Building It-tok: an Italian TikTok Corpus

Luisa Troncone[1,2,*]

[1] *University of Salerno, Via Giovanni Paolo II, 84084 Fisciano, Italy*

[2] *STL CNRS UMR 8163, University of Lille, Rue du Burreau, 59650 Villeneuve d'Ascq, France*

Abstract

This contribution focuses on the process of building a corpus for TikTok discourse. Particularly, it aims at describing the choices made during the construction of a corpus of Italian TikTok videos. The corpus It-tok was collected to individuate linguistic functional correlates of digital discourse on TikTok. It-tok includes two subsets of videos: the first one is centered on videos concerning themes of interest for the public debate (e.g. abortion, femicide, racism, internal politics); the second one is made up of videos with no specific theme, intended to constitute the control sample for the observations made for the first sub-corpus.

Keywords

TikTok, public discourse, CMC corpora, modality, linguistic functional correlates

1. Introduction

The major goal of this contribution is to present some of the choices made during the process of building It-tok, an Italian TikTok corpus. The It-tok project was born with three aims:

- to provide a first assessment of the linguistic functional correlates (LFCs) displayed by TikTok content, and, subsequently, of the modality of communication on this specific social network;
- to highlight how themes of interest for the public debate are treated on this social network;
- to compare the LFCs found in general TikTok discourse with those emerging in thematically focused content.

By functional linguistic correlates, we mean the set of features that characterize language across different modalities; consequently, spoken texts exhibit specific correlates compared to written, read, or digitally produced texts (see section 3. for a more in-depth discussion): for instance, some LFCs of spoken language with regard to written language highlighted in previous studies regarded a significantly different distribution of PoS [1], the higher count of deictics [2], or of demonstratives [3]. LFCs describe effects on language uses based on the modality the communicative event takes place in. The focus on functional linguistic correlates poses significant challenges for decisions related to the corpus design, precisely because of the platform's content multimodality. Indeed, the structure of the final product, i.e., the TikTok video, can take highly diverse forms, and can employ a number of semiotic means for conveying the message. Because of the platform's multifaceted nature, a number of choices were to be made, to achieve a selection of content for analysis which was both replicable in its methodology and suitable for the achievement of the objective. In fact, beyond the thematic relevance, the selection process is complicated by the inherently heterogeneous nature of TikTok videos. These products differ not only in terms of topic but also from a semiotic point of view: they can include spoken language, text on screen, music, sound effects, gestures, visual editing techniques, or a combination of these means. As a result, choices concerning the corpus are of various kinds and must account for the platform's multimodal nature, which challenges both linguistic analysis and methodological consistency from the outset. In this work, we chose to focus solely on verbal content, but TikTok would allow for a variety of different levels of interest.

For accomplishing the goals illustrated above, we decided to subdivide the collection stage into two phases, one concerned with general discourse (Gen) and one concerned with political and social discourse

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy

[*] Corresponding author.

ltroncone@unisa.it (L. Troncone)

0000-0003-0791-8714

(PolSo).

During the collection, a number of methodological issues arose, which will be described here, together with the solutions we opted for; but, before getting to the decision-making section, we will briefly introduce TikTok and the reasons it was chosen.

2. TikTok: Characteristics, Meaningfulness, Employment in Linguistics

TikTok is a social media platform which allows for the publication of video content. It gained much popularity in the latest years especially among the youngest parts of the population [4][5]. A *tiktok* is a (usually) short video: in fact, while the maximum duration the platform allows for a *tiktok* is 10 minutes, the mean duration of *tiktoks* stands around 50 seconds[2]. This format can be also found on other social media (e.g., *reels* on Instagram), but the reasons which led to choosing it are linked to the amount and the kind of popularity it reached lately, rather than the specific format of the content considered.

With the beginning of the post-digital era, and the intersection and overlap of the online and offline lives, digital content has begun to have a consistent effect on our *analogic* life [6]. This is especially true for public debate themes, and, in the latest years, this influence has been especially clear for TikTok content, which is getting central in the political discourse. To give an idea of the importance TikTok gained in the public opinion building process, we can provide some examples. Consider the ban imposed by Donald Trump: as neo-elected President of the US, at the end of January 2025 he imposed the closure of the platform in the US, since he held that the Chinese government was receiving sensitive data about US government and citizens through TikTok users[3]. Given the amount of public disagreement with such decision, also manifested through many Americans signing up to Xiaohongshu, another Chinese social [7][8], Trump postponed the closure, and TikTok went dark for only one day on the 20th January. Conversely, political and social topics are increasingly present on TikTok discourse: the political importance of the platform can be seen also in spreading and testifying major political events and boosting discussion about major socio-political issues, such as the Black Lives Matter protests (2020) [9], the killing of Mahsa Amini (2022)[10], the war in Gaza (2023-present)[11], the #metoo movement (2020)[12], the suspension of the unsentenced guilty raper by the University of Leuven (2025) [13].

Given the importance social media nowadays hold in our society, there is no doubt that TikTok constitutes now a fundamental political mean [14][15][16], which makes it a viable field of study for our aims.

2.1. TikTok and Linguistics

Several studies in the field of linguistics (especially acquisitional, clinical, and variational) have already considered data coming from TikTok[4]. Many studies regard specific domains, and were built through a punctual methodology, which usually is not focused on an open resource. The main application fields for these studies regard especially language learning and teaching practices enhanced through TikTok [18][19] [20][21], the study of code-switching dynamics detected on the platform [22][23][24], language creativity [25][26], or hate speech detection and moderation [27][28][29].

Still, anyways, a description of the communicative modality/ies employed on TikTok, and especially in *tiktoks*, is missing. Furthermore, corpora of Computer Mediated Communication [30] have mainly concentrated (and this is also true for TikTok studies) on thematic corpora, which on their own can provide a partial portrait of the discourse on platforms. Just to focus on some examples regarding Italian CMC corpora, the only example we were able to find of a methodology leading to a generalist corpus is the one by TWITA [31][5], while others mainly exploit thematic hashtags [32][33][34][35] or specific pages [36][37] for the extraction. The reason for a generalist ("control-like") corpus stands in the fact that, in order for assertions on specific subsections (or thematic sections) to be solid, they should be checked with respect to how language is generally used on the specific platform. It-tok aims at providing both a description of the chosen path for the creation of a generalist corpus of *tiktoks*, and a characterization of modality displayed in such a content format.

