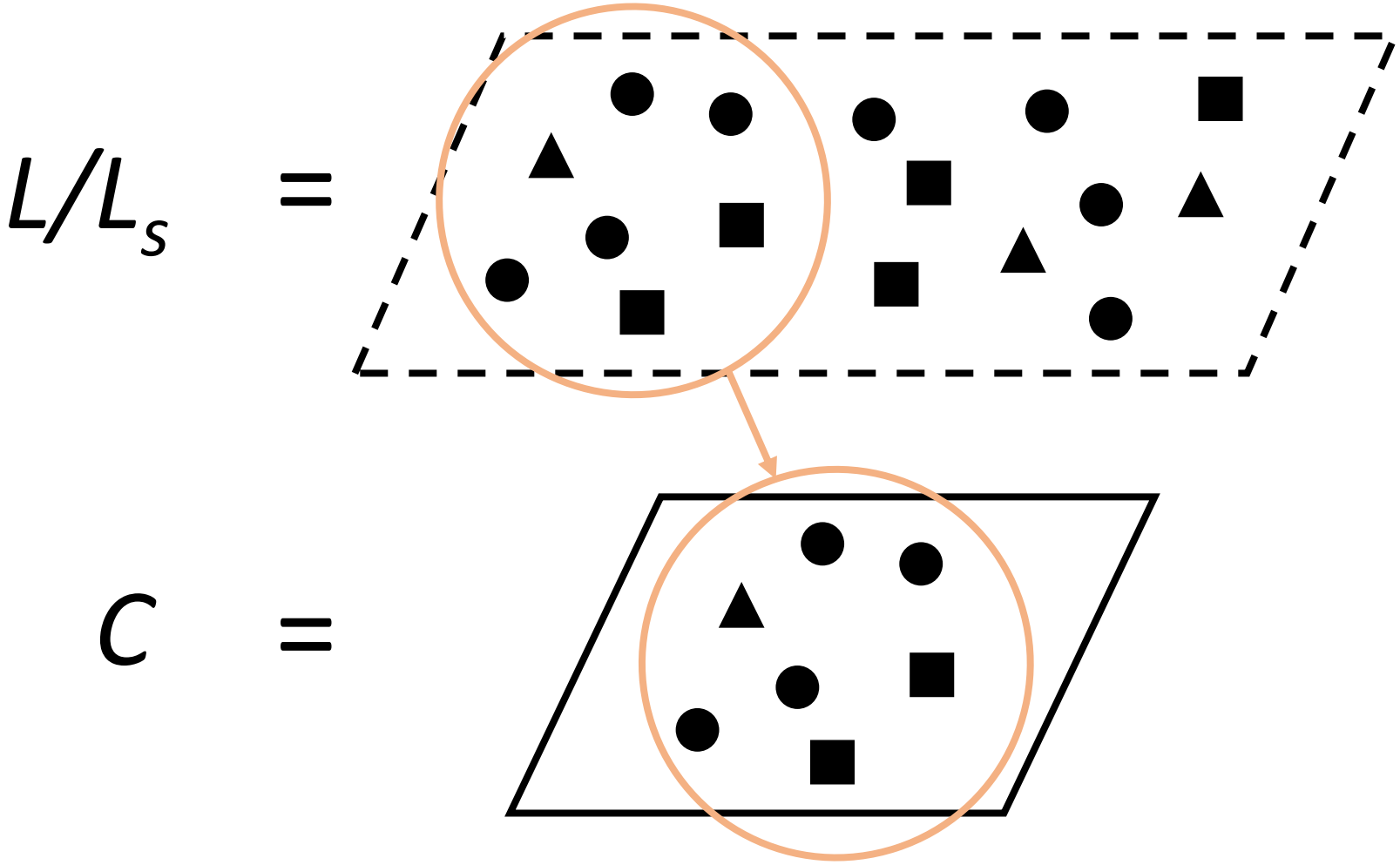

```
1  Vita d'un uomo - Tutte le Poesie, edizioni Mondadori, 1992
2
3  Poesie tratte da L'allegria 1914-1919
4
5  UN'ALTRA NOTTE
6  In quest'oscuro
7  colle mani
8  gelate
9  distinguo
10 il mio viso
11 Mi vedo
12 abbandonato nell'infinito
13
14 NOIA
15 Anche questa notte passerà
16 Questa solitudine in giro
17 titubante ombra dei fili tranviari
18 sull'umido asfalto
19 Guardo le teste dei brumisti
20 nel mezzo sonno
21 tentennare
22
23 AGONIA
24 Morire come le allodole assetate
25 sul miraggio
26 O come la quaglia
27 passato il mare
28 nei primi cespugli
29 perché di volare
30 non ha più voglia
31 Ma non vivere di lamento
32 come un cardellino accecato
33
34 [...]
```

Che cos'è un corpus?

- Una raccolta di testi in formato elettronico, opportunamente tokenizzati ed etichettati, gestibili e interrogabili informaticamente, selezionati allo scopo di essere rappresentativi della lingua presa in esame o di una sua specifica varietà.

- Un corpus è un campione rappresentativo e bilanciato di eventi comunicativi.

Campione rappresentativo e bilanciato di eventi comunicativi.



