“Ti blocco perché sei un trollazzo”. Lexical Innovation in Contemporary Italian in a Large Twitter Corpus

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

[1]Eurac Research (Institute for Applied Linguistics), Viale Druso 1, 39100 Bolzano BZ, Italy
[2]University for Foreigners of Perugia, Piazza Fortebraccio 4, 06123 Perugia PG, Italy

Abstract

This study investigates emerging vocabulary in contemporary Italian in a corpus of 5.32 M timestamped and geotagged tweets extracted from the Italian timeline throughout 2022. We automatically identify and manually distill 8 133 candidate neologisms down to 346 unattested word forms, shedding light on their spatio-temporal circulation patterns.

Keywords

twitter, social media, corpora, italian, lexical innovation, language change

1. Introduction

Lexical innovation is one of the driving mechanisms of language change [1, 2]: through the creation of new words[1] and their integration into existing lexical systems [3], languages evolve and adapt to new social and technological contexts, which are constantly and rapidly changing. The process of creating new words can be approached from different standpoints. Firstly, the choice of sources necessary to trace the process of lexical innovation has great methodological relevance. One of the main traditional sources have been newspaper texts, which have the double benefit of being easily available and quantitatively relevant [4]. Secondly, lexical innovation follows different steps and usually develops from the initial emergence of new words in specific contexts to their proliferation to wider contexts and domains. This process may end with the institutionalisation of new word forms [5, 6] through their inclusion in dictionaries and consolidation in standard use. Thirdly, the linguistic processes leading to the creation of new words can be different and can include phenomena of derivation, composition, transcategorisation, creation of portmanteau forms, semantic shifts, and borrowing from other languages.

The aim of this study is twofold. On the one hand, we present an analysis of emerging vocabulary in contemporary Italian stemming from Twitter interactions using the 2022 Italian timeline as a source; social media represents an opportunity to analyse new word forms surfacing in everyday conversation, and provide vast amounts of data produced in real time by a large, heterogeneous and representative sample of speakers. Furthermore, the availability of geotagged texts enables the investigation of possible patterns of lexical innovation related to specific geographical areas [7]. This possibility is particularly promising in languages, like Italian, characterised by deep and articulated geographical variation. On the other hand, we propose a novel methodology to process and filter word forms acquired from a sizeable Twitter corpus, with the aim of detecting those that represent the best candidates to become new words.

The result of the study is a list of 346 word forms, classified into 15 categories based on the linguistic process of lexical creation and yet unattested in two of the most up-to-date Italian lexicographic resources.

2. Related Work

Studies on lexical innovation in Italian have a long tradition [8], and have produced extensive lexicographic works dedicated to neologisms (e.g., [9], to mention one of the most recent), as well as a vast body of research (e.g., [10], [11] and [12]). One of the most widely discussed topics is the classification of the linguistic processes leading to the creation and spread of new words.

Traditionally, it is acknowledged that the means by which languages enrich their vocabulary are essentially four: the acquisition of new elements from other languages, the formation of new words from pre-existing lexical elements, the change of grammatical category and the shift in the meaning of words already in use [13]. In the last few decades, the *Osservatorio neologico della lin-*

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

[*]Corresponding author.

paolo.brasolin@eurac.edu (P. Brasolin); greta.franzini@eurac.edu (G. H. Franzini); stefania.spina@unistrapg.it (S. Spina)
https://paolobrasolin.github.io/ (P. Brasolin)
0000-0003-2471-7797 (P. Brasolin); 0000-0003-1159-5575 (G. H. Franzini); 0000-0002-9957-3903 (S. Spina)

[1] In this paper, “word” and “form” are used interchangeably.

gua italiana[2] (ONLI) [4] has been tracking new words emerging in Italian newspapers, producing a database which, to date, includes 2 986 forms with definition, date of attestation and first retrieved occurrence in the press.

More recently, several studies have highlighted the benefits of using social media to track new word forms cropping up in informal contexts, such as everyday conversation, as opposed to newspaper texts, which are more formal and draw from different registers [14, 15, 16]. Additionally, as a populous repository of conversations held in real time by a large number of speakers, social media can capture lexical creativity originating in communities of people rather than inventive journalism [17]. This use of social media has produced a number of studies [18, 7, 19] focussed on the initial and less documented phase of the lexical innovation process, right after the words' creation and first use, and well before their final institutionalisation and inclusion in dictionaries [5, 6].

It is well-known that only a small portion of the words coined in everyday language use become new entries in dictionaries and thus part of the vocabulary: many remain ephemeral but are nevertheless compelling, as they provide evidence of the linguistic mechanisms driving the lexical innovation process. Generally, social media allow researchers to extract and use an unprecedented amount of conversational data [20, 21], which can provide reliable computations of lexical innovation and thus give a significant boost to the study of language variation and change [22, 23].

3. Corpus

In order to investigate emerging vocabulary in contemporary Italian, we used a corpus of timestamped and geotagged tweets extracted from the Italian Twitter timeline throughout 2022. The corpus comprises 5.32 M tweets written by 153 k unique users, amounting to 71.5 M tokens (or 564 M characters).

To the best of our knowledge, this is the first and largest study yet to address lexical innovation in Italian Twitter. Regrettably, this could also be the last. The recent takeover of Twitter collapsed its value for academia: as of summer 2023, publicly accessible data has been severely restricted, API prices have sharply risen, and academic access has been cancelled outright.

4. Methodology

Manual annotation aside, all our procedures are implemented as code and organised into a series of modular stages. To facilitate operation, they are accompanied and coordinated by an executable dependency tree specifying the relations between them, their inputs and their outputs. Together, they constitute a cohesive and reproducible data pipeline.

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

Table 1
List of Twitter's search query language conditions defining the *Italian Twitter timeline of 2022*.

We exclusively used Open Source Software, mostly in the form of well-known Python packages and GNU[3] tools. An exhaustive list including version numbers can be found in Appendix A.

In the following, we only discuss the general implementation design. The full source code is documented and available in [24].

4.1. Acquisition

Our corpus samples the *Italian Twitter timeline of 2022*. We define this notion as the conjunction of the conditions listed in Table 1, expressed using Twitter's advanced search query language[4].

Thus, our corpus is a subset of the results given by the search combining the aforementioned conditions at the time of sampling.

4.2. Preparation

4.2.1. Geographic Data

Tweets can bear geolocation data in two independent forms: a latitude/longitude pair and an association with a place. A place is an administrative division or a point of interest and it is characterised by an id, a country code, a geographical bounding box and other metadata. In our corpus, 99.43 % of tweets bear a place, 0.04 % only bear a lat./long. pair, and 0.53 % bear neither[5]. Consequently, despite lat./long. pairs being more precise, we chose to deal with places only, as they cover the vast majority of tweets and already include the country code necessary to restrict the data exactly to Italy.

