

Pavlo Morozenko

Vinnytsia National Technical University, Vinnytsia, Ukraine

LEXICAL INNOVATION IN ENGLISH DIGITAL DISCOURSE: A REPRODUCIBLE CORPUS-BASED ANALYTICAL WORKFLOW

Abstract. Digital communication provides a productive environment for new English lexical forms, but a conspicuous online word is not necessarily a stable neologism. This paper develops a reproducible philological workflow for examining lexical innovation in public digital discourse. It integrates word-formation theory, computer-mediated discourse analysis and corpus methodology to distinguish invention, local circulation, diffusion and conventionalization. The procedure specifies sampling boundaries, candidate extraction, morphological annotation, contextual reading and ethical handling of platform content. Constructed examples show why blends, abbreviations and semantic extensions require different interpretive tests and why frequency alone cannot establish lexical institutionalization. The work is methodological rather than a report of newly harvested social-media data or measured neologism rates. It supplies an auditable approach for studying linguistic creativity while accounting for platform affordances, unstable contexts and changing communicative communities.

Keywords: neologisms; digital discourse; corpus linguistics; word formation; lexical innovation; methodology.

Introduction

English digital discourse is often discussed as an unusually fast source of new words. Online participants combine familiar roots, abbreviate recurring phrases, extend existing meanings and recycle expressions from entertainment, gaming, professional communication and interpersonal humour. Yet attention to novelty should not obscure a central lexicological distinction: the first appearance of a form is not the same as its repeated use, and repeated use within one platform is not necessarily evidence that it has entered broader English vocabulary. A transparent study needs to explain what kind of lexical change it is able to observe.

Crystal (2006) describes the diversity of linguistic practices associated with internet communication and challenges the assumption that online language is a homogeneous variety. Zappavigna (2012) studies how social-media language creates affiliation and searchable connections. Mattiello (2013) demonstrates the importance of blending, shortening and other extra-grammatical patterns for understanding English word formation. These perspectives suggest that a new form must be understood simultaneously as a morphological object, a contextual utterance and an element in a network of social practices. This paper offers a methodological synthesis of those dimensions.

The aim is to formulate a reproducible workflow for studying candidate neologisms in English-language digital texts. All examples below are constructed to illustrate coding decisions and are not offered as collected platform evidence. No corpus was downloaded, counted or analysed for this paper. The proposed contribution is a set of steps with which a future empirical project could distinguish lexical creativity from durable lexical uptake while respecting the limits of publicly accessible online data.

The problem of defining a neologism

The label 'new word' can refer to a previously unattested string, a fresh sense of an established word or an older expression newly prominent within a community. These situations should not be collapsed. A morphological innovation may be rapidly interpretable through analogy; a semantic innovation may preserve the original spelling

while changing usage conditions. Some forms remain local in a gaming group or fandom, whereas others cross genre and platform boundaries. The proposed first step is therefore to state which definition of novelty is being used and which reference materials or earlier corpora are searched for comparison.

A candidate's orthographic form can vary: capitalization, hashtags, hyphens and repeated letters may all affect the appearance of a token. Automated extraction must distinguish these variations from different lexical units. At the same time, normalizing everything to lowercase may erase meaningful differences between a proper name and a common noun. A documented procedure should preserve the original form and add a separate normalized field, rather than silently rewriting the source material. Such decisions are part of philological interpretation, not merely technical housekeeping.

Mattiello's (2013) analysis of extra-grammatical morphology supports attention to blending, clipping, acronym formation and playful reduplication. However, word-formation labels are not always mutually exclusive. A newly coined expression can simultaneously depend on analogy, abbreviation and stylistic reanalysis. When classification is uncertain, a codebook should record more than one plausible segmentation and an explicit ambiguity marker. Certainty about morphology must not be manufactured simply to fill every column in a spreadsheet.

Building a defensible digital corpus

Corpus design begins by delimiting platforms, public genres, sampling dates and language varieties. A dataset consisting only of viral posts cannot represent routine digital communication. Likewise, a sample of searchable hashtags privileges language designed for visibility and may under-represent conversation that is not organized by tags. McEnery and Hardie (2012) stress that corpus evidence depends on design decisions; the representativeness of a dataset must always be defined relative to its research purpose. Researchers should record why particular sites, communities and time windows were chosen.