Caratteristiche dei corpora

- Generalità
  - Corpora di riferimento vs. corpora specialistici
- Cronologia
  - sincronia vs. diacronia
- Lingua
  - Monolingue vs. bi/plurilingue
- Medium
  - Scritto vs. parlato vs. multimediale

Corpora di riferimento vs. corpora specialistici

- Corpora di riferimento:
  - Campione trasversale e «omnicomprensivo» rispetto alle varietà di *L*
  - Es. British National Corpus (BNC), Perugia Corpus (PEC)

- Corpora specialistici:
  - Campione di una particolare varietà o dominio d'uso
    - Linguaggio giornalistico → Repubblica Corpus
    - Apprendenti di L2 → Lessico italiano Parlato da Stranieri (LIPS)
    - patologie linguistiche → Corpus di italiano Parlato Patologico Schizofrenico (CIPPS)

Sincronia vs. diacronia

- Corpora sincronici:
  - Rappresentano la lingua *L* in un dato momento → campione di eventi comunicativi prodotti nello stesso arco temporale
  - Es. Lessico Italiano Televisivo (LIT); Corpus di Riferimento dell'Italiano Scritto (CORIS)

- Corpora diacronici:
  - Rappresentano la lingua *L* in diversi momenti → campioni di eventi comunicativi in più finestre temporali
  - Es. DIA-LIT, DiaCORIS

Monolingue vs. bi/multilingue

- Corpora monolingue
  - Eventi comunicativi di una sola lingua

- Corpora multilingue
  - Paralleli: lo stesso evento comunicativo è rappresentato (in traduzione) in più di una lingua → Europarl
  - Comparabili: eventi comunicativi in più lingue collezionati secondo le stesse specifiche → C-ORAL-ROM

Scritto vs. parlato vs. multimediale

- Scritto: campioni esclusivi della varietà scritta di una lingua
- Parlato: campioni esclusivi della varietà parlata di una lingua
- Multimediale: campioni di eventi comunicativi completi di informazione acustica e visiva

Web Corpora

- Collezionare miliardi di parole, dai testi del web, è possibile con poche righe di codice

- Il corpus design è fondamentale per assicurare un dato quanto più rappresentativo possibile

A cosa serve un corpus?

- Frequenze
- Concordanze
- Collocazioni
- Word Sketches

RIDIRE.it

Sketch

Lemma: aprire

Scegli sketch: postV_N

Calcola sketch

postV_N	135092	
porta	9801	10,34
udienza	6438	9,79
strada	5554	8,62
occhio	3906	8,61
battente	1627	8,57
inchiesta	1947	8,48
finestra	1879	8,43
fuoco	2433	8,33
bocca	1524	8,1
dibattito	1514	7,99
varco	1046	7,92
danza	1165	7,78
sipario	927	7,75
fascicolo	973	7,61
serata	1425	7,61
concerto	1981	7,59
spiraglio	812	7,58
orizzonte	913	7,51
scenario	1004	7,44
prospettiva	1204	7,42

RIDIRE.it

Corpus: Informazione

postN_V				
cucinare	19154	0	13,55	0
subire	83	0	5,81	0
raccontare	108	0	5,5	0
potere	582	582	5,18	5,23
stare	203	254	4,78	5,23
vivere	103	112	5,28	5,76
trovare	146	182	4,74	5,22
provenire	47	40	5,03	5,56
soffrire	42	34	5	5,55
appartenere	33	30	4,64	5,35
indossare	32	29	4,75	5,59
tendere	32	31	4,73	5,63
perdere	54	79	4,38	5,3
chiamare	55	86	4,47	5,5
piare	31	38	4,59	5,75
tentare	39	89	4,62	6,48
minacciare	0	31	0	5,4
picchiare	0	20	0	5,42
estrarre	0	20	0	5,44
preferire	0	45	0	5,5
aprire	0	136	0	5,9
attaccare	0	60	0	5,98
confessare	0	38	0	6,15
uccidere	0	218	0	6,7
sparare	0	91	0	6,9

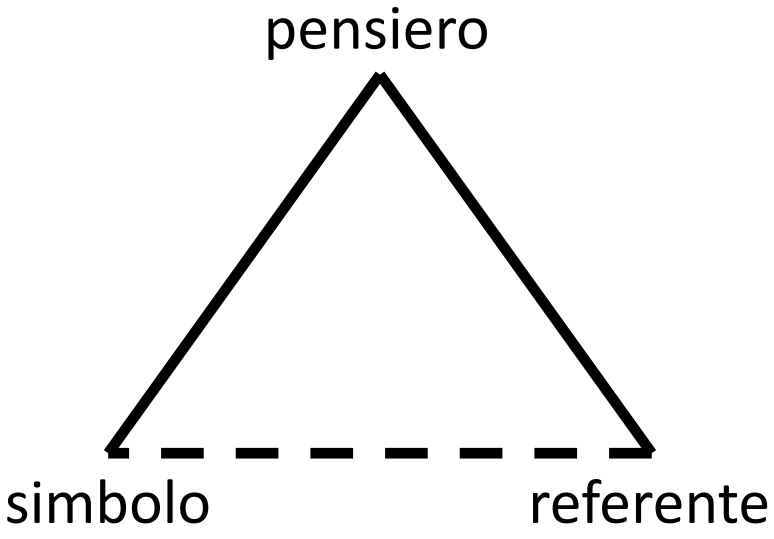
donna
-6.0
-4.0
-2.0
0.0
2.0
4.0
6.0
uomo

Corpus: Cinema

AofN				
bello	619	0	9,17	0
amato	144	0	8,77	0
occidentale	137	0	8,73	0
incinto	65	0	7,8	0
forte	136	0	7,8	0
affascinante	73	0	7,66	0
unico	131	0	7,36	0
giovane	1174	226	9,84	7,4
misterioso	81	79	7,64	7,43
maturo	46	47	7,26	7,06
anziano	95	96	8,16	7,96
moderno	50	54	6,98	6,9
sposato	76	87	8,08	8,02
perfetto	90	127	7,47	7,82
ricco	50	115	6,62	7,67
terzo	0	107	0	7,36
libero	0	81	0	7,37
medio	0	72	0	7,53
spietato	0	77	0	7,7
normale	0	85	0	7,73
invisibile	0	87	0	7,88
nero	0	166	0	7,99
qualunque	0	94	0	8,03
potente	0	120	0	8,17
comune	0	176	0	8,44

