RiCL Research in Corpus Linguistics

Detecting emerging vocabulary in a large corpus of Italian tweets

Stefania Spina[a] – Paolo Brasolin[b] – Greta H. Franzini[c]
University for Foreigners of Perugia[a] / Italy
Independent researcher[b] / Italy
Institute for Applied Linguistics, Eurac Research, Bozen[c] / Italy

Abstract – This exploratory study investigates lexical change and innovation in contemporary Italian micro-blogging using a corpus of 5.32 million timestamped and geotagged tweets sampled from the 2022 Italian *Twitter* timeline. We develop a new method to identify 720 unattested forms (347 forms and 373 hashtags) as candidate neologisms. Our results show that orthographic variation, univerbation, suffixation, loanwords and portmanteaus are the most common categories of lexical creation in the data analysed, which appears to be driven by creativity, amusement and attention-seeking behaviour rather than a need for new words to define new objects, events or situations.

Keywords – *Twitter*; social media; corpora; Italian; lexical innovation; language change

1. INTRODUCTION[1]

Lexical innovation is a productive mechanism through which languages evolve (Croft 2000; Labov 2001) and adapt to new sociocultural and technological contexts. It is a crucial process for the survival and vitality of languages, as a living language is such when it is able to accommodate the new needs of its community. Lexical innovation is, therefore, integral to the process of language change, affecting all linguistic levels — phonological, morphological, lexical and syntactic— as well as orthographic aspects of languages. Neologisms are the result of the process of lexical innovation and can be defined as new words not belonging to the vocabulary of a language and not yet recorded in dictionaries or formed by adding new meaning to an already existing word. The process of creating new words follows different steps and usually develops from their initial appearance in specific contexts to their spread to wider domains. This process may end with a final institutionalisation of new word forms (Fischer 1998; Kerremans 2015)

[1] The authors wish to thank the anonymous reviewers for their time and valuable feedback, which helped to improve the quality of the paper.

through their inclusion in dictionaries and consolidation in standard use. However, among the vast number of words that are coined in everyday language use, many remain ephemeral, and only a small number of them become new entries in dictionaries and thus part of the vocabulary. This set of emerging lexical forms, which are only occasionally used for short periods of time and do not systematically enter the vocabulary of a language, are nevertheless of linguistic interest for the insight they give into the lexical innovation mechanisms through which languages evolve.

The process of creating new words can be approached from different standpoints: lexicographical applicability, linguistic phenomena involved, and sources used. Firstly, the process of tracing emerging words has direct lexicographical applications in the creation of neologism dictionaries (e.g., Adamo and Della Valle 2003), which collect new words weaved into daily conversation over a certain period of time, officially including them in the vocabulary of a language. Secondly, the linguistic phenomena leading to the creation of new words, be those involving, among others, derivation, composition or semantic shifting, are of great interest in the field of language change, even when emerging forms are sporadic or do not make it into dictionaries. Thirdly, the choice of sources used to trace the process of lexical innovation has great methodological relevance. Traditionally, newspaper texts have commonly been adopted as reliable sources for new word forms and the study of the lexicon of a language (Marello 2020), as they provide the double benefit of being easily available and quantitatively significant (Adamo and Della Valle 2019). Moreover, newspapers are widely circulated and are commonly transmitters of lexical innovation, both for stylistic reasons and the need to refer to new concepts. Held in high regard in contemporary society, newspapers incite the acceptance and spread of new words.

This study works on the hypothesis that social media represents an opportunity to explore (new) words emerging in everyday interaction, for it provides vast amounts of data produced in real time by a large number of speakers. We test this hypothesis for contemporary Italian with an analysis of emerging vocabulary in a sizeable corpus of tweets. Specifically, we propose a methodology geared towards the detection of emerging lexis and identify 347 word forms and 373 hashtags yet unattested in two of the most up-to-date Italian lexical resources, classifying them into 14 categories of lexical creation.

2. PREVIOUS STUDIES

Research on lexical innovation has produced extensive lexicographical works dedicated to neologisms in many languages, including English (e.g., Algeo 1991; Tulloch 1991; Maxwell 2006), French (e.g., Amar 2010; Des Isnards 2014), and Spanish (e.g., Martí Antonín 1998; Alvar Ezquerra 2003; Moliner 2013). Studies on lexical innovation in Italian boast a long tradition, and have led to the production of several dictionaries or collections of new words (e.g., Migliorini 1963; Scotti Morgana 1981; Lurati 1990; Adamo and Della Valle 2003, 2006, 2008; Bencini and Manetti 2005; De Mauro 2006), as well as a substantial body of research (e.g., Lo Duca 1992; Verardi 1995; Adamo and Della Valle 2003, 2017; Marri 2006, 2018; Frenguelli 2008). The relevance of these lexicographic resources lies not only in the fact that they provide a picture of lexical innovation processes as they occur in language, but also in the role they play in the preservation and documentation of those words in a specific time interval.

One of the fundamental issues faced by lexicography in the study of lexical innovation is the distinction between the notions of 'systemic' and 'occasional' forms in vocabulary (Zgusta 1971) or between 'neologisms' and 'nonce words' (Crystal 1997), the latter denoting occasionalisms not adopted into general use. This distinction is central to lexicographic work and should, in fact, make it possible to select words that have been identified as new and eligible for inclusion in general language dictionaries. Furthermore, this distinction concerns all words hanging between acceptance and disappearance, institutionalisation, and fall into oblivion. In this phase of linguistic stasis, emerging words are placed in an "antechamber of vocabulary" (Verardi 1995:28) and are thus unstable. Neologism dictionaries make room for this instability even when the recorded forms prove to be ephemeral.

It follows that the criteria governing the identification and categorisation of emerging forms as potential neologisms are crucial albeit hard to determine. One of the most widely discussed topics in this regard is the classification of the linguistic processes leading to the creation and spread of new words. Traditionally, research on neologisms acknowledges that the means by which languages enrich their vocabulary are essentially five (e.g., Giraud *et al.* 1971; Guilbert 1975; Zolli 1989):

1. Morphological derivation, that is, the formation of new words from pre-existing lexical elements with the addition of affixes. Examples are *autoregalo* 'gift given to oneself', where the prefix *auto-* modifies the noun *regalo* 'gift'; *prosciutteria*

'ham shop', where the suffix *-eria* modifies the noun *prosciutto* 'ham', or *pigiamone*, where the augmentative suffix *-one* modifies the noun *pigiama* 'pajamas'.

2. Morphological compounding, which is the formation of new words from pre-existing separate words combined to form a new compound word. An example is *contapalle* 'fibber', where the verbal form *conta* 'tells' is coupled with the noun *palle* 'lies'.

3. Reduction or orthographic/phonetic adaptation, that is, the formation of new words through the shortening (e.g., acronyms) or the modification of pre-existing forms. Examples are the acronym *rdc* for *reddito di cittadinanza* 'universal basic income', *csx*, a short form for *centrosinistra* 'centre-left', and *tuitt*, an orthographic variation of the form *tweet* reproducing the Italian pronunciation of the English word

4. Contact, which is the acquisition of new words from other languages or dialects ('borrowing') by adapting them to the paradigms of the target language (adapted loanwords) or by preserving them in their original form.[2] Examples from our corpus are *droppare*, the adaptation of the English verb *drop* to the Italian first conjugation in *-are*, and *fallout*, which is used in its original form.

5. Grammatical or semantic shift: the acquisition of new words through a change of grammatical category or the shift in the meaning of pre-existing forms. Examples are *giornalaia* 'newsagent', used to pejoratively connote a *giornalista* 'journalist', and the verb *cuorare* 'heart', an (incorrect) derivation of the noun *cuore* 'heart'.

Another aspect of lexical innovation widely discussed in previous research concerns the sources used to collect candidate neologisms. As previously mentioned, newspapers are commonly acknowledged as reliable sources for new word forms, as well as one of the most influential agents in the acceptance and dissemination of neologisms. In the last few decades, lexicographic projects have been established to track new words emerging in newspapers. One such project is the *Osservatorio Neologico della Lingua Italiana* (ONLI

[2] While we explicitly exclude dialectal forms from our analysis, examples in our corpus of tweets include *poerannoi* 'poor us' (from the Florentine dialect), *fratm*, an abbreviation of 'my brother' (typical of southern Italy) and *giargiana*, which is used in Milan to denote people who are not from Milan.

2012; Adamo and Della Valle 2019), which has released a database now counting 2,986 new words with definition, date of attestation and first retrieved occurrence in the press.

More recently, with the popularisation of other forms of mass communication and conversational participation, research has stressed the benefits of using social media to track new words emerging in everyday conversation (Rodríguez Arrizabalaga 2021; Würschinger 2021; Tarrade *et al.* 2022). Indeed, the natural ebb and flow of conversation fostered by social media brings out vocabulary approximating the immediacy of spoken interaction (Spina 2016, 2019) and lexical creativity from ordinary users as opposed to inventive journalistic discourse (Eisenstein *et al.* 2014).

A number of recent social media-based studies (Grieve *et al.* 2016, 2018; Kershaw *et al.* 2016) have focussed on the initial phase of the lexical innovation process, that located between a word's creation and first use in a specific context, and its spread in different contexts and potential institutionalisation (Fischer 1998; Kerremans 2015). Another advantage of using social media is that it allows researchers to access unprecedented amounts of conversational data (Spina 2019; Laitinen *et al.* 2020), which can provide a reliable quantitative basis for computations of emerging word forms, thus giving a significant boost to the study of language variation and change (Nguyen *et al.* 2016; Hovy *et al.* 2019).

