UNIVERSITÀ DEGLI STUDI DI BERGAMO
Dipartimento di Lingue, Letterature e Culture Straniere

Linguistica digitale

Il lessico attraverso i corpora

Valentina Piunno

2023/2024

27 Maggio 2024

Contenuti

- Caso I: Rapporto tra frequenza e significato
  - Frequenza d’uso
  - Polisemia lessicale

- Caso II: Analisi della sinonimia

Caso I: Rapporto tra frequenza e significato

- Frequenza d’uso

- Quali sono le parole più frequenti in un corpus?
  - Il caso dell’inglese
  - Il caso dell’italiano

Parole più frequenti nel BNC

- Di che parole si tratta?
- → Parole funzione
  - → Le più frequenti a prescindere dal corpus

Rank	*Word*	*POS tag*	*N*
1	the	at0	6,187,267
2	of	prf	2,941,444
3	and	cjc	2,682,863
4	a	at0	2,126,369
5	in	prp	1,812,609
6	to	to0	1,620,850
7	it	pnp	1,089,186
8	is	vbz	998,389
9	was	vbd	923,948
10	to	prp	917,579
11	I	pnp	884,599
12	for	prp	833,360
13	you	pnp	695,498
14	he	pnp	681,255
15	be	vbi	662,516
16	with	prp	652,027
17	on	prp	647,344
18	that	cjt	628,999
19	by	prp	507,317
20	at	prp	478,162
21	are	vbb	470,943
22	not	xx0	462,486
23	this	dt0	461,945
24	but	cjc	454,096
25	's	pos	442,545
26	they	pnp	433,441
27	his	dps	426,896
28	from	prp	413,532
29	had	vhd	409,012

Parole più frequenti nel BNC

- Che tipo di verbi?

Rank	*Word*	*POS tag*	*N*
1	the	at0	6,187,267
2	of	prf	2,941,444
3	and	cjc	2,682,863
4	a	at0	2,126,369
5	in	prp	1,812,609
6	to	to0	1,620,850
7	it	pnp	1,089,186
8	is	vbz	998,389
9	was	vbd	923,948
10	to	prp	917,579
11	I	pnp	884,599
12	for	prp	833,360
13	you	pnp	695,498
14	he	pnp	681,255
15	be	vbi	662,516
16	with	prp	652,827
17	on	prp	647,344
18	that	cjt	628,999
19	by	prp	507,317
20	at	prp	478,162
21	are	vbb	470,943
22	not	xx0	462,486
23	this	dt0	461,945
24	but	cjc	454,096
25	's	pos	442,545
26	they	pnp	433,441
27	his	dps	426,896
28	from	prp	413,532
29	had	vhd	409,012

Parole più frequenti nel BNC

E per quanto riguarda le parole “piene”?

→ La frequenza delle parole piene invece dipende dal tipo di corpus

Parole più frequenti nel BNC

- Analizziamo esclusivamente la classe dei verbi

Verb lemma	*N per million words*
be	42,277
have	13,655
do	5,594
will	3,357
say	3,344
would	2,904
can	2,672
get	2,210
make	2,165
go	2,078
see	1,920
know	1,882
take	1,797
could	1,683
think	1,520
come	1,512
give	1,284
look	1,151
may	1,135
should	1,112
use	1,071
find	990
want	945
tell	775
must	723
put	700
mean	677
become	675
leave	647
work	646
need	627
feel	624
seem	624
might	614
ask	610
show	598
try	552
call	535
provide	505
keep	505
hold	481
turn	465

Parole più frequenti nel BNC

- Analizziamo esclusivamente la classe dei nomi

Noun	*N per million words*
time	1,833
year	1,639
people	1,256
way	1,108
man	1,003
day	940
thing	776
child	710
Mr	673
government	670
work	653
life	645
woman	631
system	619
case	613
part	612
group	607
number	606
world	600
house	598
area	585
company	579
problem	565
service	549
place	534
hand	532

Noun	*N per mill*
party	529
school	529
country	486
point	484
week	476
member	471
end	458
state	440
word	438
family	428
fact	426
head	402
month	398
side	398
business	394
night	393
eye	392
home	390
question	390
information	387
power	385
change	384
per_cent	384
interest	376

