A Corpus of (Annotated) Algospeak Sentences: Challenges and Opportunities

Fahad Khan
CNR-ILC,
fahad.khan@ilc.cnr.it

Ana Ostroški Anić, Ivana Brač
Institute for the Croatian Language
{aostrosk,ibrac}@ihjj.hr

Francesca Frontini
CNR-ILC,
francesca.frontini@ilc.cnr.it

Elisa Squadrito
University of Macerata,
e.squadrito@unimc.it

Abstract

This paper presents an extension of previous work on the documentation of algospeak, a form of linguistic self-censorship used by social media users to evade automated content moderation. We introduce a new multilingual resource, consisting of a corpus of annotated algospeak sentences in Croatian, English, and Italian, which has been designed to complement an existing lexicon and associated taxonomies of linguistic processes and user intentions. Our corpus provides contextualised examples linked to lexical entries and annotated for communicative intent, enabling more fine-grained qualitative analysis of the rapidly evolving phenomenon of algospeak. We discuss the methodological challenges of data collection, including the limitations of automated extraction and the need for manual curation due to the dynamic and multimodal nature of algospeak. Finally, we outline a strategy for integrating and distributing the resource within the CLARIN infrastructure, addressing issues of access, licensing, and ethical constraints related to social media data.

1 Introduction[1]

Algospeak is the name given to a form of linguistic self-censorship resorted to by social media users in order to circumvent automated content moderation. It is commonly understood as the practice of abbreviating, misspelling, or substituting words in order to avoid detection by social media platform algorithms (Steen et al., 2023). While initially associated with TikTok, algospeak is now widespread across a number of different platforms such as Instagram, YouTube, and Reddit. Examples typically involve orthographic substitutions (e.g. *p0rn*), lexical replacements (e.g. unalive for 'dead'), and the use of emojis as proxies (e.g. 🌽 for 'porn', 🍉 for 'Palestine'). In this respect, algospeak bears some resemblance to earlier forms of encoded online language such as *leetspeak*, although its emergence is directly tied to the contemporary practices of social media platforms. Algospeak can, however, also extend beyond simple orthographic disguise, drawing on sophisticated phonological, morphological, and semantic processes as well as on shared cultural references to hide the meanings of words. As such, it represents a productive site for observing language change in real time and has already attracted growing scholarly and public attention, resulting in numerous recent academic studies along with articles in the popular press, and at least one book length popular treatment (Aleksic, 2024). One scholarly treatment which has had a especially marked influence on the work which we describe below

[1] This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

is that of (Calhoun and Fawcett, 2023), thanks in particular to its proposed classification of the linguistic processes underlying algospeak. We were also strongly influenced by (Wright et. al., 2025), the follow up to the previous work, which gave a detailed lexicographic description of numerous algospeak examples. In the current article, we will describe an extension of our own prior work (Khan and Squadrito, 2025), (Khan et al, 2026) in the documentation of algospeak via the compilation of a dedicated lexical resource and the use of semantic artifacts. In particular, we will detail the addition of a new resource which is intended to complement those which we previously documented, a corpus of example sentences annotated for different, salient kinds of information. We also discuss the different kinds of challenges that the data collection process poses and how CLARIN might help overcome at least some of these challenges.

2 A Resource for Describing Algospeak

As mentioned above, in previous work (Khan and Squadrito, 2025) we introduced a resource for documenting and analysing algospeak. This initially consisted of a **multilingual lexicon** offering a lexical description of instances of algospeak in (then) three different languages (English, Italian and Arabic) -- Croatian was added subsequently -- along with two different **semantic artifact**s for use in the description of these instances: the first of these is a taxonomy for **classifying types of linguistic processes** which underlie algospeak and which is based on that given in Calhoun and Fawcett, 2023); the second is a taxonomy classifying the **intention(s)** that tend to motivate the use of a given instance of AS. Both of these taxonomies, along with a list of community guidelines which an instance of algospeak could potentially be breaching (taken from (Calhoun and Fawcett, 2023)) are used to classify the different algospeak entries in the multilingual lexicon. Take, for example, the lexicon entry *PDF file*. In our lexicon this is associated with a **base lemma**, i.e. the lemma which it is intended to replace -- in this case 'paedophile' -- and a **language**, which in this instance is English. Not all entries in the lexicon will have an associated base form or even an associated language. For instance, the aubergine emoji 🍆which stands for the male member is **translingual** (and is tagged as such) and has an associated **evoked concept** which helps to identify it, and which in this case is 'penis'. Additional information associated with entries of the algospeak lexicon includes: the **definition** of the entry in English and/or a **literal translation** to English for examples in Italian, Arabic and Croatian; a description of the processes which have **generated the form**, with descriptors taken from the taxonomy of linguistic processes, e.g., in the case of 'PDF file', this is annotated with the name of the process: *Real Homophone*. We also include information on **primary intentionality** (what intention or intentions is this instance of algospeak primarily used with), using our intention taxonomy, and including one or more examples of the algospeak instance in context.