We extracted 34.8 k unique places, keeping their id and country code (47.0 % are IT), and computed the

[2] https://www.iliesi.cnr.it/ONLI/intro.php

[3] https://www.gnu.org/

[4] Extensive unofficial documentation for the query language is available at https://github.com/igorbrigadir/twitter-advanced-search/. The user interface is found at https://www.twitter.com/search-advanced.

[5] This is possible because Twitter data can be redacted.

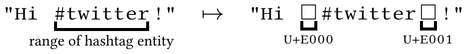

Figure 1: Schematic representation of how we inline entity range metadata as custom delimiters. This example shows how a hashtag entity is handled.

centroid of their bounding box as a reference point for geographical calculations.

4.2.2. Textual Data

Tweets are rich structures. They include an id, a user id, a timestamp, the full text, the geolocation data discussed above, a list of entities and other metadata. An entity is a character range in the full text labelled by a type (either *url*, *user mention*, *hashtag*, *symbol* or *media*) and other metadata.

First, we extracted all full texts into a flat data file to be loaded into AntConc [25] as an aid to the downstream manual annotation process.

Then, realising the entity metadata could greatly support the tokeniser at a later stage, we inlined them into the full text as delimiter markers, picking a different pair for every entity type from a set of reserved Unicode code points[6]. Figure 1 illustrates an example of how the procedure is carried out for hashtag entities.

Finally, we extracted 5.32 M tweets, keeping their id, user id, timestamp, full text with inlined entities, and place id.

91.77 % of tweets refer to places with the `IT` country code; we assigned these to Italian regions by matching their centroid with governmental data[7] on administrative boundaries in order to plot choropleth maps of Italy. Of the remaining tweets, 8.16 % refer to places with other country codes and 0.07 % refer to a generic place representing the entirety of Italy: the number of occurrences of candidate forms from these two categories are included in the choropleth maps under a legend titled "Not shown".

4.3. Cleanup and Tokenisation

We used the spaCy v3.6.1 Italian tokeniser. However, tweets are challenging for a stock tokeniser and some issues need to be addressed.

The first problem is the extensive use of Unicode (especially emojis), along with liberal usage of casing and whitespace. This can be easily addressed: we replaced all emojis with spaces, lowercased the whole text, and replaced any streaks of whitespace with a single space.

	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{A} \cap \mathcal{B}$	$\mathcal{A} \cup \mathcal{B}$
Size	6 737	21 132	979	26 890
Fraction	0.73 %	2.28 %	0.11 %	2.90 %

Table 2
Sizes of the candidate subsets as a count and as a fraction of the extracted forms.

The second, trickier, problem is the liberal usage of punctuation marks. Solving this required extending the tokeniser's default infix matcher to also match any sequence of these commonly abused punctuation marks: `?!;:,."()[]{}`.

The third and last problem is the presence of entities (urls, hashtags, etc.). This is where our previously inlined entity annotations came into play, quickly enabling us to make the tokeniser aware of them as follows:

- wrap all delimited regions in the text with spaces to nudge the tokeniser into correctly detecting their beginning,
- define a custom token matcher detecting any sequence whose extrema are our delimiter character pairs, and
- disable the tokeniser's default url matcher to avoid conflicts with our custom matcher.

The stratagems above allowed us to execute the tokeniser producing a negligible amount of spurious tokens. We then filtered its output, discarding tokens that were pure space, pure punctuation, pure numbers, broken and/or non-existent handles (i.e., tokens beginning with `@` but not marked as entities), and all entities except hashtags.

Processing all tweets as described, we extracted 71.5 M tokens, with 926 k types.

4.4. Candidate Selection

To select the candidates for annotation we applied two separate strategies, producing two subsets $\mathcal{A}$ and $\mathcal{B}$ with a slight overlap as detailed in Table 2.

$\mathcal{A}$ derives from an established method in literature, and $\mathcal{B}$ from our attempt to reach for a more interpretable and computationally lighter alternative. We now describe them both in detail.

4.4.1. Subset $\mathcal{A}$: Spearman's ρ

The first strategy follows in the steps of previous studies [18, 7] and amounts to calculating a measure of how monotonically the usage of a token increases in time in order to reject tokens below a fixed threshold. The

[6] We picked from the Private Use Area in the Basic Multilingual Plane, which is a set of code points left undefined by The Unicode Consortium [26, chapter 23.5] and reserved for special custom usage.

[7] Official ISTAT data is archived at https://www.istat.it/it/archivio/222527. We used the GeoJSON version maintained by the community, available at https://github.com/openpolis/geojson-italy/tree/2023.1.

chosen measure of monotonicity is the Spearman rank correlation coefficient between the daily occurrences of a token (normalised by daily total token count) and the day number; we denote it with ρ_O. The choice of threshold is arbitrary: while the cited studies operated on multi-billion tweet corpora picking very restrictive thresholds at 0.7 and 0.8, our corpus is much smaller so we can afford to lower the threshold until the size of the produced subset is still comfortable to annotate. We picked $\rho_O > 0.2$ selecting a subset of $4\,090$ candidates.

However, setting a positive lower bound to ρ_O penalises usage patterns we consider plausible for an emerging form (e.g., a sharp rise before midyear followed by a slow descent to a stable non-zero plateau). Therefore, we chose to extend the criteria to $|\rho_O| > 0.2$ selecting $2\,336$ additional candidates. In other words, we are discarding the central values of ρ_O, where it is less predictive. Furthermore, we decided to perform the same calculation on the daily unique users of a token; we denote the result with ρ_U. We allowed tokens with $|\rho_U| > 0.2$, selecting 311 additional candidates.

Our decision to be so permissive, at the cost of extra annotation effort, was dictated by the intention to experimentally evaluate the effectiveness of the bounds over a wide range of threshold choices.

Subset $\mathcal{A}$ is thus defined by the combined condition $\max(|\rho_O|, |\rho_U|) > 0.2$, selecting $6\,737$ candidates ($0.73\,\%$ of the total).

4.4.2. Subset $\mathcal{B}$: An Alternative Approach

ρ_O quantifies how much a form's usage increases monotonically during the year. As previously mentioned, while this complex measure correlates with the behaviour of some emerging forms, it also excludes plausible usage patterns.

We take the complementary approach and try instead to formulate *simple* criteria to *exclude* usage patterns that we would *not* expect from emerging forms:

- to reject accidental and sporadic phenomena (e.g., typos, inside jokes, etc.), we set a lower bound to the count of unique users U and occurrences O;
- to reject forms already in use from the past, we set a lower bound to the day of first occurrence A;
- to reject forms disappearing early, we set a large lower bound to the day of last occurrence Z;
- to reject ephemeral forms, we set a lower bound to the length of the usage lapse $\mathrm{Z} - \mathrm{A}$.