Parole più frequenti nel BNC

- Analizziamo esclusivamente la classe degli aggettivi

Adjective lemma	*N per million words*
other	1,336
good	1,276
new	1,154
old	648
great	635
high	574
small	518
different	484
large	471
local	445
social	422
important	392
long	392
young	379
national	376
British	357
right	354
early	353
possible	342
big	338
little	306
political	306
able	304
late	302
general	301
full	289
far	288
low	286
public	285
available	272
bad	264
main	245
sure	241
clear	239
major	238
economic	236
only	231
likely	228
real	227
black	226
particular	223

Frequenza

Word forms occurring 10 times or more		124,002
Word forms occurring 5–9 times		62,041
Word forms occurring 4 times		28,770
Word forms occurring 3 times		46,459
Word forms occurring twice		98,774
Word forms occurring once	hapax legomena	397,041
Total		757,087

Source: Based on Leech et al. (2001: 9)

Esercizio

- Cosa accade per l'italiano?

- Esercizio con Sketch Engine
    - Wordlist:
        - lemma
        - forme,
        - verbi,
        - nomi,
        - aggettivi

Frequenza e polisemia

Accezioni e polisemia

Accezioni

- una parola può avere un significato articolato in più **accezioni** o in "**famiglie di sensi**"

Polisemia

- = proprietà di una parola di avere più di un significato
- → il segno è un'entità elastica (può modificarsi e riorganizzarsi per raccogliere nuovi sensi o abbandonarne di vecchi).

- *aprire la porta* → *schiudere*
- *aprire la bottiglia* → *stappare*
- *aprire la sala* → rendere accessibile
- *aprire un dibattito* → iniziare
- *aprire una scuola* → istituire

Esercizio

Trova le diverse accezioni della parola *leggero*

Attraverso il corpus, individua i possibili contesti d’uso.

Polisemia

(Casadei 2014)

- ***leggero***

  - *un farmaco leggero* = ‘con pochi effetti collaterali’
  - *droga leggera* = ‘che dà effetti meno gravi’
  - *un vino leggero* = ‘poco alcolico’
  - *un colore leggero* = ‘tenue’
  - *un trucco leggero* = ‘che si nota poco’
  - *un suono leggero* = ‘che si sente poco’

- → Accezioni autonome di *leggero?*
- → O un’unica accezione? (=‘che produce un effetto fisico-percettivo meno intenso’)

Il significato delle parole

Semantica lessicale

- Uno dei problemi nell'analisi del lessico è quello di stabilire con precisione che cosa significhino le parole

- Il significato è soltanto una delle proprietà delle parole
- Le parole non sono gli unici elementi della lingua ad avere un significato
  - strutture sintattiche e categorie morfologiche

La polisemia

Una stessa forma lessicale ha più significati che hanno una relazione più o meno evidente tra loro;

→ Polisemia

- associazione non arbitraria di più sensi a un'entrata lessicale

Molte parole di una lingua sono polisemiche

- questo fatto è conforme alla proprietà delle lingue di essere economiche, ossia di utilizzare lo stesso materiale per più scopi.

Significati estesi

Nomi d'azione nel lessico (Simone 2000, Gaeta 2004, Melloni 2007, Jezek 2008)

parcheggio

- 'il **parcheggio** è vietato' =evento
- 'sta uscendo dal **parcheggio**' =luogo
- 'stanno costruendo un nuovo **parcheggio**' =oggetto fisico

La polisemia

- **Quali sono le parole più polisemiche? Perché?**

- Le parole più polisemiche sono i **verbi** (2,17 significati per ogni verbo)

- Perché?
- forse dovuto al fatto che il loro significato è incompleto poiché riempito dagli argomenti.

La polisemia

(Casadei 2014)

- Maggiore tendenza alla polisemia da parte dei verbi (Fellbaum 1998):
  - 2,17 significati V
  - 1, 23 significati N

- Esiste una correlazione tra la frequenza di una parola e la sua polisemia
  - principio della versatilità economica delle parole (Zipf 1949): le parole più frequenti risultano essere semanticamente più generiche e dunque più disponibili, rispetto alle parole di minor frequenza, a modularsi in un'ampia gamma di significati.

Esercizio

Trova le diverse accezioni delle seguenti parole.

