Annamária Fábián and Igor Trost

Computer-Mediated Communication to facilitate inclusion: Digital corpus analysis on disability diversity on social media

Abstract: Whereas there is a wealth of studies on computer-mediated communication (CMC), publications (e.g., Oussalah et al. 2016; Beißwenger 2016; Scheffler et al. 2019; Brookes and McEnrey 2020; Clausen and Scheffler 2020; Heritage and Baker 2022; Grieve and Woodfield 2023) specifically addressing diversity and inclusion in both, CMC and Digital Linguistics, are underrepresented. At the same time, many linguistic studies make use of data from digital media, itself an increasingly popular field of study in linguistics (Crystal 2006; Zappavigna 2012; De Decker and Vandekerckhove 2017; Bubenhofer 2017; Abel et al. 2020; Marx and Weidacher 2020; Wright 2020), but none of them focuses on disability-related diversity and inclusion. Simultaneously, inclusion has become significant in digital societies and has attracted raising awareness by the participation of people with a disability via communication on social media. This study on CMC therefore examines digital language use concerning disability and inclusion – contributed by people with and without a disability – on social media, which is in times of digital participation of diverse groups highly relevant to empowerment and inclusion in digital societies. A Twitter corpus comprising 2,559 tweets of 61,249 tokens is therefore used for this representative analysis. The corpus consists of German tweets published on *#Behinderung* ('disability') and *#Inklusion* ('inclusion') between 1st of December – 31st of December 2020. This linguistic study provides valuable first insights into the lexicon concerning disability and inclusion on social media as well as the co-occurrences of the lexical units.

Keywords: disability discourse, discourse of inclusion, sentiment analysis, Computer-Mediated Communication, Digital Linguistics

Annamária Fábián, University of Bayreuth, e-mail: Annamaria.Fabian-Trost@uni-bayreuth.de
Igor Trost, University of Passau, e-mail: Igor.Trost@uni-passau.de

https://doi.org/10.1515/9783111434018-015

1 Brief overview of the research in linguistics on Social Media discourses concerning social diversity

Whereas there is a wealth of studies on the language of discrimination, particularly within discourse studies, studies specifically addressing the linguistic practices of inclusion of diverse individuals and collectives are comparatively rare. At the same time, many linguistic studies make use of data from digital media, itself an increasingly popular field of study in linguistics (Crystal 2006; Zappavigna 2012; De Decker and Vandekerckhove 2017; Bubenhofer 2017; Abel et al. 2020; Marx and Weidacher 2020; Wright 2020). The communication of inclusion and exclusion of diverse individuals and collectives has been the focus of numerous studies within the social sciences, and discourse analysis is becoming increasingly prevalent. In recent years, linguistic studies have focused on discourses pertaining to refugees and migrants (e.g., Viola and Musolff 2019), as well as on gender-related issues (e.g., Paknahad and Baker 2016; Gnau and Wyss 2019), disability (e.g., Sties 2013; Grue 2014) and on mental health issues (e.g., Harvey 2012) in various countries and contexts. Many of these studies make use of data from digital media (e.g., Marx and Weidacher 2020; Wright 2020; Knuchel and Bubenhofer 2023). These important studies have raised awareness regarding the importance of analyzing issues related to diversity, including diverse individuals and diverse collectives from the point of view of Corpus Linguistics and Discourse Analysis. In CMC, as well as in human-centered data science, research on an inclusive digital transformation is underrepresented. Herrera (2022) argues that social media analytics tools need to be designed to support inclusive public services for all, including those with disabilities. Sinclair (2011) emphasizes the importance of paying attention to social barriers that inhibit inclusion, rather than simply technological barriers. Zelena (2020) explores, how new media platforms become the platform of communal loss for users of different ages, genders, social statuses, and diverse internet usage habits and socialization. Finally, Pan et al. (2014) examine the role of community diversity in influencing perceived inclusion of newcomers in the online community and the influence of such perception on newcomers' engagement intention. This wide range of the corpus linguistic research on language on social media as well as on, in general, inclusion in digital societies indicates not only the lack of interest in studies in terms of disability-related diversity in CMC but also in interdisciplinary studies on an inclusive digital transformation via CMC for diversity visibility.

In keeping with the scientific tradition of Corpus Linguistics, CMC (e.g., Oussalah et al. 2016; Beißwenger 2016; Scheffler et al. 2019; Brookes and McEnrey 2020; Clausen and Scheffler 2020; Heritage and Baker 2022; Grieve and Woodfield 2023)

and Computational Social Science (e.g., Brantner and Pfeffer 2018; Ralev and Pfeffer 2022; Strathern et al. 2022), this study focuses on the digital communication of disability diversity and inclusion in one month (December 2020) selected for this quantitative lexical analysis from a corpus of 14 years between 2009–2023 as well as on methodological considerations for processing with digital corpora for CMC and human-centered data science.