3. Linguistic Functional Correlates

As explained in the previous paragraph, TikTok has already undergone a number of investigations in linguistics. Still, anyways, a bottom-up description of the functional features characterizing the platform is missing. Most existing studies tend to adopt a top-down

[2] According to Statista, *Average TikTok video length in 2023 and 2024.*

[3] The Trump-TikTok controversy was already on in 2020, during his first administration.

[4] For a theoretical perspective on communication dynamics on TikTok, see [17].

[5] TWITA exploits an extraction method which would have not been much effective for tiktoks, as it is based on the extraction of tweets with Italian most frequent words, but tiktoks cannot be extracted based on words in the video, since automatic subtitles are not searcheable.

approach, focusing on specific trends or phenomena, without accounting for the underlying structural features of TikTok communication as shaped by the platform's multimodal and technologically mediated nature. Addressing this gap is one of the central aims of the It-tok project, which seeks to identify TikTok's LFCs.

The LFCs of a specific modality of communication consist in the set of features which primarily describe that specific modality and characterize it with respect to others [38]. By *modality*, we mean the combination of semiotic resources (e.g., speech, gesture, text, image, sound), interactional dynamics (e.g., synchronicity, turn-taking), and cognitive constraints (e.g., processing time, spontaneity) that shape linguistic production in a given environment. For instance, the spoken modality is typically associated with the gesture-auditory-visual channel, real-time interaction, prosody, and a high degree of context-dependence. In contrast, written modalities tend to involve planning, permanence, and syntactic density, often favoring nominal constructions over verbal ones [39].

It is important to distinguish between *modality* and *channel*. While *channel* refers specifically to the physical means of transmission (e.g., auditory, visual, tactile), *modality* encompasses the broader communicative framework that includes social conventions, technological constraints, and the multimodal configuration of the medium. In the case of TikTok, the modality is particularly complex and hybrid, since it combines features of spoken interaction (e.g., spontaneous speech, direct address to an audience) with elements of edited visual media (e.g., cuts, overlays, subtitles, background music), thereby creating a composite, dynamic communicative environment.

As noted in the literature on this topic, LFCs do not depend on sociolinguistic features of speakers, but, instead, they stay the same across diastratically and diatopically different speakers. For this very reason, the construction of It-tok could avoid taking into consideration sociolinguistic representativeness issues, focusing instead on capturing the linguistic regularities that emerge specifically from the platform's multimodal communicative modality. The primary goal was to ensure that the corpus would be suitable for identifying these modality-driven patterns, rather than for mapping speaker-based variation.

4. Building It-tok

Some issues with building a corpus from TikTok videos have already been pointed out in [40]: namely, the authors refer to different formats of the videos, necessity of manual supervising for the automatic transcriptions, ethical considerations. Throughout our work, we tried to address these issues, regarding which we tried to make choices as solid as possible.

To identify LFCs of TikTok discourse (subcorpus Gen), and to compare them to the LFCs of that sub-part of TikTok discourse which concerns themes of interest for the public debate (subcorpus PolSo), we proceeded through a double phased data collection.

4.1. Corpus Building Process

TikTok API allows for the extraction of a maximum of 100 videos per extraction, which shall be from a 30 days time period, so the extraction was carried out month by month. The affordances of this research API does not allow for queries of tokens within the automatically generated captions (which would have been the preferred path), but it allows for querying hashtags. TikTok displays several characteristics in common with other platforms. One of these, is the affordance of hashtags. Hashtags are (small strings of) words, which function as hyperlinks, and link a content directly to others which contain the same hashtag. Most hashtags are thematic, in the sense that they describe the topic of that content. But this is not the only function they have on social media. In fact, hashtags can also be exploited to gain followers, or views, and in this case their form is a bit different. While, regarding the first function, hashtags do not display particularities on TikTok, considering the second one, these hashtags usually have a very transparent form on other platforms (*#followforfollow, #followme*). This is not the case for TikTok hashtags. Here hashtags are exploited by users in a way which, according to them, would boost the algorithm, and make them gain more views[6], but their form is by far less transparent, namely we have *#foryou, #fyp*, *#perte*, which all refer to the *for you page* of the app. This type of hashtags is by far the most used on TikTok[7], compared to thematic ones[8]. The so-called *for you page* (it. *per te*), so commonly cited in the hashtags, is the main page of the app, where users get the content TikTok suggests them based on what they liked or watched for longer[9]. What distinguishes the *fyp* from the other scrollable pages is the fact that in the *fyp* users are

[6] This was also confirmed by a questionnaire we spread over March 2025, concerned with the use of hashtags on TikTok.

[7] TikTok for Business, https://ads.tiktok.com/business/creativecenter/inspiration/popular/hashtag/pc/en.

[8] The same questionnaire cited in note 3 revealed that users of TikTok are well aware there are hashtags specifically useful (or thought to be useful) for TikTok content rather than other social media.

[9] An equivalent of the TikTok *fyp* is the *search* page on Instagram, or the *tl* (*timeline*) on X.

reached by content not necessarily published by people they follow. Therefore, to get in other people's *fyp* means to get more visibility on the app. For this reason, users tend to exploit hashtags connected to *fyp*.

Another way to boost the popularity of one's content consists in using thematic trending or popular hashtags, even for videos which have nothing to do with it.

All these features consistently affected our methodology of retrieving data, which had to consider the peculiarities of the platform.

Because of the peculiarities of TikTok hashtags, we chose to pay some special attention to the hashtags used for the query, and in particular we had to avoid keyword with a scope which was too large and concurrently the ones whose scope was too restricted, since we would have risked ending up with no results. We extracted a minimum of 15 videos per month for each of the subcorpora, selecting them by duration (>60s) and region of publication ("IT"=Italy). The video extracted were all published between October 2024 and January 2025. The extraction was performed during February and March 2025. Among the videos reached, only the ones showing the *voice_to_text* feature, namely the TikTok automatic transcriptions, were considered viable for It-tok. This way we could avoid video memes (usually shorter than 60s), those videos where the message is carried by the music rather than the speech and the ones in which there is no speech at all. Note that we did not use the automatic transcription as the final transcript: its presence was solely employed as a filter to exclude videos that did not contain or feature any spoken language. This way, we isolated the materials containing spoken language, whether continuous or discontinuous, explicitly excluding content such as memes, images carousels, or other materials lacking spoken language.