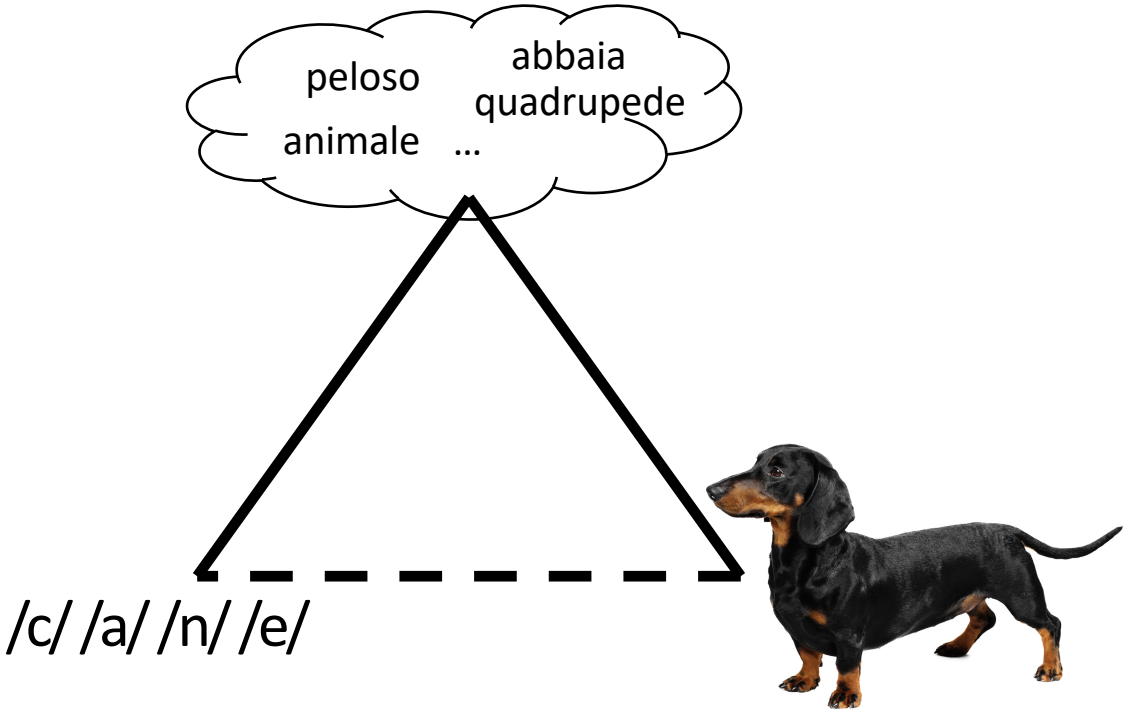
Semantica

- Studio del significato degli eventi comunicativi
- Rapporto tra simbolo, pensiero e referente



Semantica

- Studio del significato degli eventi comunicativi
- Rapporto tra simbolo, pensiero e referente