3. The corpus

To explore evolving lexis in contemporary Italian, we sampled and analysed a dataset of timestamped and geotagged tweets from the Italian *Twitter* timeline spanning the entirety of 2022. The dataset contains 5.32 million tweets authored by 153 thousand unique users, totalling 71.5 million tokens (equivalent to 564 million characters).

4. Method

With the exception of manual annotation, our procedure is structured into a reproducible modular data pipeline. Exclusively relying on Open-Source Software, primarily in the

form of widely recognised Phyton packages and GNU tools, our approach ensures transparency and accessibility.[3]

4.1. Corpus creation and preparation

Using *Twitter*'s advanced search query language,[4] we extracted tweets from the 2022 Italian *Twitter* timeline matching the conditions outlined in Table 1. Tweets can contain geolocation data in two distinct forms: 1) a latitude/longitude pair or 2) an association with a place. A place, in this context, refers to an administrative division or a point of interest and is defined by an ID, a country code, a geographical bounding box, and other metadata. Within our corpus, 99.43 per cent of the tweets are associated with a place, only 0.04 per cent have a latitude/longitude pair, and 0.53 per cent have neither. Despite the higher precision of latitude/longitude pairs, we opted to focus exclusively on places, given that they cover the vast majority of tweets and already include the country code necessary to restrict the data to Italy.

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

Tweets consist of an ID, a user ID, a timestamp, the complete text, the previously discussed geolocation data, a list of entities and additional metadata. An entity refers to a character range in the full text labelled by a type (such as url, user mention, hashtag, symbol, or media) and other associated metadata.

Firstly, we extracted all full texts into a flat file, intending to load it into the *AntConc* concordancer (Anthony 2022) to facilitate the subsequent manual annotation process. Next, we introduced entity metadata into the full text as delimiter markers to trick the downstream tokenisation process into breaking these richly structured strings correctly;

[3] The documented source code can be accessed at Brasolin (2023). For a detailed description of the computational processing of the linguistic data, see Brasolin *et al.* (2023).
[4] The official documentation of the query language is available at https://github.com/igorbrigadir/twitter-advanced-search/ and the user interface can be accessed at https://www.twitter.com/search-advanced.

for each entity type, we selected distinct pairs from a set of reserved Unicode code points.[5] Figure 1 provides an example of how this procedure was implemented for hashtag entities.

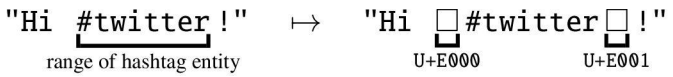

Figure 1: Schematic representation of how we inlined entity range metadata as custom delimiters. This example shows how a hashtag entity is handled

Thirdly, we extracted 5.32 million tweets, preserving their ID, user ID, timestamp, full text with inlined entities, and place ID. Of these tweets, 91.77 per cent are associated with places bearing the IT country code. By aligning their centroids with governmental data,[6] we plotted the tweets containing the emerging forms onto choropleth maps to illustrate the forms' regional distribution across Italy (see Figures 2 and 3 in Appendix A). Specifically, the maps display the simple frequency of each emerging form in the entire corpus (i.e., the sum of the number enclosed in parentheses and, if applicable, that provided in the respective legends) and the number of regional occurrences per million tokens (indicated by the colour scale to the right of the map). Of the remaining tweets, 8.16 per cent are linked to places with other country codes, and 0.07 per cent reference a generic place representative of Italy as a whole. Finally, to tokenise the corpus, we employed the *spaCy v3.6.1* Italian tokeniser.[7]

4.2. Candidate selection

To choose the candidates for annotation we used two different approaches, that is, an already established method in literature and our own attempt at a more interpretable and computationally lighter alternative. This resulted in two groups which have a few candidates in common, as shown in Table 2. The subset of candidates we annotated is the union of the two groups. We now describe both methods in detail.

[5] We picked from the Private Use Area in the Basic Multilingual Plane, which is a set of code points left undefined and reserved for special custom usage (The Unicode Consortium 2022: Chapter 22.5).
[6] 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.
[7] https://spacy.io/

	Grieve's	Alternate	Overlap	Union
Subset size	6,737	21,132	979	26,890
Fraction of total	0.73%	2.28%	0.11%	2.90%

Table 2: Sizes of the candidate subsets obtained with the two methods, both as a count and as a fraction of the extracted forms. The rightmost columns quantify the size of the overlap and of the union of the two subsets

4.2.1. Grieve's method

The first method is based on previous studies (Grieve *et al.* 2016, 2018) and amounts to calculating how consistently a word's usage increases over time and discarding any word below a certain threshold. The calculation is performed using the Spearman rank correlation coefficient comparing the daily occurrences of a word O (adjusted for the total word count of the day) and the day number. We denote this coefficient ρ_O. The choice for the threshold is somewhat arbitrary. While previous studies, which used much larger datasets, set very high levels at 0.7 and 0.8, we were able to set a lower level due to our smaller dataset and still obtain a manageable number of candidates. We chose $\rho_O > 0.2$, which gave us a subset of 4,090 candidates.

Setting a positive lower limit for ρ_O can penalise usage patterns that could represent an emerging word (for example, a sharp increase in usage before midyear followed by a slow decrease to a stable, non-zero level). Therefore, we decided to include words with $\rho_O < -0.2$ as well, which added 2,336 more potential words to our subset.

In addition, we decided to apply the same calculation to the daily unique users of a word U, obtaining the ρ_U coefficient. We included words with $|\rho_U| > 0.2$, adding 311 more potential words to our subset.

Overall, we selected 6,737 candidates (0.73% of the total) with the following criteria: $max(|\rho_O|, |\rho_U|) > 0.2$.

4.2.2. Alternative method

The measure ρ_O quantifies how much the use of a form increases steadily over the year. As previously discussed, this complex measure aligns with the behaviour of some emerging forms, but it also leaves out possible usage patterns.

We take a different approach and aim to create simple criteria to exclude usage patterns that we would not associate with emerging forms:

a) To rule out accidental and occasional phenomena (like typos, inside jokes, etc.), we set a minimum limit to the count of unique users U and occurrences O.

b) To rule out forms already in use from the past, we set a minimum limit to the day of first occurrence A.

c) To rule out forms that fade away early, we set a high minimum limit to the day of last occurrence Z.

d) To rule out short-lived forms, we set a minimum limit to the length of the usage period $Z - A$.

We chose the following thresholds: $U > 9$, $O > 9$, $A > 7$, $Z > 351$ and $Z - A > 28$. This means we are looking for 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 subset defined by the conditions above includes 21,132 candidates (2.28% of the total).

4.3. Corpus annotation

The subset for annotation comprises a total of 26,890 candidates corresponding to 2.90 per cent of the extracted forms. In an effort to streamline the manual annotation process, we used a lexicon of 514 thousand Italian forms (Spina 2014) to automatically filter out attestations from our corpus, resulting in 11,524 candidates.

4.3.1. Non-hashtags

Of the 11,524 candidates, 8133 are non-hashtag forms. The first and second authors of this paper, trained as a corpus linguist and classicist respectively, and manually annotated these forms in two stages. Firstly, we loaded the corpus into AntConc as a flat file and used its *Key Word in Context* tool to look up each form in context. At the same time, we

scanned two freely available online dictionaries, *Garzanti*[8] and *Treccani*,[9] as well as the ONLI neologism database for attestation. The *Slengo* urban dictionary was also consulted for the occasional inspection of slang forms.[10] Based on this comprehensive search, the two annotators categorised forms as either unlikely (assigning them a score of -1) or likely (assigning them a score of 1) to become new dictionary entries, resolving any inter-annotator disagreements through negotiation until consensus was achieved for all forms.

The criteria used to annotate forms as unlikely to become dictionary entries included:

- Attestation in the consulted dictionaries.
- Typos, including those caused by key proximity, e.g., *boungiorno* instead of *buongiorno* 'good morning', *cszzo* instead of *cazzo* 'dick'.
- Established popular neologisms missing from dictionaries, e.g., *bimbominchia* 'sucker'.
- Established foreign words used by the media but missing from dictionaries, e.g., *foliage*, *spending review*, *sponsorship*.[11]
- Nicknames and terms of endearment, e.g., *Gasp* for *Gasperini* or *pupone* 'big baby' for footballer Francesco Totti.
- Vowel elongation for emphasis, e.g., *amooooo* 'loveee'.
- Infrequently used foreign words, e.g., *smoothie*, *veggie*, *waffle*.
- Infrequently used foreign acronyms, e.g., *PTSD*.
- Regionalisms, e.g., *annassero* (Romanesco for *andassero*, third person plural subjunctive of *andare* 'go', *ciolla* 'dick', 'idiot' or 'drugs', depending on the context), *giargiana* (anyone who is not from central Milan)
- Gender-inclusive graphic variants, e.g., *cittadinə* 'citizens'.