Our work aims to document examples of algospeak from numerous different languages/cultural milieux in an attempt to counterbalance the current bias in algospeak research towards American social media. In addition, we are interested in comparing algospeak with examples of related phenomena, both other forms of Internet language as well as pre-Internet language. Up till now our data collection process has been manual for a number of reasons including: the **difficulties of automatically extracting data** from social media platforms and the **lack of access to social media API**s for purposes of research, as well as the **multimodal nature of much of the data.** Moreover, as algospeak is by its very nature intended to avoid algorithmic detection, it was difficult to develop scripts to automatically find examples of such data in the wild. For these and other reasons, manual data collection proved to be more efficient, at least for the first stages of our work. To start with this data collection process involved taking screenshots from the different platforms we were interested in, as well as, whenever possible, URIs.

2.1 Creating Corpora for a Qualitative Analysis of Algospeak

Quite early on in the process of the collection of example algospeak instances, we decided that it would be useful to create an annotated corpus out of the examples that we collected. The main motivation was to better attest the forms described in the lexicon. It is worth pointing out here that our lexicon is intended as a qualitative as opposed to quantitative description of algospeak in different languages. We do not intend to directly mirror the proportion of different kinds of algospeak found online in the (kinds of) entries we feature in our lexicon, nor do we (yet) give quantitative information on frequency. Instead, our focus is on documenting examples of algospeak that, for one reason or another, might be of special interest to linguists, experts in neology, social scientists, communication specialists, or even digital activists and social media content creators. In many cases, it's not just the way that a word is

censored that is interesting for our purposes, e.g., the linguistic processes involved, but *which* words are being censored and in which context. Another reason for opting for a more focused process of corpus creation was that it would help us in the development of the resource as a whole; the annotation of the sentences using the intentionality taxonomy ended up being very useful in the evolution of the latter.

Our intention is to release most of our algospeak resource (the lexicon, semantic artifacts and corpus) as open data using a standard CC-BY license; for ethical and copyright reasons, however, we are not planning on releasing all of the corpus and related documentation and materials using the same license; we describe these reasons in more detail below. We would nonetheless like to make all our materials available via a more restricted license via CLARIN. In the rest of this section we will give more details on our corpus, or rather corpora, since we are compiling a separate corpus for each of the languages we are focusing on, Croatian, English and Italian (due to other commitments, the Arabic corpus still hasn't been started).

Each corpus currently includes the following information for the context sentences: the **source of the context sentence**, with minimal metadata which includes the name of the social media platform where the example was found and the year it was published (in cases where the source was a meme image shared via internet messaging, we document this in another way); a **link** to the relevant post/video and information on a last access, as well as **source file IDs r**eferring to one of the screenshots collected to attest the examples; for each form referenced in the context sentence, the corresponding **lexical entry** (in the lexicon) and the **intentions** with which the algospeak instance has been used is given (as far as our annotators have been able to determine). In case there is more than one algospeak instance in a given sentence, the sentence is repeated for each instance documented in the lexicon. The figure below gives an example of a number of annotated sentences in Italian.

1	Sentence	Source	Source file	Lexical Entry	Intentionality	Community Guideline
2	Quando sta a p3c0r4 e ti urla: "vai più forte che sono la tua tr@i4". La sborra: devi lasciarmi andare	Instagram, 2025	p3c0r4-tr@ia-sb@rr4-p4ll3_20251010_223617.jpg	p3c0r4	Humor / Irony / Parody (HIP	Adult nudity and sexual activites (ANSA)
3	Quando sta a p3c0r4 e ti urla: "vai più forte che sono la tua tr@i4". La sborra: devi lasciarmi andare	Instagram, 2025	p3c0r4-tr@ia-sb@rr4-p4ll3_20251010_223617.jpg	tr@i4	Humor / Irony / Parody (HIP	Adult nudity and sexual activites (ANSA)
4	Quando sta a p3c0r4 e ti urla: "vai più forte che sono la tua tr@i4". La sborra: devi lasciarmi andare	Instagram, 2025	p3c0r4-tr@ia-sb@rr4-p4ll3_20251010_223617.jpg	sb@rr4	Humor / Irony / Parody (HIP	Adult nudity and sexual activites (ANSA)
5	Quando sta a p3c0r4 e ti urla: "vai più forte che sono la tua tr@i4". La sborra: devi lasciarmi andare	Instagram, 2025	p3c0r4-tr@ia-sb@rr4-p4ll3_20251010_223617.jpg	p4ll3	Humor / Irony / Parody (HIP	Adult nudity and sexual activites (ANSA)
6	PACE porcoxxio PACE!!!!	Instagram, 2025 (comments)	porcoxxio_20251008_090305.jpg	porcoxxio	Non Offensive Avoidance (NOA)	

In the case of English and Italian most of our corpus examples have been sourced from social media platforms such as Instagram, Youtube and Reddit. In Italian the examples so far have been fairly simple from the linguistic point of view, largely involving the replacement of letters by the asterisk symbol or by numbers; this is also the case for most of the English examples we were able to find, although occasionally we were also able to find other kinds of algospeak, e.g., the use of 'Moustache Man' to refer to Adolf Hitler.

