“Ti blocco perché sei un trollazzo”

Lexical innovation in contemporary Italian in a large Twitter corpus

Paolo Brasolin [1] Greta H. Franzini [1] Stefania Spina [1,2]

CLiC-it 2023: Ninth Italian Conference on Computational Linguistics
Venezia, 1st December 2023

[1]Eurac Research (Institute for Applied Linguistics), Bolzano, Italy

[2]University for Foreigners of Perugia, Perugia, Italy

Background

Lexical innovation

Lexical innovation is one of the driving mechanisms of language change.

Through the creation of new words, languages evolve and adapt to new contexts.

Different perspectives

- Linguistic processes leading to the **creation of new words**:
  - **acquisition** from other languages;
  - **formation** from pre-existing lexical elements;
  - **change** of grammatical category;
  - **shift** in meaning.
- Sources to **trace** the process of **lexical innovation**.
- Methods to **automatically identify neologisms** from large corpora.

Social media as a source of lexical innovation

Social media provides:

- an opportunity to analyse new words surfacing in **everyday conversation**;
- **vast amounts of data** produced by a large, heterogeneous sample of speakers;
- **geotagging data** to investigate geographical patterns of lexical innovation.

Research questions

Research questions

1. Are Twitter conversations a reliable source to trace lexical innovation?
2. What are the linguistic processes leading to the creation of emerging words on Twitter?

Method

Data acquisition

We sampled timestamped and geotagged tweets from the 2022 Italian timeline.

To define that, we used Twitter's advanced search query language:

Condition	Explanation
`lang:it`	written in Italian
`near:italy`	geotagged near Italy
`since:2022-01-01`	on or after 2022/01/01
`until:2023-01-01`	before 2023/01/01

The corpus includes 5.32M tweets from **153k unique users**.

The volume of content amounts to 564M characters (or 71.5M tokens).

Data preparation

Geographic metadata

$99.43\,\%$ of tweets bear a *place*[1], $0.04\,\%$ only a lat./long. pair, and $0.53\,\%$ neither. We kept places and found $34.8\,\text{k}$ unique ones; **$47.0\,\%$ are in Italy and cover $91.77\,\%$ of tweets**.

Entity metadata

Tweets include ranges locating entities in the text (e.g. urls, user mentions, and hashtags). We embedded them in the text as delimiter characters to support the tokenisation.

Textual data

We patched the spaCy v3.6.1 Italian tokeniser to improve handling of whitespace, punctuation and embedded entities. We then extracted **$71.5\,\text{M}$ tokens of $926\,\text{k}$ types**.

[1] A place is a named administrative division with a country code and a bounding box (we computed its centroid).

Candidate selection

Given a form, let U be the user count, O the occurrence count, A the first day of occurrence, and Z the last. Let ρ denote the Spearman's rank correlation coefficient[2].

Following Grieve et al. (2016, 2018), we selected $4\,090$ forms with $\rho_O > 0.2$. We then extended the condition to $\max(|\rho_O|, |\rho_U|) > 0.2$, defining a **subset $\mathcal{A}$ of 6 737 candidates**.

We also defined a second **subset $\mathcal{B}$ of 21 132 candidates** using a **novel approach** with simpler criteria: $U > 9$, $O > 9$, $\mathrm{A} > 7$, $\mathrm{Z} > 351$ and $\mathrm{Z} - \mathrm{A} > 28$.

$\mathcal{A} \cup \mathcal{B}$ comprises $26\,890$ forms ($2.90\,\%$ of the total); we discarded **15 366 candidates already attested** in the lexicon of Spina (2014) and left the **3 391 hashtags** to future work.

The remaining **8 133 candidates** were **manually annotated** by **two authors** using AntConc's KWIC feature[3] on the tweets' plain text as an aid.

[2] ρ quantifies how well two variables (e.g., occ. count and day of the year) are related by a *monotonic* function.
[3] *KeyWord in Context*. AntConc version 4.2.0.


!$#ȝ

Candidate annotation

Our annotation disregarded:

- **attestations** (Garzanti and/or Treccani online);
- **typos**, including those caused by key proximity: *boungiorno*, *cszzo*;
- already popular neologisms: *bimbominchia*;
- foreign words used in the media but absent from dictionaries: *foliage*, *sponsorship*;
- **nicknames** and terms of endearment: *pupone* for Francesco Totti;
- vowel elongation for **emphasis**: *amooooo*;
- infrequently used foreign words: *veggie*, *waffle*;
- infrequently used foreign acronyms: *PTSD*;
- gender-inclusive graphic variants: *cittadina*;
- **regionalisms** and regional variants: *annassero*, *ciolla*, *giargiana* (Slengo).

We classified the remaining **346 candidates** with an (adjusted) ONLI[4] scheme.

[4] *Osservatorio Neologico della Lingua Italiana.*

Results

Yield comparison

Let $\mathcal{A}_O^+$ be the subset of $\mathcal{A}$ with $\rho_O > 0.2$, representing the exact candidate selection strategy of Grieve et al. (2016, 2018). We compare it with $\mathcal{B}$ by yield of innovative forms[5]:

	$\mathcal{A}_O^+$	$\mathcal{A}_O^+ \cap \mathcal{B}$	$\mathcal{B}$
Innovative forms	70	**14**	281
Adjusted yield	**5.19** %	4.11 %	4.41 %
Projected yield	3.79 %	3.13 %	**4.20** %

- **Yields are comparable**. Adjusted y. favours $\mathcal{A}_O^+$ and while projected yield favours $\mathcal{B}$.
- $\mathcal{A}_O^+ \cap \mathcal{B}$ is smaller than either set, suggesting $\mathcal{B}$ **isolates different patterns** than $\mathcal{A}$.
- The criteria defining $\mathcal{B}$ are **intuitively meaningful and far less computationally expensive**[6], making it more viable for larger datasets or weaker machines.