In a second stage, we sorted likely candidates according to the ONLI category scheme with minor adjustments and integrations (see Table 3 in Section 5). Specifically, we

[8] https://www.garzantilinguistica.it/
[9] https://www.treccani.it/vocabolario
[10] https://slengo.it/
[11] Gazzardi and Vásquez (2020) provide an overview of studies on the (unnecessary) use of English words in Italian media.

focussed on categories related to formal properties, excluding, for instance, the 'expressive emphasis' category as is commonly found in *Twitter* interactions (Spina 2019) and is inherent in all other categories. Similarly, we merged multiple ONLI categories into one, namely *suffissazione* 'suffixation', *suffissoide* 'suffixoid', *alterazione* 'alteration', *deverbale* 'deverbal' and *denominale* 'denominal' into 'suffixation', and *prefissazione* 'prefixation' and *prefissoide* 'prefixoid' into 'prefixation'. Also, we introduced a new 'tmesis' category to account for forms resulting from the splitting of compounds, e.g., *facenza* from *nullafacenza* 'laziness'. Finally, and where possible, we added the part-of-speech of every form using *TreeTagger's Stein* tagset for Italian as a reference.[12]

4.3.2. Hashtags

Our 11,524 candidates also include 3,391 hashtags. Universally, hashtags appear as either single or unbroken sequences of words (including characters, numerals and underscores), and are often used in their English rendition to expose associated tweets to a wider and more diverse audience.[13] To account for the bias introduced by forced univerbation and English dominance, our hashtag analysis takes a marginally different approach to the one adopted for non-hashtag forms and follows both objective and subjective criteria. We narrow our hashtag selection by filtering out:

1) Those used by nine or fewer distinct users.

2) Proper names, including but not limited to people (e.g., *#gigidagostino*, *#vettel*, as well as portmanteaus like *#basciagoni* used to blend the surnames of Italian *Big Brother* contestants Alessandro Basciano and Sophie Codegoni), places (e.g., *#bozen*, *#regionepuglia*, *#tunisia*), organisations (e.g., *#crocerossaitaliana*, *#aeronauticamilitare*), brands (e.g., *#gucci*, *#versace*), sports teams (e.g., *#acbellinzona*) and events (e.g., *#atpfinals*), festivities (e.g., *#christmas2022*, *#carnevale22*), videogames (e.g., *#eldenring*), music bands (e.g., *#articolo31*) and concerts (e.g., *#cremoninilive22*), films (e.g., *#dontlookup*), and TV shows (e.g., *#1899netflix*, *#cepostaperte*).

[12] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/italian-tagset.txt
[13] See, for example, Hashtagify at https://hashtagify.me

3) Hashtags containing proper names, e.g., *#adaniout* (referring to football commentator Daniele Adani), *#iovotoitaliaviva* 'I support/vote for Italia Viva'.

4) (Combinations of) years, days of the week, times and numbers, e.g., *#8marzo*, *#dicembre2022*, *#anni90*, *#sundaymorning*.

5) Short-lived hashtags relating to a specific incident or time interval, e.g., *#djokovid*, *#draghistan* (referring to former prime minister Mario Draghi's leadership).

6) Univerbated hashtags that we believe have little to no probability of making it into lexical resources, e.g., *#womanlifefreedom*, *#buongiornoatutti* 'good morning everyone'.

We then separate the remaining hashtags into single and univerbated words for manual annotation. The annotation of single-word hashtags, such as *#carobenzina* 'increase in the price of petrol' or *#spiaze* 'it's a pity', is identical to that of non-hashtag forms (see Section 4.3.1), with an additional distinction between *informative* and *evaluative* hashtag function (see Section 6.1) for purposes of analysis. Instead, our annotation of univerbated hashtags, such as *#andratuttobene* 'everything will be alright' or *#booklover*, is objective with respect to ONLI and function categorisations (we do not tag for part-of-speech), but less so in regard to likelihood. In other words, we only consider those univerbated hashtags that we intuitively believe are more likely to establish themselves as new (non-hashtag) forms in Italian social media communication and/or to be acknowledged in authoritative lexical resources, for instance, *#avantitutta* 'let's go!' or *#oldschool*.

5. Results

The selection method described yields a list of 720 emerging forms (347 non-hashtags and 373 hashtags), distributed across 14 categories of lexical creation, as shown in Tables 3 and 4. The emerging forms were also labelled with zero or more part-of-speech tags, producing the distribution shown in Table 5.

The complete list is available in Appendix B, and a machine-readable dataset of the annotated candidates is freely accessible in Franzini *et al.* (2023).

ONLI category	Count	Examples
Orthographic variation	**111**	*Minkiate*, *scienzah*
Suffixation	**60**	*Cinesata*, *adorissimo*
Univerbation	**48**	*Stemmerde*, *massì*
Loanword	**39**	*Reminder*, *scammer*
Portmanteau	**33**	*Lettamaio*, *assurdistan*
Loanword adaptation	24	*Flexo*, *droppare*
Prefixation	8	*Appecoronato*, *iposcolarizzati*
Transcategorisation	7	*Cuora*, *panchinato*
Acronym	6	*Lmv* (li mortacci vostri), *vfc* (vaffanculo)
Compounding	4	*Contapalle*, *cessodestra*
Deonymic derivation	3	*Drum*, *cippalippa*
Redefinition	2	*Maranza*, *giornalaia*
Acronymic derivation	1	*Effeci*
Tmesis	1	*Facenza*
Total non-hashtag forms	**347**	

Table 3: Counts of forms by category, with examples

ONLI category	Count	Examples
Loanword	279	*#aperitif*
Univerbation	50	*#accaddeoggi*
Portmanteau	21	*#caturday*
Acronym	13	*#pdr* (Presidenza della Repubblica)
Compounding	5	*#caroenergia*
Orthographic variation	4	*#povery*
Prefixation	1	*#extraprofitti*
Total hashtag forms	**373**	

Table 4: Counts of hashtag forms by category, with examples

Part of Speech	Non-hashtag	Hashtag	Total
NOM (noun)	201	189	390
ADJ (adjective)	72	23	95
INT (interjection)	46	5	51
VER (verb)	30	17	47
ADV (adverb)	13	1	14
PRO (pronoun)	8	0	8
CON (conjunction)	7	0	7
NPR (name)	5	0	5
PRE (preposition)	2	0	2

Table 5: Counts of PoS tags by form type, and total. Note that forms can have zero or multiple tags

6. Discussion

In the following, we focus on non-hashtag emerging forms, discussing the role of hashtags in a separate section.

The results of the extraction and filtering of emerging forms in the *Twitter* timeline of 2022 allowed us to identify some noteworthy patterns in the mechanisms underlying lexical innovation in Italian. 60 new words (17% of the total number) are formed through suffixation, which is traditionally one of the most common mechanisms languages rely on to create new words (Iacobini and Thornton 1992). These 60 emerging lexical items are mainly created using the derivative suffixes that Italian resorts to in its morphological processes. Examples are the suffixes *-mento* (*impiattamento*, 'plate up'), *-ismo* (*cialtronismo*, behaviour characteristic of a slacker), *-ista* (*abilista*, 'ableist'), *-ato* (*quarantenato*, 'quaranteened'), *-ata* (*poverata*, action characteristic of a poor person), *-eria* (*prosciutteria*, 'ham shop'), *-iolo* (*legaiolo*, hostile designation of a follower of the Italian right-wing populist political party Lega), *-one* (*cazzarone*, 'big/master bullshitter'), and *-azzo* (*coglionazzo*, 'big idiot') or *-ero* (*tuitteri*, '*Twitter* users').

To create new lexical items in Italian, therefore, *Twitter* users rely on established mechanisms. Some, such as derivation through the suffixes mentioned above, are rooted in the earliest stages of the history of the Italian language, whereas others seem to emerge specifically in *Twitter* interactions. An example is the superlative suffix *-issimo*, which is very common in Italian and has the function of intensifying adjectives (Micheli 2020), as in *bellissimo* 'very beautiful'. The suffix *-issimo* has already widened its range of applications, as it can also be found applied to nouns (see Grandi (2017); e.g., *partitissima* 'very important match'.

In our corpus, this suffix finds additional applications. In two of the three emerging forms ending in *issimo* (*adorissimo* and *riderissimo*, see example (1)), the intensifying suffix does not modify an adjective but a verb (*adorare* 'adore' and *ridere* 'laugh'). These two forms represent a further extension of the possible combinations of the suffix *-issimo* and are of major interest because they not only involve lexical but also morphological innovation.

(1) *Io lo* ***adorissimo****, un genio assoluto di simpatia*.
'I adore him so much, an absolute genius in likeability'.

The third new form in *-issimo* detected in our corpus is *incantevolissimissima* 'very very enchanting'. In this case, the form is anomalous for semantic reasons because *-issimo* is applied to an inherently intensified and not gradable adjective.

The search for intensification (Spina 2019) and language economy seems to drive participants in *Twitter* interactions to create new lexical forms. Other examples are instances of the suffix *-one*, for example *cazzarone* 'big/master bullshitter', *rosiconi* (people who feel anger and/or jealousy for someone else's success), *garone* 'big competition' and *fattoni* 'unreliable individuals', 'junkies'. The shift from the original augmentative meaning of *-one* (e.g., *librone* 'big book') to the intensifying, evaluative and pejorative meaning of our examples can be explained through the extension of the suffix's core meaning 'big' to the new meaning of 'intense' (Grandi 2017), or even 'bad'. While this mechanism is not new in Italian derivational morphology, it seems to be one of the most productive ones, partly because the suffix *-one* can be applied to nouns (*garone*) as well as verbs (*rosicone* from *rosicare* 'feel envy').

