Attitudes, Communicative Functions, and Lexicogrammatical features of Anti-Vaccine Discourse on Telegram

Souad Boumechaal, Serge Sharoff,† University of Leeds

Corresponding author: † s.sharoff@leeds.ac.uk

Abstract

This paper reports the process of collecting a corpus with examples of anti-vaccine discourse and the results of its linguistic analysis. The overall aim of the project is to help public health authorities to improve their communication campaigns by better understanding the conditions for misinformation spreading via social media. More specifically, this paper analyses linguistic properties of a corpus of prominent misinformation channels in Telegram as compared against a more general COVID corpus as well as against a general purpose English corpus. For this paper, the quantitative analysis relies on corpus querying to identify the most recurrent discourse patterns related to COVID vaccines. We use the appraisal framework to analyse the patterns with respect to the attitudes conveyed in the messages. We have also applied an automatic AI classifier to predict communicative functions of these texts. This allows us to examine them more closely through the use of simple lexicogrammatical features following Biber, as well as their ideational processes following Halliday. The findings show that common collocations in the Telegram corpus containing misinformation draw on three attitudes: fear, insecurity, and mistrust in COVID vaccines which are discursively constructed to promote vaccine hesitancy among social media users. Furthermore, the misinformation messages tend to occur more often in such communicative functions as promotional texts, news reporting, and text expressed as presenting reference information.

Acknowledgements

Project has been funded via the EPSRC IAA3125 grant: "AI tracing tools for detecting COVID-19 misinformation" in collaboration with the Yorkshire&Humber Academic Health Sciences Network and the British National Health Service.

Attitudes, Communicative Functions, and Lexicogrammatical features of Anti-Vaccine Discourse on Telegram

Abstract

This paper reports the process of collecting a corpus with examples of anti-vaccine discourse and the results of its linguistic analysis. The overall aim of the project is to help public health authorities to improve their communication campaigns by better understanding the conditions for misinformation spreading via social media. More specifically, this paper analyses linguistic properties of a corpus of prominent misinformation channels in Telegram as compared against a more general COVID corpus as well as against a general purpose English corpus. For this paper, the quantitative analysis relies on corpus querying to identify the most recurrent discourse patterns related to COVID vaccines. We use the appraisal framework to analyse the patterns with respect to the attitudes conveyed in the messages. We have also applied an automatic AI classifier to predict communicative functions of these texts. This allows us to examine them more closely through the use of simple lexicogrammatical features following Biber, as well as their ideational processes following Halliday. The findings show that common collocations in the Telegram corpus containing misinformation draw on three attitudes: fear, insecurity, and mistrust in COVID vaccines which are discursively constructed to promote vaccine hesitancy among social media users. Furthermore, the misinformation messages tend to occur more often in such communicative functions as promotional texts, news reporting, and text expressed as presenting reference information.

1 Introduction

One of the manifestations of the COVID-19 pandemic is the phenomenon of ‘infodemic’, i.e., massive circulation and flow of rumours (Chou et al., 2021). While the definition of what is considered to be true or false changes with acquiring more information about COVID-19, rumours circulating through social media can undermine the activities of the public health authorities to limit the spread of the disease. This paper adopts the view of the British National Health Service (NHS), so that potentially harmful rumours are considered as misinformation, even though they might be shared with no intention to mislead. Hence in this paper we contrast this notion with a more specific concept of disinformation, which assumes creating and spreading deliberately false content (Stahl, 2006). For example, generally speaking it is very difficult to verify the truth status of a message about a close relative who died from vaccination. However, its wide circulation in social media is considered to be harmful to the vaccination efforts of the NHS, and this is also a very unlikely event from the statistical viewpoint (Hause et al., 2022). The public health authorities in the NHS do not deny the existence of vaccine side effects, but the undue focus on the vaccine side effects without acknowledging the positive value of vaccines to society as a whole is not desirable, so such rumours are referred to as misinformation in the public health discourse (Chou et al., 2021). We keep the same label in our study.

Research on analysing misinformation along with more specific deliberate disinformation is found to be gradually expanding after 2017 (Freelon and Wells, 2020). The proliferation of COVID-19 misinformation on social media platforms has often been traced back to demographic and social factors such as users' education level, beliefs, and political orientation (Osmundsen et al., 2021). Although these factors highlight crucial aspects that underlie the spread COVID-19 rumours, there is insufficient evidence to claim their role in spreading misinformation and the general trust in science (Agley and Xiao, 2021).

While there can be an interesting argument on the nature of public health communication and the reasons for its rejection by the public, as well as on the algorithmic manipulation of social media streams, the remit of this paper is purely linguistic. Here we present results of linguistic analysis of messages arguing against COVID-19 vaccination. The main aim is to understand their characteristics and the conditions for their spread via social media, even when better quality information from reputable sources is available in various forms. We address the following research questions by focusing specifically on the language against COVID-19 vaccinations in selected Telegram channels:

- What are the discourse patterns for expressing the attitudes to vaccination in this corpus?
- What is the distribution of communicative functions in messages against COVID-19 vaccinations and what are the corresponding lexicogrammatical features?
- How do attitudes and communicative functions contribute to the spread of misinformation about COVID-19 vaccinations?

In this paper we will describe:

- our corpus, including the process of its collection and annotation;
- methods for its automatic classification with respect to communicative functions of its texts;
- its lexical analysis with respect to the appraisal framework;
- its analysis with respect to lexicogrammatical features.

In the interests of Open Research, the annotated datasets, the scripts and the AI classification models are available under permissive licenses.[1] The corpora are also freely available for querying via Intellitext (Wilson et al., 2010). [2]

2 Related research

The COVID-19 pandemic has demonstrated how social media platforms can be a double-edged communication tool. While they played a key role in keeping people socially connected, the massive increase in COVID-19 misinformation put vulnerable people at higher risks and created panic and anxiety among all communities (Chou et al., 2021). Several recent studies have explored the amplification of misinformation on social media with the aim to understand the factors behind their rapid spread and the reasons behind trusting such content. For example, Memon and Carley (2020) investigate the nature of social networks characterising social media users who share misinformation and compare them to those who are well informed and share true information (Memon and Carley, 2020). Their findings demonstrate that the former tends to be denser and more structured than the latter as the circulated misinformation was part of disinformation campaigns. Interestingly, their findings covered the linguistic style used in misinformation posts. Misinformed users were found to use less narratives and more analytical language compared to well-informed communities. However, their linguistic analysis was limited to narrative style and did not consider other communicative functions of Twitter posts, such as those aimed at achieving persuasive effect. Other studies have attempted to investigate sociolinguistic factors driving the spread of misinformation. For example, Huang and Carley (2020) focus on the characteristics of Twitter users sharing misinformation such as locations, social identities, and political orientation (Huang and Carley, 2020) . Their study revealed that both regular users and bot accounts play a crucial role in spreading fake content about COVID-19 and these mainly appeared to be generated from the US, UK and Philippines.