Finally, we got a total of 196 viable videos. Those videos were automatically downloaded, transcribed through the tool Open-AI Whisper in Python [41], both in aligned .txt and .eaf files. The transcriptions were annotated using the CLIPS [42] standard [43]. We decided to add some tags, which we thought would be useful for detecting specific sections of the texts. Table 1 summarizes the tags to be found in the annotated transcription.

Finally the .txt files were automatically tagged (through spaCy [44]), ending up in a .conllu file.

To sum up, for each video It-tok provides:

- a .mp4 file;
- a .txt file;
- an antr.txt file;
- a .conllu file;
- a .eaf file.

The CoNLL-U file PROPN tags were exploited to carry out the anonymization of the files.

Table 1
List of tags used in the annotated transcription, respectively from CLIPS and added specifically for It-tok

Tag	Meaning	Source
<sp> <lp>	*short and long pauses*	CLIPS
<ehm> <eeh>	*full pauses*	CLIPS
<MUSIC> <NOISE> <inspiration> <breath>	*non verbal noises*	CLIPS
[foreign_word] [dialect]	*other languages (start)*	CLIPS
[/foreign_word] [/dialect]	*other languages (end)*	It-tok
<READ> </READ>	*read section (start and end)*	It-tok
<CLIP> </CLIP>	*clip (start and end)*	It-tok
<MASK_IL> </MASK_IL>	*masking inappropriate language (start and end)*	It-tok
<OTHER_SP> </OTHER_SP>	*other speaker speech (start and end)*	It-tok
<CHUNKING> <KISS>	*non-verbal noises*	It-tok

As for now, the CoNLL-U files were checked just for the PoS and lemma columns. Here we also tagged discourse markers (DMs), in order to make them easily retrievable. We chose to tag DMs because we thought they could provide a measure of the extent to which TikTok discourse could be compared to spoken language, and since they are also included in the features which make up LFCs [39].

During the extraction, both in the process for PolSo and Gen, we noticed that the number of minimum extraction necessary for reaching the minimum of 15 viable videos per month differed sensibly from month to month, as can be seen in Table 2. We supposed it depended on the period of the year the videos we were extracting belonged to. Particularly, we decided to extract videos from October, November and December of 2024 and January of 2025. As it is well known, the amount of posting, and the quality of posts on social media is very much dependent on the time of posting. In particular, during the last months of the year more "seasonal" posting happens [45][46][47], which may be due to specific festivities (Halloween, Christmas, New

Years' Eve) or the whole period of "end of the year" wrapped. This seasonal posting primarily consists of videos that likely do not meet our extraction criteria, as they are probably shorter than 60 seconds and/or lack spoken language. Consequently, to make sure it was a contingency of the peculiarities of the months considered, we attempted a subsequent extraction of February and March 2025, which showed a piece of evidence favoring our hypothesis, as they show a rate of videos featuring *voice_to_text* similar to the one displayed by January. This happens because the trends usually developing or spreading at the end of the year are trends that usually do not produce videos that would have been considered viable for our data (i.e., they are usually short, with songs or media carrying the message rather than the words and consequently not featuring *voice_to_text*).

Table 2
Number of extacted videos compared to viable videos, per month

Subcorpus	Month	videos extracted	voice_to_text videos	
PolSo	Oct	678	75	11%
	Nov	521	37	7%
	Dec	605	36	6%
	Jan	186	61	*33%*
	Feb	*180*	*64*	*35%*
	Mar	*185*	*69*	*37%*
Gen	Oct	521	65	12%
	Nov	808	40	5%
	Dec	847	35	4%
	Jan	250	76	*30%*
	Feb	*258*	*88*	*34%*
	Mar	*267*	*97*	*36%*

4.2. PolSo

Our thematic section was collected by extracting videos whose description included (at least) one in a list of hashtags. Due to the original thematic nature of hashtags, they usually have a general form, which made us prefer them with respect to keywords in video descriptions, as these last would have needed a broader consideration of their flected and/or derived forms (i.e., *femminista* 'feminist.SG', *femministe* 'feminist.PL.F', *femministi* 'feminist.PL.M', *femminismo* 'feminism').

The selection of the hashtags was carried out based on our common sense of users, and on the most recent surveys about what worries Gen Z[10] the most (especially compared to GenX), carried out by IPSOS in 2022[11] [49]. Furthermore, to ensure that this preference aligned with the interests of Italian youth, we distributed a brief online questionnaire via Google Forms to a random sample of individuals under the age of 27, selected through cluster sampling. The responses confirmed the primary areas of interest and, to some extent, introduced additional hashtags related to foreign policy, an area that, anyways is not currently taken into consideration for It-tok. The themes regard mainly civil rights and internal politics issues and can be subdivided in four groups: environmental and ecological crisis, national identities and policies, politicians, and social intersectional rights. Table 3 shows the hashtag names selected, together with their category.

Following the questionnaire, we plan on realizing an expansion of the PolSo section with themes from foreign politics[12].

4.3. Gen

As regards the generalist section, our *modus operandi* was completely different. Since we could not find any TikTok generalist corpus building methodology which would have been somewhat exhaustive, we opted for a format-based extraction strategy. Specifically, we selected three widely used formats on the platform, which are very common on the platform and that differ primarily in their varying degree of (perceived) interactionality: *storytimes, answers, stitches.* Particularly, the extraction of these last two types was based on the external caption TikTok automatically produces when creating a video in these formats: namely, *risposta a* 'answer to' and *#stitch con* 'stitch with'. Storytimes were extracted through hashtags.

In order to maintain internal consistency and comparability, we also aimed to keep the duration of the videos across the three formats as uniform as possible.

4.3.1. Storytime

The first format we exploited was the *storytime*, extracted through the corresponding hashtag. A storytime video displays a person usually speaking directly to the camera, telling a story, usually from their personal life, and unsolicited by anyone. Therefore, storytimes are strongly monological. They make a format of their own on a number of platforms[13], as they were also published on YouTube since 2015.

Table 3

[10] GenZ is the most present generation on TikTok [48].