Attraverso il corpus, individua il possibile contesto sintattico e semantico di ciascuna accezione

- Bello
- Chiaro
- Verde

Caso II: La sinonimia

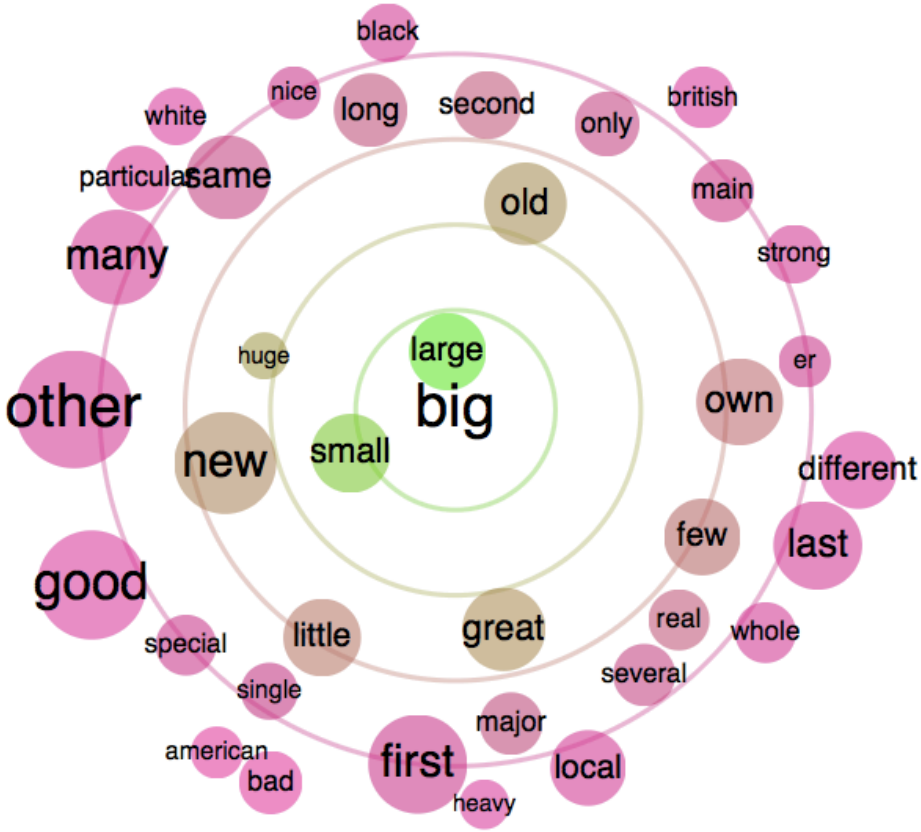
La sinonimia

Sinonimia

- Due parole sinonimiche dovrebbero essere identiche dal punto di vista
  - Delle proprietà semantiche
  - Del valore connotativo
  - Del comportamento grammaticale

- → sinonimia parziale

“Sinonimi” di *big*



NB: Funzione Theasaurus in Sketch Engine → parole che condividono gli stessi contesti sintattici

English Oxford Dictionary

Big

= Of considerable size or extent

Large

= Of considerable or relatively great size, extent, or capacity

Collocati di *big*

Nomi concreti & individui
→ grandezza

1. Antonia fantasised about chauffeur-driven cars, wearing **big** hats
2. Dad was so pleased he brought me a **big** cup of tea
3. I, I agree you don't need that great **big house**.
4. He was a **big man** with a square red face
5. Milan was one of the **big cities** where she felt very much at home.
6. So you have a big **big thing** of chocolate
7. …big heavy knitted cardigan and this great **big fish** on the back

Collocati di *big*

Nomi astratti
→ Grandezza metaforica

1. Was there a **big difference** between your meals and his meals?
2. Family biographies have become **big business**
3. The **big race** this weekend is at Sandown Park
4. The **big companies** surely miss a chance by doing nothing for the club

Collocati di *big*

Nomi concreti & astratti
→ Intensificatore = importante

1. So she had a **big time** as a young women running around the ship
2. The secrets of a successful wedding since he married Vera -- the one **big mistake** in his life
3. # **Big day** can be so very expensive
4. When you were at drama school' you were a **big fish** -- now you're a tiddler’
5. The **big question** was: Who is this girl –
6. Going to live in another country is a pretty **big step** to take