Another productive suffix for lexical innovation in *Twitter* is *-ata*, which is "one of the most semantically fragmented Italian suffixes" (Grossmann and Rainer 2004: 253). Among the emerging forms in *-ata*, with the exception of those classified as adapted loanwords such as *cringiata* (something embarassing) or *blastata* 'humiliation', 'derision', four cover at least two of the multiple senses of the suffix: in *cinesata*/*cinesate* (to indicate Chinese products), *mandrakata* 'ingenious find', or 'scam' and *poverata* (to denote an action characteristic of a poor person) *-ata* is attached to a nominal animate subject (a Chinese product, Mandrake, a poor person) to connote an action and a negative/pejorative meaning. Example (2) shows this of *cinesata*.

(2) *Beh l'originale è sempre meglio della* ***cinesata****, si sa.*
'Well, everybody knows that the original is always better than the Chinese version'.

The borrowing of foreign words, whether adapted to Italian morphology or not, is another driver of lexical innovation, covering 18 per cent of all of the new forms. The 63 loanwords come from English, with the only exception of *selca* (see example 3), which is a Korean word for *selfie* (*self* + *camera*), and of *matcha*, used to indicate a variety of Chinese green tea or, as the adaptation of the English 'match' to the Italian third person of the present indicative.

(3) *Se non posta un **selca** con i capelli mossi faccio la pazza.*
'If (s)he doesn't post a selfie with wavy hair I'll act crazy'.

English forms imported into Italian can belong to specific lexical domains, such as music (*e.g., djset*, *soundbar*, *soundcheck*) and online environments (e.g., *admin*, *reel*, *twitstar*, *twitterino* '*Twitter* user', *trollino* 'little troll', *trollazzo* 'big nasty troll'), or be part of general everyday use (e.g., *fail*, *flu*, *reminder*, *shoutout*). The abundance of these commonly used words is a notable advantage of using social media conversations among large and diverse groups of ordinary users as a source for lexical innovation. Indeed, while newspapers do contain features of informal everyday speech (Pulcini *et al.* 2012; Marello 2020), articles penned by a limited number of journalists typically employ a more formal vocabulary associated with politics, news reporting or foreign affairs, often detached from everyday use.

One of the differences between direct and adapted loanwords relates to grammatical categories. With the exception of two interjections (*bollox* and *burp*), the former are mainly nouns and adjectives, whereas adapted loanwords —excluding the few nouns adapted through the alterating suffixes *-ino* (*trollini*) or *-azzo* (*trollazzo*), or through the productive suffix *-ata* (*blastata*, *cringiata*), are mainly verbs (*switchare*, *stalkero*, *ghosta*, *flexo*, *droppare*)— conjugated in the first conjugation in *-are*, as is the case for *ghosta* in example (4):

(4) *Ho perso una persona così immatura che **ghosta** invece di dire che non vuole sentirmi più.*
'I have lost a person so immature they'd rather ghost me than say they no longer want to speak to me'.

This difference lies in the fact that the Italian verbal morphology is much more articulated than its nominal morphology, so a verb borrowed from another language must necessarily undergo adaptations in order to become part of the Italian vocabulary. However, in other collections of Italian neologisms based on newspaper articles, such as the ONLI, loanword adaptation does not even exist as a category. Again, adaptations of foreign words to Italian morphology are familiar in register, and thus not suitable to the more formal journalistic style. Two interesting examples of a noun deriving from an adapted loanword are *cringiata* (5) and *blastata* (6), where *cringe* and *blast* become nouns through the addition of the suffix *-ata*.

(5) *La casa di carta coreana la **cringiata** del secolo ora mi dovete spiegare perché.*
'The Korean house of cards is so cringy now you have to tell me why'.

(6) *Mamma mia che **blastata**, me la sto davvero facendo sotto.*
'My goodness what an attack, I was truly scared'.

32 per cent of all of the detected emerging forms were labelled as orthographic variation, which is the most productive category of lexical creation in our corpus. Related research (Grieve *et al.* 2016:110) reports that:

> spelling variation is not generally considered a standard word formation process, as it is not an option in spoken language. From an orthographic perspective, however, these are new linguistic forms.

While our 111 lexical forms mostly align with this observation and are treated as candidates for dictionary inclusion, there are exceptions. Some of their functions are closely tied to the peculiar context of social media interactions, including the need to write quickly and within limited character counts, which often leads to word shortening (e.g., *rix* for *risposta* 'answer'; *sll* for *sullo/a* 'on'; *snx* for *sinistra* 'left'; *csx* for *centrosinistra* 'centre-left'). Similarly, in an effort to conceal potentially offensive or sensitive words, online users often resort to leetspeak to trick automatic censoring filters without altering the words' readability (e.g., *f4scist4* for *fascista* 'fascist', or *merd@* and *merxa* for *merda* 'shit'). However, there are cases of forms labelled as orthographic variation that serve other functions and reveal some interesting driving mechanisms for the creation of new words. An example is orthographic variation used as a joke (e.g., *gomblotto* for *complotto* 'conspiracy', *graduidamende* for *gratuitamente* 'free', *kultura* and *kompagni* for *cultura* 'culture' and *compagni* 'companions/comrades'), or for emphasis (e.g., *coolo* 'arse', *minkiate* 'bullshit (talk/things)', *pikkolo* 'small', *pazzeska* 'crazy'). In all of these cases, the replacement of one or more characters is capable of conveying nuances of meaning that the original spelling could not convey. In *gomblotto*, for instance, the initial *g* alludes to a regional pronunciation of the word; in *kompagni* and *kultura* the letter *k* replaces the *c* to allude to German spelling, and thus to the country's stereotypical authoritarian regime. Moreover, as voth *gomblotto* and *graduidamende* mimic the mispronunciation in spoken Italian of the correct form (be that out of ignorance or dialectal influence), their use moves beyond the confines of written language.

While on the subject of mispronunciation, orthographic variation is also used to mock the Italian pronunciation of foreign words, such as *biutiful* 'beautiful', *singol* 'single', *vairus* 'virus' or *vaucher* 'voucher', and, in a small number of cases, to convey sarcasm. In example (7), the orthographic variation of *scienza* 'science' with the final *-h* serves as a sarcastic expression of scepticism towards scientific advances.

(7) *Credete ciecamente nella **scienzah** anche contro l'evidenza.*
'You blindly believe in pseudoscience against all evidence'.

Univerbation is another productive category of lexical innovation involving the graphic representation of words. In this category, we include all sequences of two or more forms merged by *Twitter* users into a single word through blank space removal, e.g., *buonagiornata* 'goodday' and *ierisera* 'lastnight'. Univerbation has been integral to the evolution of the Italian language over the centuries, leading to the formation of new lexical items in common use today by joining two existing words together (e.g., *invece* 'instead', from the forms *in* and *vece*). Online conversations make frequent use of univerbated forms, partly for a need to economise on the number of characters, and partly owing to hashtags, which —when consisting of two or more words— are necessarily univerbated forms. However, a number of univerbation occurrences in our *Twitter* corpus serve, once more, as an emphatic device, as is the case for *eddaiii* (from *e dai*, 'come on'), *evvaiiiiii* (from *e vai*, 'go/yes'), *opperbacco* (from *o perbacco*, 'my goodness'), *stemmerde* (from *(que)ste merde*, literally 'these shits' to mean 'these arseholes'). The emphatic forms are often characterised by the syntactic doubling of the initial consonant of the second word (e.g., *massì* from *ma* + *sì* 'but yes of course', where the initial *s-* is duplicated).

Portmanteau words or blends (Micheli 2020) also constitute a category in our list of candidate neologisms. In this case, the emerging form is a word combining two or more existing words, as in *presiniente* from *presidente* and *niente* 'a nobody' (referred to a president), *intertristi* from *interisti* and *tristi* 'sad Inter (football club) supporters', or *nazipass* from *nazi* and *greenpass*. in our corpus, portmanteaus mostly relate to politics and are usually used as ironic wordplay (e.g., *lettamaio*, the fusion of politicians Enrico Letta's and Luigi Di Maio's surnames resembling the word *letamaio* 'pigsty'). Additionally, they differ from candidates categorised as compounds: while portmanteaus combine forms where at least one is part of a word (*presi* for *presidente*), compounds result from the juxtaposition of two full words, as is the case of *contapalle* 'fibber' in example (8):

(8) *Grazie è 1 pagliaccio infame **contapalle**, per quello fa ridere.*
'Thanks he's a hateful fibbing clown, that's why he's funny'.

Our list of new forms only includes four compounds (e.g., *fotocazzo* 'dick pic'). This is consistent with the general spread of compounds in Italian, which tends to favour

derivational rather than compositional morphological processes in the formation of new words (Micheli 2020). The ONLI, for instance, includes 430 neologisms obtained through compounding but more than 1,500 obtained through derivation.

Another category used to classify emerging forms is prefixation. Some of the words included in this category are parasynthetic, that is, they involve the addition of both a prefix and a suffix (Micheli 2020), e,g., *appecorato* from *ad-* + *pecora* 'sheep' + *-ato*, used to denote a servile person. We labelled these forms as 'prefixation' since the prefix semantically trumps the suffix (as is also the case for *iposcolarizzati* 'undereducated', where the *ipo-* prefix connotes the low level of education). *Autoregalo* in (9) is one of the eight forms in this category.