We chose the following thresholds: $U > 9$, $O > 9$, $\mathrm{A} > 7$, $\mathrm{Z} > 351$ and $\mathrm{Z} - \mathrm{A} > 28$. They read out as: we want forms that are used at least ten times by at least ten people, appear from the second week of January, do not disappear before mid December and last more than four weeks.

The specific values were tuned to cut off the markedly heavier tails from the distributions of the respective variables. This furthers the intention underlying our criteria to exclude the most common behaviours expected from non-emerging forms.

Appendix D contains charts showing how $\mathcal{A}$ and $\mathcal{B}$ partition the dataset and comparing the effect of their defining criteria over the parameter space.

Subset $\mathcal{B}$ defined by the conditions above includes $21\,132$ candidates ($2.28\,\%$ of the total).

4.5. Annotation

The subset for annotation $\mathcal{A} \cup \mathcal{B}$ amounts to $26\,890$ candidates ($2.90\,\%$ of the total extracted forms). To reduce the amount of handiwork, we used a lexicon of $514\,\mathrm{k}$ Italian forms specifically built for part-of-speech tagging tasks [27] to automatically tag already attested forms as uninteresting (including hashtags, to be analysed separately at a later stage) and thus excluding $18\,757$ candidates. This left us with $8\,133$ candidate forms for manual annotation, which was performed in two stages by the second and third author of the present paper, trained as a classicist and a corpus linguist respectively. Firstly, we loaded the corpus into AntConc [25] to look up each form's context (*KWIC - KeyWord in Context* format), while concurrently cross-checking two freely available online dictionaries and the ONLI neologisms database for attestation[8]. As a result of this search, the annotators rated forms as either *innovative* or *non-innovative*. Inter-annotator disagreement was settled with a negotiating phase until agreement could be reached for all forms. Examples of discarded entries include forms attested in at least one of the consulted dictionaries; mistypes caused by key proximity; popular terms, e.g., *bimbominchia*; foreign words well attested in the media but not in dictionaries (yet), e.g., *foliage*, *spending review*, *sponsorship*; adapted loanwords, e.g., *followo*, *crashare*; infrequently used foreign words, e.g., *smoothie*, *veggie*, *waffle*; infrequently used foreign acronyms, e.g., *PTSD*; regionalisms and regional variants, e.g., *annassero*, *ciolla*, *giargiana*; gender-inclusive graphic variants, e.g., *cittadinə*; nicknames, e.g., *pupone* for footballer Francesco Totti, and the unfriendly portmanteau *Cessica* (*cesso* + *Jessica*).

Next, and as shown in Table 3, we grouped innovative forms into one or more categories according to the ONLI typology scheme with minor adaptations and integrations. Specifically, we only relied on categories referring to formal properties, and thus ignored the *expressive*

[8] Garzanti at https://www.garzantilinguistica.it/ and Treccani at https://www.treccani.it/vocabolario/. The Slengo https://slengo.it/ urban dictionary was also used for the occasional look-up of slang forms.

Category	Forms	Examples
orthographic variation	109	*minkiate*, *rix*, *scienzah*
univerbation	48	*lho*, *miraccomando*
suffixation	45	*cinesata*, *sfanculamento*
loanword	40	*fancam*, *scammer*
portmanteau	33	*gintoxic*, *nazipass*
loanword adaptation	24	*flexo*, *droppare*
alteration	17	*fattoni*
prefixation	8	*bidosati*, *pregirata*
acronym	6	*lmv*, *sgp*
transcategorisation	6	*cuora*
compounding	3	*contapalle*
deonymic derivation	3	*drum*
redefinition	2	*maranza*
acronymic derivation	1	*effeci*
tmesis	1	*facenza*
Total form count	**346**	

Table 3
Categories used with respective candidate form counts and examples.

emphasis category used in the ONLI: emphasis is very common in Twitter interactions [21] and falls under all other categories. In addition, we merged multiple ONLI categories into one: e.g., *suffissazione*, *suffissoide*, *deverbale* and *denominale* were merged into *suffixation*, while *prefissazione* and *prefissoide* were merged into *prefixation*. Finally, a new *tmesis* category was added to account for forms deriving from the splitting of compounds (e.g., *facenza* from *nullafacenza*). Appendix C provides the complete list, and a machine-readable dataset of annotated candidates is available in Franzini et al. [28].

5. Results and Discussion

5.1. Emerging Forms

The most productive categories of lexical innovation in our corpus are:

- orthographic variation, often used either for emphasis (e.g., *minkiate*), to shorten existing words (e.g., *rix* for *risposta*), to conceal online conversation (also known as "leetspeak", e.g., *f4scist4*), for fun (e.g., *gomblotto*) or for sarcasm (e.g., *scienzah* with a final *-h* expressing scepticism towards scientific advances);
- univerbation, with forms such as *miraccomando*, *lho* or *senzapalle*;
- suffixation, featuring many forms ending in *-ato/a* (e.g., *cinesata*, *quarantenato*), *-mento* (e.g., *sfanculamento*) or with the intensifying *-issimo/a* applied to verbs (e.g., *riderissimo*) and to inherently intensified adjectives (e.g., *incantevolissimissima* from *incantevole*);
- (adapted) loanword, chiefly borrowed from English, with forms like *flexo*, *loser* and *trollazzo*;
- portmanteau, mostly relating to politics, with words such as *cessodestra*, *sinistronzi* and the amusing *lettamaio* (the combination of politicians Enrico Letta's and Luigi Di Maio's surnames reading as "pigsty"), but also *gintoxic* and *maxipass*.

Overall, the 346 forms give insights into the most common means by which potential new words are created by Italian speakers. Some of these are those traditionally detected in neologism studies: the *-ata* (*poverata*), *-ismo* (*cialtronismo*) and *-mento* (*sfanculamento*) suffixes, for example, are among the most common morphological resources used to derive new words from existing ones [12]. However, other forms seem particularly productive as potential sources of lexical innovation. Adapted loanwords, for instance, draw on the broad mechanism of inclusion of foreign verbs in the first conjugation in *-are* (*droppare*, *followo*, *switchare*), but also on less common phenomena, such as alteration through the suffixes *-ino* (*trollini*) or *-azzo* (*trollazzo*). Moreover, the widespread attitude towards evaluative language in social media interactions is witnessed by the presence of several emphatic and intensifying forms relying on different expressive means: in addition to the superlative suffix *-issimo/a* applied to verbs (*adorissimo*, *riderissimo*) or even employed as an autonomous word, particularly noteworthy is the use of augmentative suffixes like *-one* (*personaggione*, *garone*), univerbated forms (*opperbacco*, *eddaiii*, *masticazzi*), or portmanteaus such as *nazipass* and *sinistronzi* where emphasis blends with wordplay. Indeed, ironic and catchy wordplay frequently leads to lexical innovation and is typical of social media conversations.