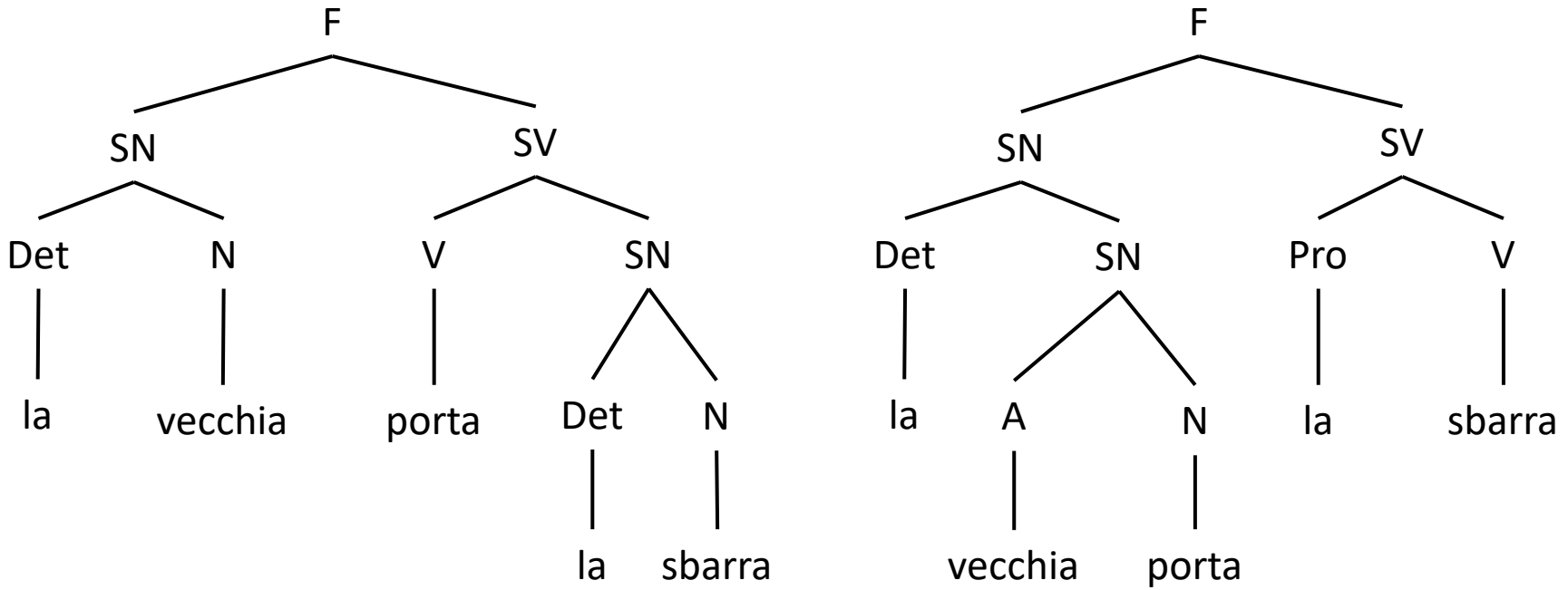


Ambiguità del senso e polisemia

La vecchia porta la sbarra

Ambiguità del senso e polisemia

La vecchia porta la sbarra



Ambiguità del senso e polisemia

- Come si risolve? → Word Sense Disambiguation
- Reti semantiche e ontologie

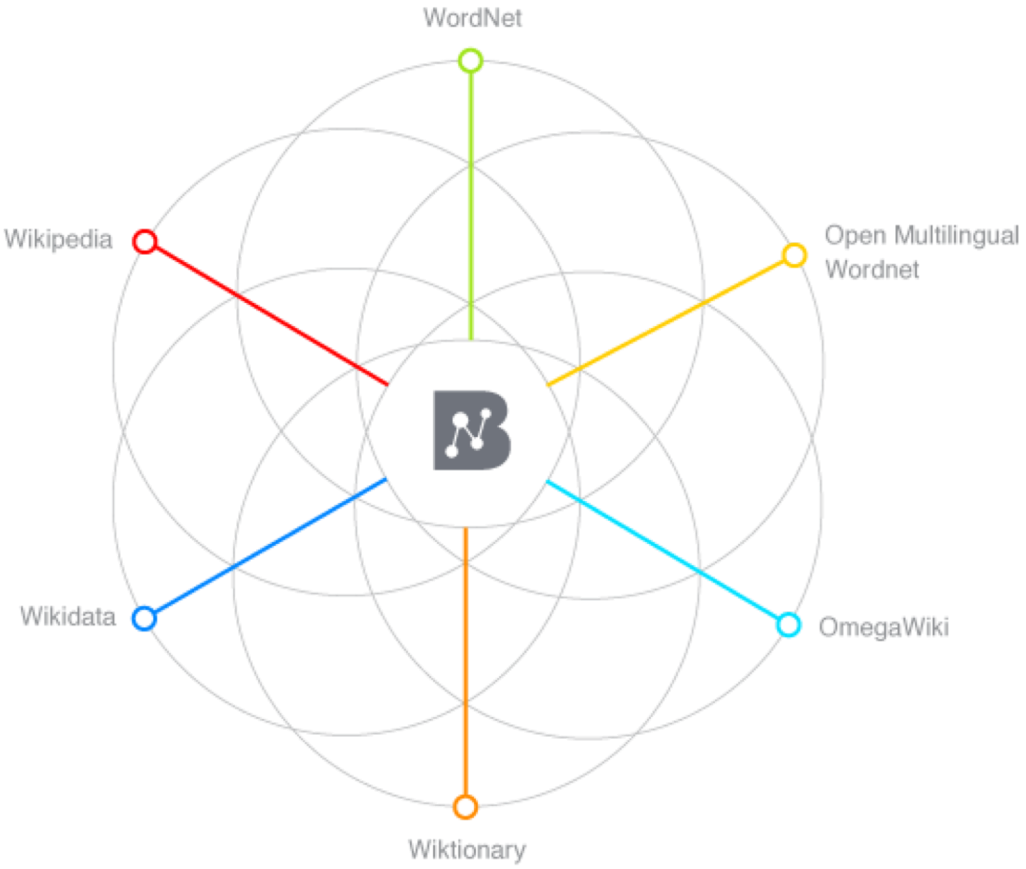
WordNet

- Più importante rete semantica della lingua inglese → 117.000 synsets
- Synset: insieme di lemmi che condividono lo stesso significato
- I synset sono collegati attraverso relazioni ontologiche e semantiche

S: (v) **put**, **set**, **place**, **pose**, **position**, **lay** (**put into a certain place or abstract location**) ***"Put your things here"; "Set the tray down"; "Set the dogs on the scent of the missing children"; "Place emphasis on a certain point"***
S: (v) **put** (cause to be in a certain state; cause to be in a certain relation) *"That song put me in awful good humor"; "put your ideas in writing"*
S: (v) frame, redact, cast, **put**, couch (formulate in a particular style or language) *"I wouldn't put it that way"; "She cast her request in very polite language"*