Although similar studies have highlighted social, political and stylistic characteristics detected in misinformation posts shared by the users, there is still a need to better understand the language characterising

[1] https://github.com/ssharoff/covid19misinfo

[2] http://corpus.leeds.ac.uk/it/

Table 1: COVID corpora from social media

#K **messages**	#M **Words**	**Mean length**	**Source**
62,976	639.49	10.16	Twitter, General, 08/2020 – 07/2021
51	1.33	26.08	Twitter, misinformation , 08/2021 – 03/2022
2	0.07	34.98	Facebook, misinformation, 12/2021 – 03/2022
421	14.27	33.90	Telegram, misinformation, 10/2021 – 03/2022

COVID-19 misinformation using mixed methods with more advanced quantitative methodologies. The approach of corpus linguistics has proven successful in investigating change and patterns in language through time. Studies such as (Baker et al., 2008) on discourses about refugees and asylum seekers in UK press and (Baker et al., 2013) on Muslim representations in the British press have demonstrated how corpus assisted discourse analysis can be utilised to understand social issues 'caused or exacerbated by public text and talk, such as various forms of social power abuse (domination) and their resulting social inequality' (Van Dijk, 2009, p. 63). Health issues have also become an area of scrutiny for many corpus linguists. In particular, recent corpus research studies have explored language patterns characterising healthcare communication. For example, (Baker et al., 2019) examined 29 million words of online patient feedback and 11 million words of responses to the feedback from the national health service providers. Their analysis does not merely give evaluation of the NHS service provision, but it was combined with critical discourse analysis to reveal the underlying institutional ideologies and practices from patients' perspectives. Similarly, Brookes et al (2022) use a range of corpora from adolescent websites and other newspaper articles to evaluate the representation of illnesses such as HIV, AIDS, and Dementia (Brookes et al., 2022). Their case studies analysed frequent collocations to understand how these health issues are discussed. Systemic-Functional Linguistics (SFL) provides a rich meta-language to analyse discourse, which led to such studies as (Lecompte-Van Poucke, 2022) which offers analysis of Facebook posts on the topic of endometriosis awareness with the aim of uncovering ideology behind the discursive moves. Another related study in the SFL tradition used a range of computational methods to investigate differences in the ideological positions behind the concept of austerity (Bateman and Paris, 2020).

Despite the plethora of research undertaken on health issues, COVID-19 is still an under-explored area that requires a combination of quantitative and qualitative methods. Our paper attempts to deploy corpus linguistic techniques to uncover the discourses shaping the language against COVID-19 vaccines in Telegram messages. We particularly draw on a functional approach to examine how specific text criteria can be put together to fulfil communicative and persuasive effects and appeal to the audience. This approach aligns with Martin and White's (2005) appraisal framework which is a sub-branch of Systemic Functional Linguistic with a focus on the interpersonal dimension between the text producer and the reader. Martin (2000, p. 143) argues that this framework allows the researcher to examine "how speakers can exploit different ranges of appraisal to construct particular personae for themselves". Within this framework, we will focus on the expression of attitudes towards COVID-19 vaccines in misinformation texts. We adopt Martin and White's (2005, p. 35) definition of attitudes as expressions "concerned with our feelings, including emotional reactions, judgements of behaviour and evaluation of things".

Table 2: Topical categories: manual annotation in Twitter and automatic prediction in Telegram

Common categories	Tweets	Telegram
Anti-vaccine	401	139592
COVID-19 Politics	310	91432
Conspiracy theories	155	112482
Unreliable Scientific Sources	90	12268
Unreliable Latest News	160	104625
Common Concerns	60	12941
Irrelevant to COVID-19	797	239026

3 Methodology for this study

3.1 Data scraping from social media

This paper is a report from a bigger project which is aimed at better understanding the conditions for information dissemination through development of AI classification models for such properties as the communicative function of messages or the age and gender of users forwarding them. This involves collecting COVID-related posts from several social media platforms, such as Facebook, Telegram and Twitter using the official APIs of the respective platforms.

In the first stage of the overall project we collected general tweets on the topic of COVID-19, following the framework of (Lamsal, 2020), which provided a set of hashtags and keywords for collecting COVID-relevant information. In the second stage of the project we focused on collection of misinformation from our three social media platforms by seeding our search with user accounts known to be prominent for publishing misinformation, according to the NHS Comms team. We added related accounts which (1) are active in sharing messages from the seed accounts or (2) are often shared by the seed accounts. Then, we started monitoring the expanded set of accounts via the respective APIs. Since both Facebook and Twitter are active in policing misinformation, while Telegram is more lenient in this respect, we ended up having considerably more messages from Telegram, see the message counts in Table 1. The focus of the project on channels known for spreading COVID-19 rumours implies that we have collected a relatively small subset of all channels talking about COVID-19.

In this table, the number of messages from each source is given in thousands, the corpus size in millions of words, the mean message length is given in words. The corpus size is the main reason we focus our linguistic analysis on the Telegram corpus in this paper. Also, the messages in Telegram are slightly longer in comparison to Twitter which can be also helpful for linguistic analysis, see Column 3 in Table 1.

3.2 Manual annotation of samples

Then we proceeded to manual annotation of a sample of messages to better understand the most common and recurrent topics. We focused on messages that were (1) about COVID, (2) not supported by scientific evidence and (3) potentially harmful to the NHS public information campaing. In this annotation we were guided in consultation with the partners of our project in the NHS Communications department. While the NHS do not offer a specific definition of misinformation or harmfulness, they offer public guidance on authoritative sources to prevent the spread of misinformation (NHS, 2023). The examples we use below

(1)-(21) provide family resemblance in Wittgensteinian terms (Gert, 1995) to the notion of harmfulness to public health communication.

