New Anglicisms and their currency in Italian corpora: a comparison between itTenTen16 and CORIS

Virginia Pulcini (Università degli Studi di Torino, Italy)
Marek Łukasik (Pomeranian University in Slupsk, Poland)

X International Conference on Corpus Linguistics, Cáceres, 9-11 May 2018

Background

- Corpora for loanword lexicography
- For cross-linguistic investigation (GLAD) comparable “national corpora” should be available
- How can corpora help us to establish frequency?
  - * = less frequent, ** = frequent, *** = highly frequent)

Italian corpora: itTenTen and CORIS

CORIS 2017: 150 million words of written Italian (1980- 2016)

Genres: press, narrative, academic, miscellaneous, ephemera

PRESS - 38 million words (newspapers, periodic, supplement)

FICTION - 25 million words (novels, short stories)

ACADEMIC PROSE - 12 million words (human sciences, natural sciences, physics, experimental sciences)

LEGAL AND ADMINISTRATIVE PROSE - 10 million words

MISCELLANEA -10 million (words books on religion, travel, cookery, hobbies, etc.)

EPHEMERA - 5 million words (letters, leaflets, instructions)

Italian Web 2016 (itTenTen): 4.9 billion word corpus made up of web-based texts (end of May – mid-August)

Corpus CORIS, annotated version (2017, 150Mw)
- Corpus query form -

User Authentication

CORIS access is now free for research purposes
(Please, read the footnote carefully).

Query

(Query Language Help).

"app"

Subcorpus All ▾ **Section** All ▾

- STAMPA
- NARRATIVA
- PROSA ACCADEMICA
- PROSA GIUR.AMMIN.
- MISCELLANEA
- EPHEMERA
- MONITORs
- All

Concordance Options

Show ◉ 30 ○ 100 ○ 300 ○ 1000 **lines.**

Sort positio

Collocations

Get Collocates? ○ NO! ◉ Yes.

Sort using ◉ Log-Likelihood Ratio. ○ Mutual Information. ○ T-score. ○ Raw frequency.

Esegui Cancella

The data

- **410 new Anglicisms** recorded in 3 recent editions of the Italian general dictionary Zingarelli, namely 2014, 2017 and 2018.
- **three time spans**: the first in 2010-2013 (2014 edition, 146 new items) the second in 2014-2016 (2017 edition, 141 items), and the third in 2017 (2018 edition, 123 items)

Research questions

1) Which of the **2 corpora** is more suitable to provide reliable frequency scores?

2) Are **Anglicisms** recorded between 2010 and 2017 current enough and representative of “general, modern, commonly used” type of discourse (see GLAD guidelines for contribution to the Anglicism database)?

3) Do corpus data confirm that the most affected semantic fields are **IT**, **economy** and **sport** (Pulcini 2017)?

4) Do differences emerge among **the 3 time spans**?

The pilot study (wordlist #1)

- New Anglicisms recorded in the 2014 edition of Zingarelli dictionary (compared to 2010)
- Anglicisms recorded in 2011, 2012 and 2013
- Total number: **146**
- **hashtag** 2009, **microblog** 2007, **paywall** 2010
- **bloodhound** 1861, **dumping** 1914, **company** 1926
- 70.5% general meanings vs 37% specialized meanings

Procedure

- Anglicisms were looked up in itTenTen and CORIS
- Items were searched for in both lowercase and uppercase
- Items were searched for in singular and plural forms
- Multi-words were searched for in their solid, separate and hyphenated forms
- Multi-words were also searched for in both lowercase and uppercase
- Figures were summed up and a lemma list was created
- Lemmas feature in the final list in the form attested by the reference dictionary