Overall, a non-negligible part of the detected innovative forms are tied to the online sphere, and, in specific cases, are not expected to be used in different contexts or to establish themselves as new Italian words (e.g., *f4scista* or *mer*a*, which are mainly used to conceal content). Nevertheless, their emerging use in Twitter interactions evidences the linguistic mechanisms underlying lexical innovation in Italian. For each form we produce a choropleth map showing its usage. Appendix E presents the maps of all emerging forms mentioned in the article, while Figure 2 illustrates four notable examples from different categories. The map of *gomblotto* shows that orthographic variation, when used for emphasis or ludic purposes, is widespread in almost all regions, though predominantly in Lombardy. Conversely, when orthographic variation is not primarily intended as a joke (e.g., *poki* or *qndo*), the spread of new forms is not as far-reaching. Similar considerations can be made for univerbated forms, which appear to be evenly –albeit thinly– spread out with the

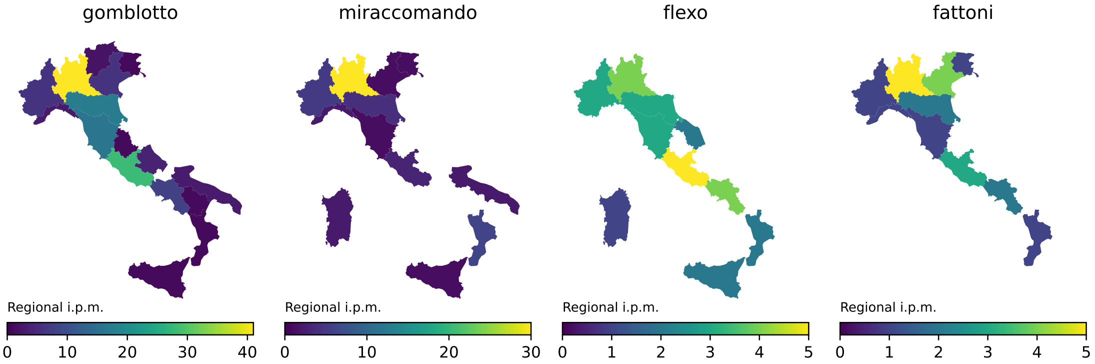

Figure 2: Choropleth maps showing the number of instances per million tokens at a regional level for the following forms: *gomblotto* (139 total instances), *miraccomando* (58), *flexo* (29) and *fattoni* (21). As previously mentioned, instances of forms found in tweets without an IT place association are not mapped: *gomblotto* (10), *miraccomando* (4) and *flexo* (1).

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

Table 4
Comparison of innovative form counts and yields between $\mathcal{A}_O^+$, $\mathcal{B}$ and their intersection.

occasional regional peak: *miraccomando*, for instance, is popular in Lombardy but less so in other regions. Other words reveal different patterns: the loanword *flexo*, for instance, meaning "to flaunt", is mostly used in the western part of the country with little to no attestation in the lower eastern regions; *fattoni*, an alteration of "fatto" to denote unreliable individuals and junkies, appears to be in use in the northern regions of Lombardy and Veneto but not so in either the eastern part of the country or the islands. Although, intuitively, spatial variation in social media has different characteristics from traditional geographical variation in relation to language use, previous research has detected a broad alignment between regional lexical variation in Twitter corpora and traditional survey data [29]. The geographical patterns revealed by the data, therefore, provide curious insight into the analysis of lexical innovation in Italian.

5.2. Yields Comparison

To evaluate our $\mathcal{B}$ strategy, we compare subset $\mathcal{B}$'s yield with $\mathcal{A}_O^+$, which is defined as the partition of $\mathcal{A}$ with $\rho_O > 0.2$, in order to fairly represent the approach of previous studies [18, 7]. Table 4 shows the results.

The adjusted yield, computed excluding attested forms and hashtags, favours $\mathcal{A}_O^+$. However, the projected yield, computed including hashtags and assuming the previous yield on them, favours $\mathcal{B}$.

Even without hashtags, $\mathcal{B}$ is noteworthy: its intersection with $\mathcal{A}_O^+$ yields less than the other two, indicating non-redundancy and hence the success of $\mathcal{B}$ in isolating behaviours excluded by $\mathcal{A}_O^+$.

Despite requiring five thresholds, $\mathcal{B}$'s are intuitively meaningful, unlike Spearman's more abstract ρ. Additionally, ρ is computationally expensive[9], making our approach more suitable for data exploration on weaker machines or larger datasets.

5.3. Limitations

Although the one-year time frame considered is both effective in the context of Twitter, where linguistic phenomena appear and spread in a short span of time, and coherent with our objective to investigate the initial emergence of new words, it could well fail to detect new forms that spread more slowly albeit at a constant rate.

Annotation with AntConc revealed the sporadic presence of tweets in French and Spanish. These had no impact on the identified forms but on the selection of the subsets. However, we expect this impact to be negligible and refrain from quantifying the effect at this time. Conversely, the `lang:it` filter most likely excluded some tweets in Italian, but no further assessment is possible with our dataset; there is also no public information about Twitter's proprietary language identification algorithm. Some instances of local Italian varieties were also noticed, confirming previous work [30], but they had no bearing on our analysis as we discarded regionalisms.

[9] A full-fledged time/space analysis is beyond the scope of this work, but we estimate our approach to be upwards of 50 times faster. More details are provided in Appendix B.

6. Conclusions and Future Work

Lexical innovation in Twitter seems to stem mostly from creativity, amusement and attention-seeking behaviour rather than a need for specific new words to indicate new objects, events or situations. The sense of belonging to a large and cohesive community such as Twitter plays a key role in the creation and dissemination of new words. The possibility of being adopted and reused in traditional oral conversation, in large (online) communication streams or, in a trans-medial perspective, by the press, makes at least some of these forms reliable candidates to become institutionalised neologisms.

Next steps in this ongoing study, to appear in Spina et al. [31], will focus on refining the list of candidate neologisms with additional dictionary look-ups (e.g., Zingarelli [32]) and on extending the analysis to the hashtags we put aside by virtue of their multi-functional and natively univerbated nature. Furthermore, we intend on leveraging our annotation data to examine how the yields of the two methods vary in restricting the threshold choices, in the hope of locating *sweet spots* to use as a rule of thumb in future studies. Finally, we will experiment with an estimator for the convexity of the cumulative usage, which, while computationally comparable to ρ, has better interpretability.

Should Twitter die out, planned efforts to scale-up our analysis to multiple Italian timelines will be redirected to other text-based microblogging and social networking platforms, namely Mastodon[10], Bluesky[11] and Threads[12].

Acknowledgments

The authors wish to thank the reviewers for their thorough and constructive feedback.

References

[1] W. Croft, Explaining Language Change: An Evolutionary Approach, Pearson Education, 2000.