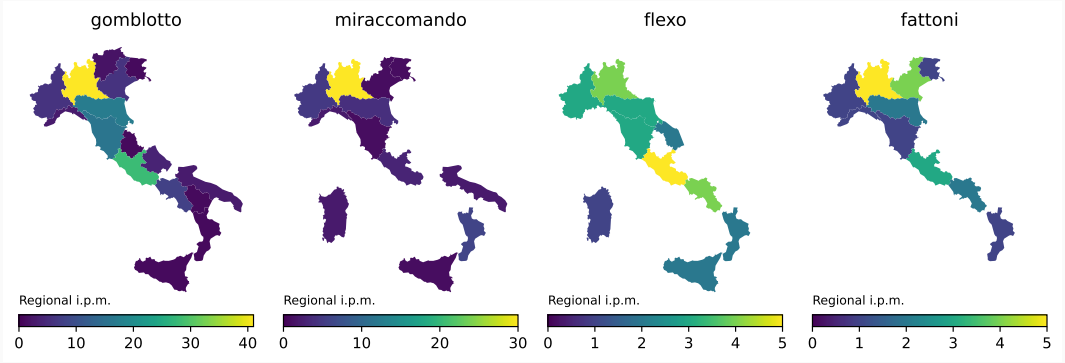
[5] Adjusted yield excludes hashtags; projected yield includes them assuming uniform yield.
[6] More details are in the full paper; we estimate our approach to be upwards of 50 times faster.

Emerging forms

ONLI category	Forms (346)	Examples
orthographic variation	**109**	*minkiate*, *scienzah*
univerbation	**48**	*massì*, *stemmerde*, *miraccomando*
suffixation	**45**	*cinesata*, *pisellata*, *adorissimo*
loanword	**40**	*reminder*, *scammer*
portmanteau	**33**	*lettamaio*, *assurdistan*
loanword adaptation	24	*flexo*, *droppare*, *trollazzo*
alteration	17	*fattoni*
prefixation	8	*appecoronato*, *iposcolarizzati*
acronym	6	*lmv* (li mortacci vostri), *vfc* (vaffanculo)
transcategorisation	6	*cuora*
compounding	3	*contapalle*
deonymic derivation	3	*cippalippa*
redefinition	2	*giornalaia*
acronymic derivation	1	*effeci*
tmesis	1	*facenza*

Geographical distribution

Here are sample choropleth maps showing the number of instances per million tokens at a regional level for the forms *gomblotto* (139 total instances), *miraccomando* (58), *flexo* (29) and *fattoni* (21).



Conclusion

Findings

1. Are Twitter conversations a reliable source to trace lexical innovation?
   - Many forms are tied to the online sphere, not expected to be used in other contexts
   - Nevertheless, their emergence evidences the linguistic mechanisms underlying lexical innovation in Italian
   - Geographic patterns broadly align with regional variation patterns of classical survey data
2. What are the linguistic processes leading to the creation of emerging words on Twitter?
   - Orthographic variation, univerbation, suffixation, loanword and portmanteau are the dominating categories and account for 80 % of the innovative forms
   - Innovation seems to be driven by creativity, amusement and attention-seeking behaviour, rather than a need for new words to indicate new objects, events or situations

Future work

On the corpus:

- Additional dictionary look-ups
- Reproduce analysis on hashtags
- Publish the corpus (compliantly)
- Extend study to additional timelines
- Extend study to other social platforms

On the methodology:

- Search for yield *sweet spots* over the mapped parameter space and produce *rules of thumb* for both methods
- Refine our method introducing a new *convexity* parameter

Repository

github.com/breviloquia-italica

Thank you *very* much!
Any questions?

Paolo Brasolin
paolo.brasolin@gmail.com

Greta H. Franzini
greta.franzini@eurac.edu

Stefania Spina
stefania.spina@unistrapg.it

References I

Grieve, J., Nini, A., & Guo, D. (2016). **Analyzing lexical emergence in Modern American English online.** *English Language & Linguistics*, *21*(1), 99–127. https://doi.org/10.1017/S1360674316000113

Grieve, J., Nini, A., & Guo, D. (2018). **Mapping Lexical Innovation on American Social Media.** *Journal of English Linguistics*, *46*(4), 293–319. https://doi.org/10.1177/0075424218793191

Spina, S. (2014). **Il Perugia Corpus: Una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione.** In R. Basili, A. Lenci, & B. Magnini (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014* (pp. 354–359, Vol. 1). Pisa University Press.

Backup slides

Limitations

- The one-year timeframe is sufficient for quickly emerging forms, but slower phenomena are excluded.
- The language filter is opaque:
  - the implementation of `lang:it` is proprietary;
  - occasional non-Italian tweets were observed (French and Spanish);
  - Italian tweets might have been excluded;
  - we expect the effect to be negligible, but no assessment is possible.