2 Lexical study and sentiment analysis of the language use regarding inclusion and disability on social media as a key research objective

CMC encompasses various forms of communication, which take place by way of digital devices and networks. The language used in CMC can vary depending on the platform, context, and participants involved. According to Barbaresi (2019: 29-30), "specialized corpora of the language of CMC and social media are increasingly vital for the analysis of diversity in terms of speakers and settings in digital contexts". As it is important to notice that people with a disability face ableism in education internationally and have consequently limited access to academia, this study wants to contribute to speakers' and platform diversity in CMC by listening to the perspectives of disabled persons. Minorities (including individuals with a disability) contribute to the visibility of social diversity, and with this, to inclusion by raising awareness to different diversity dimensions, their own personal situation and their perspectives on inclusion, discrimination, and exclusion as well as on everyday life. Furthermore, the digital communication of individuals with a disability evoke digital conversations between people with and without a disability essential to inclusion. This significant digital activism of disabled individuals often leads to a social transformation through the shift of perspectives in society via CMC. Moreover, individuals with a disability, have been successfully engaged on social media for inclusion through visibility for more than 10 years. For a study on significant voices and perspectives on disability and inclusion as a result of the communicative co-construction of both, diversity and inclusion, on German Twitter, we set up a corpus along *#Behinderung* ('disability') and *#Inklusion* ('inclusion'), mainly but not exclusively written by individuals with a disability and their representatives. The corpus underlying this research consists of 2,559 German tweets, together with 61,249 tokens, as part of a large corpus made up of 14,926 tweets in total with 5,663,504 tokens. The large corpus however includes mainly German tweets published 2009–2023 under the hashtags 'inclusion' and 'disability', while the small corpus was pub-

lished in a time period of one month, from the 1st to the 31st of December 2020 UTC. This paper therefore provides an analysis of the communication of disability diversity on social media. For the analysis, we chose to examine the data for a single month, in order to gain first insights into the lexicon and the sentiment of the entire corpus. The outcome of this corpus-driven study contributes to decision-making on data processing for further qualitative and quantitative CMC-related studies and facilitates effective navigation of large-scale data by introducing a methodological design combining the user friendly tools AntConc and SentiStrength for the initial evaluation of digital discourses and corpora.

Before processing with the quantitative examination of the corpus on *#Behinderung* ('disability') and *#Inklusion* ('inclusion') on Twitter, we would like to introduce our decision for processing with Twitter (rebranded to X in July 2023) data. Before Twitter's acquisition by Elon Musk, the platform with several members of the German former and current government, journalists, and other significant public figures was broadly used for disability agenda setting by individuals with a disability in Germany as well as in other countries.[1] The selection of the corpus from December 2020 is based on the progress of the German words 'Behinderung' and 'Inklusion' in the corpus of 14 years as those words show a particularly high frequency in this one month compared to the time due to and after 2020. The reason of this comparatively high frequency is associated with COVID19 as a serious threat to human life, in particular to those with health impairments, which has clearly resulted in this heightened interest in disability and inclusion. Similar to other digital social movements, (e.g., Dang-Anh 2013; Fábián 2020), language and computer-mediated-communication are verified essential keys for activism concerning inclusion in digital societies. As this corpus consequently is of high significance to people with health impairments, also including many individuals with a disability, we conduct a computer-driven lexical examination of relevant parts of the German discourse on disability and inclusion. This quantitative CMC-study was prepared to gain insights into the sentiment of self-representation of people with disabilities as well as of inclusion on social media based on the investigation into the language and communication used when discussing disability and inclusion on Twitter under the participation of people with a disability.

1 As Elon Musk has refused the free use of the API to scientists since the end of April 2023, the data gathered prior to this time is also historically relevant to German society as well as to people with a disability in Germany.