[2] W. Labov, Principles of Linguistic Change, volume 2, Wiley-Blackwell, Oxford, 2001.

[3] E. Jezek, Lessico. Classi di parole, strutture, combinazioni, Itinerari, 2 ed., Il Mulino, 2011.

[4] G. Adamo, V. Della Valle, Osservatorio Neologico della Lingua Italiana. Lessico e parole nuove dell'italiano, volume 1 of *Temi e Strumenti*, ILIESI Digitale, 2019.

[5] R. Fischer, Lexical change in present-day English: a corpus-based study of the motivation, institutionalization, and productivity of creative neologisms, number 17 in Language in performance, G. Narr, Tübingen, 1998.

[6] D. Kerremans, A Web of New Words, Peter Lang, Frankfurt am Main, 2015.

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

[8] G. Adamo, V. Della Valle, Che cos'è un neologismo, Carocci, Roma, 2017.

[9] AA.VV., Neologismi (parole nuove dai giornali 2008-2018), Istituto dell'Enciplopedia Treccani, Roma, 2018.

[10] G. Adamo, V. Della Valle (Eds.), Innovazione lessicale e terminologie specialistiche, Olschki, Firenze, 2003.

[11] G. Adamo, V. Della Valle, Neologismi quotidiani. Un dizionario a cavallo del millennio, Olschki, 2003.

[12] F. Marri, I neologismi dentro e fuori dei repertori recenti, Quaderns d'Italià 23 (2018) 11–26. URL: https://doi.org/10.5565/rev/qdi.238.

[13] P. Zolli, Come nascono le parole italiane, Rizzoli, 1989.

[14] B. Rodríguez Arrizabalaga, Social Networks: A Source of Lexical Innovation and Creativity in Contemporary Peninsular Spanish, Languages 6 (2021). URL: https://www.mdpi.com/2226-471X/6/3/138. doi:10.3390/languages6030138.

[15] L. Tarrade, J.-P. Magué, J.-P. Chevrot, Detecting and categorising lexical innovations in a corpus of tweets, Psychology of Language and Communication 26 (2022) 313–329. URL: https://www.sciendo.com/article/10.2478/plc-2022-15. doi:10.2478/plc-2022-15.

[16] Q. Würschinger, Social Networks of Lexical Innovation. Investigating the Social Dynamics of Diffusion of Neologisms on Twitter, Frontiers in Artificial Intelligence 4 (2021). URL: https://www.frontiersin.org/articles/10.3389/frai.2021.648583/full. doi:10.3389/frai.2021.648583.

[17] J. Eisenstein, B. O'Connor, N. A. Smith, E. P. Xing, Diffusion of Lexical Change in Social Media, PLoS ONE 9 (2014). URL: https://dx.plos.org/10.1371/journal.pone.0113114. doi:10.1371/journal.pone.0113114.

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

[19] D. Kershaw, M. Rowe, P. Stacey, Towards Modelling Language Innovation Acceptance in Online Social Networks, in: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM '16, Association for Comput-

[10] https://joinmastodon.org/

[11] https://blueskyweb.xyz/

[12] https://www.threads.net/

ing Machinery, New York, NY, USA, 2016, pp. 553–562. URL: https://doi.org/10.1145/2835776.2835784. doi:10.1145/2835776.2835784.

[20] M. Laitinen, M. Fatemi, J. Lundberg, Size Matters: Digital Social Networks and Language Change, Frontiers in Artificial Intelligence 3 (2020). URL: https://www.frontiersin.org/article/10.3389/frai.2020.00046/full. doi:10.3389/frai.2020.00046.

[21] S. Spina, Fiumi di parole. Discorso e grammatica delle conversazioni scritte in Twitter, Aracne, 2019.

[22] D. Nguyen, A. Seza Doğruöz, C. P. Rosé, F. De Jong, Computational Sociolinguistics: A Survey, Computational Linguistics 42 (2016) 537–593. URL: https://direct.mit.edu/coli/article/42/3/537-593/1536. doi:10.1162/COLI_a_00258.

[23] D. Hovy, A. Rahimi, T. Baldwin, J. Brooke, Visualizing Regional Language Variation Across Europe on Twitter, in: S. D. Brunn, R. Kehrein (Eds.), Handbook of the Changing World Language Map, Springer International Publishing, Cham, 2019, pp. 3719–3742. URL: http://link.springer.com/10.1007/978-3-030-02438-3_175. doi:10.1007/978-3-030-02438-3_175.

[24] P. Brasolin, Breviloquia italica: data pipeline, 2023. URL: https://doi.org/10.5281/zenodo.10010427. doi:10.5281/zenodo.10010427.

[25] L. Anthony, AntConc (Version 4.2.0) [Computer Software], https://www.laurenceanthony.net/software, 2022. Tokyo, Japan: Waseda University.

[26] The Unicode Consortium, The Unicode Standard, Technical Report Version 15.0.0, Unicode Consortium, Mountain View, CA, 2022. URL: https://www.unicode.org/versions/Unicode15.0.0/.

[27] S. Spina, 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, volume 1, Pisa University Press, Pisa, 2014, pp. 354–359.

[28] G. H. Franzini, S. Spina, P. Brasolin, Breviloquia italica: annotations, 2023. URL: https://doi.org/10.5281/zenodo.10010528. doi:10.5281/zenodo.10010528.

[29] J. Grieve, C. Montgomery, A. Nini, A. Murakami, D. Guo, Mapping Lexical Dialect Variation in British English using Twitter, Frontiers in Artificial Intelligence 2 (2019). URL: https://www.frontiersin.org/articles/10.3389/frai.2019.00011/full. doi:10.3389/frai.2019.00011.

[30] A. Ramponi, C. Casula, DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy, in: Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 187–199. URL: https://aclanthology.org/2023.vardial-1.19. doi:10.18653/v1/2023.vardial-1.19.

[31] S. Spina, P. Brasolin, G. H. Franzini, Mapping emerging vocabulary in a large corpus of italian tweets, Research in Corpus Linguistics (in preparation).

[32] N. Zingarelli, lo Zingarelli 2022, I grandi dizionari, 2022.

Name	Version	Webpage
jq	1.6	jqlang.github.io/jq
GNU Parallel	20230622	gnu.org/software/make
GNU Bash	5.1.16	gnu.org/software/bash
GNU Make	4.3	gnu.org/software/parallel
Python	3.10.8	python.org
NumPy	1.25.2	numpy.org
SciPy	1.11.1	scipy.org
Pandas	2.0.3	pandas.pydata.org
Modin	0.23.0	modin.readthedocs.io
JupyterLab	4.0.4	jupyterlab.readthedocs.io
topojson	1.5	mattijn.github.io/topojson
Shapely	2.0.1	shapely.readthedocs.io
GeoPandas	0.13.2	geopandas.org
emoji	2.7.0	github.com/carpedm20/emoji
spaCy	3.6.1	spacy.io
Matplotlib	3.7.2	matplotlib.org
seaborn	0.12.2	seaborn.pydata.org

Table 5
Software and Python packages used in our data pipeline.