The early annotation stages involved several discussions among the project members on how to code data with relevant annotation labels. We followed a bottom-up approach starting with an initial set of data to define the categories most useful to develop the automatic classifier. We started with the Twitter corpus because we had the Twitter pipeline ready before we started collecting Telegram data.

After obtaining a bigger sample of potential misinformation sources from Twitter, we manually annotated 1700 Tweets with respect to the topics identified at the early annotation stage. The recurrent concepts were grouped under topics that formed the categories. For example, such expressions as *big pharma, 5G, Bill Gates foundation, biological weapon* were common to the conspiracy theory category. Similarly, texts with harmful content and which exploit names of medical doctors in the context of unverifiable studies and websites were annotated under *Unreliable Scientific Sources*. To ensure the quality of the annotations, the project members discussed the annotations in several iterations and agreed to proceed with a multilabel annotation scheme in which each texts can be annotated with more than one label, as in the following example:

(1) Project veritas fourth video in its covid vaccine investigative series quotes two Pfizer scientists who said natural immunity is better than Pfizers vaccine,

As the source "Project veritas" lacks credibility and the intention against Pfizers vaccine is clearly expressed, we annotated this text as *Unreliable Scientific Sources* and *Anti-vaccine*.

Our bottom-up manual annotations eventually resulted in 7 categories, see Table 2, which lists the frequency of the topic categories from the manually annotated Twitter sample. Annotating this sample helped in development of annotation guidelines that define each category and examples of what each category covers. The guidelines were also followed in annotating a small sample of Telegram data. To train the automatic classifier, the same labels had to be used across different corpora.

After that we started exploring recurrent patterns from the three categories in our Telegram misinformation corpus. We started with the initial list of keywords identified via the log-likelihood score (Rayson and Garside, 2000) for the Telegram corpus against the general Twitter corpus as the reference. Then, we proceeded to manually exploring the context of their occurrence in concordance lines to verify to what extent these keywords refer to COVID-related topics in our misinformation corpora.

While analysing the concordance lines, we expanded the list with more keywors. The criteria we considered important to select a keyword were semantic links and frequency. First, the keyword has to be semantically and pragmatically relevant to the domain COVID-19 vaccines. This criterion also included named entities such as famous people and organisations, and locations, which regularly occur in anti-vaccination messages. The following examples illustrate the context in which the keywords are used:

(2) Governor destroys totalitarian *Biden Regime Vaccine Mandates*, Weaponized Treasure Department! use the constitution!

(3) S***** vaccines and *big pharma* drugs... This is what vaccines do to children ... this is what is happening to people behind closed doors ... little by little, more and more people are getting a *toxic dose*.

(4) US CDC admits that it has no record of an unvaccinated person spreading COVID after recovering from COVID and Bayer executive- mRNA shots are ‘gene therapy’ marketed as ‘vaccines’

As shown above, searches with words such as *vaccine mandates, vaccines* have led to highlighting further keywords such as *Biden Regime, big pharma, toxic dose, US CDC* these have served to expand the

Table 3: Keywords for the three most common topics in the Telegram misinformation corpus

Topics	**Keyword in the corpus**
Anti-vaccine	*booster, injections, jabs, mandates, transmission, forced, side effects, medical choice, free, freedom, humanity, resistance, control*
COVID-19 politics	*Democracy, Biden, Trump, US Government, measures, Democratic Party, Fauci*
Conspiracy	*agenda, plans, media, corporate, big pharma, new world order, natural immunity herd immunity, question everything, tyrants, shocking, harming, evil*

list of keywords. The second criterion relates to frequency. Given the size of the Telegram misinformation corpora, a keyword had to occur more than 3 times across the data to be considered crucial for the analysis.

This resulted in reliable manual annotation, which was then used to fine-tune a topical classifier to apply it to the entirety of our Telegram corpus, see the last column in Table 2 with the final list of the keywords per topic see Table 3.

We also searched for frequent collocations for these concepts starting with such nouns as *vaccine, agenda, jab* with a precise display of 1 collocate to the right or to the left of the entered keywords while restricting the part of speech of the collocation candidates (verbs, nouns and adjectives). The collocation lists were analysed for the most significant collocates according to the log-Dice score (Gablasova et al., 2017), while the respective concordance lines were checked to verify whether the collocations remain COVID-related, for example, most of the politics keywords do not lead to relevant collocations, even when the messages themselves are COVID-related.

Similarly to the Twitter corpus, the anti-vaccine topic is more common in the Telegram corpus that any other topic, so this provides better potential for further linguistic analysis. Therefore, this paper focuses on the representation of COVID vaccines in the Telegram corpus to identify the discourse patterns and attitudes common in such messages.

The full list of keywords for the qualitative analysis within the Anti-vaccine topic is as follows:

> *vaccine, jabs, forced, side effects, medical, choice, free, freedom, humanity, resistance, control, agenda, tyranny, evil, experimental, poison, death.*

We have also attempted contrastive analysis by comparing the frequency of collocations in the Telegram misinformation corpus, in our general COVID-19 corpus from Twitter and in a general-purpose Web corpus, more specifically, enTenTen (Jakubíček et al., 2013). The general Twitter corpus has been collected from the overall Twitter feed following the hashtags and keywords from (Lamsal, 2020). Our inspection shows that it contains only a small proportion of what is considered as misinformation according to the NHS Guidelines, while topically it is no different from the misinformation corpora. In contrast, the enTenTen corpus is much larger and it provides a reliable window into English language use, so it is used to check the general use of collocations with our topical keywords.

The next step analysed the collocations across the three corpora. We extracted the collocations for the selected topical keywords from each corpus (ranking them by the log-Dice score). The results of the queries were then filtered to focus on the collocations related to COVID vaccines. Below in section 4 we discuss the significant differences in collocation frequency between the corpora and the implications to our research.