(9) *Beh un **autoregalo** per tirarmi un po' su il morale.*
'Well, a self-gift to cheer me up a little'.

In addition to being less common, prefixed forms are not as informal and are less tied to emphasis or irony: the words *biolaboratori* 'biolaboratories', *iposcolarizzati* and *pregirata* 'prerecorded', for instance, pertain to health, education and videomaking respectively, their prefixes used to form domain-specific lexical items rather than wordplay.

The seven forms labelled as 'transcategorisation' (*cuora*, *cuorare*, *cuoro*, *issima*, *issimo*, *panchinato* and *vaffanculi*) relate to three lemmas (*cuorare* 'heart', *panchinare* 'bench', and *vaffanculo* 'fuck you') and to the superlative suffix *-issimo*, used here as an actual word. The verb *cuorare* in example (10) is derived from the noun *cuore* 'heart' to mean 'like' or 'love' and is thus strictlly used in online conversation.

(10) *Non ti **cuoro**, perché non sono d'accordo.*
'I won't heart you because I don't agree'.

In line with the propensity of *Twitter* interactions to use emphatic and intensified forms, *issimo* in its word form occurrence can both strengthen a preceding superlative, as in example (11), or intensify a preceding adjective, as in example (12).

(11) *Ma come fa ad essere bellissimo **issimo issimo** pure vestito da Aladdin?*
'How can he be so so handsome even when dressed as Aladdin?'

(12) *Il prototipo della sinistra intelligente.... direi anche **issima**.*
'The prototype of the intelligent left.... extremely [intelligent], I would add'.

6.1. The role of hashtags

Hashtags are a form of social tagging that allows social media users to incorporate metadata in their posts (Zappavigna 2015). As such, hashtags are able to convey a range of meanings, for they are part of the linguistic structure of online texts whilst providing additional information about them. Owing to this aggregating role, the present study treats these 'super words' as a separate set of emerging lexical items. The total number of hashtags extracted from our *Twitter* corpus as emerging forms is 373, 75 per cent of which are loanwords (single and univerbated). As well as grouping them into the ONLI categories, we tagged hashtags according to their particular function. This has been described by Spina (2019) as either informative if they serve as topic-marker devices (e.g., *#spuntablu* 'blue tick' in example (13)), or as interpersonal/evaluative if they convey the subjective stance of the author (e.g., *#facciamorete* 'together' in example (14)).

(13) *Trovo incomprensibile la polemica per gli 8$ chiesti in cambio della* ***#spuntablu***.
'I really don't understand the controversy surrounding the $8 charge for a #bluetick'.

(14) *Lo diciamo da liberi e pensanti cittadini attivi!* ***#facciamorete****: tutti a votare, senza disperdere voti!*
'We say this as free and rational active citizens! #together: let's all go out and vote without wasting votes!'

The majority of emerging hashtags has an informative function (63%). They are mostly single (*#christmas*, *#olympics*) or univerbated English words (*#weddingday*, *#photooftheday*) used to tag topics. A few widespread acronyms can also be spotted, both from English (*#ootd* for *outfit of the day*) and Italian words (*#rdc* for *reddito di cittadinanza* 'universal basic income'). The rare informative one-word hashtags are compounds, built with the two productive forms *caro-* (*#carobenzina* 'increase in the cost of petrol', *#carobollette* 'increase in household bills' and *#caroenergia* 'increase in energy costs'), and *toto-* (*#totoministri* 'minister pools'). Among the informative and univerbated hashtags based on Italian words, *#allertameteo* 'weatherwarning', *#pausapranzo* 'lunchbreak', and *#biancoenero* 'blackandwhite' are particularly interesting, since they are not restricted to the social media sphere but are used in much more general contexts. Evaluative hashtags are those added to the tweet to comment on its content. They are therefore more creative, starting with their spelling. While we found no instances of orthographic variation in informative hashtags, for evaluative hashtags we

count four, all mostly conveying nuanced ironic meaning. Examples are *#spiaze* 'pity' (15) and *#povery* 'poor people' (16). The former literally means 'feel sorry', but it is used to ironically comment on an unpleasant situation; the original *-c* (*spiace*) becomes *-z* (*spiaze*) to graphically represent the northern pronunciation of a well-known Italian football celebrity from whom the irony originated.

(15) *SerieA: il Cagliari con 1 solo tiro in porta voleva vincerla;* ***#Spiaze****.*
'SerieA: Cagliari wanted to win it with 1 single goal kick; #Pity'.

Similarly, *#povery* (orthographic variation of *#poveri*) adds a touch of British snobbery to the meaning of 'poor':

(16) *Da quello che vedo è più ricco Zhang di voi* ***#povery****.*
'From what I can tell Zhang is richer than you #poorpeople'.

Among the univerbated evaluative hashtags, a number of forms emerge as exhortations (*#andratuttobene* 'everythingwillbealright'), greetings (*#buonagiornata* 'goodday') and interjections (*#buonavita* '[have a] goodlife').

6.2. Institutionalisation in Zingarelli

23 out of the total 347 emerging forms used in *Twitter* in 2022 have been included in *Lo Zingarelli 2024* (Zingarelli 2023), the monolingual dictionary of Italian published in 2023, which incorporates 250 new words and 750 new multi-word forms compared to the previous year's edition. *Lo Zingarelli 2024* can be considered the most up-to-date lexicographical collection of neologisms, partly because the dictionary releases a new edition every year with a section specifically dedicated to neologisms. The 22 forms shared between our candidate neologisms and the last edition of the dictionary (listed in Table 6, below the mid rule) are therefore those that have completed their process of neologisation, from their initial occasional appearance in specific contexts to their spreading in wider situations and, finally, their institutionalisation.

The institutionalised neologisms in our corpus are created through suffixation (12), adapted (4) and direct borrowing (4), prefixation (1), transcategorisation (1), and blending (1). No emerging form created through changes in spelling is accepted into the dictionary the year after its recurring appearance in *Twitter* conversations. This might suggest that orthographic variation is not regarded as a lexicographic criterion which is strong enough for institutionalisation, although the variability in their graphic form is the most common

source of lexical innovation in social media interactions. From a grammatical point of view, the majority of these forms are nouns (12), nouns and adjectives (4), verbs (4) and adjectives (3). It follows that, in the context of *Twitter*, a noun obtained through suffixation seems to be the most likely candidate for dictionary inclusion and, thus, institutionalisation.

Candidate	**Category**	**PoS**
Abilista	Sufixation	ADJ; NOM
Appecoronati	Prefixation	ADJ
Blastata	Loanword adaptation	NOM
Coglionazzo	Sufixation	ADJ; NOM
Condizionalità	Loanword adaptation	NOM
Docuserie	Portmanteau	NOM
Fail	Loanword	NOM
Fallout	Loanword	NOM
Falsona	Sufixation	ADJ; NOM
Fisicati	Sufixation	ADJ
Misunderstanding	Loanword	NOM
Paccare	Sufixation	VER
Paccotto	Sufixation	NOM
Panchinato	Transcategorisation	VER
Pigiamone	Sufixation	NOM
Pigiamoni	Sufixation	NOM
Pirlotto	Sufixation	ADJ
Posturologo	Sufixation	NOM
Rosiconi	Sufixation	ADJ; NOM
Soggettone	Suffixation	NOM
Soundbar	Loanword	NOM
Stalkero	Loanword adaptation	VER
Switchare	Loanword adaptation	VER
#breaking	Loanword	ADJ
#breakingnews	Loanword	N/A
#carobenzina	Compounding	NOM
#crossfit	Loanword	NOM
#genderfluid	Loanword	ADJ
#graphicdesign	Loanword	N/A
#greenwashing	Loanword	NOM
#mindfulness	Loanword	NOM
#omg	Acronym	INT
#reel	Loanword	NOM
#reels	Loanword	NOM
#street	Loanword	NOM
#totoministri	Compounding	NOM

Table 6: Hashtags and non-hashtag forms acknowledged in *Lo Zingarelli 2024* with their respective ONLI category of lexical creation and part(s)-of-speech

7. CONCLUSION

This exploratory study represents the most extensive investigation into lexical innovation in Italian *Twitter* yet. Our findings show that the emergence of new words in *Twitter* appears to be driven more by creativity, entertainment, and a desire for attention rather

than a necessity to introduce novel terms to describe new objects or events. Indeed, the 347 emerging forms mainly perform functions related to irony (*povery*, *presiniente*), intensification (*adorissimo*) and emphasis (*massi*). As has been consistently highlighted in previous studies on social media discourse (e.g., Zappavigna 2012; Spina 2019), the sense of belonging to a large (online) community significantly influences the generation and spread of new words. Some of these coined expressions have the potential of being adopted and reused not only in spoken discourse but also in online communication streams and, in a trans-medial perspective, by the media. The dynamics of their diffusion and a deeper investigation into their probability of becoming institutionalised neologisms could be the focus of future research.

The one-year time frame we adopted proves effective for the detection of emerging usage patterns in the dynamic context of *Twitter*, where linguistic phenomena surface and disseminate rapidly, supporting us in our goal to explore the initial emergence of (novel) words. Nonetheless, it may not capture forms that spread more slowly, maintaining a consistent but slower rate of propagation.