[11] The survey showed that while GenX members are more interested in themes such as taxes, (un)employement levels and job market, GenZers care more about the environment, education and civil rights.

[12] Since LFCs do not depend on the theme, they shall not be interested in the specific topic of the video, so they shall remain the same as the ones we will extract from PolSo as it is.

[13] The massive presence of such a format is linked to its efficiency in being an instrument for creating online communities [50].

List of the hashtags chosen for the extraction

Category	hashtags
environmental and ecological crisis	*ecologismo* ‘ecologism’, *ecoansia* ‘ecoanxiety’, *ecoterrorismo* ‘ecoterrorism’, *overtourism, antispecismo* ‘antispeciesism’, *specismo* ‘speciesism’, *ecofemminismo* ‘ecofeminism’
national identities and policies	*capitalismo* ‘capitalism’, *anticapitalismo* ‘anticapitalism’, *migrante* ‘migrant’, *migranti* ‘migrants’, *rifugiati* ‘refugees’, *antifascista* ‘antifascist’, *antirazzista* ‘antiracist’, *razzismo* ‘racism’
social intersectional rights	*femminismo* ‘feminism’, *feminist, metoo, femminicidio* ‘femicide’, *patriarcato* ‘patriarchy’, *violenzadigenere* ‘gender-based violence’, *aborto* ‘abortion’, *misoginia* ‘misogyny’, *omofobia* ‘homophobia’, *transfobia* ‘transphobia’, *omolesbobitransfobia* ‘homo- lesbo- bi- trans- phobia’, *dirittiLGBT* ‘LGBT rights’, *abilismo* ‘ableism’, *grassofobia* ‘fatphobia’, *femminismointersezionale* ‘intersectional feminism’, *intersezionale* ‘intersectional’, *privilegio* ‘privilege’, *woke*
politicians’ names	*giorgiameloni, governomeloni, matteosalvini, giuseppeconte, ellyschlein, antoniotajani*

4.3.2. Answer

The second format we selected is composed by *answers.* In these videos, the creator selects a comment to one of their videos and *answers* to the comment through a *tiktok*, rather than through writing another comment. This affordance is exploited for two reasons: either because writing would have costed too much time, or because a video answer is content in itself, whereas a written answer is not, and TikTok algorithm is said to boost accounts posting frequently. Since answers directly refer to one comment, they can be considered *more* interactional, compared to storytimes. This is also to be seen in the linguistic features answers they display: more deictic expressions or second person pronouns and verb forms.

4.3.3. Stitch

Stitches are somewhere in the middle between answers and storytimes. A *stitch* is a video which starts with/from another video section, usually lasting no more than 10 seconds, by another creator[14]. The *stitched* video (i.e., the original one) can be either used as a base for speaking one’s mind on a subject introduced by said stitched video itself, making it just a clarifier for the context, or can have the same function as the base comment in the *answer* videos, and in this case, as happens for answer videos, linguistic features include second person pronouns and verb forms. So, because of their very nature, stitches can be seen as formats more interactional than storytimes, but less interactional than answers.

4.4. Semi-supervised Automatic Collection

All the processes the videos went through were to be checked manually, as automatized processing turned out to be not always completely reliable. As for the transcriptions, this could be due to the quality of the sound, the presence of dialect or of strong regional markedness, or the (highly variable) speech rate. Regarding the tagging, as we mentioned earlier (4.1), at the moment we did not check the syntactic tagging yet, but the PoS tagging and lemmatization outputs were to be manually checked, as they presented some inaccuracies. Lemmatization was to be corrected especially for cases such as:

- less frequent verbs or verb forms, e.g. *tatuo* ‘tattoo.PRS.1SG’ was lemmatized as *tatuere* instead of *tatuare,* or future forms like *ripeterai* ‘repeat.FUT.2SG’ lemmatized as *ripeterai* instead of *ripetere*;
- irregular verb forms, e.g., *sai* ‘know.PRS.2SG’ got lemmatized as *saare* instead of *sapere*;
- verb forms displaying suppletivism, e.g., *vai* ‘go.PRS.2SG’ got lemmatized as *vai* instead of *andare.*

Regarding PoS tagging, main issues pertained:

- loans, marked as proper names;
- big numbers, such as years, which in the transcription were written in words (e.g. not *20* but *twenty*) and got marked as proper names;
- deverbal nouns, e.g. *(il) ritorno* ‘(the) come back’, tagged as *VERB* instead of *NOUN.*

5. Current Status and Future Perspectives

5.1. Current status of It-tok

As for April 2025, we extracted a total of 196 videos for a total duration of more than 7 hours (see Table 4 for a deeper insight). We are in the process of analyzing some

[14] According to TikTok support: “Stitch allows you to combine a video on TikTok with one you're creating” [51].

LFCs, focusing on particular lexicosyntactic traits. Anyways, we are well aware that a number of sociophonetic and morphological features could be analyzed, but we chose to focus only on some levels. In the meantime, It-tok shall be published online by the beginning of 2026. Concerning the publication and the treatment of data, we considered what was done during other studies based on TikTok data. Particularly, since the text data cannot be traced back directly to the original videos, and since the content we extracted is publicly accessible online, the data is to be considered of public domain [40][52].

Table 4
Current status of It-tok

Subcorpus	Total duration	Total word count
PolSo	3:43:49	32 581
Gen	4:06:04	35 254
It-tok	*7:50:54*	*67 835*

5.2. Preliminary Observations

Table 5 shows the distribution of videos found for groups of hashtags in PolSo (notice that a video can present more than one hashtag, and hashtags belonging to different subsections).

Table 5
Distribution of the hashtags types through PolSo

Subsection	Videos
Ecologic crisis	6
Political and national identities	11
Civil rights	76
Politicians' names	35

Table 6 shows the top ten most frequent hashtags in the whole It-tok corpus, excluding the ones we used for the extraction.