3.3 Appraisal analysis of collocations

Our dataset shows that a large proportion of messages in the misinformation corpus expresses negative attitudes towards COVID vaccines. To explore the attitudes more closely, we relied on the appraisal framework with the theoretical underpinnings related to the Systemic-Functional Linguistics (Halliday and Matthiessen, 2004), as this is also relevant to our analysis of the communicative functions and the transitivity, see below. The appraisal framework provides a theory to study how a text can be constructed to position itself within the discourse and evoke solidarity among readers. The authors' positionings are situated within a socio-cultural context and are realised using appropriate lexicogrammatical resources since every 'utterance enters into processes of alignment or misalignment with others, helping us to understand the levels and types of ideological solidarity that authors maintain with their potential readers/listeners' (Oteíza, 2017, p. 457). Moreover, the appraisal framework allows a systematic study of the way how messages express the meanings to "establish solidarity between writers and readers legitimate certain positions and social values over others" (Oteíza, 2017, p. 469). The appraisal framework has three subsystems: Attitude, Engagement, and Graduation whereby each one focuses on a specific aspect in categorising evaluative language. The Attitude focuses on the speaker/writer's position and types of attitudes expressed, Engagement and Graduation concentrate more on the degree of alignment and strength or emphasis of the evaluated language respectively. Our initial analysis of data indicated the predominance of negative attitudes towards Covid-19 vaccines. Thus, focusing on the component of attitudes seemed crucial to understand the types of attitudes expressed in the messages. This subsystem offers a systematic framework to identify language patterns of attitudes towards COVID-19 vaccines. This in turn can give an insight on the way misinformation messages instil different attitudinal strategies to present a convincing narrative, gain credibility, and elicit support from readers.

Consequently, in this paper we will focus on the category of *Attitude* and its subcategories *Affect, Judgment* and *Appreciation*, which reflect the positive and negative attitudes as expressed in a text. They are sometimes expressed directly or evoked indirectly by implicit references and metaphoric expressions. Our analysis below follows Martin and White' (2005) framework which divides the attitudes expressed in texts using a system of opposition (dis/inclination, un/happiness, in/security, and dis/satisfaction), each of these is explicitly or implicitly expressed in messages. Secondly, the domain of Judgment involves Veracity (how truthful someone is) and Propriety (how ethical someone is), see (Martin, 2000). Finally, Appreciation can be divided into: our reactions to things (do they catch our attention? do they please us?), their composition (balance and complexity), and their value (was it worthwhile?), see (Martin and White 2005: 56).

Keeping this system as a guide, we attempt to map out Martin and White's (2005: 50-51) list of lexical items and their meaning with the examples from our Telegram corpus to establish the way specific attitudes are realised using the collocations most frequent in the corpus. By identifying the common adjectives, nouns, and verbs used with our list of frequent collocations and comparing these to Martin and White's (2005) list, we then grouped these collocations under three attitudinal themes: Fear, Insecurity, and Mistrust. Through these systems of Attitudes, we examine the structure of messages against Covid-19 vaccines by focusing on the choices of verbs and noun phrases. We present the results along with the discussion in section 4.

3.4 Analysis of communicative functions

In addition to the lexical analysis of collocations, we also analyse the distribution of communicative functions, such as academic research vs authoritative advice vs informative news reporting vs expressions of opinions. We have developed AI models for predicting various properties of the messages and the accounts of users which promote them, such as the presence of misinformation in a message, the generalised com-

Table 4: Cross-validation performance of our automatic classification model

	Precision	Recall	F1-score
A1.argument	0.65	0.85	0.74
A4.fiction	0.82	0.82	0.82
A7.instruct	0.79	0.76	0.77
A8.news	0.81	0.77	0.79
A9.legal	0.79	0.73	0.76
A11.personal	0.66	0.74	0.70
A12.promotion	0.84	0.89	0.86
A14.academic	0.78	0.83	0.80
A16.information	0.77	0.62	0.69
A17.review	0.88	0.72	0.79

municative function, for example, providing advice, reference information or argumentation.

To achieve this on our large corpora, we have fine-tuned a Deep Learning pre-trained transformer model for classification of communicative functions (Sharoff, 2021), as available through the HuggingFace transformer framework (Wolf et al., 2019). For fine-tuning we used the annotated data from (Sharoff, 2018). The F1 cross-validation scores of our fine-tuned model are listed in Table 4, this offers a moderate improvement to the original model from (Sharoff, 2021), which unlike our experiment did not use the specific Roberta model (Liu et al., 2019).

The fine-tuned model has been applied to the Telegram misinformation corpus to assess the distribution of communicative functions in this corpus. We also used the tools from (Sharoff, 2021) to link these functions to lexicogrammatical features initially introduced for describing register variation by Biber (1988). The features include such categories as:

Lexical features such as:

B05,[3] time adverbials = *afterwards, again, earlier, early, eventually, formerly, immediately,...*

H36, concessives = *although, nonetheless, though, tho*

L52, possibility modals = *can, may, might, could*;

L53, necessity modals = *ought, should, must*;

Part-of-speech (POS) features such as:

A03, present tense verbs

I42, adverbs

Syntactic features such as:

F18, passives with the explicit agent expressed by *by*;

I40, attributive adjectives;

N60, *that* deletions (for verbs which can express projection with *that*);

Text-level features such as:

J43, Type-Token Ratio (TTR)

J44, average word length

Each text in a corpus can be computationally represented as a vector of features obtained from the respective counts (normalised by text length if necessary). The values of features can be compared across

[3] The codes follow (Biber, 1988).

texts of respective communicative functions, for example:

A03	B05	F18	I42	L52	L53	N60	Function	Text
0.014	0.0084	0.0056	0.039	0.0000	0.00000	0.01404	Personal	Message (5)
0.008	0.0029	0.0029	0.028	0.0059	0.00147	0.00590	Academic	Message (6)

(5) I have witnessed the "Jab's" deleterious affects firsthand as some of my friends and family members have been harmed by it: A man in our community had Covid, then got the Vaxx- he has been sick for over a month now, with chronic fatigue, constant headaches and nausea- he hasn't had energy to work during this entire time…

(6) The COVID vaccines in all their variants, AstraZeca, Pfizer, Moderna, Sinovac, Janssen, Johnson & Johnson, etc., also contain a considerable dose of graphene oxide nanoparticles. This has been the result of their analysis by electron microscopy and spectroscopy, among other techniques used by various public universities in our country. Graphene oxide causes alteration of the immune system. By decompensating the oxidative balance in relation to the gulation reserves. If the dose of graphene oxide is increased by any route of administration, it causes the collapse of the immune system and subsequent cytokine storm…

The set of features used by Biber is fairly different from the grammatical framework by Halliday, which is based on such parameters as the functional types of processes (material, mental, etc), the properties of interpersonal exchange or thematic text development (Halliday and Matthiessen, 2004). However, each feature in the Biber set can be computed efficiently for a large corpus. The Biber set of features has been shown as being able to capture the parameters of variation in a number of other studies (Biber and Conrad, 2009), so we have applied the tools from (Sharoff, 2021) to our corpus.