Follow-up work will extend the analysis to additional timelines but, owing to the lately takeover of *Twitter*, which has significantly undermined its value for academic research, will likely have to be redirected to other openly accessible micro-blogging platforms, such as *BlueSky*,[14] or *YouTube* (comments).[15] Furthermore, we will investigate the geographical distribution of emerging forms and hashtags with the aim of identifying regional patterns of lexical creation across Italy. Finally, we will leverage our annotated data to explore how the outcomes of the two methods adopted differ when adjusting threshold choices, aiming to identify optimal points as practical guidelines for future research.

REFERENCES

Adamo, Giovanni and Valeria Della Valle. 2003. *Neologismi Quotidiani. Un Dizionario a Cavallo del Millennio*. Firenze: Leo S. Olschki.

Adamo, Giovanni and Valeria Della Valle. 2006. *Che Fine Fanno i Neologismi? A Cento Anni dalla Pubblicazione del Dizionario Moderno di Alfredo Panzini*. Firenze: Leo S. Olschki.

[14] https://bsky.app/
[15] https://www.youtube.com

Adamo, Giovanni and Valeria Della Valle. 2008. *Le Parole del Lessico Italiano*. Roma: Carocci.

Adamo, Giovanni and Valeria Della Valle. 2017. *Che Cos'è un Neologismo*. Roma: Carocci.

Adamo, Giovanni and Valeria Della Valle. 2019. *Osservatorio Neologico della Lingua Italiana: Lessico Parole Nuove Dell'italiano*. Roma: ILIESI Digitale.

Algeo, John ed. 1991. *Fifty Years Among the New Words. A Dictionary of Neologisms, 1941–1991*. Cambridge: Cambridge University Press.

Alvar Ezquerra, Manuel. 2003. *Nuevo diccionario de voces de uso actual*. Madrid: Arco Libros.

Amar, Yvan. 2010. *Les Mots de L'actualité*. Paris: Éditions Belin.

Anthony, Laurence. 2022. *AntConc (Version 4.2.0)* [Computer software]. https://www.laurenceanthony.net/software.

Bencini, Aandrea and Beatrice Manetti. 2005. *Le Parole Dell'Italia che Cambia*. Grassina: Le Monnier Università.

Brasolin, Paolo. 2023. *Breviloquia Italica: Data Pipeline* (Version 1.1.1) [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10010427

Brasolin, Paolo, Greta H. Franzini and Stefania Spina. 2023. "Ti blocco perché sei un trollazzo": Lexical innovation in contemporary Italian in a large Twitter corpus. In Federico Boschetti, Gianluca E. Lebani, Bernardo Magnini and Nicole Novielli eds. *Proceedings of the Ninth Italian Conference on Computational Linguistics.* Venice: CEUR-WS. https://ceur-ws.org/Vol-3596/paper12.pdf

Croft, William. 2000. *Explaining Language Change: An Evolutionary Approach*. Harlow: Pearson Education.

Crystal, David. 1997. *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

De Mauro, Tullio. 2006. *Dizionarietto di Parole del Futuro*. Roma: Editori Laterza.

Des Isnards, Alexandre. 2014. *Dictionnaire du nouveau Français*. Paris : Allary Éditions.

Eisenstein, Jacob, Brendan O'Connor, Noah A. Smith and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLoS ONE* 9/11: e113114. https://doi.org/10.1371/journal.pone.0113114

Fischer, Roswitha. 1998. *Lexical Change in Present-day English: A Corpus-based Study of the Motivation, Institutionalization, and Productivity of Creative Neologisms*. Tübingen: Gunter Narr Verlag.

Franzini, Greta H., Stefania Spina and Paolo Brasolin. 2023. *Breviloquia Italica: Annotations (Version 1.0.1)* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10010528

Frenguelli, Gianluca. 2008. Come si studiano le parole nuove. In Maurizio Dardano and Gianluca Frenguelli eds. *L'Italiano di Oggi. Fenomeni, Problemi, Prospettive*. Roma: Aracne, 99–120.

Gazzardi, Antonella and Camilla Vásquez. 2020. A taxonomic approach to the use of English in the Italian media. *World Englishes* 41: 1–14.

Giraud, Jean, Pierre Pamart and Jean Riverain. 1971. *Les Mots dans le Vent*. Paris : Larousse.

Grandi, Nicola. 2017. Intensification processes in Italian: A survey. In Maria Napoli and Miriam Ravetto eds. *Exploring Intensification: Synchronic, Diachronic and Cross-Linguistic Perspectives*. Amsterdam: John Benjamins, 55–77.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2016. Analyzing lexical emergence in modern American English online. *English Language and Linguistics* 21/1: 99–127.

Grieve, Jack, Andrea Nini and Diansheng Guo. 2018. Mapping lexical innovation on American social media. *Journal of English Linguistics* 46/4: 293–319.

Grossmann, Maria and Franz Rainer. 2004. *La Formazione delle Parole in Italiano*. Tübingen: Max Niemeyer Verlag.

Guilbert, Louis. 1975. *La Créativité Lexicale*. Paris: Larousse.

Hovy, Dirk, Afshin Rahimi, Timothy Baldwin and Julian Brooke. 2019. Visualizing regional language variation across Europe on Twitter. In Stanley D. Brunn and Roland Kehrein eds. *Handbook of the Changing World Language Map*. Cham: Springer, 3719–3742.

Iacobini, Claudio and Anna M. Thornton. 1992. Tendenze nella formazione delle parole nell'italiano del ventesimo secolo. In Bruno Moretti, Dario Petrini and Sandro Bianconi eds. *Linee di Tendenza Dell'italiano Contemporaneo. Atti del XXV Congresso Internazionale della Società di Linguistica Italiana*. Roma: Bulzoni, 25–55.

Kerremans, Daphné. 2015. *A Web of New Words*. Bern: Peter Lang.

Kershaw, Daniel, Matthew Rowe and Patrick Stacey. 2016. Towards modelling language innovation acceptance in online social networks. In Paul N. Bennet ed. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. New York: ACM, 553–562.

Labov, William. 2001. *Principles of Linguistic Change*. Malden: Wiley-Blackwell.

Laitinen, Mikko, Masoud Fatemi and Jonas Lundberg. 2020. Size matters: Digital social networks and language change. *Frontiers in Artificial Intelligence* 3. https://doi.org/10.3389/frai.2020.00046

Lo Duca, Maria G. 1992. "Parole nuove," regole e produttività. In Bruno Moretti, Dario Petrini and Sandro Bianconi eds. *Linee di Tendenza Dell'italiano Contemporaneo. Atti del XXV Congresso Internazionale della Società di Linguistica Italiana*. Roma: Bulzoni, 57–81.

Lurati, Ottavio. 1990. *3000 Parole Nuove: La Neologia Negli Anni 1980–1990*. Bologna: Zanichelli.

Marello, Carla. 2020. New words and new forms of linguistic purism in the 21st century: The Italian debate. *International Journal of Lexicography* 33: 168–186.

Marri, Fabio. 2006. Parole nuove, meno nuove, troppo nuove (I). *Lingua Nostra* 57/3–4: 113–122.

Marri, Fabio. 2018. I neologismi dentro e fuori dei repertori recenti. *Quaderns d'Italià* 23: 11–26.

Martí Antonín, María A. 1998. *Diccionario de Neologismos de la Lengua Española*. Barcelona: Larousse.

Maxwell, Kerry. 2006. *From Al desko to Zorbing. New Words for the 21st Century*. London: Macmillan.

Micheli, M. Silvia. 2020. *La Formazione delle Parole. Italiano e altre Lingue*. Roma: Carocci editore.

Migliorini, Bruno. 1963. *Parole Nuove: Appendice di Dodicimila Voci al "Dizionario Moderno" di Alfredo Panzini*. Milano: U. Hoepli.

Moliner, María. 2013. *Neologismos del Español Actual*. Madrid: Gredos.

Nguyen, Dong, A. Seza Doğruöz, Carolyn P. Rosé and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42/3: 537–593.

Osservatorio Neologico della Lingua Italiana (ONLI). 2012. *Parole Nuove dai Giornal*. https://www.iliesi.cnr.it/ONLI/BD.php.