Usi di *big*

big

Nomi **concreti** → grandezza **fisica**

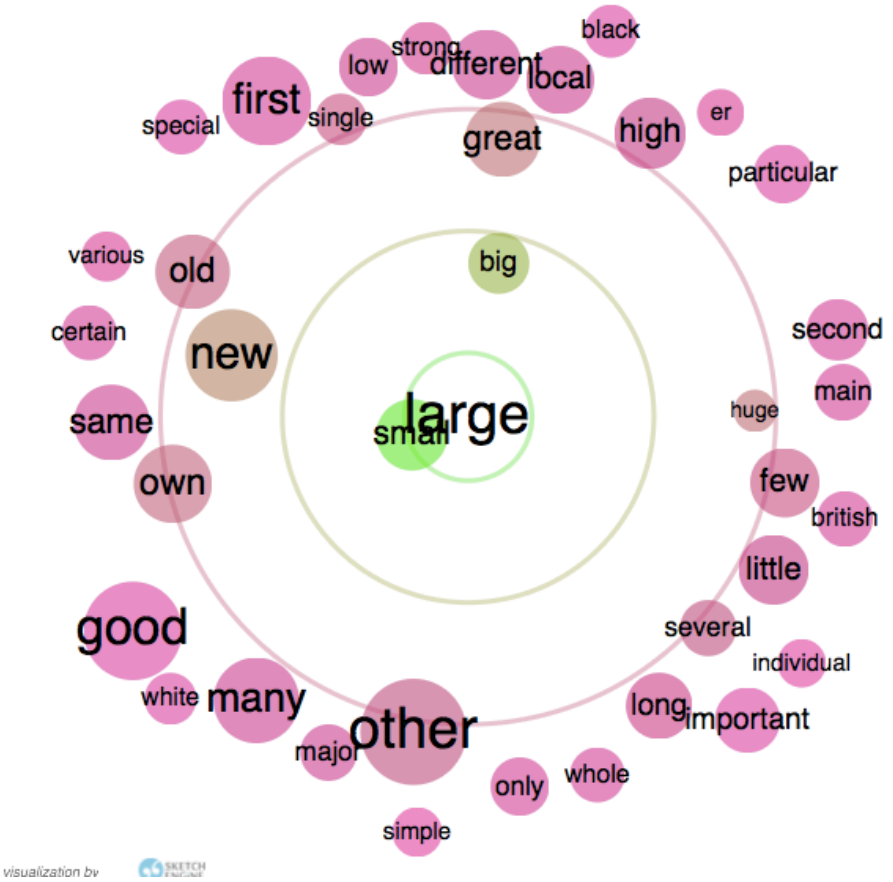
↓

Nomi **astratti** → grandezza **metaforica**

↓

Nomi **astratti/eventi/concreti** → **intensificatore** di qualità: “important”

“Sinonimi” di *large*



NB: Funzione Theasaurus in Sketch Engine → parole che condividono gli stessi contesti sintattici

Collocati di *large*

Nomi concreti & luoghi
→ Grandezza

1. It is a **large house** and has two entrances, each leading into an atrium.
2. The office was a **large room** on the first floor
3. Place all the ingredients together in a **large bowl**
4. …excellent job in bringing it to peoples' attention that **large areas** of tropical forest are being destroyed

Collocati di *large*

Nomi collettivi
→ Numero di individui

1. The figures were derived by estimating separately for small companies i.e. those with up to 200 employeesand **large companies** i.e. 200 plus employees
2. **Large families** often live in just one room
3. Businessmen are in daily contact with **large groups** of people
4. Even in a **large population**, very few individuals may be free of any deleterious mutations

Collocati di *large*

Nomi di quantità
→ Grande quantità

1. There are parish organizations represented here but not a very **large number**
2. We have a **large proportion** of ethnic minority citizens
3. We have a very **large amount** of information and facilities to offer
4. The universities have attracted **large sums**

