EXPANDING THE FRENK CORPUS OF SOCIALLY UNACCEPTABLE DISCOURSE TO FRENCH

Kristina PAHOR DE MAITI TEKAVČIČ,[1,2] Nikola LJUBEŠIĆ,[3] Darja FIŠER[2]

[1]Faculty of Arts, University of Ljubljana
[2]Institute of Contemporary History, Ljubljana
[3]Jožef Stefan Institute, Ljubljana

This paper outlines the creation of the French part of the FRENK corpus, which contains socially unacceptable comments posted in response to news articles on the topics of LGBT and migrants which were published on Facebook by prominent media outlets. These comments were manually annotated for the type and target of socially unacceptable comments. Out of 10,239 comments with around 300,000 tokens in FRENK-fr, around a third of the comments represent socially unacceptable discourse, of which around 1% are violent. These are most often aimed at migrants, who together with the LGBT community and their supporters represent the most prominent target group of socially unacceptable comments. FRENK-fr is fully comparable to other language-specific parts of the FRENK corpus, and can serve as a valuable resource for cross-cultural qualitative analyses of disrespectful online communication which can also inform actions by civil society and political institutions. Additionally, FRENK-fr provides essential data for training more generalizable language models to identify socially unacceptable discourse.

Keywords: socially unacceptable discourse, hate speech, French, migrants, LGBT

1 INTRODUCTION

The last two decades have seen a visible rise in the importance of social media platforms in influencing public opinion and actions. Social media have become a powerful hybrid space merging the public and private sphere in previously unseen ways which blurs the boundaries between information of general relevance, gossip and verified facts. The ease of content production enabled by social media platforms, allows for large amounts of messages to enter the virtual space. This saturated media landscape makes verifying and filtering information

difficult, which is why the end users can be quickly faced with disinformation and socially unacceptable content. Such messages are highly problematic because they influence our reasoning and decision-making processes, but also because they negatively impact social cohesion and thus the possibilities for a better future.

Efforts to understand socially unacceptable discourse (SUD) are crucial in order to limit its propagation and nonconstructive or even dangerous effects on the recipients (López & López, 2017), but since SUD is a highly heterogenous phenomenon without clear register characteristics (Zhang & Luo, 2019), the task proves especially complicated. Moreover, it has been shown that negative impact can be triggered even by implicit inappropriate messages (Kopytowska & Baider, 2017) which is why the efforts cannot remain limited to the investigation of *hate speech* or the most explicit violent content alone. In fact, the researchers of SUD increasingly shift their attention to non-violent, but nevertheless SUD messages and explore the role of various rhetorical devices and strategies in the construction of SUD (Despot et al., 2023).

If the shift from explicitly violent to implicitly offensive messages has gained momentum, the research on SUD still shows an important limitation due to its bias toward English datasets (Piot et al., 2024).[1] This proves largely inadequate in the current state of events when different national internet safety agencies regularly report on the propagation of SUD online, and national and EU regulations are becoming more stringent in reference to online content moderation.[2] Imbalanced representation of SUD corpora in multiple languages hinders the development of robust and generalizable models for the moderation of SUD online which makes the efforts towards a less toxic online environment on the long run less successful. Moreover, where SUD datasets in languages other than English exist, the comparative research and model development is made difficult due to differences in definitions, terminology and annotation guidelines (Carneiro et al., 2023).

[1]This is not to say that hate speech datasets in other languages, in particular French and Slovene do not exist, for example Chiril et al. (2020); Vanetik and Mimoun (2022) or Kralj Novak et al. (2021).
[2]See, for example, the EU Digital Services Act, retrieved May 10, 2024, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

The FRENK dataset (Ljubešić et al., 2021) represents one of the rare multilingual comparable corpora of socially unacceptable online comments. So far, common data collection and annotation methodology has been used to create datasets of Croatian, English, Slovene and Dutch online comments that include socially acceptable and unacceptable content (Ljubešić et al., 2019; I. Markov et al., 2021). This paper is concerned with the extension of the FRENK corpus to French. A set of online comments in French was annotated according to the FRENK methodology thus creating a dataset that is comparable to other languages already included in the FRENK corpus. The new corpus FRENK-fr 1.0 is available for download from the CLARIN.SI repository (Pahor de Maiti et al., 2024).

The paper is structured as follows: in Section 2, we provide an overview of the corpus creation process with the explanation and examples related to the annotation schema, some information about the annotation campaign and the level of inter-annotator agreement. Section 3 outlines the characteristics of the new FRENK-fr focusing on the structure of the corpus with regard to the distribution of types of socially unacceptable discourse and its targets. When relevant, the results are put into context by comparing them to data from the Croatian, English or Slovene part of the FRENK corpus. The paper ends with Section 4 which summarizes the main features of FRENK-fr and addresses its comparative potential in relation to other language-specific parts of the FRENK corpus.