Pulcini, Virgina, Cristiano Furiassi and Félix Rodríguez González. 2012. The Lexical influence of English on European languages: From words to phraseology. In

Cristiano Furiasi, Virginia Pulcini and Félix Rodríguez González eds. *The Anglicization of European Lexis*. Amsterdam: John Benjamins, 1–24.
Rodríguez Arrizabalaga, Beatriz. 2021. Social networks: A source of lexical innovation and creativity in contemporary peninsular Spanish. *Languages* 6/3: 138. https://doi.org/10.3390/languages6030138
Scotti Morgana, Silvia. 1981. *Le Parole Nuove*. Bologna: Zanichelli.
Spina, Stefania. 2014. Il Perugia Corpus: Una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In Roberto Basili, Alessandro Lenci and Bernardo Magnini eds. *Proceedings of the First Italian Conference on Computational Linguistics*. Pisa: Pisa University Press: 354–359.
Spina, Stefania. 2016. Le conversazioni scritte dei social media: Un'analisi multidimensionale. In Francesca Bianchi and Paola Leone eds. *Linguaggio e Apprendimento Linguistico: Metodi e Strumenti Tecnologici*. Milano: Associazione Italiana di Linguistica Applicata, 83–102.
Spina, Stefania. 2019. *Fiumi di Parole. Discorso e Grammatica delle Conversazioni Scritte in Twitter*. Canterano: Aracne editrice.
Tarrade, Louise, Magué, Jean-Philippe and Jean-Pierre Chevrot. 2022. Detecting and categorising lexical innovations in a corpus of tweets. *Psychology of Language and Communication* 26/1: 313–329.
The Unicode Consortium. 2022. *The Unicode Standard* (Version 15.0.0). Unicode Consortium. https://www.unicode.org/versions/Unicode15.0.0/
Tulloch, Sara. 1991. *The Oxford Dictionary of New Words. A Popular Guide to Words in the News*. Oxford: Oxford University Press.
Verardi, Giuseppe Marco. 1995. *Le Parole Veloci. Neologia e Mass Media Negli Anni 90*. Locarno: Armando Dadò.
Würschinger, Quirin. 2021. Social networks of lexical innovation: Investigating the social dynamics of diffusion of neologisms on Twitter. *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.648583
Zappavigna, Michele. 2012. *Discourse of Twitter and Social Media. How We Use Language to Create Affiliation on the Web*. London: Continuum.
Zappavigna, Michele. 2015. Searchable talk: The linguistic functions of hashtags. *Social Semiotics* 25/3: 274–291.
Zingarelli, Nicola. 2023. *Lo Zingarelli 2024: Vocabolario della Lingua Italiana*. Bologna: Zanichelli.
Zgusta, Ladislav. 1971. *Manual of Lexicography*. The Hague: Mouton De Gruyter.
Zolli, Paolo. 1989. *Come Nascono le Parole Italiane*. Milano: Rizzoli.

Corresponding author
Stefania Spina
University for Foreigners of Perugia
Department of Italian Language, Literature and Art in the World
Piazza Fortebraccio, 4
06123 Perugia
Italy
E-mail: stefania.spina@unistrapg.it

received: November 2023
accepted: July 2024

APPENDIX A: CHOROPLETH MAPS

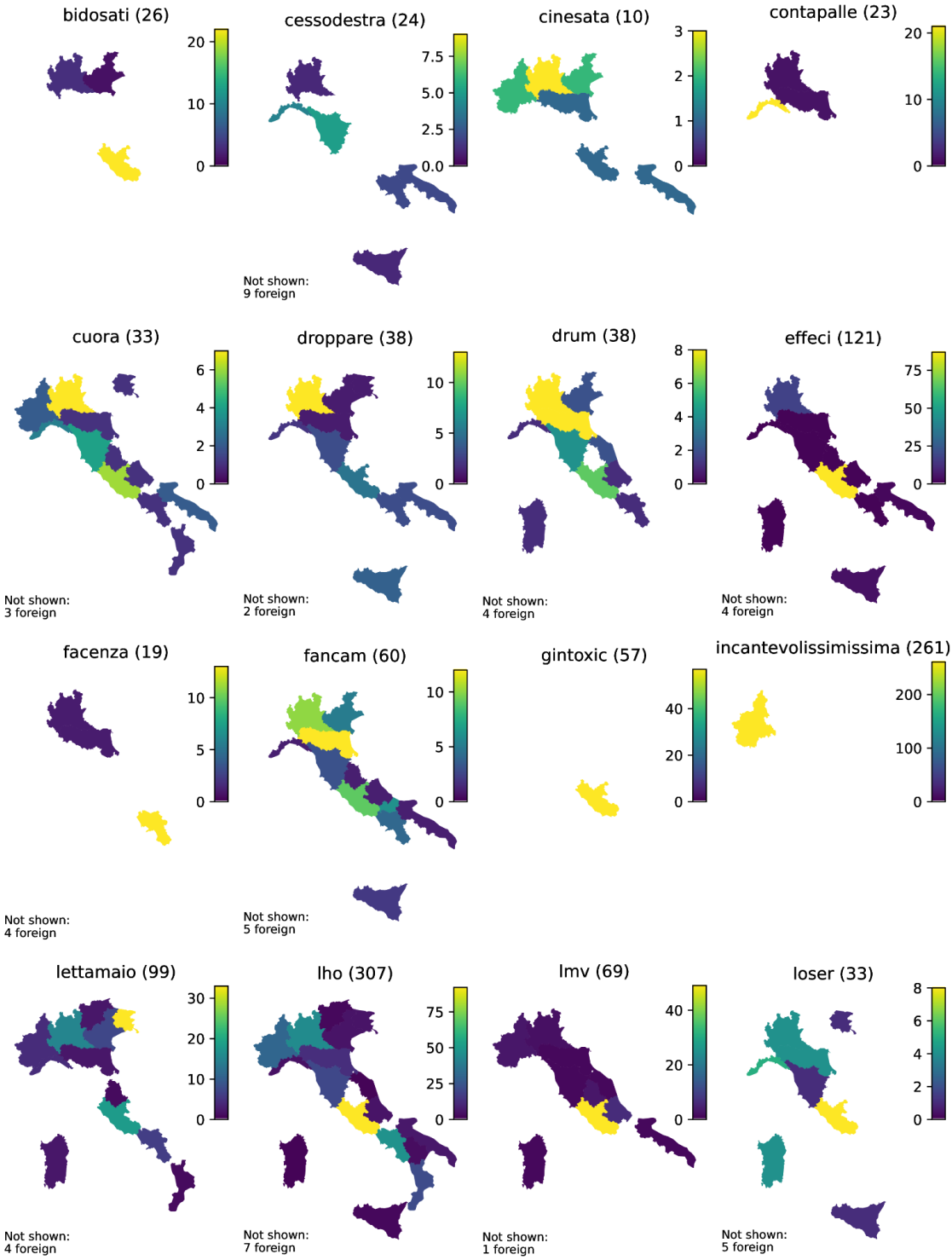

Figure 2: Choropleth maps of candidate neologisms from A to L. The colour scale represents instances per million tokens at the regional level. Total occurrences in Italy are provided with the titles. Occurrences outside Italy are not shown and counted in the legends.

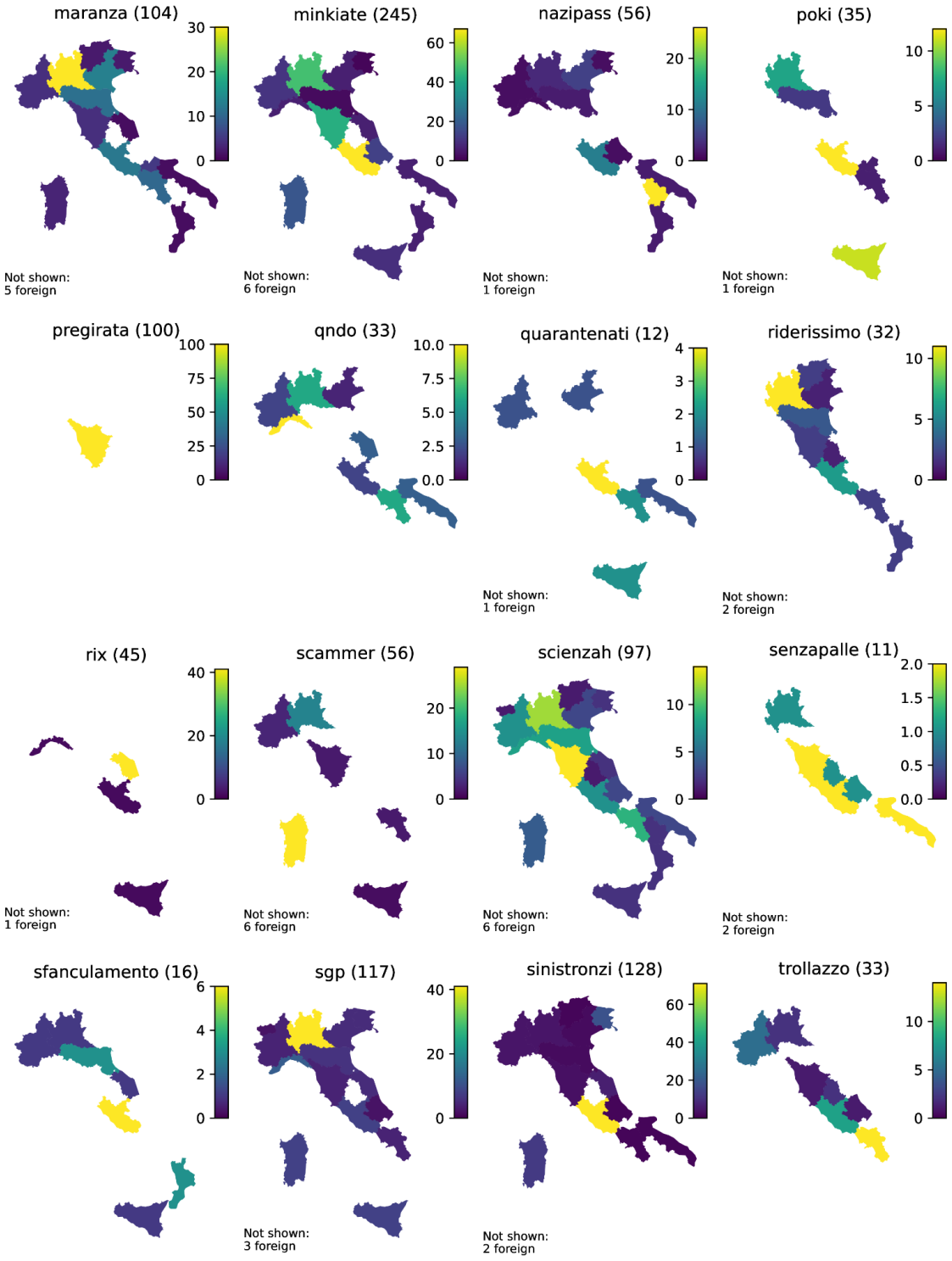

Figure 3: Choropleth maps of selected candidate neologisms from M to Z. The colour scale represents instances per million tokens at the regional level. Total occurrences in Italy are provided with the titles. Occurrences outside Italy are not shown and counted in the legends