4 Results and discussion

4.1 Comparison of Telegram against other corpora

Our analysis focuses on the Telegram corpus which we compare to two more corpora, our general COVID-19 Twitter corpus (which mostly consists of news and discussions with little misinformation) and the Sketch Engine enTenTen (using it as a general-purpose Web corpus). We used a collocation query to identify the strongest collocates with these keywords (ranking them by the log-Dice score). The result of the queries was then filtered to focus on the collocations related to COVID vaccines. To achieve this, we had to closely examine a sample of concordance lines for each noun phrase to understand how they are used to describe COVID vaccines. Table 5 lists the keywords that we first searched in the Telegram misinformation corpus to identify the recurrent noun phrases with their frequencies compared to the general COVID-related Twitter corpus and to enTenTen. To make the frequencies comparable, they are reported in terms of the instances per million words (ipm). For large corpora, even small ipm values mean a large number of examples, for example, *vaccine injury* has 249 examples in our Telegram corpus and 4,608 examples in enTenTen (often in the context of *The National Vaccine Injury Compensation Program* established in the US in the 1980s), so the log-likelihood keyness score (Rayson and Garside, 2000) is above the critical value of 10.83 ($p < 0.001$) for all cases in Table 5, for example, the log-likelihood keyness score for *vaccine injury* in the Telegram corpus against enTenTen is 1949.75. The 2020 version of enTenTen which we used has been extracted over the period of 2019-2020; it has some references to COVID-19, but proportionally far fewer than our corpora, for example, the frequency of COVID in enTenTen is 4.41 ipm as opposed to 488.94 ipm in our Telegram corpus.

Table 5: Common noun phrases and their ipm frequencies

Common phrases	**Telegram**	General Twitter	enTenTen
vaccine injury	17.66	1.47	0.11
vaccine damage	3.90	0.25	0.02
vaccine narrative	1.19	0.00	0.01
vaccine harm	2.04	0.00	0.01
vaccine apartheid	1.42	0.00	0.01
vaccine agenda	22.18	0.54	0.01
tyrannical **vaccine**	1.23	0.00	0.00
vaccine death	6.23	0.00	0.01
evil vaccine	0.57	0.14	0.01
evil agenda	2.42	0.34	0.02
evil government	0.52	0.21	0.02
experimental **vaccine**	5.95	0.71	0.06
poison vaccine	1.19	0.21	0.01
poisonous jab	0.80	0.00	0.01
death jab	15.14	0.00	0.01
death shot	7.09	0.00	0.01
free speech	10.80	3.71	4.14
free society	1.23	0.79	0.65
free choice	0.95	0.00	0.66
freedom fighter	10.52	1.15	0.66
freedom rally	2.42	0.05	0.02
freedom movement	4.57	0.08	0.26
mind **control**	8.71	1.70	1.15
government **control**	2.61	1.88	0.83
controlled population	2.38	0.62	0.09
medical apartheid	10.42	0.00	0.01
medical tyranny	8.80	0.00	0.01
medical fascism	1.23	0.00	0.01

With respect to the COVID vaccines, the Telegram corpus shows several misinformation themes, which relate to 1) dangers from the vaccines, 2) compulsory vaccine mandates, and 3) conspiracy stories about the hidden agenda behind the vaccines. We analyse these themes in subsection 4.2 below.

In comparison to the general Twitter corpus and enTenTen, the collocations of *vaccine* in the Telegram corpus are much more frequent. For example, the collocations of *vaccine* with *narrative, harm, apartheid* and *death* do not occur at all in the general COVID Twitter corpus and are found less than 0.01 in enTenTen. In such cases, their use in the general corpus does not seem to be related to the fear of vaccines, for example, *vaccine apartheid* in enTenTen only refers to the lack of vaccines in the developing world, rather than to the differences in the access rights for vaccinated and unvaccinated persons, as how it is most commonly used in the Telegram corpus. The rest of the keywords *experimental, free, freedom, medical* and *control* do share some similar noun phrases as shown in Table 5, while they are used in very different contexts in comparison to the misinformation corpus.

When viewed in the opposite direction, the common collocations with *vaccine* in enTenTen, are general and not related to COVID, but to other diseases, such as rubella, ebola .... etc. However, nouns collocating with *control, freedom, free, poison, experimental* are similar to the nouns identified in the Telegram misinformation corpus, while they are less frequent. Moreover, we could not identify any collocations with *medical, evil* that are relevant to vaccines. These two keywords are used in enTenTen with nouns from other areas, i.e., *medical care, treatment, condition* or *evil spirits, deeds, empire*.

4.2 Analysis of attitudes in noun phrases

In the appraisal framework, the system of Attitude focuses on how emotions are constructed through language to negotiate ‘solidarity between writers and readers‘ (Oteiza, 2017: 469). This negotiation forms the basis for legitimising certain discourses around COVID vaccine misinformation. Below we analyse how attitudes in the subsystems of *Affect, Appreciation* or *Judgement* are expressed in the Telegram corpus by identifying the common adjectives, nouns, and verbs used with our list of frequent collocations. These were compared to Martin and White’s (2005) list to decide the type of attitude triggered. This mapping resulted in three recurrent categories *Fear, Insecurity* and *Mistrust*. We discuss below these three common themes in turn to explain how they construct discourses around COVID vaccines.