2 CORPUS ANNOTATION

The data for the FRENK corpus, including FRENK-fr, contain comments that were posted under Facebook posts of most popular national mainstream media outlets which often share their news articles via social media platforms. The three most trending news media outlets in France were selected according to the *Alexa* service, and include *Le Monde, Le Figaro* and *20minutes*.[3] The data were collected during the FRENK project (ARRS J7-8280; 2017–2020) and cover the period between 2010 and 2017 with the majority of data posted from 2015 on. Since the corpus was intended to be manually annotated, the first objective

[3]The Alexa service, retrieved in 2017/2018, from https://www.alexa.com/topsites/countries, is no longer available.

was to collect data that will likely include a relatively high share of potential SUD. Given the (then) prominent and controversial societal events related to stronger migration flows and several initiatives addressing the discrimination of the LGBT community, the decision was made to create a classifier that filtered the harvested Facebook posts of media outlets for two specific topics, namely LGBT/homophobia and migrants/Islamophobia (see Ljubešić et al. (2019) for more details on classifier creation).

2.1 Annotation schema

The French dataset is annotated according to the project-specific typology (Ljubešić et al., 2021) which includes *Types* and *Targets* of SUD. More specifically, SUD is defined as any form of communication that is vulgar, offensive or violence-evoking and/or that represents a disruption in communication by propagating negative claims that can hurt the addressee and do not represent any added value to the argument put forth (Vehovar et al., 2020). Thus, SUD differs from an argumented critique or literary/journalistic genres, like satire, by recurring to unfounded claims and language use that can reasonably be expected to trigger psychological or physical harm.

The *Type* level of the annotation scheme consists of four classes that indicate the type of discourse found in the comment, as listed below:

- **acceptable speech** (the comment does not contain vulgar lexis nor any claim that could be judged offensive or violent)
  - *Il me semble qu'il y ait la séparation de l'Eglise et de d'Etat en France. Que l'Eglise prenne des positions ne me choque pas, au final c'est l'Etat qui décide.* [I think there is a separation of Church and State in France. I'm not shocked that the Church takes a stand, in the end it's the State that decides.]
- **inappropriate speech** (the comment contains vulgar language that is not directed at anyone in particular or this connection cannot be clearly established)
  - *conneries!* [bullshit!]

- **offensive speech** (the comment contains discriminatory and defamatory claims that target an identifiable entity and is potentially punishable by civil law (Vehovar et al., 2020))
  - *Et c'est vrai que le mec au milieu qui ressemble a papillon de Tchernobyl a dus prendre son déguisement dans la derniere gay pride !!!* [And it's true that the guy in the middle, who looks like a Chernobyl butterfly, must have gotten his costume at the last gay pride parade!!!]
- **violent/threatening speech** (the comment contains calls to physical violence or threats and is potentially punishable by criminal law (Fišer et al., 2017))
  - *Faut les écrasés ces mecs* [They need to be run over those guys]

Furthermore, *Offensive* comments and *Violent/threatening* comments are given an additional tag according to the type of discrimination. Two tags are available to this end: *Background* – when the basis for discrimination is the target's background or any of the protected characteristics, such as gender, sexual orientation, nation, race, ethnicity, religion, disability, etc.; and *Other* – when the basis for discrimination are the target's interests or professional group affiliations, such as political groups, media publishing houses, civil protection groups, etc. In case the annotator finds several SUD elements that indicate different SUD types, the harshest type of SUD is selected as the final category. See the guidelines published in the CLARIN.SI repository at http://hdl.handle.net/11356/1462 (Ljubešić et al., 2021) for more details.

The *Target* level of the annotation scheme consists of five categories which were applied in a hierarchical order, meaning that when a comment offended or attacked multiple targets, the comment was assigned the first relevant target from the list below:

- **migrants/LGBTQ**
  - *OH, pauvre petit Syrien....* [OH, poor little Syrian....]
- **supporters of migrants/LGBTQ**
  - *Qu'ils arrêtent de nous faire chier avec cette connerie qui ne concerne qu'une minorité.* [Let them stop bothering us with this nonsense that

only concerns a minority.] (as a response to a news item post about a pro-LGBT marriage activist)

- **journalist/media**
  - *20minutes@ liberté d'expression juste quand ça vous arrange !! mais bon être honnête ne vous tuera pas croyez moi, vous devrez l'essayer* [20minutes@ freedom of expression just when it suits you!! But well, being honest won't kill you, believe me, you should try it]
- **commenter** (used when the comment targets the other commenter but the pro or against position of the targeted commenter cannot be clearly established)
  - *On voit bien ton intelligence povre mek* [Your intelligence is clearly evident, poor guy]
- **other** (anyone else, including the opponents of migrants/LGBTQ)
  - *Mais l'ONU ne sert à rien !* [But the UN is useless !]