Table 6
Most frequent hashtags in It-tok

Rank	Hashtags' names	Frequency
1	***perte***	21
2	***fyp***	14
3	*meloni*	11
4	*politica*	10
5	*fratelliditalia*	8
6	*donne*	8
7	*diritti*	7
8	***viral***	7
9	***neiperte***	7
10	*salvini*	6

Table 6 confirms what we asserted about the way hashtags are used on TikTok: on the one hand, it is well true that some of the most used hashtags, excluding the ones we explicitly searched videos for, are thematic, but this is a very specific way of using hashtags, since the only ones we find are connected to PolSo themes. Videos from PolSo, in fact, display an overabundance of hashtags, compared to the ones in Gen (i.e., 9,37 vs. 3,2 hashtags per video, in average). Nonetheless, the most frequent hashtags remain *#perte* and *#fyp*, both very TikTok-specific and directed towards gaining views and/or boosting the content through the algorithm.

Turning to LFCs, we preliminarly looked at the distribution of PoS in the It-tok corpus. PoS were chosen for first assessment of some modality characteristics because it has been shown that they correlate with the spokenness of texts. In particular, nouns and verbs, and their respective modifiers, adjectives and adverbs, have been said to act as pivotal units in the construction of a text. Their frequency offers significant insights into how different types of texts are syntactically structured and how modality influences linguistic composition. Specifically, nouns tend to be more prominent in written texts, while the frequency of verbs increases progressively as one moves towards more natural spoken language [39]. Such a tendency is seen also for It-tok: Table 7 shows PoS occurrence percentages of It-tok, along with a comparison with the corpora:

- *Lessico dell'Italiano Parlato* (LIP), a corpus of spoken Italian [53];
- *Primo Tesoro della lingua italiana letteraria del Novecento* (PTLLI), a corpus of literary Italian, from the 1900s [54]
- *Corpus Scritto* (CS), a corpus of written Italian [55].

Table 7
Nouns, adjectives, verbs and adverbs occurrence in It-tok, compared to other Italian corpora.

	It-tok	*LIP*	*PTLLI*	*CS*
NOUN	17,9%	15,7%	20,0%	21,7%
ADJ	5,6%	8,8%	7,9%	17,0%
VERB	21,2%	20,0%	18,7%	10,4%
ADV	10,6%	10,1%	6,1%	3,8%

The data for LIP, PTLLI and CS in Table 7 is taken from [40]. The corpora shown above are the ones previous research was performed on, and for which the tendencies named had been observed. However, similar tendencies emerge also from the CORIS [56], ItWac [57] and Paisà [58] written corpora, see Table 8. Notice that these last two corpora are collected from online sources. Also with respect to these, It-tok alignes more closely to

LIP.

Table 8
Nouns, adjectives, verbs and adverbs occurrence in It-tok, compared to CORIS, ItWac and Paisà.

	It-tok	*CORIS*	*ItWac*	*PAISÀ*
NOUN	17,9%	24,7%	22,3%	18,3%
ADJ	5,6%	9,3%	8,6%	7,4%
VERB	21,2%	12,5%	14,0%	12,1%
ADV	10,6%	4,6%	3,4%	2,6%

It must be considered, anyways, that the two subcorpora differ greatly by linguistic features displayed. As an example, Figure 1 shows the difference in the occurrence of DMs, which are strongly associated with spoken modality, with respect to the total of tokens in the different texts[15].

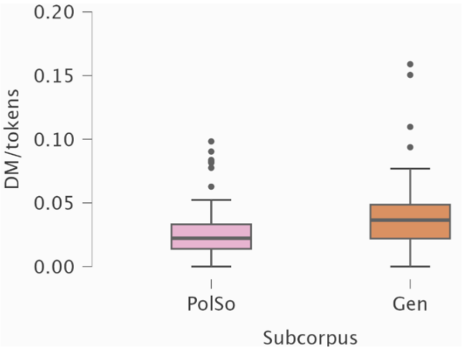

Figure 1: DM occurrence in It-tok, subdivided per subcorpora.

As can be seen in Figure 1, even though the two subcorpora are equally spoken, it seems that the thematic one employs a kind of speech which is probably less hesitant. Further research will include a differentiation based on the different themes, within PolSo, and based on different features, between the PolSo and Gen.

5.3. Future perspectives

Next advancements in It-tok building involve exploring the language features findable on TikTok (e.g., newly imported constructions or neologisms), broading It-tok and its scope, and the building of a treebank of TikTok discourse.

Though It-tok being a still very small corpus, it displays the potentiality to show a number of linguistic uses hardly findable in traditional corpora, like creative uses or newly registered loans (see 1-3), for which It-tok could also be searched.

(1) venire blastato da mio nipote è stato meraviglioso (0125_S)

‘to be *blasted* by my nephew was wonderful’

In (1), *blastato* < *blastare* < en. *to blast* stands for ‘getting humiliated’ through words. An adapted loan can be seen also in (2), where *flexare* sth. < en. *to flex* stands for ‘show one’s ability in sth.’.

(2) [...] possibilità di flexare un po' di statistica (G0125_16)

‘[...] opportunity to *flex* some statistics’

In (3) it’s the whole passive construction to get borrowed.

(3) Un calciatore della Juventus è stato fatto outing (1224_F)

‘A Juventus player was outed’

Another set of features, phonetics in nature, that could be thus investigated regards the so-called “influencer accent”, which was noticed around the internet but still never assessed [59][60].

Furthermore, due to its informal nature, TikTok could provide naturalistic data for a number of linguistics areas of interest, e. g., neologisms and gendered neologisms [61][62][16], or code switching/mixing phenomena [63].

The expansion we foresee for It-tok regards Gen, but also partially PolSo. Nonetheless, based on the methodology applied for the extraction, Gen could be easily systematically broadened, making it a potential monitor corpus for Italian TikTok discourse through an yearly update. Furthermore, PolSo will be widened to include themes of foreign policy. A further expansion could pertain a set of videos that were systematically excluded with the present methodology: particularly, the video memes, and that could constitute the base for studies on innovative linguistic forms, could be extracted through hashtags such as #*memetok*. The pseudosuffix -*tok* can apply to any word X, i.e. *X-tok*, standing for ‘section of TikTok regarding X’. Examples of usage involve *booktok* ‘section of TikTok regarding books’, *cattok* ‘section of TikTok regarding cats’, *footballtok* ‘section of TikTok regarding football,

[15] $p < 0.001$ for the Mann-Whitney test.

[16] In fact, some gendered neologisms, such as *girl dinner*, actually were born from a trend on TikTok, and then spread all over other social networks.

feministtok 'section of TikTok dedicated to feminism', *lefttok* 'section of TikTok filled with leftists. From this pseudosuffix, It-tok takes its name.