4.2.1 Fear of hidden agenda behind COVID-19 vaccines

Noun phrases constructing this theme often occur in messages where people’s negative experiences are used. The noun phrases *vaccine injury, vaccine harm, injection damage* were analysed within the Affect subcategories of Disinclination and Fear. The typical verbs surrounding these noun phrases provide such contexts as *suffered, put at risk, died* and *endured.* These explicitly convey negative feelings that trigger both Disinclination and Fear of COVID-19 vaccines. For example:

(7) There is no ethical justification for superfluous vaccination that will *put* children *at elevated risk* of *vaccine harm*

(8) This verified Australian emergency department doctor is so *upset* by the quantity and type of vaccine injuries he’s witnessed professionally

Fear is conveyed in the expression *put at elevated risk* used with the collocation *vaccine harm* as intended for concerned parents, thus cultivating an attitude of fear and "disinclination" (Martin and White, 2005) towards COVID vaccines as presenting more risk and harm for their children.

In the second example, the affective mental process *upset* is presented as a feeling caused by the collocation *vaccine injuries*. According to Martin and White's (2005) model, the adjective *upset* falls within dissatisfaction as a category. This further promotes fear of COVID vaccines despite the fact in the example above the specific nature of vaccine injuries is unclear and left open for readers' interpretations.

Similarly, the noun phrase *jab deaths* is common with mental processes in which such verbs as *solicit* are used:

(9) Further to the MHRA's latest Tweet ‘ Every Report Counts ’, where it is *soliciting* information regarding jab deaths and injuries; it looks as if this government agency is indeed running for cover.

The Affect types of Disinclination and Fear can be expressed covertly and overtly in the choices of verbs and adjectives. While in several cases discourses about the COVID vaccine side effects are charged with negative feelings of Fear, in other messages it seems that authors maintain a positioning of Disinclination towards the COVID vaccines by sharing false content about their side effects.

4.2.2 Insecurity and vaccine mandates

Keywords *free, freedom, medical* and their related noun phrases are often found in discourses around criticism of mandatory vaccines. These noun phrases tend to occur as the object of *push, enforce, bring* and *impose*. These verbs represent a material process whereby the actors are often the government and its institutions, which are seen as imposing unreasonable restrictions on freedom.

Combined with the noun phrases such as *medical apartheid, government control* the triggered emotional response concerns Insecurity. This attitudinal Affect according to Martin and White (2005) covers emotions dealing with threats to eco-social well-being — anxiety, and lack of confidence. Our collocations overwhelmingly focus on how human freedom and rights are threatened by vaccine mandates. This can be illustrated with this message:

(10) Latest data confirms vaccines have limited effect for very limited period on reducing transmission & infection. So vaccine passports *pointless* as is *medical apartheid*.

According to the appraisal framework, the lexical choices of *pointless* and *medical apartheid* can be analysed through the semantic domain of *Judgment*. As *Judgment* is concerned with human behaviour vis-a-vis the social norms (Oteíza, 2017), this message reflects a ‘social sanction’ involving both values of Veracity (how truthful someone is) and Propriety (how ethical someone is). The noun phrase *medical apartheid* raises concerns about the unethical nature of mandatory vaccines by referring to human freedom, whereas the adjective *pointless* conveys distrust in the government's decision. These lexical choices are chosen to establish solidarity between the text producer and readers. Other examples that use the same rhetoric to convey this solidarity are found to deploy the discourses of *freedom fighter, free choice*.

(11) University students have chosen not to comply with the tyrannical vaccine mandates. Join the students, freedom fighters

(12) IT'S A LIE When they tell you it's necessary to protect ‘public health’, IT'S A LIE When they tell you free choice is selfish, IT'S A LIE When they tell you everyone else is doing it

These two messages convey a stance and manage interpersonal relations. In the first example, the material process is realised by such verbs as *comply* and *join* that are used in the imperative mood. The second example uses the ‘power of three’ (Mooney and Evans, 2018) as the statement *IT'S A LIE* is repeated three times. This is one persuasive discourse strategy that is used to emphasise the message. According to the *Judgment* subcategory, the message overtly represents the veracity of COVID vaccine importance as dishonest and deceptive. Moreover, the message also presents the COVID vaccine Propriety as unethical and as something that violates *free choice*. These positionings portray the text producer as a 'human rights' promoter who seeks to protect and defend social values, whereas the government and health institutions' decisions are represented as not morally correct. As such these noun phrases fuel narratives of Insecurity towards and doubt in science and medical institutions.

4.2.3 Mistrust and conspiracy stories about COVID vaccines

The collocations also provide an illustration on how this topic is also exploited by conspiracy theorists. The noun phrase *evil vaccines* is used in such contexts as *fight against* where the author is drawing on conspiracy stories. Similarly, when *evil* is describing the object *tyrants*, while the requests for actions such as *take back, look at, and listen* are used with personal pronouns *I, we you* as the actor, putting the readers into the position of active agents or active observers. Similarly, *vaccine agenda* is another frequent collocation that occurs as a direct object for *oppose, expose, don't give in* as shown in the message below.

(13) Dr. Judy Mikovits who has also been a guest on London Real *exposing the vaccine agenda* and the vested interests behind it…

The use of material processes in the passive voice is also common with the noun phrases listed in Table 5. This construction foregrounds such phrases as *vaccine injuries* as the subject:

(14) Sadly, there are many cases of *vaccine injury caused by* the Gates Foundation or those they funded.

Specifically the verb *to cause* tends to express negative consequences, such as *damage, death, pain* or *troubles*, which are caused by undesirable risk factors, so that *The Gates Foundation* is treated as one of them, see also (Hunston, 2011).

This also calls for analysis of the transitivity patterns. Specifically for the verb *to control* we compared the distribution of its actors vs its goals, i.e., who *controls* vs what is *controlled* in our misinformation corpus as compared against enTenTen. In the general purpose corpus, the typical direct objects are *to control diseases, pests, costs, weeds*. Analysis of examples of seemingly neutral collocations, such as *to control access*, indicates the typical requirement of limitations which have to be imposed by the Actor of *control*. In general-purpose corpora, this is presented as a desirable outcome:

(15) This Oracle Directory Server Enterprise Edition 11g training will teach you how to perform routine maintenance, *control access* and monitor and tune servers.

The aspect of *to control* as imposing limitations is amplified in the Telegram misinformation corpus, but this time it is presented is undesirable, as its use involves the positive direct objects *to control our lives, the world* or *everything* while the typical actors of *to control* are mass media, government bodies or *parasites*, so that *control* is presented as undesirable:

(16) The faces of the *parasites that controlled our lives* and destiny for far too long.