For this examination, we therefore undertake a combined software-based lexical and sentiment analysis with AntConc and SentiStrength,[2] in particular of the nouns *Inklusion* ('inclusion') and *Behinderung* ('disability') and associated lexical entities, which is prevalent for a CMC-based linguistic study of minority languages reporting on issues and agenda of individuals with a disability, but not exclusively of those with a disability. Our research is guided by the hypothesis that the German discourse on *#Inklusion* ('inclusion') and *#Behinderung* ('disability') can be lexically classified and characterized on social media, and that the discourse is highly positive from the point of view of the discourse participants, which we will demonstrate on the corpus. First, corpus linguistic insights from an excerpt of a digital discourse on disability and inclusion on social media is essential as Fábián et al. (2024: 24) demonstrate the participation of individuals with a disability on Social Media based on the German Twitter (X) example and their community organization by using the hashtags 'disability' and 'inclusion'. This kind of human-centered studies contribute to gathering information on self-empowerment of diverse individuals and collectives often facing discrimination in society, essential for inclusion. Although our first CMC study (Fábián et al. 2024) provides the first information on disability participation in a digital society via computer-mediated communication, the semantic evaluation of digital discourses on disability and inclusion has not been covered, neither in CMC nor in human-centered data science, making this study unique, and simultaneously essential for first insights into data on disability self-empowerment and public disability visibility for an inclusive transformation in society. A semantic classification of tweets into the categories negative, neutral and positive with SentiStrength as part of a Sentiment Analysis will supplement this investigation (e.g., Kiritchenko et al. 2014; Dai et al. 2017; Palomino et al. 2020) on the lexicon by AntConc. AntConc was developed by Anthony Lawrence (Waseda University/Japan), SentiStrength by Mike Thelwall (University of Wolverhampton/UK). Both of them are at no cost available for nonprofit goals and can also be easily used by students, early-career scientists as well as by scientists without knowledge of Computational Linguistics engaged in qualitative studies on corpora, which convinced us to use these tools. Our research design includes quantitative research methods, while pursuing the following goals:

1. We observe the lexicon (incl. collocations) in the Twitter discourse on disability and inclusion in order to arrive at a first impression on the semantic and emo-

2 SentiStrength was only developed for the "sentiment strength detection for short informal text" but not for large corpora.

tional aspects of communication in digital discourse concerning disability and inclusion.
2. We provide a lexical analysis including the analysis of collocations (Corpus-driven lexical Analysis) on disability and inclusion in our Twitter corpus.
3. We classify the tweets as part of our digital corpus in negative, neutral and positive (Sentiment Analysis).

In addition, the project aims to gain insights into effective digital linguistic methods (tools, software etc.) adaptable for the communicative analysis of data on social media. Dai et al. (2017) propose a word embedding-based clustering method for tweet classification that achieves good accuracy without requiring labeled training data. Lui and Baldwin (2014) but also Heaton et al. (2023) evaluate off-the-shelf language identification systems for tweets and their usability for linguistic analysis. Lui and Baldwin (2014) find that simple voting over three specific systems consistently outperforms any specific system. Yang and Srinivasan (2014) propose a methodology for translating surveys into social media surveillance, which achieves better precision and recall than standard methods using lexicons or classifiers. While Yurchenko and Ugolnikova (2021) focus on linguistic methods in social media marketing, the paper highlights the relevance of simple linguistic methods for a short overview of corpora before processing with further and more detailed analysis of communication in corpora. We decided to combine therefore AntConc often used for a quick analysis of the lexicon and the collocations, and SentiStrength, which is far less widespread among corpus linguists making this paper useful for a corpus linguistic sentiment analysis. According to Palomino et al. (2020: 8), SentiStrength has the methodological advantage of simple application for the identification of "the polarity of tweets as positive, negative or neutral, though SentiStrength can also work as a binary classification tool – positive or negative.", which is the main reason for using this specific tool for a semantic evaluation of the analyzed digital corpus.

3 A Data-Driven Semantic Study of 'disability' and 'inclusion' in a Digital Corpus on Twitter

3.1 Conducting a Data-Driven Semantic-Analysis with SentiStrength and AntConc – methodological Considerations for corpus linguists

As highlighted in chapter 3, the quantitative background of this digital linguistic study is twofold:

1. First, we conduct a lexical analysis of the corpus on *#Behinderung* ('disability') and *#Inklusion* ('inclusion')[3] by using AntConc, a tool often used by digital linguists. We chose AntConc as a tool as the adaptability of AntConc is useful for capturing and visualizing the lexical units and their collocates.
2. Second, we carry out a sentiment analysis with SentiStrength. SentiStrength is a sentiment classification tool which does not need proficiency in Machine Learning and can also be easily used by digital linguists without a background in Computational Linguistics.