Finally, we will be building a treebank of at least 10% of It-tok, based on the methodology implemented for the KIPARLA forest project [64][65]. This will allow for syntactic queries, and make visible LFCs which are proper of the syntactic level of analysis (e.g., types of clauses, syntactic dependencies, subordination, syntactic heaviness).

6. Conclusions

With this contribution, we aimed to provide a brief overview of the methods adopted and the decisions made during the construction of an Italian TikTok corpus. Our choices were guided both by the specific communicative dynamics of the platform and by our research objectives, namely, to assess certain LFCs of TikTok discourse and, where applicable, to distinguish between generalist and thematic subtypes.

It-tok is structured to represent the first generalist corpus of spoken Italian on TikTok, and besides its main aims about LFCs and characteristics of political and social discourse online, it can represent a way to open TikTok to linguistic systematic studies, because of its replicable methodology, also applicable to create comparable corpora for other languages.

Due to the aim to describe LFCs in both general TikTok discourse and discourse on political and social topics, we adopted a split extraction strategy. Manual supervision was required at all stages of the automated processing to ensure consistency, accuracy, and compliance with the criteria established for corpus inclusion.

Importantly, the multimodal nature of TikTok, as a platform where language coexists with visual, auditory, and gestural elements, means that its texts are inherently complex and shaped by multiple interacting variables. These characteristics pose unique challenges for corpus design and analysis, they also provide valuable insights into modern digital communication practices, both in terms of parallel communication channels and the simultaneous use of multiple semiotic modes to construct a message (e.g., use of emojis, or particular visual rendering of verbal language, such as the SpongeBob Mocking meme to convey derision [66]).

Once completed, It-tok will provide a linguistically annotated corpus of Italian TikTok discourse, featuring transcriptions formatted according to the CLIPS conventions and annotated at multiple levels, including PoS tagging, morphological features, and syntactic dependencies in UD.

The corpus will also include a small but representative treebank, offering structured syntactic analyses of selected texts that reflect the linguistic complexity of this emerging multimodal variety.

Acknowledgements

I would like to thank prof. Voghera, and the whole linguistics nucleus of the University of Salerno, for the suggestions and the discussions about the project.
This project is part of the program of education and sensibilization towards online public discourse of the *OLiNDiNUM*[17] *- Observatoire LINguistique du DIscours NUMérique* project.

References

[1] G. Policarpi, M. Rombi, M. Voghera, Nomi e verbi in sincronia e diacronia: multidimensionalità della variazione, in: A. Ferrari (Ed.), Sintassi storica e sincronica dell'italiano. Subordinazione, coordinazione, giustapposizione. Atti del X Congresso della Società Internazionale di Linguistica e Filologia Italiana (Basilea, 30 giugno - 3 luglio 2008), Cesati, Firenze, vol. I (543-560), 2009.

[2] C. Sammarco, Il contributo delle costruzioni senza verbo nell'espressione delle relazioni spaziali nel Parlato, in Testi e Linguaggi, 14 (2020), 91-124.

[3] L. Gaudino-Fallegger, I dimostrativi nell'italiano parlato. Wilhelmsfeld: Egert, 1992.

[4] *TikTok user demographics: what's the average age of TikTok users?* SOAX. URL: https://soax.com/research/average-age-of-tiktok-users. Last accessed on 28th April 2025.

[5] *TikTok: distribution of global audience, by age and gender.* Statista. URL: https://www.statista.com/statistics/1299771/tiktok-global-user-age-distribution/. Last accessed on 28th April 2025.

[6] I. Bhatt, and L. Gourlai, Postdigital / More - Than - Digital Meaning-Making, Postdigital Science and Education (2024) 6:735–742. doi.org/10.1007/s42438-024-00512-1

[7] *American Defiance Against TikTok Ban Fuels Surge in Alternative Social Media Platforms*, Legal News Feed, Last accessed on 26th April 2025. URL: https://legalnewsfeed.com/2025/01/14/american-defiance-against-tiktok-ban-fuels-surge-in-alternative-social-media-platforms/?

[8] *TikTok users in US flock to 'China's Instagram', RedNote, ahead of ban*, Al Jazeera, Last accessed on

[17] Website of the observatory available at URL: https://olindinum.huma-num.fr.

26th April 2025. URL: https://www.aljazeera.com/amp/economy/2025/1/15/tiktok-users-in-us-flock-to-chinas-instagram-ahead-of-ban

[9] *TikTok serves as hub for #blacklivesmatter activism*, CNN. Last accessed on 26th April 2025. URL: https://edition.cnn.com/2020/06/04/politics/tik-tok-black-lives-matter/index.html

[10] T. Walsh, "TikTok as a site of social protest in Iran's Gen-Z uprising." Discourse & Society, 35.5 (2024): 625-650. doi.org/10.1177/09579265241234351

[11] T. Abu Laban, "The Role of TikTok in Disseminating the Palestinian Narrative during the War on Gaza from the Perspective of Palestinian University Students." Advances in Journalism and Communication, 11 (2023): 394-408. doi.org/10.4236/ajc.2024.123021

[12] A. Boyd, and B. McEwan, Viral paradox: The intersection of "me too" and #MeToo, New Media & Society, 26.6 (2022): 3454-3471. doi.org/10.1177/14614448221099187

[13] *Leuven Public Prosecutor appeals verdict of medical student rape case.* The Brussel Times. Last accessed on 28th April 2025. . URL: https://www.brusselstimes.com/1518910/leuven-public-prosecutor-appeals-verdict-of-medical-student-rape-case

[14] I. Literat, N. Kligler-Vilenchik, TikTok as a Key Platform for Youth Political Expression: Reflecting on the Opportunities and Stakes Involved. Social Media + Society, 9.1 (2023): doi.org/10.1177/20563051231157595

[15] *From Viral Dances to Political Movements: The Impact of TikTok Challenges and Memes*, Medium. Last accessed on 28th April 2025. URL: https://medium.com/%40wilsonrolypaul/from-viral-dances-to-political-movements-the-impact-of-tiktok-challenges-and-memes-609632842f3e

[16] *The Weapon of the Century: Contemporary Politics Through the TikTok Algorithm,* The Harvard Political Review. Last accessed on 28th April 2025. URL: https://theharvardpoliticalreview.com/tiktok-politics-algorithm/

[17] G. Marino, B. Surace (Eds.), TikTok. Capire le dinamiche della comunicazione iper-social. Hoepli editore, Milano, 2023.