2.2 Annotation campaign

Similarly to the campaigns for other languages in FRENK, the annotation campaign for FRENK-fr included a training period, i.e., training sessions with an expert annotator from the FRENK project, written guidelines, and an annotation period with expert support which included a mailing list exchange and regular feedback sessions to resolve difficult cases. The comments were annotated in context, which means they were delivered to the annotators in entire threads and linked to the original post. It should be noted that the entire thread does not mean a complete string of comments ever posted in absolute terms, since it was only possible to collect the data that existed at the time of collection. Thus, the comments and posts removed by Facebook (or even the information about the number of such posts) could not be retrieved.

As with other languages, the potential bias in annotation was addressed not only with training sessions, but also with ensuring multiple annotations for each comment and the support of an expert annotator that provided regular feedback. While the Slovene and English data were annotated by approximately eight annotators (Ljubešić et al., 2019), the French data, similar to Dutch data (I. Markov et al., 2021), received two annotations, with all disputed cases resolved by an

expert annotator. The comments were annotated in a spreadsheet editor where they were listed in the original thread order and were linked to the relevant news item post and the information of their status as a reply or not. Each annotator worked independently of the other annotator, and needed around 100 hours each to complete the annotation task (this amount does not include the time needed to disambiguate the disputed cases, since this task was not timed).

A lower number of annotations per comment was due to difficulties in engaging the already trained annotators and lack of funding, but this annotation setting was deemed appropriate for two reasons. First, the student annotators were well familiar with the guidelines, the French annotation campaign being already their third or even fourth annotation task (after Slovene, English and Croatian); and second, all disputed cases were resolved by the expert annotator. The two annotators were paid for their work and were advanced students of French, but non-native speakers. To mitigate the potential influence of their lack of linguistic and sociohistorical knowledge on the interpretation of the comments, contact was established with a native French speaker for consultations when needed. While the sophistication of arguments in the comments was rather low, the main difficulty for comprehension was certainly the use of slang expressions, and occasional severe orthographic or syntactic errors that hampered the processing of the comment (which, however, is a common characteristic of the FRENK corpus).

The encountered difficulties in engaging the annotators warrant a short note. Annotating SUD proved to be a highly challenging task not only at the technical level with a complex annotation schema, but also at the psychological level, which was first anecdotally shown by relatively high withdrawal of annotators between the annotations campaigns, but later also confirmed empirically (Pahor de Maiti & Fišer, 2021). The dropout from one annotation campaign to the other. i.e., from Slovene to French, was of course influenced by the project team's judgment of annotator's performance, their self-reported level of linguistic knowledge and personal circumstances not related to the annotation task as such, but the analysis on the annotator's perception of this particular work clearly showed that the annotation of SUD is psychologically burdening and can lead to adverse effects on the people involved, but can also negatively impact the task at hand if these aspects are not addressed.

2.3 Inter-annotator agreement

Table 1 shows inter-annotator agreement for the French dataset which was assessed with Krippendorff's Alpha. The calculations were performed with the statistical package K-Alpha Calculator (Marzi et al., 2024) and considered SUD *Type* as an ordinal variable and SUD *Target* as a nominal variable as per observations reported in Ljubešić et al. (2019). The coefficient (2 annotators, 20694 values) for SUD *Type* is 0.651 [CI (95%, 1000 iterations): 0.635, 0.668], and 0.634 [CI (95%, 1000 iterations): 0.618, 0.648] for SUD *Target*. The result is slightly below the threshold for moderate agreement (0.67) (Krippendorff, 2019), but low-agreement is expected in SUD annotation because of the complexity of the phenomenon (Waseem, 2016).

Table 1: Inter-annotator agreement scores for SUD *Type* and SUD *Target* by topic and combined with 95% confidence interval over 1000 iterations for FRENK-fr.

Subset	*SUD Type*	*CI*	*SUD Target*	*CI*
LGBT	0.504	[0.472, 0.538]	0.488	[0.462, 0.515]
Migrants	0.810	[0.786, 0.834]	0.811	[0.785, 0.834]
Combined	0.651	[0.635, 0.668]	0.634	[0.618, 0.648]

There is, however, a noticeable difference in the coefficient between the two topics in the French dataset with a lower agreement on the annotations in the *LGBT* subset. Ljubešić et al. (2019) observe a similar pattern for English, i.e., lower agreement in the *LGBT* subset, and suggest that this can be possibly explained by the fact that the three most frequent annotation combinations in the *LGBT* dataset account for 91% of the annotations (compared to 75–80% in other subsets) which increases the possibility for the agreement by chance.