Before processing with our corpus linguistic study with SentiStrength, it was necessary to prepare the corpus for processing with SentiStrength as SentiStrength was developed to analyse shorter texts line by line especially for business purposes. First, it was necessary to eliminate all line breaks in the corpus on the hashtags *Inklusion* ('inclusion') and *Behinderung* ('disability') for an overall analysis at sentence level. In addition, SentiStrength does not output the results in a separate file but puts them to a txt-UTF-8 corpus file, which slightly doubles in size as a result. While these framework conditions imply that the program cannot analyse large corpora and is therefore not useful for studies on large-scale data, SentiStrength enables first insights into the sentiment along lexical items in selected parts of a large-scale corpus. The outcome of this kind of first analysis supports scientists involved in studies on CMC with navigating through large-scale corpora and making decisions on how to process with the data for further examinations of communication as part of a research project. This triggered our decision to reduce our corpus for this paper and provide a Sentiment analysis on the communication of one month. For the analysis, however, we chose December 2020, which was in the midst of the Covid lockdown in German-speaking countries, exposing many indi-

3 We developed a register with keywords for the data collection. Our main keywords for the collection were *#Behinderung* ('disability') and *#Inklusion* ('inclusion').

viduals, particularly with those with health impairments and/or a disability, at a high risk. This international health emergency prompted our choice to process with the data for this time period. This part of our large corpus consists of 2,559 tweets, 950 full sentences,[4] 61,249 tokens and 11,251 types. Our corpus choice consequently has an impact on the Sentiment Analysis in the corpus as 'COVID' is quite frequent.[5]

The German sentiment strength dictionary file, *EmotionLookupTable_v5_fullforms*, for the program SentiStrength was provided by SentiStrength (http://sentistrength.wlv.ac.uk, last accessed 14 February 2025) and Hannes Pirker, Interaction Technologies Group at the Austrian Research Institute for Artificial Intelligence (OFAI) with additions from Elias Kyewski of the University of Duisburg-Essen.

SentiStrength performs the sentiment analysis using a sentiment strength dictionary, in which lexemes are assigned a sentiment rating. Positive sentiment ratings are marked with a scale of 1 to 5, negative ones with a scale -1 to -5. Each lexeme is rated with a maximum of 4 or -4, only repeated occurrences can result in a rating of 5 or -5 for a phrase. A neutral sentiment of a lexeme is marked with 0. In this paper, the positive numbers are always marked with a plus sign, i.e., the positive scale is +1 to +5.

Pertaining to sentences, the rating is always made up of a negative and a positive rating, e.g., -2/+3. These two ratings of a sentence are the results of the addition of the positive ratings and the addition of negative ratings. The sum is capped at +5 or -5. When the overall sentiment rating of a sentence is calculated, the maximum values which can result are +4 (=+5-1) or -4 (=+1-5).

While using SentiStrength, our first considerations were that this dictionary file *EmotionLookupTable_v5_fullforms* is very extensive for negative words such as insults. We also considered that the negative ratings are occasionally inconsequent as serious verbal insults such as *Scheiße* ('shit', 'fuck' or 'fucking') are rated at -3, but *leider* ('unfortunately') at -4. In light of this consideration, we decided to implement the necessary corrections: In our new sentiment strength dictionary file, *EmotionLookupTable_v6_fullforms*, *Scheiße* ('shit', 'fuck' or 'fucking') is rated at -4, and *leider* ('unfortunately') at -3. Another observation on SentiStrength was that the sentiment strength dictionary v5 contains only few positive words. Positive foreign words and positive word formations (very frequent in German morphology) are highly underrepresented in the lexicon of SentiStrength. Particularly in the Ger-

4 Not every tweet contains a full sentence.

5 Individuals with health impairment and/or disability often used 'COVID' as a lexeme, also combined with a hashtag, for protection by governmental regulations.

man-speaking countries, non-partisan recognized political words which express a high level of positivity ('Hochwertwörter') such as *gerecht* ('just') or *sozial* ('social') – also often occurring in corpora on social issues such as disability, and inclusion – are missing and, as a consequence, classified by SentiStrength as neutral (0). In this respect, the sentiment strength dictionary v5 had to be significantly revised for a sentiment analysis of public communication in the social and political sphere. In addition, we realized that strongly discourse-relevant keywords for our study, which are associated with a positive semantic, have not been included in the old sentiment strength dictionary file v5. Keywords in our study with a positive semantic include words such as *Inklusion* ('inclusion'), *Teilhabe* ('participation'), and *Barrierefreiheit* ('accessibility'), and the adjective *barrierefrei* ('accessible'). After recognizing the inadequately trained vocabulary of SentiStrength in German, we developed a register essential to our corpus linguistic analysis and finalized the list with – from the point of view of our CMC study on disability and inclusion – words not registered in the SentiStrength vocabulary. We therefore conducted a corpus-linguistic analysis of the lexicon key to the discourse on disability and inclusion along the hashtags #Inklusion ('inclusion') and #Behinderung ('disability'), which built the basis for detecting the key words in the corpus. Consequently, we developed a core register for the Sentiment Analysis with SentiStrength only after detecting the vocabulary by using AntConc. In this way, we augmented our register with the most important lexemes highly relevant to the discourse on disability and inclusion.