A mixture of positive and negative feelings can be detected in the suggestion below:

(17) Going to see an unvaccinated doctor is *best*. Doctors who are not about *the evil agenda*. Just be doctors and nurses and do your job which is to help *save lives* not *kill people* with *evil covid19 vaccines* and *evil breathing treatments*

On the one hand, the positive feeling is indicated in the use of the adverb *best* and verb *save lives* to describe Inclination toward consulting an unvaccinated doctor. On the other hand, the negative feeling is demonstrated in Disinclination towards vaccinated doctors represented as part of the *evil agenda* of killing people with *evil vaccines*. This distinction constructs a judgment whereby the positive feeling induces the Veracity and Truthfulness of unvaccinated doctors, whereas the negative feeling around vaccinated doctors portrays dishonesty.

Based on the attitudinal system of the appraisal framework, these findings show that conspiracy messages about COVID vaccines rely on negative feelings that are realised as a relational process or a material process. In several cases, the type of *Judgment* echoed social sanction of mistrust in COVID vaccines and dishonesty of measures promoting them. Concurrently, positive feelings are expressed in messages to trigger veracity towards antivaccine movement.

Table 6: Comparison of communicative functions in two COVID-19 corpora

Telegram, misinfo		Twitter, general	
33.74%	A1.argument	58.84%	A1.argument
4.17%	A11.personal	10.10%	A11.personal
32.95%	A12.promotion	15.72%	A12.promotion
0.39%	A14.academic	0.26%	A14.academic
4.98%	A16.information	0.77%	A16.information
2.51%	A7.instructive	3.76%	A7.instructive
20.94%	A8.news	10.18%	A8.news

4.3 Analysis of communicative functions

Table 6 compares the distribution of the communicative functions as detected by the classifier in the two social media corpora on the topic of COVID-19, the Telegram misinformation corpus vs the general Twitter corpus. The general purposes of sharing messages via Twitter and Telegram are similar, and both corpora have been collected from similar social media platforms devoted to discussing the same topic. This leads to the fact that the function of argumentation is the most frequent category in both corpora. At the same time, the Telegram corpus contains a higher proportion of promotional texts which, in addition to obvious spam, include numerous click-bait news items, which are more likely to be predicted as A12 by the classifier, for example, the following message has been predicted as promotion with the secondary function of argumentation:

(18) Woman's husband gets myocarditis from vaccine and his doctor can't write him an exemption because he will get his licensed threatened. Her realizations are so basic it makes me want to pat her on the head like a little bunny. Better late than never. Follow us for more info on COVID vax injuries.

In comparison to the general Twitter corpus, the Telegram misinformation corpus also contains a higher proportion of texts classified as either news reporting or as texts pretending to provide reference information, for example:

(19) This Pfizer patent application was approved on 31 August 2021 and is the first patent that appears in a list of more than 18500 patents that serve the remote tracking of all vaccinated people worldwide, which is a quantum connection with pulsating microwave frequencies of 2.4 GHz or higher - from mobile masts and satellites directly to the graphene oxide in the fatty tissue of all persons who have received the vaccine.

In contrast, bona fide argumentation is more common in the Twitter corpus, compare the values in the first line in Table 6.

With respect to variation in the lexicogrammatical features, the misinformation corpus demonstrates a considerable rise in the rates of C11 (*anyone, everything*) and F18 (passives with an explicit agent), see the top part of Table 7. Following (Biber, 1988), the numbers in this table represent the rates, i.e., the total count of a respective feature in a text divided by the length of this text in words.

The rise in the use of the indefinite pronouns indicates the preference for making very general claims in the misinformation corpus, while the rise of the passives with an explicit agent indicates the preference

Table 7: Comparison of the lexicogrammatical features in two COVID-19 corpora

Features	**Telegram, misinfo**	**Twitter, general**
Over-used:		
C11.indefinite pronouns	0.082760	0.002239
F18.BY-passives	0.000358	0.000248
Under-used:		
B05.timeAdverbials	0.001989	0.003659
C09.impersProns	0.002817	0.005088
H35.causative	0.000404	0.000920
H36.concessives	0.000009	0.000225
H37.conditional	0.001262	0.002068
H38.otherSubord	0.000365	0.000843
I40.attrAdj	0.032500	0.043610
K47.generalHedges	0.000187	0.000388
K50.discoursePart	0.001380	0.002518
K58.seemappear	0.000181	0.000370
L52.possibModals	0.001861	0.003468
L53.necessModals	0.000684	0.001253

of shifting the onus of the predicate (the grammatical subject) from the agent (often health authorities) to the ‘victims’ affected by their action. The most frequent verbs occurring in such constructions are *advise, affect, cause, drive, own, target* with examples like:

(20) I’m super proud to be an awake, pure blood healthy vaccine free awake human, just the way we are born and designed to be … not *owned* by big pharma and not enslaved by them.

(21) We want AHPRA Registered Professionals who have been *affected* by the jab mandates.

In comparison, the most frequent verbs used in this construction in the general Twitter corpus are *cause, interview, produce, stop, tell*, where the passive use in *interview* and *tell* merely present the focus on what is being reported rather than the reporter.

At the same time, the Telegram misinformation corpus demonstrates a marked drop in the frequency of other features, see the bottom part of Table 7. Most of these features manifest the preference of the authors of misinformation messages to avoid hesitancy or hedging (H36, H37, K47, K58, L52). Fewer time adverbials (B05) and attributive adjectives (I40) indicate less specific information provided in the message in comparison to messages from the Twitter corpus.

5 Conclusions

The wide spread of COVID misinformation during the pandemic has led to what is often described as ‘infodemic’, a term that refers to massive circulation and flow of rumours related to COVID-19. In this paper, we focused on understanding the linguistic patterns that are frequent in discourses about COVID-19 vaccines. We explored these discourses through corpus analysis while relying on concepts from the appraisal theory and the SFL framework.