Ethical requirements are equally central. The fact that a message is technically visible does not automatically make extensive quotation or redistribution appropriate. Platform rules, expectations of privacy, potential re-identification and the presence of minors must shape collection and publication decisions. A corpus can preserve analytic features such as date range, platform genre and morphological context while minimizing personally identifying details. Quotations should be necessary for linguistic explanation rather than included merely because they are memorable. These precautions are especially important when an expression originates in a small online community.

A minimal reproducibility package would include sampling criteria, extraction scripts or search expressions, normalization rules, exclusion criteria and an anonymized annotation schema. Where source data cannot be redistributed, the research should explain the restriction and provide the procedure and aggregate descriptions that are legally and ethically shareable. Reproducibility does not require violating participant expectations; it requires communicating how evidence was selected, processed and interpreted within stated limits.

Identifying and classifying lexical candidates

The workflow can be organized into six stages. First, compile a temporally bounded sample and preserve metadata needed for contextual interpretation. Second,

create a candidate list through a documented combination of automatic comparison against a reference vocabulary and close reading. Third, remove technical strings, accidental misspellings and repeated reposts according to recorded rules. Fourth, annotate candidate morphology and possible sources. Fifth, inspect concordance lines to identify meaning, stance and discourse function. Sixth, compare distribution across time, genres or communities before making any claim about diffusion.

Consider an invented digital form such as 'scrolltired', intended here to describe fatigue associated with extended feed browsing. Its apparent meaning can be inferred from its components, but a single occurrence would establish only creative use in one context. To argue that the expression is spreading, a study would need independent attestations across users and settings. To claim institutionalization, it would need stronger evidence that the form is recognizable and relatively stable beyond a temporary exchange. The example illustrates the logic of the method; it is not a claim that this form is newly coined or absent from actual dictionaries.

The same logic applies to semantic change. A familiar word used ironically may signal a local evaluative meaning, but irony in one sentence does not establish a new dictionary sense. Concordance reading should separate literal, quoted, metalinguistic and innovative uses. Zappavigna's (2012) account of searchable affiliation makes this especially relevant to hashtags, whose communicative function can be categorization, stance-taking or community alignment as well as ordinary lexical reference. Treating every hashtag as a lexical innovation would inflate estimates and confuse pragmatic function with word-formation status.

Frequency, dispersion and contextual evidence

Raw frequency is an inadequate measure of lexical establishment. A form may appear thousands of times because a single original post was copied, while a less frequent expression may be distributed across independent authors and genres. Analysts should therefore report document frequency, temporal persistence and dispersion alongside token counts. They must also define the observation window: a temporary spike after a news event has a different interpretation from recurrent use over many months. Gries (2009) supplies general statistical principles for transparent linguistic analysis, although any chosen measure must be matched to the research question.

Concordance tools can reveal recurring grammatical frames and collocations; corpus-analysis environments such as Sketch Engine provide ways to inspect such patterns (Kilgarriff et al., 2014). Automated summaries nevertheless require careful review when data contain sarcasm, quotation, code-switching or platform-specific formatting. A candidate may occur in syntactic slots typical of nouns, verbs or adjectives, and those distributional clues help describe possible conventionalization. They do not settle the social status of the expression without consideration of speakers and contexts.

A well-designed analysis consequently combines numerical and interpretive evidence. Quantitative measures answer how often and how widely forms appear under stated sampling rules. Qualitative reading asks what the words do for participants: express solidarity, label an experience, mock an institution or reframe an established concept. The two modes support one another. An unexamined wordlist misses meanings, while memorable screenshots provide a poor foundation for generalization.

Boundary cases: memes, hashtags and recycled words

Lexical innovation becomes especially hard to identify when expressions function simultaneously as words and as signals of social alignment. A phrase repeated with minimal changes can become recognizable in a particular community without developing a stable lexical meaning outside that setting. Its success may depend on a visual template, shared event or recognizable speaker. If an analyst strips away those contexts and extracts only a token string, the item may appear more independent than it actually is. For this reason, annotation should record whether a candidate operates as a stand-alone lexical unit, an allusive quotation or part of a repeatable multimodal pattern.

Hashtags reveal this issue particularly clearly. Some connect multiple posts on a topic, some encode evaluative positions, and others become conventional written expressions outside their original tagging function. Zappavigna (2012) explains why searchable discourse can organize affiliation. A study that treats every new hashtag as a word would confuse metadata practice with a lexical category. A stronger test asks whether the unit can be used without the tagging context, how speakers inflect or modify it, and whether independent users apply it to new situations. These tests do not provide an infallible boundary, but they help separate communicative novelty from enduring changes in the lexicon.