wordnet.princeton.edu

BabelNet

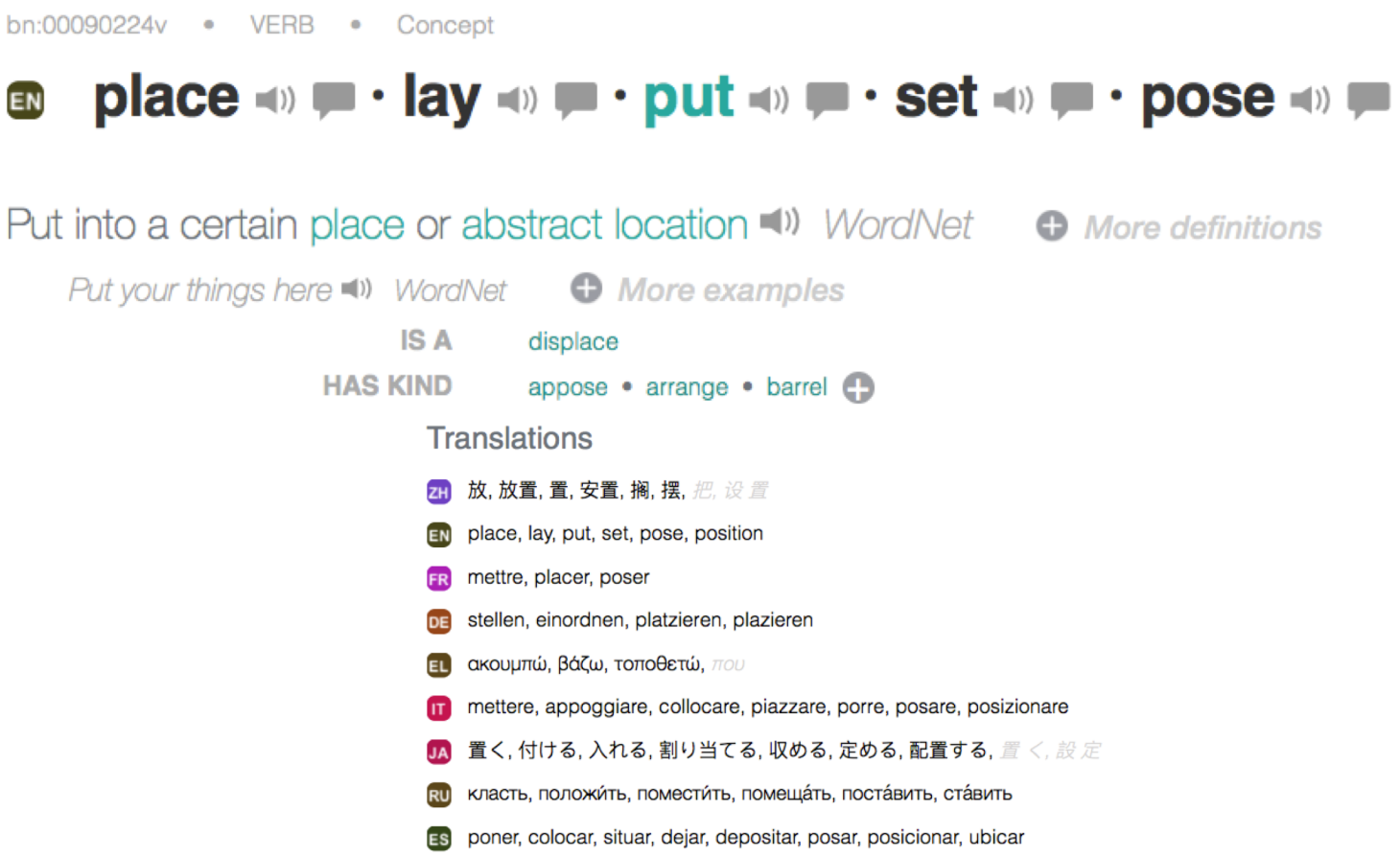


BabelNet 4.0: General statistics

Number of languages:	284
Total number of Babel synsets:	15,788,626
Total number of Babel senses:	832,469,391
Total number of concepts:	6,117,108
Total number of Named Entities:	9,671,518
Total number of lexico-semantic relations:	1,307,706,673
Total number of glosses (textual definitions):	72,542,300
Total number of images:	53,879,884
Total number of Babel synsets with at least one domain:	2,637,414
Total number of Babel synsets with at least one picture:	10,524,280
Total number of sources:	47

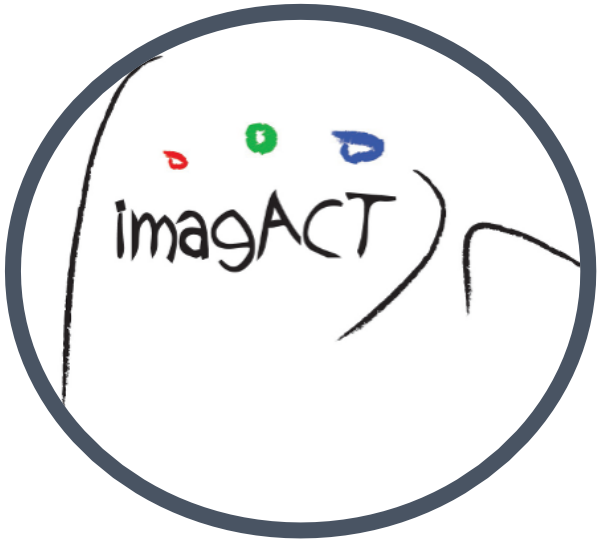
babelnet.org

BabelNet



babelnet.org

IMAGACT



www.imagact.it

UNIVERSITÀ DEGLI STUDI FIRENZE

UNIVERSITÀ DI SIENA 1240

- **Ontologia Multimediale e Multilingue dell'azione**
- **Video-based Translation**
- **Framework di disambiguazione dei verbi d'azione**

Finanziato da:

FAS Fondo Aree Sottoutilizzate 2007-2013

REGIONE TOSCANA

Costruzione della risorsa

1 - Corpora di parlato ITA-EN

2 - Analisi delle occorrenze

spargere	spread
attaccare	stick
indossare	wear
prendere	**take**
portare	bring
mettere	put