For a number of reasons to do with the (perceived) state of censorship on Croatian social media, the Croatian corpus is somewhat different. Indeed, most of the data in the Croatian corpus consists of examples found in the comments section of the Croatian online news portal `index.hr`. Apart from the usual intention of using algospeak to avoid censorship for harmful and offensive comments, other common intentions include humor, i.e., irony and parody used both as a means to offend someone, and to create a humorous effect for other readers. The most common linguistic mechanism applied is the deletion of letters or their replacement by the asterisk symbol (e.g., *j....e* 'fucking', *k....c* 'dick', *pi*** 'pussy', *ku*ac* 'dick', *od*bi* 'fuck off'), while using numbers for letters is a less common strategy (e.g., *j3bac* 'fucker', *ped0f1li* 'pedophiles', *7meće* 'trash'). The most interesting examples are those that reflect complex metaphorical mappings, and require an in-depth knowledge of the context. Those are also usually directed as insults towards individuals or groups, e.g., towards an eminent politician (*Zorangutan* 'Zoran + orangutan', *Anemični* 'the anemic one'), or a political party (*kradeze* - combination of 'to steal' and the acronym of the Croatian Democratic Party (HDZ); *left@rdelije* - the combination of 'leftists', 'retard' and a colloquial word delije 'the braves, tough guys'). Some of these examples have developed from individual instances of algospeak into common nicknames used in online communication, e.g. *kradeze* for the largest political party, and *Anemični* to refer to the current

prime minister.

3 How CLARIN Can Provide Access to Our Resource

Our planned publication strategy is closely aligned with CLARIN infrastructure best practices and with current initiatives around **computer-mediated communication (CMC)** corpora. We intend to deposit the textual corpora and the associated semantic artifacts in ILC4CLARIN, the Italian CLARIN B-centre repository, which offers persistent, machine-readable metadata, long-term preservation, and Single Sign-On access for protected resources. ILC4CLARIN also supports the publication of restricted datasets when full public release is not possible, something which is especially relevant for our materials given the copyright, privacy, and ethical issues surrounding social-media data, in our case largely screenshots. Through this strategy, the openly shareable parts of the resource will be made available under a public licence, while more sensitive components could be distributed under controlled access conditions, in line with the CLARIN licensing framework and its PUB/ACA/RES categories. More specifically, our resource is a strong fit for the CLARIN ecosystem around CMC and social-media data. Since we are already working within CLARIN, this initiative will be developed in collaboration with the relevant CLARIN bodies and services, including the CLARIN centre, the CLARIN Knowledge Centre for **Computer-Mediated Communication and Social Media Corpora (CKCMC)**, and the **CLARIN Legal Issues Committee (CLIC)**. Inclusion in the CLARIN CMC Resource Family is of course envisaged. Related CLARIN-hosted CMC resources, including work such as Ljubešić et al. (2023), also provide useful precedents for reflecting on how socially sensitive and legally constrained online data may be documented, exposed, or restricted in different ways.

References

Aleksic, A. (2025). *Algospeak: How social media is transforming the future of language*. Knopf Publishing.

Calhoun, K., & Fawcett, A. (2023). "They Edited Out her Nip Nops": Linguistic Innovation as Textual Censorship Avoidance on TikTok. *Language@Internet*, *21*, 1–30. https://doi.org/10.14434/li.v21.37371

Ljubešić, N., Erjavec, T., Batanović, V,. Miličević, M & Samardžić, T. (2023), Croatian Twitter training corpus ReLDI-NormTagNER-hr 3.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1793.

Khan, F., Gugliotta, E., Squadrito, E., Tarquini, M., & Frontini, F. (2026). A New Semantic Artifact Based Framework for Studying and Documenting Algospeak and Related Phenomena. *Proceedings of LREC 2026 (to Appear)*. Language Resources and Evaluation Conference (LREC).

Khan, F., & Squadrito, E. (2025). Linguistic Creativity in the Age of Content Moderation and Filter Word Lists. *Book of Abstracts of the International Conference on Historical Lexicography and Lexicology (ICHLL15).* The International Conference on Historical Lexicography and Lexicology (ICHLL15).

Steen, E., Yurechko, K., & Klug, D. (2023). You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, *9*(3), 20563051231194586. https://doi.org/10.1177/20563051231194586

Wright, K. E., Brewster, E., Carson, C. E., Hughes, B., McLean, J., Zhang (张笛菲), L., & Zimmer, B. (2025). Among the New Words. *American Speech*, *100*(2), 259–276. https://doi.org/10.1215/00031283-11868193