Recycled expressions create the reverse problem: the visible word may be old while its online use is new. An ordinary adjective can acquire an ironic community-specific evaluation, and an established noun can become a verb in a limited technical setting. Lexical novelty thus requires a reference point in time, sense and speech community. Corpus comparisons should retain sufficient context to distinguish a new function from an unusual but conventional occurrence. A dictionary's absence of a sense may provide a useful clue but not decisive proof of innovation, especially when dictionaries vary in coverage and update frequency.

Validation, reporting and lexicographic value

For a future study, the codebook should begin with operational rather than impressionistic rules. A candidate could be marked as formally novel when it is not present in designated reference resources during a recorded search, semantically innovative when a contextual meaning differs from the established sense inventory, and socially diffused when attestations cross a prespecified number of independent contexts. Any numerical threshold must be justified and reported as an analytic convention, not as a natural law of language. It is also necessary to record when a candidate cannot be evaluated because archives are incomplete or relevant conversations are inaccessible.

Human review remains necessary even when automatic tools suggest candidate strings. A concordance line can show that a form occurs several times, but only contextual reading can explain whether users are quoting, rejecting, parodying or productively extending it. Independent annotation of a sample can reveal systematic disagreements about sense boundaries and morphological structure. Instead of resolving every disagreement through majority vote, a project can preserve an uncertainty category and publish the conditions that created it. Gries (2009) provides a methodological reminder that statistical summaries are meaningful only after the relevant constructs and units have been defined consistently.

Publication of findings should include both productive patterns and negative evidence. If a candidate receives intensive attention for one week but disappears from subsequent samples, its trajectory may be more informative than a list of apparently successful inventions. Conversely, a form with modest frequency across independent writers may deserve closer examination than a rapidly copied meme. The analyst should avoid interpreting disappearance from the sampled data as definitive extinction: users may have migrated to other platforms, adopted alternative spellings or restricted the expression to private contexts. Conclusions must remain proportionate to what the sample actually covers.

The workflow has implications for lexicography and language education. A descriptive dictionary can benefit from contextual examples that document usage, grammatical distribution and time-sensitive meaning without treating online popularity as sufficient for entry. In teaching, students can compare the morphology of candidate blends and shortenings while debating which evidence would demonstrate wider uptake. Such activities reveal lexical categories as scholarly judgments supported by sources and methods. Their purpose is not to predict which fashionable expressions will survive, but to teach how words acquire functions through repeated social use and why transparent observation is preferable to confident speculation.

Interpretation, limitations and conclusions

A major analytical hazard is treating all digital English as one speech community. Different platforms reward different forms of visibility, editing and interaction, and language practices vary across age, expertise and social network. A form popular in a narrowly specialized group may be functionally stable there without becoming common in broader public discourse. It is therefore better to speak of degrees and domains of circulation than to force a binary distinction between 'slang' and 'standard English'. Such caution is part of a descriptively responsible lexicology.

The proposed workflow has not been empirically evaluated in this paper, and it cannot establish present-day neologism rankings or diffusion rates. A next step would be a small, ethically collected and carefully stratified corpus with independent coding of morphology and sense. Longitudinal comparison could then test whether a candidate persists beyond an initial event, and human reader judgments could explore recognizability. Any observed change would need to be reported with full attention to missing data and differences in platform access.

Lexical innovation in English digital discourse is best studied as a process rather than an isolated novel word. The proposed sequence—sampling, extraction, normalization, morphological coding, contextual interpretation and analysis of dispersion—creates a path from suggestive examples to verifiable evidence. It preserves the distinction between coinage, circulation and conventionalization and makes ethical and methodological choices visible. For philology, this approach connects fine-grained interpretation of language use with research practices robust enough to be checked and extended.

REFERENCES

1. Crystal, D. (2006). *Language and the internet* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9780511487002>.

2. Gries, S. T. (2009). *Statistics for linguistics with R: A practical introduction*. De Gruyter Mouton. <https://doi.org/10.1515/9783110216042>.
3. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1, 7–36.
4. Mattiello, E. (2013). *Extra-grammatical morphology in English: Abbreviations, blends, reduplicatives, and related phenomena*. De Gruyter Mouton. <https://doi.org/10.1515/9783110295399>.
5. McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511981395>.
6. Zappavigna, M. (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. Continuum.