[18] A. Nu'man, R. Indriana, Z. Ahmad, A. Ainul & R. D. Hasti. Improving Verbal Linguistic Intelligence in Early Childhood Through the Use of Tiktok Media, Jurnal Obsesi Jurnal Pendidikan Anak Usia Dini, 6.3 (2022) 2316-2324.

[19] T. N. Fitria, Value Engagement of TikTok: A Review of TikTok as Learning Media for Language Learners in Pronunciation Skill. EBONY, Journal of English Language Teaching, Linguistics, and Literature, 3.2 (2023), 91–108. doi.org/10.37304/ebony.v3i2.9605

[20] T. N. Fitria, Using TikTok application as an English teaching media: a literature review, Journal of English Language Teaching, Applied Linguistics, and Literature, 6.2 (2023), 109–124. doi.org/10.20527/jetall.v6i2.16058

[21] G. Leon Liu, X. Zhao & M. T. Feng, TikTok Refugees, Digital Migration, and the Expanding Affordances of Xiaohongshu (RedNote) for Informal Language Learning, International Journal of TESOL Studies (2025), 250123. doi.org/10.58304/ijts.250123

[22] S. H. Daulay, A. H. Nst, F. R. Ningsih, H. Beretu, N. R. Irham & R. Mahmudah, Code Switching in the Social Media Era: A Linguistic Analysis of Instagram and TikTok Users, Humanitatis: Journal of Language and Literature, 10.2 (2024), 373-385. 10.30812/humanitatis.v10i2.3837

[23] Z. Li & L. Wang, Investigating translanguaging strategies and online self-presentation through internet slang on Douyin (Chinese TikTok), Applied Linguistics Review, 15.6 (2024), 2823-2855. doi.org/10.1515/applirev-2023-0094

[24] E. Nurchurifiani, I. Hanum, A. Damiri, O. Oktariyani, A code mxing usage on social media: a linguistic analysis of video on TikTok, KLAUSA: Kajian Linguistik, Pembelajaran Bahasa, dan Sastra – Journal of Linguistics, Literature, and Language Teaching, 9.1 (2025), 90-101. doi.org/10.33479/klausa.v9i1.1194

[25] B. Ugoala, Generation Z's lingos on TikTok: analysis of emerging structures, Journal of Language of Communication, 11.2 (2024), 211-224. 10.47836/jlc.11.02.08

[26] M. Tomenchuk & T. Tiushka, The impact of TikTok on the English language: slang and trends, Věda a perspektivy, 11.42 (2024), 441-447. doi.org/10.52058/2695-1592-2024-11(42)-441-447

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

[28] J. S. Rauchberg, #Shadowbanned: Queer, trans, and disabled creator responses to algorithmic oppression on TikTok, in: P. Paromita (Ed.), LGBTQ digital cultures: A global perspective (196–209). Routledge. 2022.

[29] N. Fadhilah, B. Suswanto & Y. P. Utami. Forensic Linguistics: Netizens' Hate Speech Implicature on the Issue of the 2024 Presidential Election (TikTok Social Media Case Study), Technium Social

Sciences Journal, 50 (2023), 204-210.

[30] C. Thurlow, L. Lengel & A. Tomic, Computer Mediated Communication, Sage publications, London, 2004.

[31] V. Basile & M. Nissim, Sentiment analysis on Italian tweets, in: A. Balahur, E. van der Goot & A. Montoyo (Eds.), Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (100-107), ACL, 2013.

[32] M. Donati, L. Polidori, P. Vernillo, G. Gagliardi, Building a corpus on Eating Disorders from TikTok: challenges and opportunities, in Proceedings of the Ninth Italian Conference on Computational Linguistic (CLiC-it 2024), 2023.

[33] M. Palermo, La rappresentazione multimodale dei dialetti su TikTok, Italiano LinguaDue, 14.2 (2023), 131–139. doi.org/10.54103/2037-3597/19652

[34] I. Caiazzo, G. M. Dimitri & L. Tronci, IncluInstIT: Un nuovo corpus per lo studio di linguaggio inclusivo su Instagram, in: S. Rebora, M. Rospocher, S. Bazzaco, (Eds.), Diversità, Equità e Inclusione: Sfide e Opportunità per l’Informatica Umanistica nell’Era dell’Intelligenza Artificiale, Proceedings del XIV Convegno Annuale AIUCD 2025 (35-39). Verona: AIUCD, 2025.

[35] A. T. Cignarella, C. Bosco & V. Patti, TWITTIRO`: a Social Media Corpus with a Multi-layered Annotation for Irony, in: R. Basili, M. Nissim & G. Satta (Eds.), Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-it 2017, 11-12 December 2017, Rome (101-106), Accademia University Press, 2017.

[36] C. Ferrini, Il parlato-digitato dell'italiano come heritage language nei gruppi Facebook: riflessioni e modellizzazioni da un corpus multilingue. Italica, 98.1 (2021): 112–128. doi.org/10.5406/23256672.98.1.08

[37] Y. Martari, Come scrivono i politici italiani su Facebook Appunti per un’analisi linguistica comparative, L’Analisi Linguistica E Letteraria, 26.2 (2018), 81-114.

[38] M. J. Luzón, Forms and functions of intertextuality in academic tweets composed by research groups, Journal of English for Academic Purposes 64 (2023), 101254. doi.org/10.1016/j.jeap.2023.101254.

[39] M. Voghera, Dal parlato alla grammatica. Costruzione e forma dei testi spontanei, Carocci, Roma, 2017.