The findings show that COVID misinformation messages about vaccines can be categorised into three discourse patterns: vaccines side effects, vaccine mandate, and conspiracy stories on the hidden agenda behind vaccines. Within each discourse pattern, the Appraisal analysis of common noun phrases and their context of use shows the kinds of Attitude expressed in the misinformation messages. Predominantly, the messages convey the negative feeling with common Affect types of Fear, Disinclination, and Insecurity. These were found to represent COVID vaccine side effects and vaccine mandates. Moreover, the Judgment type conveyed in various messages demonstrates social sanction of mistrust. These Judgment types were overtly expressed in conspiracy narratives that seek to cast mistrust in government and health institution measures. The distribution of the communicative functions expressed via misinformation messages in our corpus shows the presence of a higher proportion of promotional texts, news reporting, and text expressed as presenting reference information in comparison to a more general social media corpus devoted to the same topic.

These findings are helpful to the public health authorities as they provide an insight into the linguistic mechanisms deployed in misinformation messages such as expression of attitudes, emotive language and communicative functions and how they are harnessed to create fear and mistrust. Understanding these mechanisms will allow countering misinformation more effectively. These could help in improving the design of intervention campaigns and messages with the aim of exposing these mechanisms and of restoring trust and security in society. To our knowledge, this study is the first of its kind to use an original COVID misinformation corpus to conduct qualitative and quantitative analysis. Access to the annotated corpus will be made publicly available at the time of publication of this paper. Our mixed method approach allowed us to better understand the language of COVID misinformation about vaccines and the types of attitudes underlying the messages spreading on Telegram. The findings revealed that COVID misinformation about the vaccines trigger specific attitudes (Fear, Insecurity, and Mistrust). These attitudes tend to draw on ideological positionings such as *medical and human freedom, science fallibility*, and *government conspiracy plans*. COVID misinformation messages about vaccines are ideologically constructed to negotiate solidarity with readers who support the respective views. This explains their widespread impact and viral distribution among the readers willing to share their stance.

References

Agley, J. and Xiao, Y. (2021). Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health*, 21(1):1–12.

Baker, P., Brookes, G., and Evans, C. (2019). *The language of patient feedback: A corpus linguistic study of online health communication*. Routledge.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., and Wodak, R. (2008). A useful methodological synergy? combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the uk press. *Discourse & society*, 19(3):273–306.

Baker, P., Gabrielatos, C., and McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press.

Bateman, J. A. and Paris, C. L. (2020). Searching for 'austerity': Using semantic shifts in word embeddings as indicators of changing ideological positions. In *Multimodal Approaches to Media Discourses*, pages 11–41. Routledge.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Biber, D. and Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Brookes, G., Atkins, S., and Harvey, K. (2022). Corpus linguistics and health communication: using corpora to examine the representation of health and illness. In *The Routledge Handbook of Corpus Linguistics*, pages 615–628. Routledge.

Chou, W.-Y. S., Gaysynsky, A., and Vanderpool, R. C. (2021). The COVID-19 misinfodemic: Moving beyond fact-checking. *Health Education & Behavior*, 48(1):9–13.

Freelon, D. and Wells, C. (2020). Disinformation as political communication. *Political Communication*, 37(2):145–156.

Gablasova, D., Brezina, V., and McEnery, T. (2017). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, 67(S1):130–154.

Gert, H. J. (1995). Family resemblances and criteria. *Synthese*, 105:177–190.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2004). *Introduction to Functional Grammar*. Arnold, London.

Hause, A. M., Marquez, P., Zhang, B., Myers, T. R., Gee, J., Su, J. R., Blanc, P. G., Thomas, A., Thompson, D., Shimabukuro, T. T., et al. (2022). Safety monitoring of bivalent COVID-19 mRNA vaccine booster doses among persons aged$\geq$12 years – United States, August 31–October 23, 2022. *Morbidity and Mortality Weekly Report*, 71(44):1401–1406.

Huang, B. and Carley, K. M. (2020). Disinformation and misinformation on Twitter during the novel coronavirus outbreak. *arXiv preprint arXiv:2006.04278*.

Hunston, S. (2011). *Corpus approaches to evaluation: phraseology and evaluative language*. Routledge.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., and Suchomel, V. (2013). The tenten corpus family. In *Proc Corpus Linguistics Conference*, pages 125–127, Lancaster.

Lamsal, R. (2020). Coronavirus COVID-19 tweets dataset. https://dx.doi.org/10.21227/781w-ef42.

Lecompte-Van Poucke, M. (2022). 'You got this!': A critical discourse analysis of toxic positivity as a discursive construct on Facebook. *Applied Corpus Linguistics*, 2(1):100015.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Martin, J. (2000). Beyond exchange: APPRAISAL systems in English. In Hunston, S. and Thompson, G., editors, *Evaluation in Text*. Oxford University Press, Oxford.

Memon, S. A. and Carley, K. M. (2020). Characterizing COVID-19 misinformation communities using a novel Twitter dataset. *arXiv preprint arXiv:2008.00791*.

Mooney, A. and Evans, B. (2018). *Language, society and power: An introduction*. Routledge.

NHS (2023). Covid-19 misinformation. Online. https://library.hee.nhs.uk/patient-information/health-information-online/covid-19-misinformation.

Osmundsen, M., Bor, A., Vahlstrup, P. B., Bechmann, A., and Petersen, M. B. (2021). Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review*, 115(3):999–1015.

Oteíza, T. (2017). The appraisal framework and discourse analysis. In Bartlett, T. and O'Grady, G., editors, *The Routledge Handbook of Systemic Functional Linguistics*, pages 457–472. Routledge, London.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *Proc Comparing Corpora Workshop at ACL 2000*, pages 1–6, Hong Kong.

Sharoff, S. (2018). Functional text dimensions for the annotation of Web corpora. *Corpora*, 13(1):65–95.

Sharoff, S. (2021). Genre annotation for the web: text-external and text-internal perspectives. *Register studies*, 3:1–32.

Stahl, B. C. (2006). On the difference or equality of information, misinformation, and disinformation: A critical research perspective. *Informing Science*, 9:83–96.

Van Dijk, T. A. (2009). *Society and discourse: How social contexts influence text and talk*. Cambridge University Press.

Wilson, J., Hartley, A., Sharoff, S., and Stephenson, P. (2010). Advanced corpus solutions for humanities researchers. In *Proc Advanced Corpus Solutions, PACLIC 24*, pages 36–43, Tohoku University.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.