A. Data Pipeline Software Stack

The broad strokes of how we used Open Source Software to build our data pipeline are as follows: jq for bulk JSONL data manipulation parallelised with GNU Parallel; NumPy, SciPy and Pandas for general data manipulation and analysis parallelised with Modin; JupyterLab for data exploration; topojson, Shapely and GeoPandas for geographical data manipulation; emoji and spaCy for textual data cleanup and tokenisation; Matplotlib and seaborn for visualisation. All logic and glue code is written using Python and GNU Bash. GNU Make is used to codify an executable dependency tree between the pipeline stages, inputs and outputs. The versions of all stand-alone software and Python packages we used are listed in Table 5. Indirect Python dependencies are listed in the `requirements.txt` file of [24].

B. Computational Complexity

A full-fledged time/space complexity analysis is beyond the scope of this work, as it would require delving into

Figure 3: Code benchmarking the two methods we used and returning the speedup at various dataset sizes.

the implementation details of NumPy, SciPy, Pandas and Modin. However, we can still provide some general considerations and empirical measures on the behaviour of the two proposed methods on a dataset with c columns and r rows. In our case, $c = 365$ (days of the year) and $r \simeq 926k$ (token types).

Calculating Spearman's ρ for a row involves ranking two time series and calculating their Pearson correlation coefficient, so it is safe to assume its *best-case run-time is linear in* c (and probably log-linear on average depending on implementation details). Applying our method to a row involves (cumulative) sums and finding minima/maxima, so its *worst-case run-time is linear in* c. Naïve implementations using either method would simply iterate on the rows of the dataset, so they have *linear run-time in* r.

Given this rough time complexity analysis, we can expect our method to have *some* advantage regardless of implementation details. To quantify it, we ran a benchmark abstracting the core computations of the two methods and comparing their run-times for $c = 365$ and values of r up to the scale of our dataset. The code is presented in Figure 3 and the results are charted in Figure 4: we observe that our method is more than 50 times faster on bigger datasets.

The benchmark was run on a single core and expressed only as a speedup ratio to give a sense of what to *generally* expect. The implementation in Brasolin [24] is parallelised using Modin because we could run it on a hefty Intel Xeon E5-2690 v4 CPU with 128 GB RAM: we traded heavy memory usage for a further speedup, essentially making data exploration in a Jupyter notebook not only viable but pleasant. As a result, performing a detailed space complexity analysis is a particularly delicate matter and one that we do not address here. However, we should stress that our alternative method was initially developed because our means at the outset were much more limited (memory in particular proved to be a bottleneck at 16 GB), and that the initial, sequential, memory-aware implementation is still present in a comment alongside the parallelised one for use on smaller machines.

C. Full List of Innovative Forms

See Figure 5.

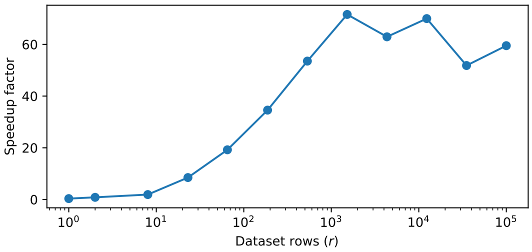

Figure 4: Chart showing the speedup of our method compared to calculating Spearman's ρ.

D. Comparison Charts for $\mathcal{A}$ and $\mathcal{B}$

See Figure 6.

E. Choropleth Maps of Examples

See Figures 7 and 8.

Orthographic variation

accaunt
adovo
affan
amerika
amiketti
amio
amïo
ancielo
anzia
assaj
azzzzz
babbà
benza
biutiful
c4zz0
caiser
cazxi
cazza
cme
collab
comple
coolo
csx
cuxo
dll
duddi
eu4ia
f4scist4
f4scista
fassisti
feffettissimo
gaz
gomblotto
graduidamende
graduidamente
graduido
gretina
grin
incaxxano
incaz
incazz
kaffè
kaimano
kazzate
kompagni
kultura
laik
leccac
lvi
madreh
mbeh
*mer*a*
merd@
merxa
minkiate
minkione
neanke
nerah
norde
nsomma
okk
okok
ovvove
pazzeska
pienah
pikkolo
pk
plis
poki
qlcosa
qlcuno
qlk
qndo
qnt
qt
qulo
qusto
reposta
rimba
rix
rubba
scienzah
sexi
sexo
singol
sinix
sll
snx
stronxate
stronz
tks
troya
trq
tuitt
ubri
urka
vafancul
vaff
vaffan
vaffanc
vairus
vaucher
vergonya
xazzo
xe
xhe
xsino
yessa
zola

Univerbation

ammiocuggino
anchio
buonagiornata
buonamattina
buontutto
cho
ciaobuogiorno
daltronde
demmè
diobono
dioca
diocan
dioporco
eddaiii
eropd
essu
estigrancazzi
evvaiiiiii
flattax
fuoriluogo
gintonic
graziealcazzo
ierisera
instagramstory
lho
lowcost
massí
masticazzi
mavalà
mavattelapijànd'erculo
miocuggino
miraccomando
ncazzo
nculo
noeuro
nowar
opperbacco
porcaputtana
porcodd
senzapalle
serietv
sottocasa
stemmerde
stica
streetart
terzopolo
tuttappost
ziocane

Suffixation

abilista
accannate
accannato
adorissimo
amorina
baguettari
benissimamente
cazzarone
ciacchera
cialtronismo
cinesata
cinesate
coglionazzo
ducessa
estaters
fisicati
godicchio
gretini
impiattamento
incantevolissimissima
inverners
legaiolo
mandrakata
memiamo
paccare
panchinato
pddizzato
piddini
pisellate
posturologo
poverata
presidenta
prosciutteria
quarantenati
riderissimo
rosiconi
senzadubbiamente
sfanculamento
sierare
sierata
tuitteri
twettini
twitteri
zanzarologi
zanzarologo

Loanword

admin
af
baller
banger
bollox
burp
champ
cishet
dilf
djset
drip
fail
fallout
fanbase
fancam
flu
horny
locals
loser
mentor
misunderstanding
reel
reminder
rimming
scammer
selca
shoutout
showrunner
slim
solution
soundbar
soundcheck
stats
terf
throwback
tier
topping
twitstar
venue
recap

Portmanteau

5stalle
assurdistan
cessodestra
deltacron
docuserie
fasciocomunista
fascioleghista
fascioleghisti
flurona
gintoxic
giornalanza
grillioti
grillopiddini
grillopitechi
intertristi
lettamaio
nazipass
naziucraini
pdiota
pdioti
piddiota
piddioti
pidiota
pidioti
presiniente
putler
renziota
renzioti
scansuolo
sinistronzi
tecnopolo
tridosato
triplodosati