Although this is a possible explanation which might be due to overly detached annotators, a highly plausible risk in annotation of distressing content (Pahor de Maiti & Fišer, 2021), this result might simply reflect the specifics of the annotation schema which makes certain combinations impossible and other more likely. For instance, the *Acceptable speech* label can only be paired with *No target* label; *Background offensive* is most naturally paired with *LGBT/Migrants* as the target, and *Other offensive* is very likely to be paired with *Other* as a target. In FRENK-fr, the three most frequent *Type-Target* combinations represent 86% of all annotations in the *LGBT* subset and 89% in the *Migrants* subset, and in

both subsets include *Acceptable speech – No target*, *Other Offensive – Other* and *Background offensive – LGBT/Migrants*. These combinations are expected both in terms of *Type-Target* pairs, but also in terms of their frequency, since violent comments are more likely to get removed by the platform. Therefore, the observed distribution might be a reflection of actual dataset characteristics rather than due to the chance.

In comparison to Slovene or English, as reported in Ljubešić et al. (2019), the outlier seems to be the high score for the French *Migrants* subset which shows high agreement both for *Type* and *Target*, whereas the French *LGBT* subset and other languages exhibit low to moderate agreement. Although the three most frequent annotation combination in the *Migrants* subset account for a large proportion (89%) of the annotations, this score can be, as with the *LGBT* dataset, rather than by chance, explained by annotators' experience, more easily interpretable comments (given their relative shortness, see Section 3), but also the annotators' personal convictions which might have been more aligned with regard to rights of migrants than that of the LGBT community. This also supports the observation that in French as well as in Slovene and English, the agreement is higher for the *Migrants* subset.

3 KEY CHARACTERISTICS OF FRENK-fr

This section quantitatively describes the structure of FRENK-fr focusing first on the overall size of the dataset and then more specifically on the *Type* and *Target* distribution. Table 2 indicates the number of Facebook posts posted by media outlets, the number of comments and tokens. FRENK-fr contains comments in entire threads posted under 66 posts made by the three selected media outlets on Facebook (see Section 2). The corpus consists of more than 10,000 annotated comments with around 300,000 tokens which all relatively evenly cover both topics. The number of comments is comparable to the Croatian (10,970 comments), Dutch (10,732 comments), English (11,661 comments) and Slovene dataset (10,164 comments) (Ljubešić et al., 2021; I. Markov et al., 2021).

Table 2: The number of media outlet posts, comments and tokens per topic in FRENK-fr.

	LGBT	*Migrants*	*Total*
Posts	31	35	66
Comments	5,182	5,057	10,239
Tokens	172,418	128,426	300,844

3.1 Type of socially unacceptable discourse

Table 3 gives information about the distribution of comments by their SUD *Type* in both topics. The data show that around two thirds of comments in FRENK-fr contain *Acceptable* discourse, and one third SUD with around 1% of the content representing *Violent* propositions. A low number of *Violent* comments observed is expected given the generally low share of violent comments on social media. The estimations vary, but are usually below 7% of the observed dataset (Berglind et al., 2019; Vidgen & Yasseri, 2020), and are probably the result of platform moderation efforts and of societal pressure which makes publishing violent content less appropriate.[4]

Table 3: The absolute and relative number of comments in FRENK-fr per topic reflecting the distribution of SUD *Types*.

	LGBT		*Migrants*		*Total*	
	#	*%*	*#*	*%*	*#*	*%*
Acceptable	3,476	33.95	3,692	36.06	7,168	70.01
Inappropriate	24	0.23	19	0.19	43	0.42
Offensive	1,664	16.25	1,248	12.19	2,912	28.44
–*Background*	437	4.27	308	3.01	745	7.28
–*Other*	1,227	11.98	940	9.18	2,167	21.16
Violent	18	0.18	98	0.96	116	1.13
–*Background*	3	0.03	83	0.81	86	0.84
–*Other*	15	0.15	15	0.15	30	0.29
Total	5,182	50.61	5,057	49.39	10,239	100.00

[4]See for example Facebook community standards on hate speech, retrieved May 10, 2024, from https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/

Table 4: The absolute and relative number of tokens in FRENK-fr per topic reflecting the distribution of SUD *Types*.

	LGBT		*Migrants*		*Total*	
	#	*%*	*#*	*%*	*#*	*%*
Acceptable	104,146	34.62	78,473	26.08	182,619	60.70
Inappropriate	579	0.19	513	0.17	1,092	0.36
Offensive	67,216	22.34	47,796	15.89	115,012	38.23
–*Background*	16,417	5.46	11,656	3.87	28,073	9.33
–*Other*	50,799	16.89	36,140	12.01	86,939	28.90
Violent	477	0.16	1,644	0.55	2,121	0.71
–*Background*	56	0.02	1,289	0.43	1,345	0,45
–*Other*	421	0.14	355	0.12	776	0.26
Total	172,418	57.31	128,426	42.69	300,844	100.00