ID_scena: be0c573d

LEMMA: prendere

TIPO 1
- BE1: Marco prende la matita dal tavolo

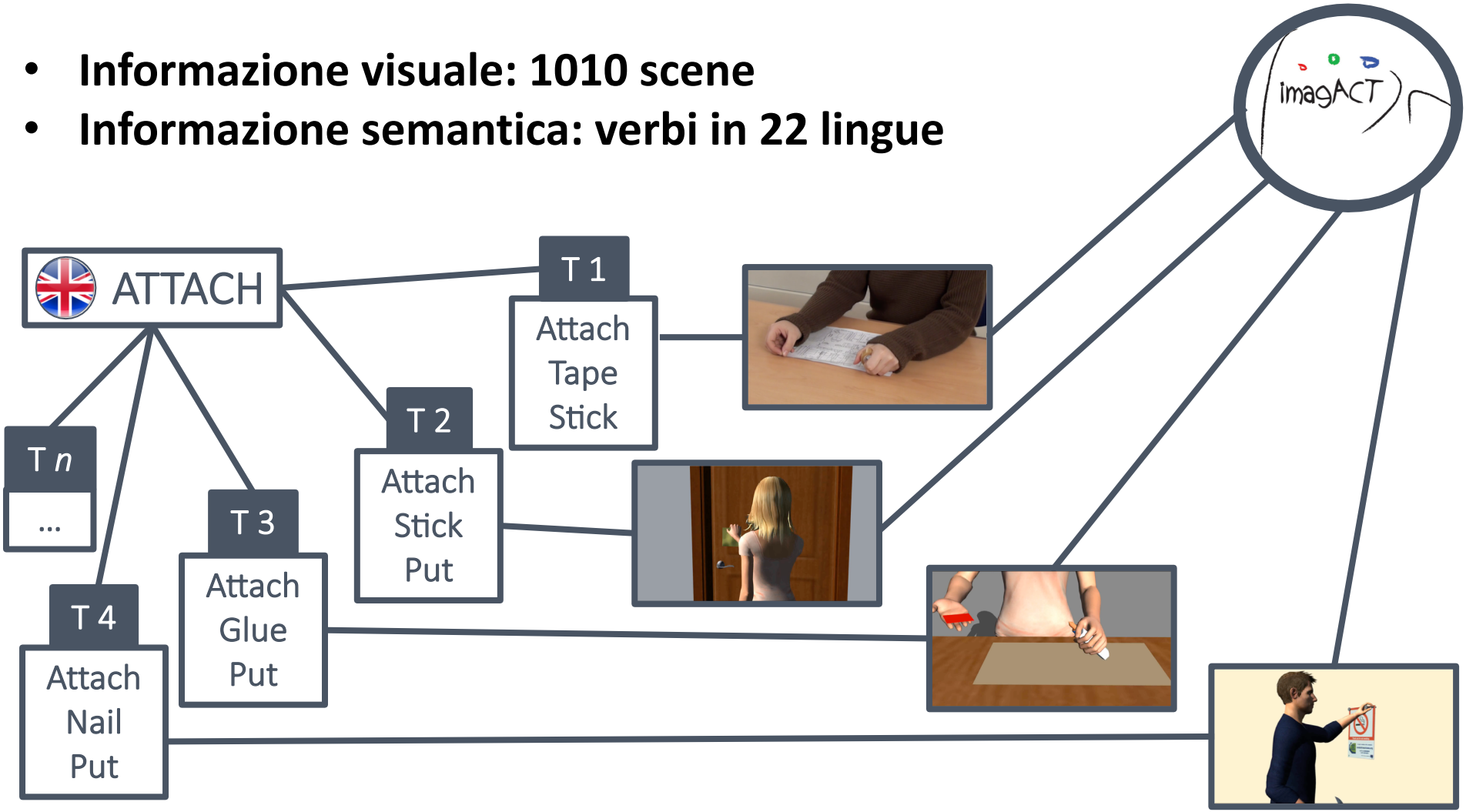
TIPO 2
- BE1: Marco prende il fazzoletto dal pacchetto

TIPO 3
- BE1: Marco prende Cristina per un braccio
- BE2: Marco prende il dito di Cristina
- BE3: Marco prende il volante con le mani