Loanword adaptation

blastata
blessata
boyz
broder
condizionalità
cringiata
droppare
eppi
flex
flexo
followo
ghosta
matcha
pullato
schip
squirtare
stalkero
switchare
trollata
trollazzo
trolling
trollini
twerka
twitterino

Alteration

busoni
cazzaroni
eurini
falsona
fattoni
garone
paccotto
patati
patatino
personaggione
piagnina
pigiamone
pigiamoni
pirlotto
prezzemolina
ridolini
soggettone

Prefixation

appecorato
appecoronati
autoregalo
bidosati
biolaboratori
intrasezioni
iposcolarizzati
pregirata

Acronym

afc
lms
lmv
rdc
sgp
vfc

Transcategorisation

cuora
cuorare
cuoro
issima
issimo
vaffanculi

Compounding

contapalle
fotocazzo
fregacazzi

Deonymic derivation

cippalippa
drum
lippa

Redefinition

giornalaia
maranza

Acronymic derivation

effeci

Tmesis

facenza

Figure 5: Exhaustive list of the innovative forms we found, grouped by category.

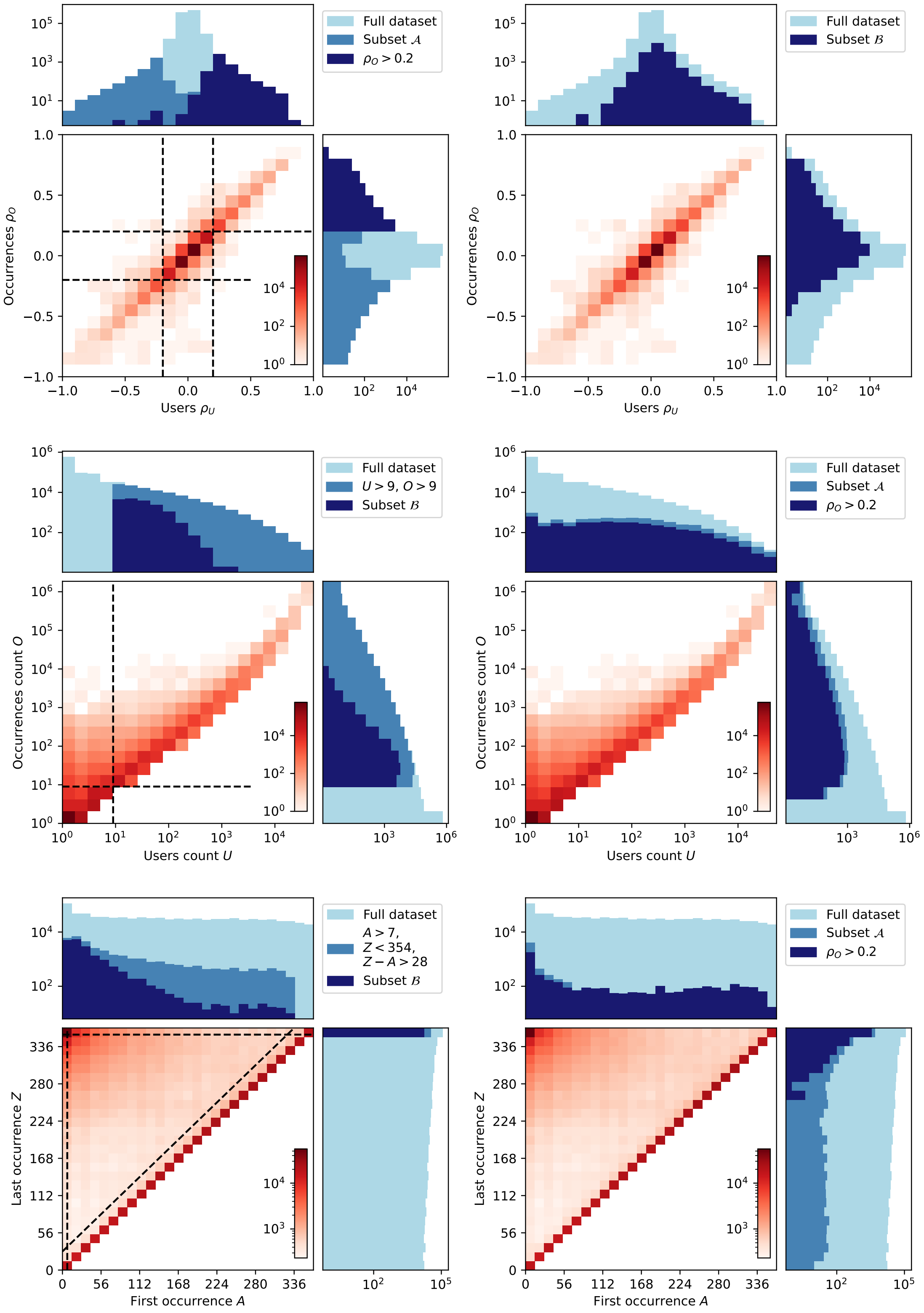

Figure 6: Charts comparing how $\mathcal{A}$ and $\mathcal{B}$ partition the dataset. Unlabeled axes are token counts. The dashed lines highlight how the thresholds act effectively discarding the densest areas. The last two charts reveal an intriguing pattern: a dense diagonal with tokens that appear and disappear quickly, and an opposite-facing dense corner with tokens that occur throughout the year.