[40] M. Donati, P. Vernillo, La linguistica dei corpora nell’era dei social media: Le nuove sfide poste da TikTok, in: S. Mattiola, M. Miličević Petrović, CLUB Working Papers in Linguistics, volume 8, University of Bologna, Bologna, 2024, doi.org/10.6092/unibo/amsacta/8065

[41] A. Radford, J. W., Kim, T., Xu, G., Brockman, C., McLeavey, I., Sutskever, Robust Speech Recognition via Large-Scale Weak Supervision, International Conference on Machine Learning (2022). doi.org/10.48550/arXiv.2212.04356

[42] F. Albano Leoni, A. Sobrero, and A. Paoloni, Corpora e lessici di italiano parlato e scritto (CLIPS), Bollettino di italianistica, Rivista di critica, storia letteraria, filologia e linguistica 2 (2007): 121-0, doi: 10.7367/71826

[43] R. Savy, CLIPS. Specifiche per la trascrizione ortografica annotata dei testi raccolti. Università del Salento. URL: https://www.unisalento.it/documents/20152/5018562/Specifiche+per+la+trascrizione+ortografica.pdf/414d183f-fd6a-2d31-7fbe-44ac7ff63772?version=1.0

[44] M. Honnibal, I. Montani, S. Van Landeghem, & A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303

[45] N. Kumar, G. Ande, J. S. Kumar, M. Singh, Toward Maximizing the Visibility of Content in Social Media Brand Pages: A Temporal Analysis, Social Network Analysis and Mining 8.11 (2018). doi: 10.1007/s13278-018-0488-z

[46] Y. Sano, H. Takayasu, S. Havlin, M. Takayasu, Identifying long-term periodic cycles and memories of collective emotion in online social media, PLoS ONE 14.3 (2019): e0213843. doi.org/10.1371/journal.pone.0213843

[47] N. Okano, M. Higashi, A. Ishii, The Influence of Social Media Writing on Online Search Behavior for Seasonal Events: The Sociophysics Approach, 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, 4339-4345, doi: 10.1109/BigData.2018.8622186.

[48] *TikTok leads time spent on social for most US adults,* E-marketer. Last accessed on 27th April 2025. URL: https://www.emarketer.com/content/tiktok-leads-time-spent-on-social-most-us-adults#

[49] IPSOS, Elezioni politiche 25 settembre 2022: il confronto tra Generazione Z e Millennials. Last accessed on 29th April 2025. . URL: https://www.ipsos.com/it-it/millenials-generazione-z-rapporto-giovani-politica-italia

[50] Z. Papacharissi, Affective Publics: Sentiment, Technology, and Politics, Oxford University Press, Oxford, 2014.

[51] *What is a stitch*, TikTok support. Last accessed on 20th April 2025 . URL: https://support.tiktok.com/en/using-tiktok/creating-videos/stitch

[52] S. S. C. Herrick, L. Hallward , L. R. Duncan, "This is just how I cope": An inductive thematic analysis

of eating disorder recovery content created and shared on TikTok using #EDrecovery, Int J Eat Disord, 54.4 (2021): 516-526. doi:10.1002/eat.23463.

[53] T. De Mauro, F. Mancini, M., Vedovelli, and M. Voghera, Lessico di frequenza dell'italiano Parlato (LIP), Etaslibri, Milano, 1993.

[54] T. De Mauro, Primo Tesoro della lingua italiana letteraria del Novecento (PTLLI), UTET, Torino, 2007.

[55] F. Mancini, L'elaborazione automatica del corpus, in: T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera (Eds.), Lessico di frequenza dell'italiano Parlato (LIP), Etaslibri, Milano, 1993.

[56] R. Rossini Favretti, F. Tamburini & C. De Santis, CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model, in: S. Wilson, P. Rayson & T. McEnery (Eds.), A Rainbow of Corpora: Corpus Linguistics and the Languages of the World (pp. 27-38), München, LINCOM-Europa, 2002.

[57] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora, Language Resources & Evaluation, 43, 209–226 (2009). doi.org/10.1007/s10579-009-9081-4

[58] V. Lyding, E. Stemle, C. Borghetti, M. Brunello, S. Castagnoli, F. Dell'Orletta, H. Dittmann, A. Lenci, & V. Pirrelli, The PAISÀ Corpus of Italian Web Texts, in: F. Bildhauer & R. Schäfer (Eds.), Proceedings of the 9th Web as Corpus Workshop (WaC-9), 36–43, Gothenburg, Sweden. Association for Computational Linguistics. 2014.

[59] *How TikTok created a new accent – and why it might be the future of English*, BBC. Last accessed on 2nd May 2025. . URL: https://www.bbc.com/future/article/20240123-what-tiktok-voice-sounds-like-internet-influencer

[60] N. Adomaitis, L. Hoang, M. Shama, S. Trieu, K. Zhao, The TikTok Influencer Voice: Do Sociolinguistic Features Influence the Success of TikTok Videos?, Languaged Life - Studies in language and society, UCLA (2024).

[61] O. Foubert, and M. Lemmens, Gender-biased neologisms: the case of man-X, Lexis Journal in English Lexicology (Lexical and Semantic Neology in English), 12, 2018, 1–26.

[62] M. Szymańska, Gendered Neologisms Beyond Social Media: the Current Use of Mansplaining, Research in Language, vol. 20.3, 2020, 259–276.

[63] D. P. Wardhani, and Y. Arifin, Code switching and code mixing in Ritsuki's vlog on Digita Media TikTok: a study of sociolinguistics, Esteem Journal of English Education Study Programme, 8.1 (2025), 200-205. doi: doi.org/10.31851/esteem.v8i1.18124.

[64] L. Pannitto, Towards the first UD Treebank of Spoken Italian: the KIParla forest. ArXiv, abs/2410.04589. doi.org/10.48550/arXiv.2410.04589

[65] L. Pannitto, C. Mauri, The KIPARLA Forest treebank of spoken Italian: an overview of initial design choices. ArXiv, abs/2411.06554, doi.org/ 10.48550/arXiv.2411.06554

[66] B. Yazell & A. Wohlmann, Memes in the Literature Studies Classroom, Narrative Works. Issues, Investigations, & Interventions, 12.1 (2023), 1-17. /doi.org/10.7202/1111279ar.

A. Online Resources

The corpus repository, which documents the corpus and treebank construction processes and the challenges encountered in the syntactic annotation of spoken data, is available on GitHub (https://github.com/cabinsix/It-Tok). The repository will host the transcribed and anonymized files, along with their corresponding CoNLL-U formatted versions. The treebank is currently under development and can be accessed via the Arborator platform (https://arborator.grew.fr/?#/projects/It-tok).