Although a 30% share of SUD is much smaller compared to, for example, the Slovene dataset where it amounts to a worrying 50% of all comments (Pahor de Maiti et al., 2020), this is still a clear indicator that an important share of discriminatory speech online is tolerated and not recognized as problematic enough to be reported and consequently removed from the platform. Furthermore, this observation is interesting from an interlingual perspective: we see that in the FRENK corpus, the French comments contain less SUD than Slovene, and the same is true of English comments when compared to Slovene comments in FRENK (Ljubešić et al., 2019), as well as to immigration-related online posts in Polish (Lewandowska-Tomaszczyk, 2017) or Cypriot Greek (Baider & Kopytowska, 2017). A contributing factor could be that in some of these studies, non-native speakers were employed for the annotation, and so their comprehension of comments might have lacked in precision which could have led them judge comments more or less harshly than a native speaker would. Nonetheless, a certain consistency across languages points towards an interesting avenue for future work since it is not clear whether this is just a result of potential annotation bias, more efficient algorithms for content moderation or if it indeed highlights a trait of communication culture linked to speakers of these languages.

Topic-wise, FRENK-fr contains a rather evenly distributed number of comments on the topic of LGBT and migrants which is a consequence of the data selection

process. The interesting point, however, is that commenters post slightly more SUD comments on the LGBT topic than on migrants topic which is different from, for example, Slovene and English data where we can observe the opposite situation (Ljubešić et al., 2019). The *Violent* comments, however, are in all three languages more frequent in the *Migrants* subset. In FRENK-fr, in particular, we see five times more violent SUD in comments discussing migrants-related issues as opposed to those discussing LGBT-related issues.

This result is possibly impacted by the choice/availability of the media outlet posts, since certain subtopics related to LGBT or migrants trigger more disagreement then other subtopics, but it is also likely that it shows a lower level of tolerance of French commenters in FRENK toward LGBT issues in comparison to migrants. It should be noted, however, that this result cannot be interpreted as a clear indicator of more prominent discrimination of the LGBT community alone, since the topic subset of data includes SUD comments targeting different individuals or groups (see Section 3.2).

Table 4 provides information on the shares of tokens per topic. Length-wise, based on the comparison of the number of comments (see Table 3) and tokens produced (see Table 4) and their median value for the number of tokens (indicated in brackets), *Offensive* (26) comments appear to be longer than *Acceptable* (14) and *Violent* (13) ones, and comments on the LGBT topic (20) tend to be longer than those related to migrants (15). The two-tailed Mann-Whitney U statistics, evaluating whether the comment length differs between the *Offensive* and *Violent* comments on the one hand, and between *LGBT-related* and *migrants-related* comments on the other hand, is statistically significant in both cases at medium effect size (*LGBT* vs. *MIGR* subset: U = 15138751.0, p = .00 (3.07e-42), LGBT = 5,182, MIGR = 5,057, r = 0.135; *Violent* vs. *Offensive* comments: U = 240941.0, p = .00 (6.03e-15), Offensive = 2,912, Violent = 116, r = 0.142).[5] A statistically significant difference in the length of *Violent* and *Offensive* comments was also observed in the Slovene FRENK (Pahor de Maiti et al., 2020), and suggests a certain complexity inherent to *Offensive* comments. Based on qualitative analyses of Slovene data (Pahor de Maiti et al., 2023), it seems that in the *Offensive* comments, the commenters especially often use

[5]Calculated with the SciPy library created by Virtanen et al. (2020).

face-saving strategies in order to preserve their social image in view of possible damage brought about by their discriminatory propositions.

Long comments are less frequent in the *Violent* and *Acceptable* comments since there is less need to save face (Brown & Levinson, 1987). By posting an acceptable, i.e., a neutral/positive message, a person's public image is not in danger, while those commenters that post violent comments, which usually consist of short calls or allusion to violent actions, and thus intentionally break societal norms, do so by a lack of concern for their face, or more likely, to strengthen their position in a selected in-group. Authors of offensive comments, on the other hand, often attempt to present their views in an elaborate manner, and although they may genuinely be trying to address the complexity of the issue at hand, these does not override their discriminatory content. However, due to their complexity, such comments may give the impression of solid and acceptable argumentation, allowing them to be perceived as legitimate contributions to democratic discourse, despite being fundamentally socially unacceptable. Given that LGBT-related comments tend to be longer and possibly more complex than migrants-related ones, this might also contribute to the explanation why the LGBT-related comments might appear more difficult for the annotators, and hence a lower inter-annotator agreement for this topic across languages.

3.2 Target of socially unacceptable discourse

Table 5 provides information about the distribution of comments by their SUD *Target* in both topics. The shares are given for each topic separately. The data is provided only for the *Offensive* and *Violent* subset of data, since the *Acceptable* and *Inappropriate* comments did not receive the *Target* label.