Usi di *large*

large

Nomi concreti e luoghi → grandezza fisica

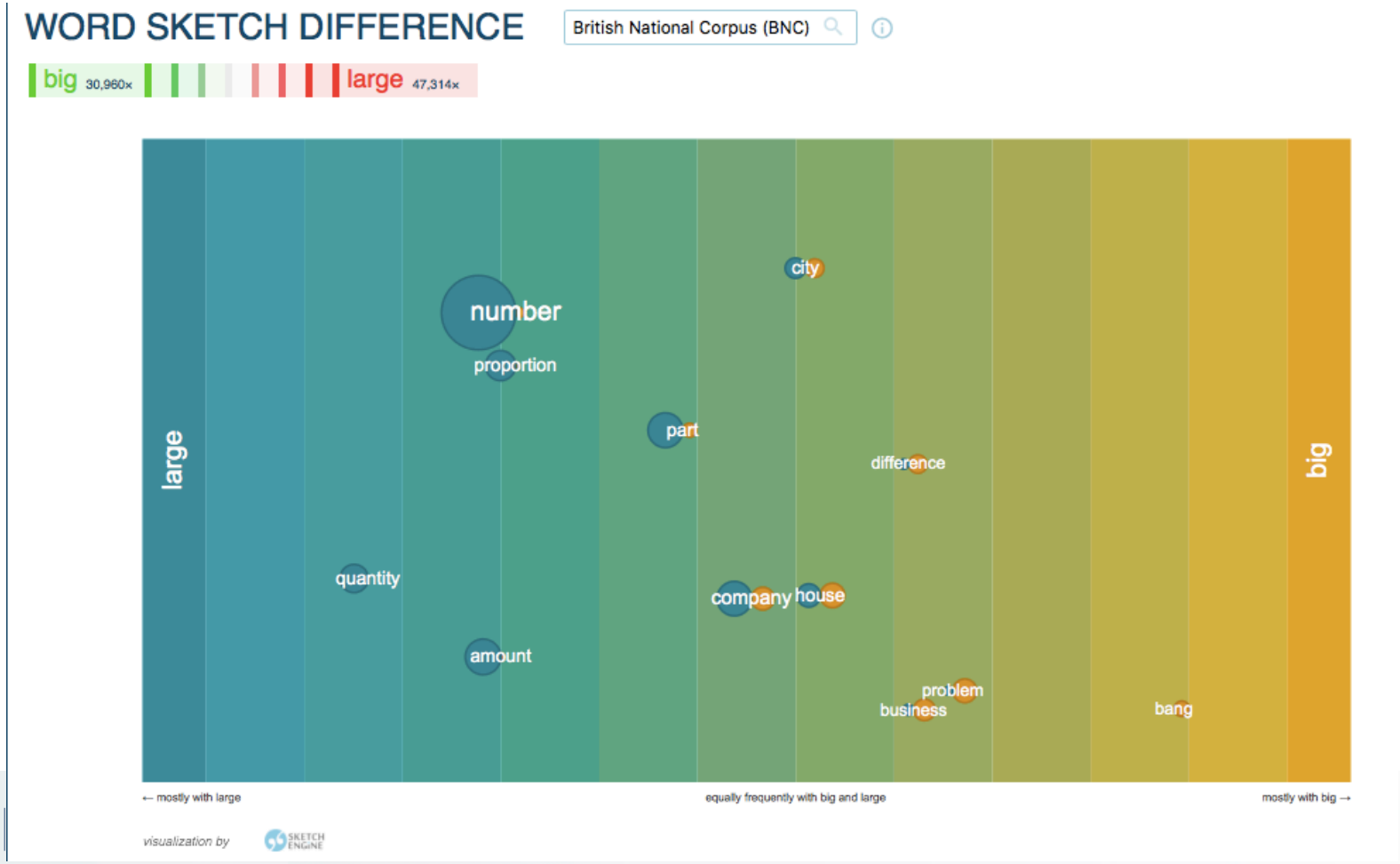
↓

Nomi collettivi → numero di individui

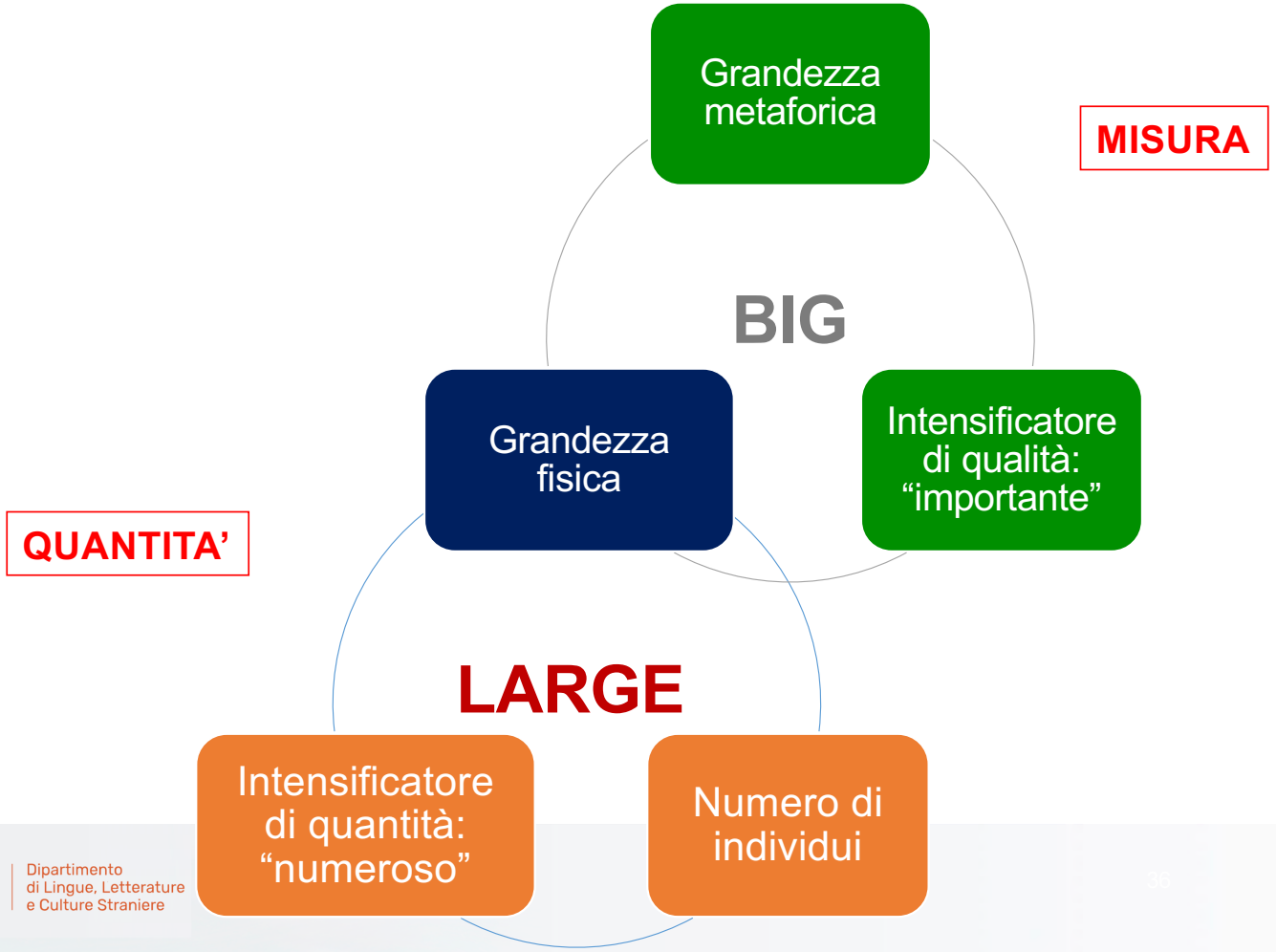
↓

Nomi di quantità → intensificatori di quantità: “numerous, copious”

Collocati di *big* e *large*



Sovrapposizione di significati



Esercizio

Analizza le proprietà semantiche delle seguenti coppie di parole, individuando eventuali differenze combinatorie:

- Profumo – Odore
- Ingresso - Entrata
- Utilizzo – Uso
- Chiaro – Preciso
- Fare – Effettuare

Conclusioni

- Un approccio basato sul corpus porta a risultati diversi per quanto riguarda la **frequenza** e la **tipicità dei pattern** (Bianchi 2012).

- L'inaspettatezza dei risultati derivati dai corpora porta alla conclusione che **l'intuizione non è completamente affidabile** come fonte di informazioni sulla lingua (Tognini-Bonelli 2001: 86).

- I dati del corpus non sono solo esempi illustrativi, ma anche una **risorsa teorica**.
    - È sempre necessaria una **teoria del linguaggio** per sapere cosa cercare in un corpus e per spiegare ed interpretare I dati estratti.

Grazie per l'attenzione!

valentina.piunno@unibg.it

UNIVERSITÀ DEGLI STUDI DI BERGAMO | Dipartimento di Lingue, Letterature e Culture Straniere

visualization by SKETCH ENGINE