	A	B	C	D	E	F	G
1	FINAL LEMMA	TOTAL FREQ. (itTenTen)	RELATIVE FREQUENCY (base=1M)		FINAL LEMMA	TOTAL FREQ. (Coris)	RELATIVE FREQUENCY (base=1M)
2	mobile	501898	100,59		mobile	4048	26,99
3	app	242457	48,59		app	785	5,23
4	iPad	77506	15,53		company	778	5,19
5	cloud	71552	14,34		iPad	724	4,83
6	premium	47588	9,54		chino	576	3,84
7	company	47418	9,50		tweet	344	2,29
8	tutorial	44889	9,00		indie	326	2,17
9	tweet	33420	6,70		camp	275	1,83
10	camp	31945	6,40		memorial	238	1,59
11	memorial	25892	5,19		networking	218	1,45
12	update	23576	4,72		runner	198	1,32
13	Photoshop	22817	4,57		premium	171	1,14
14	walking	22318	4,47		spending review	149	0,99
15	networking	21763	4,36		Photoshop	148	0,99
16	template	21559	4,32		framework	146	0,97
17	hashtag	20944	4,20		template	141	0,94
18	outfit	18763	3,76		gender	122	0,81
19	framework	18629	3,73		setting	113	0,75
20	spending review	14904	2,99		duty free	107	0,71
21	best practice	14857	2,98		cloud	100	0,67
22	green economy	14009	2,81		follower	100	0,67
23	gender	13863	2,78		tutorial	97	0,65
24	follower	12362	2,48		dumping	94	0,63
25	setting	12207	2,45		hashtag	94	0,63
26	screenshot	11622	2,33		megastore	79	0,53
27	widget	11562	2,32		walking	74	0,49
28	chino	11241	2,25		finger	70	0,47
29	problem solving	11017	2,21		update	68	0,45
30	runner	9817	1,97		direct marketing	66	0,44
31	primer	9667	1,94		meme	66	0,44
32	finger	9585	1,92		glam	63	0,42

Foglio di lavoro
di Microsoft Excel

Comparison among the top 50 Anglicisms

- Items featuring in itTenTen and not in CORIS:

outfit, widget (IT), primer, lifestyle, regular season, Dropbox (IT), torrent, snippet (IT), slideshow (IT), anti-age, veg, multitouch (IT)

- The items featuring in CORIS and not in the itTenTen:

duty free, dumping, megastore, direct marketing, private banking, melting pot, peer review, premiership, downsizing, celebrity, backdoor (IT), Neet.

Relative frequency

Anglicisms are low-frequency lexical items

Frequency is calculated out of 1M words

app 5.25 (CORIS) vs 48.59 (itTenTen)

outfit and **snippet** (very high score in itTenTen, very low or absent in CORIS)

premiership and **downsizing** (very high score in CORIS, very low in itTenTen)

Field labels

itTenTen:

no label 28 (56%)

IT= 13

Internet=4

IT and Internet= 34%

econ.=2

sport=1

cinema/theatre=1

psychology=1

CORIS:

no label= 32 (64%)

IT=8

Internet=3

IT and Internet= 22%

economy=3

cinema/theatre=1

econ./autom.=1

psychology=1

sport=1

Zero occurrences in CORIS

snippet	1.26
adware	0.42
counsellor	0.35
Segway	0.22
mockumentary	0.14
paintball	0.11
Blu-ray Disc	0.08
blurb	0.07
ski cross	0.06

trashware	0.05
fit box	0.04
overruling	0.02
retrorunning	0.02
freegan	0.01
overdesign	0.01
websurfing	0.01
bling-bling	0.00
dedendum	0.00

Discussion and conclusions

1) Which of the **2 corpora** is more suitable to provide reliable frequency scores?
   - itTenTen (but a large, balanced corpus would be better)
   - Corpus data must be filtered by speakers' perceptions and experience

2) Are **Anglicisms** recorded between 2010 and 2017 current enough and representative of "general, modern, commonly used"?
   - No

3) "Do corpus data confirm that the most affected semantic fields are **IT**, **economy** and **sport**?
   - IT and Internet are the top donor fields in the new millennium, followed by economy and economic-related fields (marketing, business). Sport is on the decline.

Thank you.

virginia.pulcini@unito.it

marek.lukasik@apsl.edu.pl