We see that both *LGBT* and *Migrants* subset exhibit a similar distribution of SUD *Targets* across the comments. Namely, the topic-related minority group (LGBT or migrants) and their supporters represent the main SUD *Target* referenced in slightly more than 40% of the comments. Although this observation is not unique to FRENK-fr, and is in fact less pronounced than, for instance, in the Slovene or Croatian FRENK where the minority groups and their supporters are targeted in 50–70% of SUD comments (Pahor de Maiti, n.d.), we can see this as a clear indicator that more needs to be done to limit SUD aimed at

vulnerable social groups in order to prevent the spread of discriminating ideas often advocated as free speech.

Table 5: The relative and absolute number of comments in FRENK-fr per topic and SUD Type reflecting the distribution of SUD *Targets*.

		Offensive	*Violent*	*Total*	
		%	%	%	#
LGBT	LGBT/related	44.05	0.59	44.65	751
	Other/opponent	24.02	0.42	24.44	411
	Commenter	21.34	0.06	21.40	360
	Media	9.51	/	9.51	160
MIGR	Migrant/related	35.14	6.24	41.38	557
	Other/opponent	29.12	0.89	30.01	404
	Commenter	21.77	0.15	21.92	295
	Media	6.69	/	6.69	90

Although SUD aimed at the minority groups is prevailing, an important share of SUD comments is also aimed at other individuals, namely the opponents and other commenters (20–30% each). This suggests that measures are needed at the level of general public in order to promote social cohesion and limit the back and forth of SUD comments that only aggravate the atmosphere. Moreover, migrants-related discussion in FRENK-fr appears slightly more polarizing for the French-speaking commenters then the LGBT-related topic which has a two times larger difference between comments targeting the minority group versus those targeting the opponents (approx. 10-point difference in the *Migrants* subset, and 20-point difference in the *LGBT* subset). In comparison to other languages, however, the French dataset appears more polarized than, for example, the Slovene and Croatian dataset where the commenters tend to be much more united in producing SUD mainly against the minority groups and their supporters (e.g., the Croatian dataset contains 70% of the comments targeting LGBT/supporters and only less than 10% of the comments targeting the opponents) (Pahor de Maiti, n.d.).

Media/journalists, on the other hand, are referenced in a relatively small amount of the comments, namely in around 8% of the French SUD comments with a higher count of such comments in the *LGBT* subset. This is still a higher share than in the Slovene or Croatian dataset where SUD targeting media/journalists amounts to around 5% of all comments, and is more frequent in the *Migrants*

subset (Pahor de Maiti, n.d.). Therefore, this shows that the commenters more frequently express their dissatisfaction with the work of journalists in French than in Slovene or Croatian, and are more critical in the case of reporting on the topic of LGBT. Furthermore, despite the small frequency of the comments targeting media/journalists, and although expressing critique can, in general, be a positive thing, our observation pertains to disrespectfully-communicated criticism which can be much more damaging to journalists and the general perception of journalistic integrity than an argumented opposition, and should, therefore, be limited (Č. Markov & Đorđević, 2024).

On the positive note, however, media/journalists are never targeted in FRENK-fr with violent propositions. These are, as observed in Section 3.1, mostly found in the *Migrants* subset where they are in the great majority of cases aimed at migrants and their supporters. The same observation can be made for the Slovene dataset, but not for the Croatian one where the LGBT community and their supporters are the central group receiving violent verbal attacks. This indicates also a cultural difference in the perception of the two minority groups, and shows French commenters in FRENK as more violently intolerant towards migrants. Opponents and other commenters are less often targeted in violent comments both in the *Migrants* subset as in the *LGBT* subset, where *Violent* comments are altogether rather rare (around 1% of data). Unsurprisingly, the violent comments are especially rare in direct addresses of other commenters since the level of othering and dehumanization is usually higher for an external group of the discussion, like migrants, than for one of the active discourse participants.

4 CONCLUSION

This paper presented the creation process of the French part of the FRENK corpus of socially unacceptable comments, and its characteristics. FRENK-fr includes comments posted as a reaction to news posts related to the topics of LGBT/homophobia and migrants/Islamophobia which were published by well-known national media outlets on Facebook. These comments were manually annotated according to the FRENK project-specific schema and contain the type of socially unacceptable discourse, varying from acceptable to inappropriate, offensive and violent content, and the target of socially unacceptable

comments which included the topic-related minority groups (LGBT, migrants), their supporters and opponents, other commenters and journalists.

The main aim of the creation of FRENK-fr was twofold: first, obtaining data for a qualitative comparative analysis of socially unacceptable communication practices on Facebook, and second, expansion of the FRENK corpora wit a new language for the needs of the development of a robust and generalizable model for data classification and hate speech detection.