3.2 Findings of the corpus-driven analysis with AntConc and SentiStrength

A log-likelihood[6] analysis with the corpus linguistic tool AntConc of the collocates of the #-words *Inklusion* ('inclusion') / *inklusiv* ('inclusive') and *Behinderung* ('disability')/*behindert* ('disabled') illustrates the lexicon mostly significant and consequently highly-frequent in the discourse:

6 Standard settings: threshold $p<0.05$ (3.84 with Bonferroni), effect measure size: MI, search window span from five words left to five words right.

Table 1: Collocates of *inklusi**.

Collocates of ***inklusi****	FreqLR	FreqL	FreqR	Likelihood
Inklusion (inclusion)	335	172	163	369.051
Hilfe (help, aid, assistance)[7]	347	11	336	247.531
Deutschland (germany)	366	21	345	238.594
News[8]	358	25	333	215.490
Berlin	322	37	285	169.196
Teilhabe (participation)	223	97	126	111.421
mit (with)[9]	301	164	137	80.026
Barrierefreiheit (accessibility)	149	77	72	68.312
Menschen (humans)	192	93	99	56.250
SARS	18	13	5	38.405
barrierefrei (accessible)	74	47	27	35.526
CoV	20	14	6	34.365
Behinderung (disability)	624	476	148	33.453
Pflege (care)	79	17	62	26.221
Menschenrecht (human right)	29	6	23	20.427

7 The lexeme *Hilfe* ('help') is mainly used by one of the mostly 'visible' actors around disability and inclusion, which is a professional organization. The productivity of this organization in terms of the production of tweets has an impact on the evaluation of the entire corpus. Other frequently posting users – especially individuals with disabilities without institutional background – however do not use 'help' very often.

8 see comment above

9 This frequency is related to the frequent usage of the inclusive reference *Menschen mit Behinderung* ('people with disability').

Table 2: Collocate of *behinder**.

Collocate of ***behinder****	FreqLR	FreqL	FreqR	Likelihood
Menschen (humans)	732	669	63	781.086
mit (with)	865	797	68	610.112
Deutschland (Germany)	375	23	352	436.349
Hilfe (help)	333	13	320	377.870
News	316	16	300	279.251
Tag[10] (day)	188	165	23	198.911
Berlin	261	27	234	178.352
Behinderung (disability)	144	61	83	162.125
internationalen[11] (international)	65	58	7	83.670
der[12]	478	340	138	61.514
internationaler[13] (international)	42	36	6	60.523
internationale[14] (international)	39	35	4	51.498
Welttag (World Day)	35	30	5	48.299
von (of)	210	159	51	40.894
es (it, e.g., in *es braucht = it is necessary*, also there: *es gibt = there is*)	48	12	36	38.631
vielen (many)	5	2	3	35.567
Gesundheit (health)	31	19	12	35.339
Inklusion (inclusion)	701	171	530	31.376

10 The noun *Tag* occurs in the corpus as part of the phrase Internationaler *Tag der Menschen mit Behinderung* (Ínternational Day of People with Disability, the 3rd of December) very often.

11 The lexeme *international* occurs in our corpus in many different forms as the German grammar system has a complex flexion system with many different endings. This leads, however, to frequent appearance of the same word with different endings which are recognized as different findings by programs for processing with language data.

12 *der* can be understood as a definite article in German (masculinum), for instance in the collocation ***der** internationale Tag* ('the international day'), but also the pluralform with genitive, for instance in the collocation *der internationale Tag **der** Menschen mit Behinderung* ('The International Day of People with Disability')

13 See Footnote 11

14 See Footnote 11

Collocate of *behinder**	FreqLR	FreqL	FreqR	Likelihood
ich (I)	36	5	31	31.161
Corona (COVID 19)	116	68	48	29.127
SARS	12	7	5	28.751
CoV	14	9	5	24.455
das	80	22	58	23.811
Beschäftigung (employment)	21	19	2	23.027
veröffentlicht (published)	4	4	0	20.749
Teilhabe (participation)	113	63	50	19.797
Erinnerungen (memories)	8	6	2	19.710