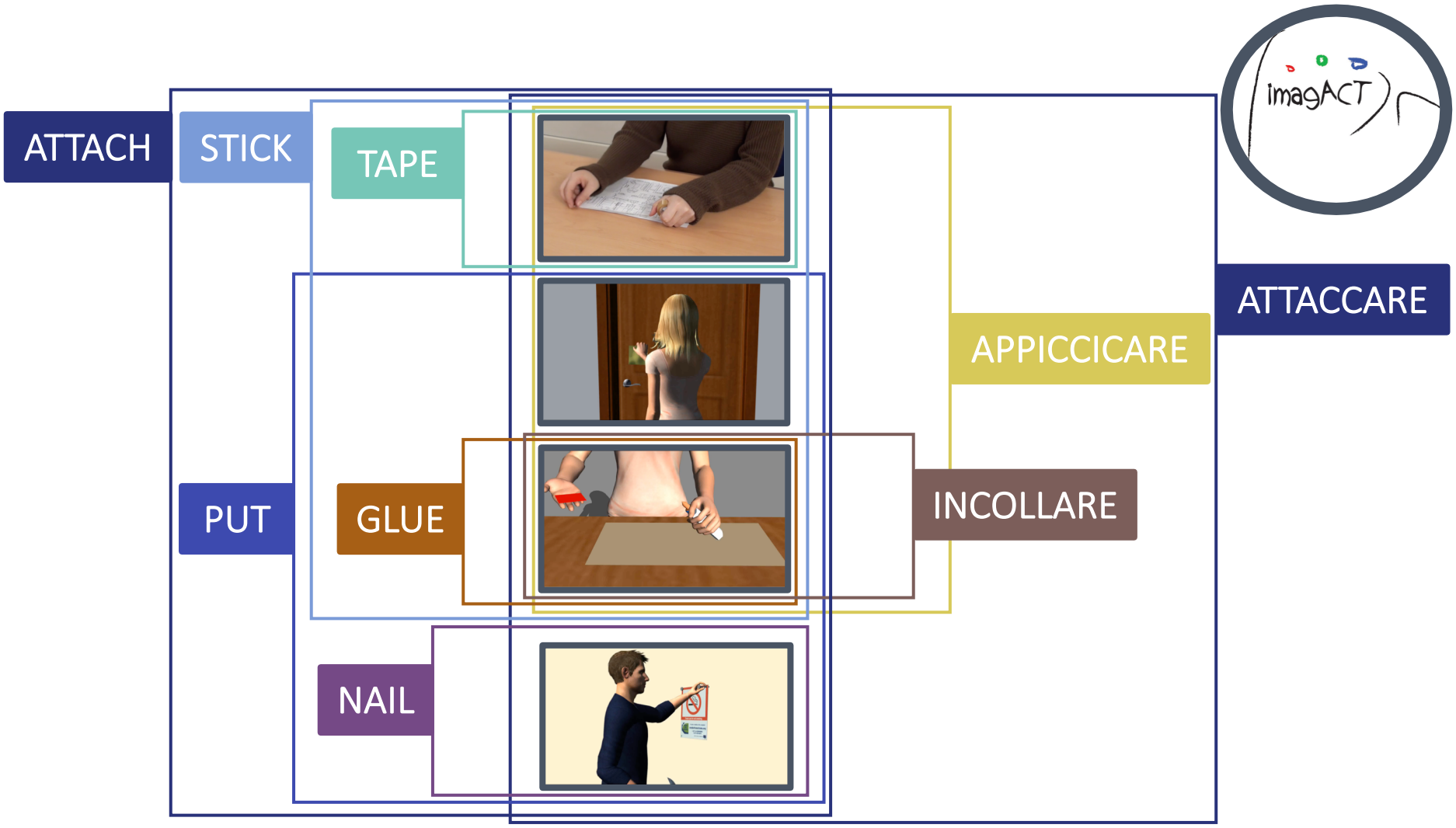
3 – Creazione dei tipi azionali

IMAGACT

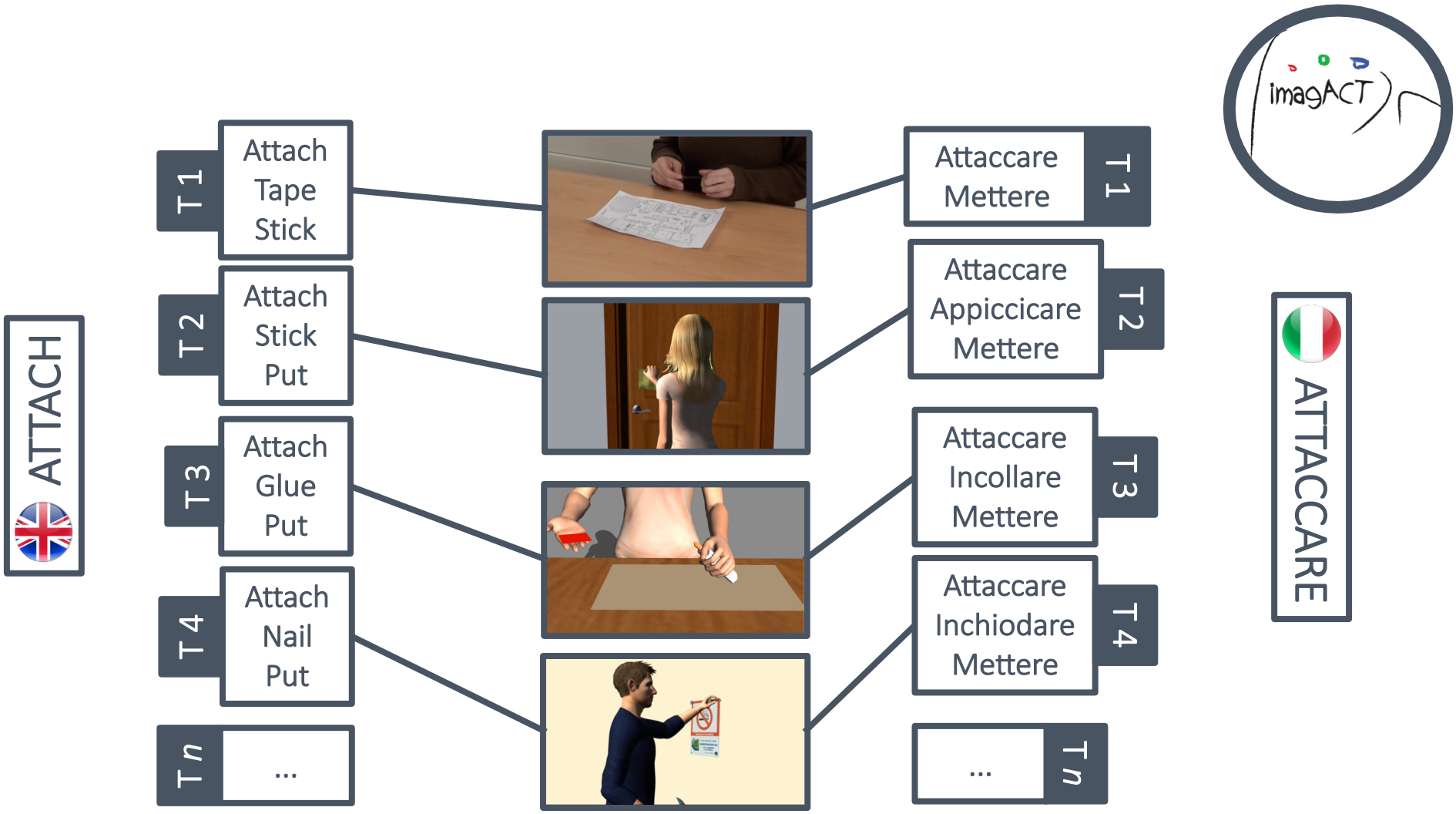
- **Informazione visuale: 1010 scene**
- **Informazione semantica: verbi in 22 lingue**