The here presented FRENK-fr is a dataset that is fully comparable to other language-specific parts of the FRENK corpus, i.e., Croatian, English, Slovene and Dutch, since the data collection and filtering process, data size, as well as the annotation schema and annotation procedure were adopted from the Slovene FRENK which was created first. The main difference compared to the Croatian, English and Slovene, but not the Dutch FRENK dataset, is in the number of the annotations per comment. This means that the final annotation in the case of Croatian, English and Slovene was calculated as the mode of all the annotations, while in for Dutch and French, the final annotations are either the input of two annotators in the case of agreement, or the expert annotator label in the case of dispute. The inter-annotator agreement assessment shows similar scores across languages.

FRENK-fr is comparable to other languages both in the selection of media outlets as well as the relevance of the period covered. The selection of media sources, encompassing mainstream liberal, conservative and a more sensationalist outlet, is consistent across both French and other languages. Furthermore, the time frame of the comments coincides with similar societal changes across all languages/countries.

In particular, the 2015–2017 period was a time of an increased migration flow from the conflict zones in the Middle East and Africa that spread across Europe. France, like the other countries, faced difficulties processing asylum applications, had problems solving issues inside refugee camps, and faced increased intolerance towards the Muslim population in the country. Regarding the LGBT issues, all countries saw the organisation of the Pride Parade and other LGBT-related awareness-raising but also anti-LGBT events which received important media coverage, but also witnessed different campaigns advocating for legislative changes to address LGBT-discriminatory laws (e.g., *mariage pour*

tous [marriage for all] campaign in France, or *čas je za* [time for yes] in Slovenia, advocating for the legalization of the same-sex marriage).

The analysis of the FRENK-fr annotation campaign showed that similarly to other language-specific FRENK datasets, the LGBT topic appears more difficult for the annotators which resulted in a lower inter-annotator agreement on that topic. FRENK-fr is, in general, characterized by a lower number of SUD comments compared to some of the other languages included in the FRENK corpus, with 30% of the comments representing SUD, of which only 1% is labeled as *Violent* speech. Furthermore, we observed that both topics attract a similar amount of SUD comments, and that *Offensive* comments tend to be longer than *Acceptable* or *Violent* ones, and LGBT-related comments longer than migrants-related comments. Target-wise, SUD comments are most often aimed at the two topic-specific minority groups, migrants and LGBT, and their supporters. With a lower, but not negligible share, SUD comments also target opponents of minority groups and other commenters. Violent comments, in particular, are in most cases aimed at migrants.

These results but also public reports on the current situation of communication culture online, clearly indicate that there is still much work to do in order to promote inclusive behaviour in public online environments. FRENK-fr can importantly contribute to this objective. It expands the FRENK corpus of socially unacceptable discourse to French, and because of its corpus creation design, enables fully comparable analyses with the other parts of the FRENK corpus. As such, it represent a highly valuable resource for inter-cultural qualitative analyses of disrespectful communication practices online that can inform the actions of the civil society and political institutions, but also provides crucial material for training language models created for the classification of socially unacceptable discourse.

5 ACKNOWLEDGMENTS

This work has been supported by the Slovenian-Flemish bilateral basic research project "Linguistic landscape of hate speech on social media" (ARRS N06-0099), Digital Humanities research programme (Digital Humanities: resources, tools and methods; ARRS P6-0436) and the ARENAS project (EU Horizon Europe research and innovation programme; GA No. 101094731).

REFERENCES

Baider, F., & Kopytowska, M. (2017). Conceptualising the Other: Online discourses on the current refugee crisis in Cyprus and in Poland. *Lodz Papers in Pragmatics*, *13*(2), 203–233.

Berglind, T., Pelzer, B., & Kaati, L. (2019).Levels of Hate in Online Environments. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 842–847).

Brown, P., & Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage* (Vol. 4). Cambridge University Press.

Carneiro, B. M., Linardi, M., & Longhi, J. (2023). Studying Socially Unacceptable Discourse Classification (SUD) Through Different Eyes:" Are we on the same page?". *arXiv preprint arXiv:2308.04180*.

Chiril, P., Moriceau, V., Benamara, F., Mari, A., Origgi, G., & Coulomb-Gully, M. (2020). An Annotated Corpus for Sexism Detection in French Tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1397–1403).

Despot, K. Š., Anić, A. O., & Veale, T. (2023). "Somewhere Along Your Pedigree, a Bitch Got Over the Wall!" A proposal of Implicitly Offensive Language Typology. *Lodz Papers in Pragmatics*, *19*(2), 385–414.

Fišer, D., Erjavec, T., & Ljubešić, N. (2017). Legal Framework, Dataset and Annotation Schema for Socially Unacceptable Online Discourse Practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 46–51).

Kopytowska, M., & Baider, F. (2017). From Stereotypes and Prejudice to Verbal and Physical Violence: Hate Speech in Context. *Lodz Papers in Pragmatics*, *13*(2), 133–152.