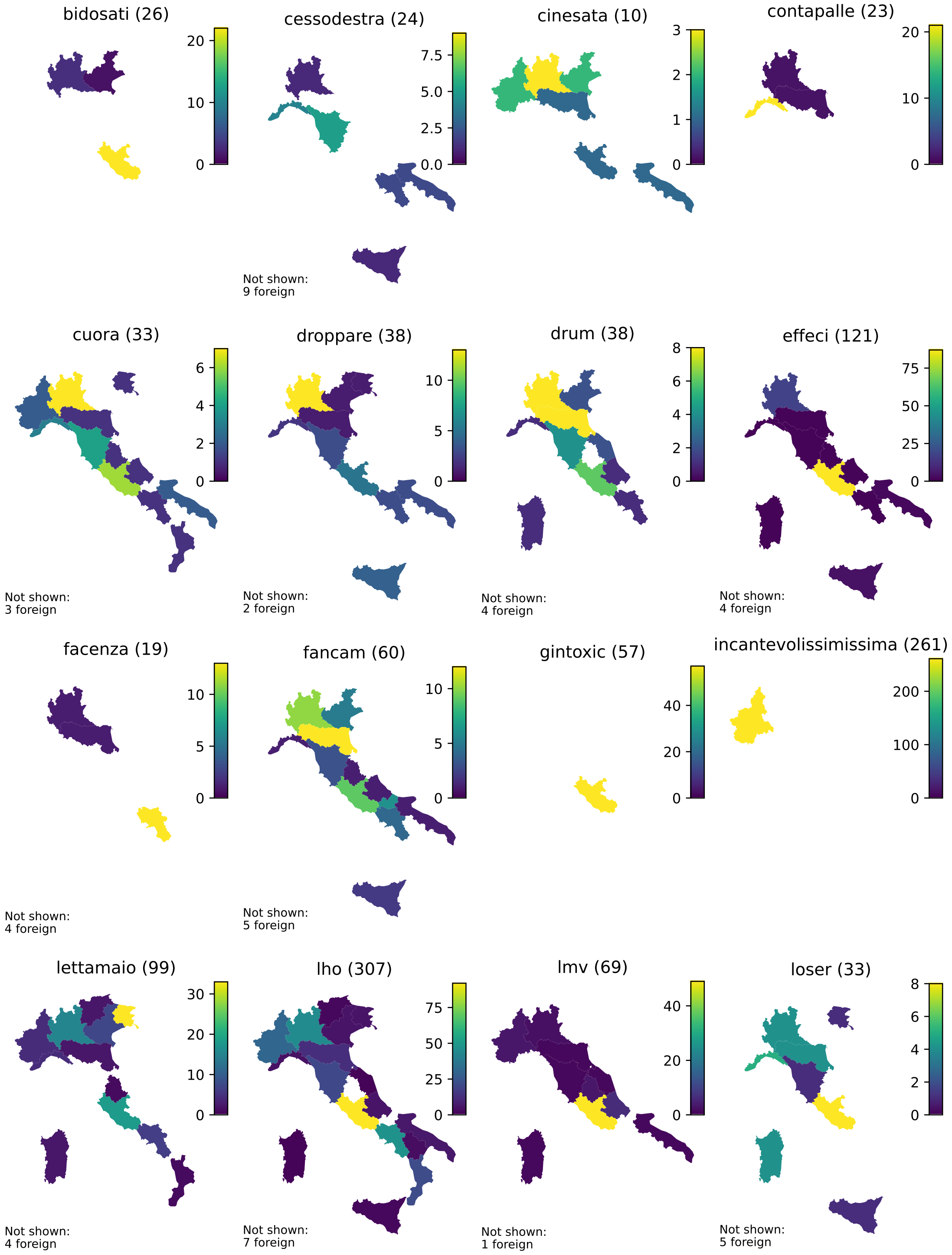

Figure 7: Choropleth maps of innovative forms mentioned as examples, from A to L. The colour scale represents instances per million tokens at the regional level. Total occurrences are provided with the titles, foreign ones in the legends. We omit *f4scist4* as it occurs outside of Italy only.

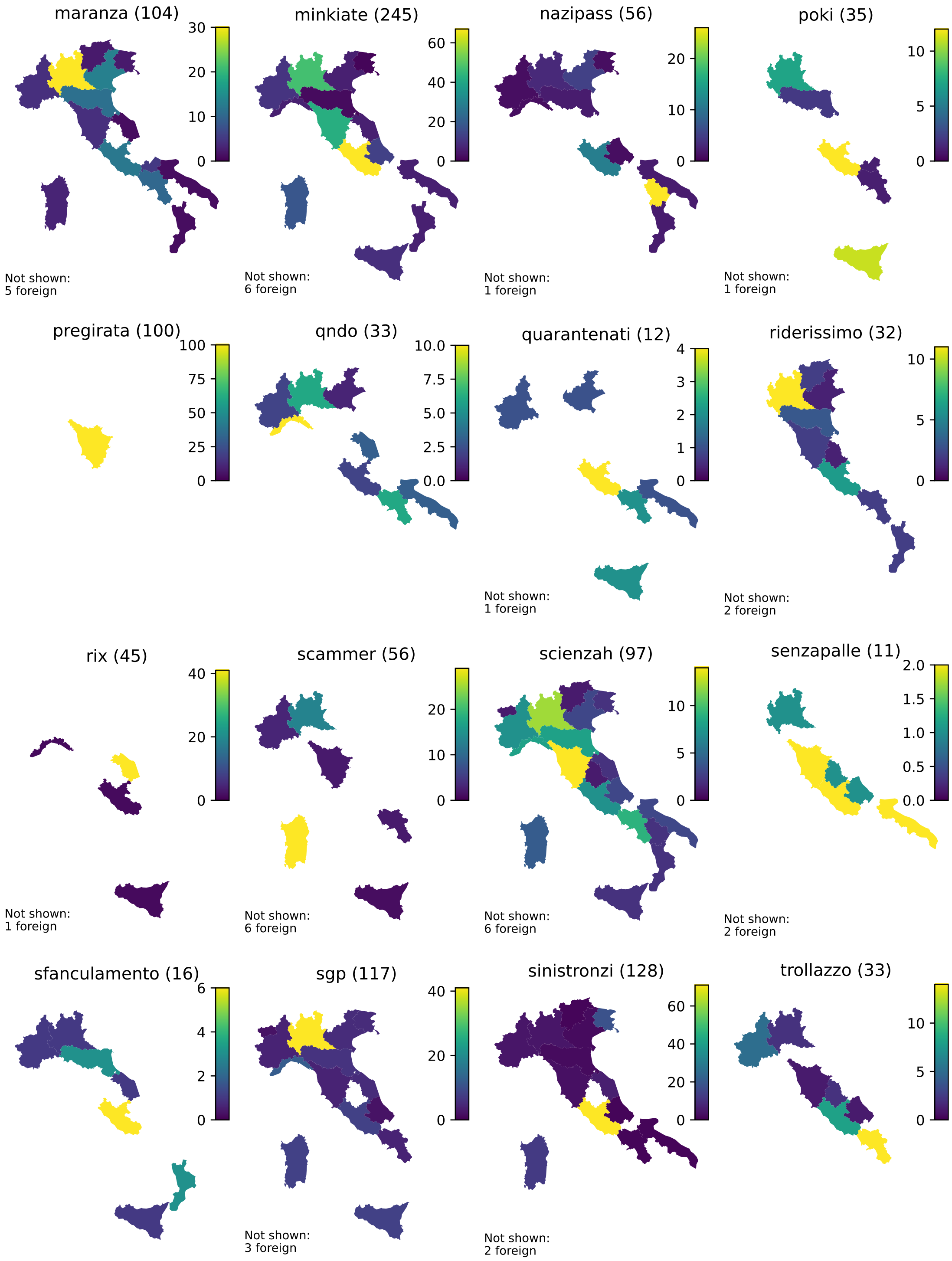

Figure 8: Choropleth maps of innovative forms mentioned as examples, from M to Z. The colour scale represents instances per million tokens at the regional level. Total occurrences are provided with the titles, foreign ones in the legends.