Our quantitative lexical analysis shows the highest frequency of *Behinderung* ('disability') in a collocation with *inklusi** (FreqLR) and the highest significance (log-likelihood) of *Inklusion* ('inclusion') in a collocation with *inklusi**. Although *Hilfe* ('help') and *News* occur with high frequency in the corpus and are significant for *inklusi**, both tokens are mainly used only by one of the mostly 'visible' actors around disability and inclusion, which is a professional organization. The productivity of this organization in terms of the production of tweets has an impact on the evaluation of the entire corpus. Other frequently posting users – especially individuals with disabilities without institutional background – however do not use 'help' very often. Also, the toponyms *Deutschland* ('Germany') as well as *Berlin* occur in the collocation with *inklusi** quite frequent in the corpus, which depends on the use of these tokens on the one hand for setting the local context of disability- and inclusion-related topics referred to in the digital discourse, on the other hand as part of a metonym [*Berlin*] for the German government. This frequent use is in the context of policies necessary for more inclusion. While *Teilhabe* ('participation') and *Barrierefreiheit* ('accessibility')/*barrierefrei* ('accessible') show medium frequency and significance in the corpus, *Menschenrecht* ('human right') occurs infrequently in the time period of one month. This gives rise to the hypothesis that individuals with a disability focus more on inclusion and accessibility and their practical transformation in everyday life.

The second table includes the highest collocations (FreqLR) and mostly significance (log-likelihood) with *behinder** ('disabled'). *Menschen* ('humans') and the preposition *mit* ('with') were the most frequently used tokens examined. This frequent use of both tokens is associated with the self-reference of people with disabilities (*Menschen mit Behinderung*/'people with disability'). The reference is often used for agenda setting by people with a disability and civil society fostering inclu-

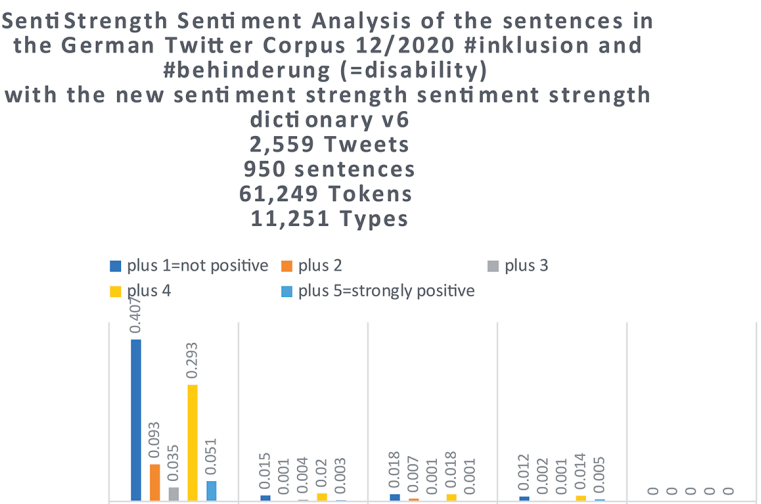
sion in German society. Although *Inklusion* ('inclusion') is particularly frequent in the collocation profile, the log-likelihood ratio of this token is comparatively low, because *Inklusion* is to be expected in the discourse and therefore idiomatic. The token *der* occurs often in the collocation ***der*** *internationale Tag* ***der*** *Menschen mit Behinderung* ('the International Day of People with Disability') or in *Welttag* ***der*** *Menschen mit Behinderung* ('World Day of People with a Disability'). While *Barrierefreiheit* ('accessibility')/*barrierefrei* ('accessible') occurred frequently in a collocation with *inklusi**, they remained statistically insignificant with *behinder**. This finding depends on the propensity that people with a disability seek *inclusion*. In a second step, as a consequence of the demand for *inclusion*, individuals with a disability and their representative organizations seek *accessibility*, which is reflected in the corpus. This is why *accessibility*/*accessible* appears as a collocate of *inclusion*/*inclusive* and not of *disability/disabled*.

The corpus-linguistic AntConc analysis has shown that the collocates of the lexemes *Inklusion* ('inclusion') and *Behinderung* ('disability') in the German discourse on inclusion occur in the corpus with positively framed words (*Hochwertwörter*) for instance *Hilfe* ('help', 'aid', 'assistance'), *Menschenrecht* ('human right'), *Teilhabe* ('participation'), and *Welttag* ('World Day'). In addition, they are associated with individuals (*Menschen* 'humans'), places (for instance *Berlin* as a metonym for the German federal government), professional lexicon (for instance social: *Beschäftigung* 'employment' or medical *Corona*) and function words (for instance *mit* 'with').