Kralj Novak, P., Mozetič, I., & Ljubešić, N. (2021). *Slovenian Twitter Hate speech dataset imsypp-sl*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1398

Krippendorff, K. (2019). *Content analysis: An introduction to its methodology (4th Ed.)*. Sage publications. https://doi.org/10.4135/9781071878781

Lewandowska-Tomaszczyk, B. (2017). Incivility and confrontation in online conflict discourses. *Lodz papers in pragmatics*, *13*(2), 347–367.

Ljubešić, N., Fišer, D., & Erjavec, T. (2019). The FRENK datasets of socially unacceptable discourse in Slovene and English. In *Text, Speech, and Dialogue: 22nd International Conference, TSD 2019, Ljubljana, Slovenia, September 11–13, 2019, Proceedings 22* (pp. 103–114).

Ljubešić, N., Fišer, D., Erjavec, T., & Šulc, A. (2021). *Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.1*. Slovenian language resource

repository CLARIN.SI. http://hdl.handle.net/11356/1462

López, C. A., & López, R. M. (2017). Hate Speech in the Online Setting. In S. Assimakopoulos, F. H. Baider, & S. Millar (Eds.), *Online Hate Speech in the European Union – A Discourse-Analytic Perspective* (pp. 10–12). Springer.

Markov, Č., & Đorđević, A. (2024). Becoming a target: Journalists' perspectives on anti-press discourse and experiences with hate speech. *Journalism Practice*, *18*(2), 283–300.

Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the eleventh workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 149–159).

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *MethodsX*, *12*, 102545.

Pahor de Maiti, K. (n.d.). *Metaphor in socially unacceptable discourse online [doctoral dissertation]*. University of Ljubljana.

Pahor de Maiti, K., Fišer, D., & Erjavec, T. (2020). Grammatical footprint of socially unacceptable facebook comments. *Language Technologies & Digital Humanities*.

Pahor de Maiti, K., & Fišer, D. (2021). Working with socially unacceptable discourse online: Researchers' perspective on distressing data. In *Proceedings of the 8th Conference on CMC and Social Media Corpora for the Humanities (CMC-Corpora 2021)* (pp. 78–82).

Pahor de Maiti, K., Franza, J., & Fišer, D. (2023). Haters in the spotlight: gender and socially unacceptable Facebook comments. *Internet Pragmatics*, *6*(2), 173–196.

Pahor de Maiti, K., Ljubešić, N., & Fišer, D. (2024). *Offensive language dataset of French comments FRENK-fr 1.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1947

Piot, P., Martın-Rodilla, P., & Parapar, J. (2024). MetaHate: A Dataset for Unifying Efforts on Hate Speech Detection. *arXiv preprint arXiv:2401.06526*.

Vanetik, N., & Mimoun, E. (2022). Detection of racist language in french tweets. *Information*, *13*(7), 318.

Vehovar, V., Povž, B., Fišer, D., Ljubešić, N., Šulc, A., & Jontes, D. (2020). Družbeno nesprejemljivi diskurz na Facebookovih straneh novičarskih portalov. *Teorija in praksa*, *57*(2), 622–645.

Vidgen, B., & Yasseri, T. (2020). Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, *17*(1), 66–78.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C.J., ... & Contributors, S. (2020). SciPy 1.0.: Fundamental algorithms for

scientific computing in python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138–142).

Zhang, Z., & Luo, L. (2019). Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, *10*(5), 925–945.

RAZŠIRITEV KORPUSA DRUŽBENO NESPREJEMLJIVEGA DISKURZA FRENK NA FRANCOŠČINO

Ta prispevek predstavlja francoski del korpusa FRENK, ki zajema družbeno nesprejemljive komentarje, napisane kot odziv na novice s tematiko LGBT in migracij, ki so jih na Facebooku objavile priljubljene medijske hiše. Ti komentarji so bili ročno označeni glede na vrsto in tarčo družbeno nesprejemljivega govora. Od 10.239 komentarjev z okoli 300.000 pojavnicami v korpusu FRENK-fr je približno tretjina komentarjev označena kot družbeno nesprejemljiva, od tega pa le 1% predstavljajo nasilni komentarji. Ti so najpogosteje usmerjeni proti migrantom, ki skupaj z LGBT skupnostjo in njihovimi podporniki predstavljajo napogostejšo ciljno skupino družbeno nesprejemljivih komentarjev. Korpus FRENK-fr je v celoti primerljiv z drugimi jezikovno-specifičnimi deli korpusa FRENK in lahko služi kot pomemben vir za medkulturne kvalitativne analize nespoštljive komunikacije na spletu, ki lahko podajo pomembne uvide za oblikovanje ukrepov na ravni civilne družbe in političnih institucij. Poleg tega korpus FRENK-fr zagotavlja tudi kakovostne podatke za treniranje jezikovnih modelov, namenjenih za prepoznavanje družbeno nesprejemljivega diskurza.

Keywords: družbeno nesprejemljivi diskurz, sovražni govor, francoščina, migranti, LGBT