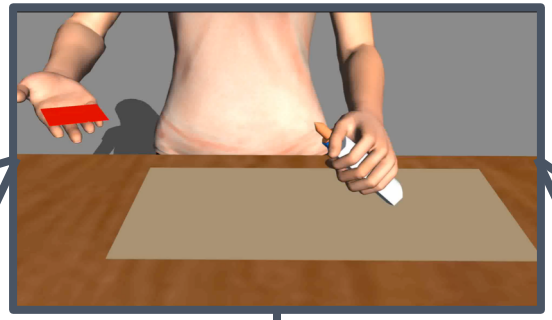
IMAGACT


ATTACH
STICK
TAPE
PUT
GLUE
NAIL
APPICCICARE
INCOLLARE
ATTACCARE
imagACT

IMAGACT


ATTACH
T 1
Attach
Tape
Stick
T 2
Attach
Stick
Put
T 3
Attach
Glue
Put
T 4
Attach
Nail
Put
T n
...
Attaccare
Mettere
T 1
Attaccare
Appiccicare
Mettere
T 2
Attaccare
Incollare
Mettere
T 3
Attaccare
Inchiodare
Mettere
T 4
...
T n
ATTACCARE

IMAGACT



ATTACH
STICK
PUT
GLUE

UNIR
PEGAR
FIJAR

贴 (tiē)
粘 (zhān)

ATTACCARE
APPICCICARE
INCOLLARE

...

Dizionario multilingue

Verb

prendere

Input language

Italian

Output language

English

Show types

IT

prendere

pigliare

EN

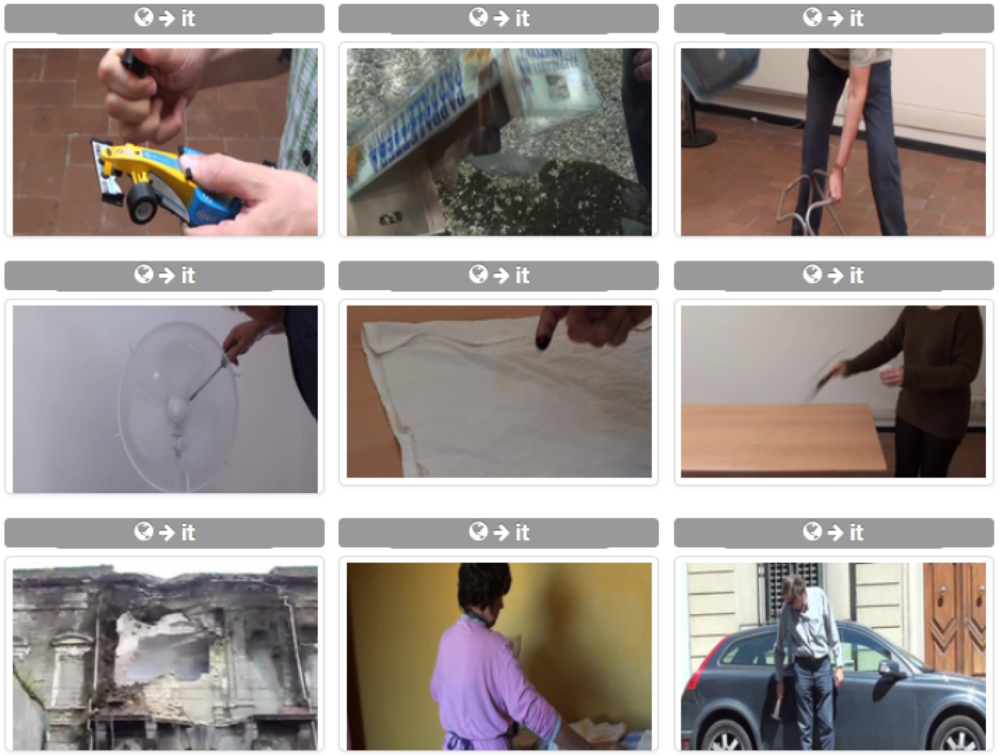
take

Galleria Multimediale

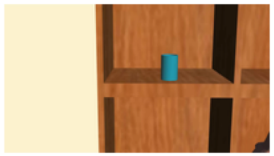
Output language

Italian

Deterioration of an object: 28



Strumento di comparazione

prendere	prendere ⇄ take	take
en →		it ←
		
Fabio prende Marta per un braccio	Fabio prende la tazza dal ripiano	John takes the chair over to the table
en →		it ←
		
Fabio prende il coniglio con la trappola	Fabio prende la pallina cadute a terra	The police take the prisoner with them
en →		it ←
		
Marta prende la farfalla con il retino	Marta prende il libro dal tavolo	John takes the blind man across the street

Estensione di IMAGACT

Estensione di IMAGACT- Statistiche

	Italian		English		Chinese		Spanish		Portuguese		Danish	
Num Verbs (Total)	522		554		414		**736**		**793**		646	
>30 scenes	6	**1,1%**	7	**1,3%**	0	0,0%	8	**1,1%**	9	**1,1%**	1	0,2%
11-30	29	**5,6%**	30	**5,4%**	1	0,2%	43	**5,8%**	46	**5,8%**	21	3,3%
5-10	94	18,0%	98	17,7%	8	1,9%	80	10,9%	90	11,3%	57	8,8%
2-4	195	37,4%	222	40,1%	80	19,3%	243	33,0%	241	30,4%	179	27,7%
1 scene	198	37,9%	197	35,6%	325	**78,5%**	362	49,2%	407	51,3%	388	**60,1%**

	German		Serbian		Polish		Hindi		Arab	
Num Verbs (Total)	**992**		**1124**		**1193**		466		574	
>30 scenes	0	0,0%	2	0,2%	0	0,0%	2	0,4%	4	**0,7%**
11-30	22	2,2%	31	2,8%	19	1,6%	17	3,6%	32	**5,6%**
5-10	61	6,1%	101	9,0%	119	10,0%	37	7,9%	90	15,7%
2-4	294	29,6%	363	32,3%	382	32,0%	128	27,5%	221	38,5%
1 scene	615	**62,0%**	627	**55,8%**	673	**56,4%**	282	**60,5%**	227	39,5%

Distribuzione spaziale delle azioni