The quantitative lexical analysis and its findings facilitated in conducting the sentiment analysis as the evaluation of SentiStrength's register is based on the outcome of the quantitative analysis. While the lexical analysis confirmed that the collocates *Inklusion* 'inclusion', *Teilhabe* 'participation', *Barrierefreiheit* 'accessibility', *barrierefrei* 'accessible' (in bold in the Tables 1 and 2) are keywords of the discourse, we discovered that these tokens are not included in the original register of SentiStrength. We therefore added these words to our register to make the SentiStrength program more sensitive to our discourse study on disability and inclusion. In terms of these lexical findings, we would like to point out that the corpus-linguistic program AntConc recognizes all German morphological forms as separate types. Due to the variety of forms of the German adjective inflection with up to 17 endings (including the Ø-ending and the combinations with the endings of comparative forms), the log-likelihood analysis for adjectives by AntConc is often incorrect as AntConc does not recognize the morphological relations. The adjective *inklusiv* ('inclusive') has 87 tokens with eight morphological forms in the corpus and therefore AntConc rated *inklusiv* falsely as non-significant. This prompted our decision to add it to our extended register. Further proof checks on SentiStrength's register revealed that not only the lexemes with positive ranking such as *Inklusion* ('inclusion')/ *inklusiv* ('inclusive') and *Behinderung* ('disability')/ *behindert* ('disabled')

were not recognized by the program, but also the most frequent words with a negative sentiment rating in the corpus such as *Exklusion* ('exclusion'), *exklusiv* ('exclusive'), *Diskriminierung* ('discrimination'), and *diskriminierend* ('discriminatory').

As a result, we trained SentiStrength by adding the above vocabulary to the evaluation list of the sentiment strength dictionary: We rated the positive words *Inklusion* ('inclusion')/ *inklusiv* ('inclusive')/ *Teilhabe* ('participation')/ *Barrierefreiheit* ('accessibility'), and *barrierefrei* ('accessible') with +4 and the negative words *Exklusion*/ ('exclusion'), *exklusiv* ('exclusive'), *Diskriminierung* ('discrimination'), and *diskriminierend* ('discriminatory') with -4. These rating values were concluded from the point of view of the discourse participants – mainly individuals with disabilities – as inclusion is essential for users with a disability, while discrimination and exclusion have an enormous negative impact on the lives of many people with a disability. In its default settings, the program SentiStrength will rate these lexemes with +4 or -4 for single occurrences and with the maximum rate +5 or -5 for multiple occurrences within a sentence. These rating values also help to provide us



	minus 1= not negative	minus 2	minus 3	minus 4	minus 5 = strongly negative
plus 1=not positive	0.407	0.015	0.018	0.012	0
plus 2	0.093	0.001	0.007	0.002	0
plus 3	0.035	0.004	0.001	0.001	0
plus 4	0.293	0.02	0.018	0.014	0
plus 5=strongly positive	0.051	0.003	0.001	0.005	0

Diagram 1: Results of the SentiStrength Sentiment Analysis

with an insight into the polarized positions on inclusion, discrimination and exclusion in the analysed discourse.

The chart (see diagram 1) illustrates the results of the sentiment analysis with the corrected sentiment strength dictionary file *EmotionLookupTable_v6_fullforms* and the corpus without the # characters.

The overall sentiment rating of a sentence is calculated from the positive and the negative sentiment rating. Overall, a neutral or a positive sentiment rating of the discourse on inclusion can be seen:

- **40.7 %** of all 950 sentences are somewhat **neutral (+1-1=0)**, they have a neutral positive (+1) and a neutral negative (-1) sentiment rating.
- **47.2 %** of all 950 sentences have a **positive sentiment without negative sentiments,** i.e., they primarily consist of positive words:
- 29.3 % of all 950 sentences are **very positive (+4-1=+3)**, they have a highly positive (+4) and a neutral negative (-1) sentiment rating.
- 9.3 % of all 950 sentences are **slightly positive (+2-1=+1)**, they have a positive (+2) and a neutral negative (-1) sentiment rating.
- 5.1 % of all 950 sentences are **highly positive (+5-1=+4)**, they have a strongly positive (+5) and a neutral negative (-1) sentiment rating.
- 3.5 % of all 950 sentences are **positive (+3-1=+2)**, they have a very positive (+3) and a neutral negative (-1) sentiment rating.

Only 4.5 % of the sentences show a **negative sentiment rating**, i.e., they mainly consist of negative words:

- 1.8 % of all 950 sentences are **negative (+1-3=-2)**, they have a neutral positive (+1) and a very negative (-3) sentiment rating.
- 1.5 % of all 950 sentences are **slightly negative (+1-2=-1)**, they have a neutral positive (+1) and a negative (-2) sentiment rating.
- 1.2 % of all 950 sentences are **very negative (+1-4=-3)**, they have a neutral positive (+1) and a highly negative (-4) sentiment rating.