APPENDIX B: FULL LIST OF EMERGING FORMS

B.1. Non-hashtag forms by category

Orthographic variation (111): *5s, accaunt, adovo, affan, amerika, amiketti, amio, amïo, ancielo, anzia, assaj, azzzzz, babbà, benza, biutiful, c4zz0, c@@@o, 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.*

Suffixation (60): *abilista, accannate, accannato, adorissimo, amorina, baguettari, benissimamente, busoni, cazzarone, cazzaroni, ciacchera, cialtronismo, cinesata, cinesate, coglionazzo, ducessa, eurini, falsona, fattoni, fisicati, garone, godicchio, gretini, impiattamento, incantevolissimissima, legaiolo, mandrakata, memiamo, paccare, paccotto, patati, patatino, personaggione, piagnina, piddini, pigiamone, pigiamoni, pirlotto, pisellate, posturologo, poverata, presidenta, prezzemolina, prosciutteria, quarantenati, riderissimo, ridolini, rosiconi, senzadubbiamente, sfanculamento, sierare, sierata, soggettone, tridosato, triplodosati, tuitteri, twettini, twitteri, zanzarologi, zanzarologo.*

Univerbation (48): *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', miocuggino, miraccomando, ncazzo, nculo, noeuro, nowar, opperbacco, porcaputtana, porcodd, senzapalle, serietv, sottocasa, stemmerde, stica, streetart, terzopolo, tuttappost, ziocane.*

Loanword (39): *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.*

Portmanteau (33): *5scemi, 5stalle, assurdistan, deltacron, docuserie, estaters, fasciocomunista, fascioleghista, fascioleghisti, flurona, gintoxic, giornalanza, grillioti, grillopiddini, grillopitechi, intertristi, inverners, lettamaio, nazipass, naziucraini, pdiota, pdioti, piddiota, piddioti, pidiota, pidioti, presiniente, putler, renziota, renzioti, scansuolo, sinistronzi, tecnopolo.*

Loanword adaptation (24): *blastata, blessata, boyz, broder, condizionalità, cringiata, droppare, eppi, flex, flexo, followo, ghosta, matcha, pullato, schip, squirtare, stalkero, switchare, trollata, trollazzo, trolling, trollini, twerka, twitterino.*

Prefixation (8): *appecorato, appecoronati, autoregalo, bidosati, biolaboratori, intrasezioni, iposcolarizzati, pregirata.*

Transcategorisation (7): *cuora, cuorare, cuoro, issima, issimo, panchinato, vaffanculi.*

Acronym (6): *afc, lms, lmv, rdc, sgp, vfc.*

Compounding (4): *cessodestra, contapalle, fotocazzo, fregacazzi.*

Deonymic derivation (3): *cippalippa, drum, lippa*

Redefinition (2): *giornalaia, maranza.*

Acronymic derivation (1): *effeci.*

Tmesis (1): *facenza.*

B.2. Emerging hashtag forms by category

Loanword (279): *#actor, #adoptdontshop, #adventure, #airport, #amazing, #aperitif, #archaeology, #artist, #artistic, #artwork, #attitude, #autumn, #autumnvibes, #award, #awards, #babyboy, #baroque, #beard, #behappy, #bestfriends, #bicycle, #biodiversity, #birds, #black, #blackandwhite, #booklover, #breaking, #breakingnews, #budgetcap, #burger, #butterfly, #cancer, #cathedral, #catlife, #catlover, #chess, #chill, #circulareconomy, #cityscape, #climate, #climateaction, #climatechange, #clubbing, #coffeelover, #colorful, #colour, #colours, #comedy, #communication, #couple, #cousins, #creativity, #crossfit, #cryptocurrency, #culturalheritage, #curvy, #cycling, #dad, #dancers, #daughter, #dawn, #daytime, #devotion, #digitalart, #dinnertime, #documentary, #doglover, #drama, #dress, #dusk, #earth, #earthquake, #ebike, #elegance, #euphoria, #fail, #fairplay, #fall, #familyfirst, #fashionstyle, #finance,*

#followme, #followme (unicode homograph of the previous entry), #foryou, #freetime, #fridayvibes, #fuck, #fuckcancer, #gameday, #genderfluid, #getoutthere, #glasses, #goalkeeper, #goat, #gold, #goodevening, #goodtimes, #graphicdesign, #grateful, #gratitude, #greenwashing, #gymlife, #hair, #hairstyle, #happybday, #happyholidays, #happyness, #hat, #health, #heart, #holiday, #homedecor, #homedesign, #hospitality, #icecream, #ink, #innovation, #instore, #interior, #interiordesign, #interview, #investing, #investment, #iphonography, #italiansdoitbetter, #journalism, #journey, #joy, #kids, #landscapes, #life, #lighting, #lights, #likeforlikes, #lunchtime, #luxury, #macteanimo, #marathon, #medieval, #meditation, #menstyle, #mentalhealth, #midnights, #migrants, #mindfulness, #mirror, #mondaymood, #monochrome, #monument, #musiclover, #naturalbeauty, #naturelovers, #newbook, #newcollection, #newlife, #newlook, #nextgen, #nightlife, #nomask, #noracism, #novax, #nowar, #nowars, #nowplaying, #nowwatching, #oldschool, #olympics, #onelove, #onfire, #partytime, #peaceandlove, #peacenotwar, #philosophy, #photoart, #photograghy, #photographer, #photooftheday, #picoftheday, #pictures, #pizzatime, #pontifex, #portrait, #portraits, #positivevibes, #prayforpeace, #president, #pricecap, #production, #proud, #quality, #quoteoftheday, #quotes, #rain, #raw, #recording, #reel, #reels, #relaxing, #remember, #renaissance, #rescue, #respect, #roadtrip, #roses, #sad, #sand, #saturdayvibes, #savetheplanet, #seafood, #seascape, #see, #shadows, #shame, #ship, #shoes, #shoot, #singer, #sisters, #slavaukraini, #slavaukrainii, #slavaukraïni, #song, #songs, #songwriter, #space, #specialguest, #spring, #springtime, #steak, #stopwar, #street, #summercamp, #sunglasses, #supergreenpass, #tatoo, #tattooart, #theater, #thebadguy, #thoughts, #throwbackthursday, #tourism, #town, #trail, #trailrunning, #travel, #travelgram, #traveller, #travelling, #tree, #trees, #tuesdayvibe, #tuscanygram, #vacation, #vanlife, #vibes, #vintagestyle, #viral, #voice, #volcano, #waiting, #wakeup, #walking, #wall, #wanderlust, #war, #waterfall, #waves, #weather, #webmarketing, #weddingday, #whatelse, #wildlife, #win, #window, #wine, #winetime, #winteriscoming, #woman, #women.

Univerbation (50): *#accaddeoggi, #allertameteo, #amoremio, #andratuttobene, #aperitivotime, #avantitutta, #avantiunaltro, #bellavita, #biancoenero, #buonacena, #buonagiornata, #buonappetito, #buonascuola, #buonaserata, #buonavita, #buonefeste, #buonenotizie, #buonevacanze, #buonlavoro, #buononomastico, #buonpranzo, #casadolcecasa, #cessateilfuoco, #ciaociao, #dallapartegiusta, #dalleparoleaifatti,*

#facciamorete, #governodegliorrori, #governodeimigliori, #governodeipeggiori, #governodellavergogna, #governodipagliacci, #grandebellezza, #grazieatutti, #idearegalo, #iomivaccino, #ionondimentico, #iononmollo, #maimollare, #neiperte, #nonato, #nopos, #oggicosi, #pausapranzo, #perte, #qrcode, #romanzoquirinale, #spuntablu, #sulserio, #unovaleuno.

Portmanteau (21): *#bookstagram, #catstagram, #caturday, #chilhavister, #fantacitorio, #farsopoli, #foodstagram, #instaart, #instabook, #instacat, #instadog, #instagood, #instamoment, #instamood, #instaphoto, #instapic, #instatravel, #lettamaio, #pfizergate, #sapevatelo, #sivax.*

Acronym (13): *#bnw, #fyp, #ia, #ig, #mma, #omg, #ootd, #otnba, #pdr, #rdc, #tb, #tbt, #wwiii.*

Compounding (5): *#carobenzina, #carobollette, #caroenergia, #cinesalvini, #totoministri.*

Orthographic variation (4): *#anala, #chesucc3de, #povery, #spiaze.*

Prefixation (1): *#extraprofitti.*