Some sentences are contradictory regarding their sentiment analysis, e.g.,:

- 1.8 % of all 950 sentences are **confrontational and positive in the result (+4-2=+2)**. These sentences include as well a highly positive (+4) as a negative (-2) sentiment rating, as they contain many positive words but also some negative words.

These contradictory results of positive and negative sentiment ratings in a sentence are partly due to controversies in the discourse, but above all, they are attributed by the program SentiStrength to sentences with negations of positively rated lexemes, e.g., *keine* (=-2) *Inklusion* (=+4) (‘*no inclusion*’).

While the first column indicates an enormous positive evaluation of the German discourse on disability and inclusion on Social Media regarding the example of Twitter (X), the second, third, fourth, and the fifth column illustrate that the corpus is barely associated with a negative sentiment. As the fifth column does not include any result with the lowest and highly negative sentiment (-5) in the corpus, a highly negative evaluation of the discourse can be excluded. In summary, the positive evaluation of the discourse dominates significantly over the negative evaluation. More negative sentiments occur mainly associated with #Barrierefreiheit (accessibility) as people with a disability and their families report on their experience with discrimination and exclusion on social media requiring inclusion and accessibility. Furthermore, the log-likelihood values in the collocation analysis have already provided an indication that the discourse related to inclusion is positive. This outcome is particularly significant as many digital discourses are conducted in a confrontational and polarizing style due to high polarization such as the German discourse (cf. Trost 2023) on COVID19, which also depends on the discourse participants. While the principal participants involved in the discourse on disability and inclusion for inclusion are individuals with a disability, the discourse on COVID19 is often dominated by members and voters of the German Radical-Right-Party "Alternative für Deutschland" (AfD) targeting democratic decisions, government, politicians affiliated with democratic parties, but also diversity and inclusion. From the point of view of human-centered data science and social sciences, this positive sentiment verifies the high level of acceptance of the discourse on disability and inclusion among the digital discourse participants who – according to Fábián et al. (2024) – predominantly are individuals with a disability, their digital community, their allies, and their representatives (representative organizations). In addition, the positive evaluation also reflects the emotional value of this discourse to individuals with disabilities and their allies, which makes it even more emergent to present additional digital data on the digital self-empowerment of individuals with a disability regarding information essential to inclusive agenda setting in society.

Our quantitative analysis based on this integrative research design consisting of AntConc and SentiStrength illustrated that this combined method provides first insights into the lexicon and the sentiment of a particular discourse. This kind of initial corpus linguistic studies on diversity-related discourses can serve CMC as well as HCDS with choosing a particular focus for further research. This combination enables an analysis, which takes both keywords and non-keywords into account as a concise keyword analysis can be carried out with AntConc supplemented by an analysis with SentiStrength adding non-keywords to the results conducted with AntConc. In addition, a sentiment analysis thus enables the validation of log-likelihood values by a detailed analysis of the framing at the level of individual lexemes and sentences. Studies in Digital Linguistics consequentially allow a

concise analysis of CMC corpora revealing contents of substantial significance to politics and society, which can contribute to research in Computational Social Science, Social Science, and Political Science essential to society. In addition, methodological considerations from Digital Linguistics can contribute to the development of programs and tools for data-driven language processing. Even though our findings indicate the relevance of our corpus linguistic study for human-centered data science with valuable information on disability participation in digital society, more accurate and, in particular, large-scale studies between Corpus Linguistics, CMC, Human-centered Data Science, and Computational Social Science on disability and inclusion in digital society are necessary for further research. These initial insights demonstrate solely the contribution of data analysis to disability empowerment, which could support communities of individuals with disability, their representative organizations as well as institutions and representatives for anti-discrimination as first indications reveal topics, content and views on the inclusion of individuals with a disability in an online or even offline society.

4 Conclusion

In terms of methodological impact, SentiStrength developed by computational scientists needs to be adapted and sometimes also trained for Corpus Linguistic Studies. This paper highlights that tools and methods of Digital Linguistics and Computational Science, also relevant to Computational Social Science, can be integrated in a research design for the analysis of digital discourses on diversity, disability and inclusion, including discriminatory phenomena such as discrimination, and exclusion. For a more concise study of social and political language use on social media, we would like to advocate for more interdisciplinary collaborations between Corpus Linguistics and Social and Political Science as well as Human-centered Data Science, Computational Social Science and, in general, Computational Science. Our methodological findings indicate that SentiStrength is an important tool with significant potential for Corpus Linguistics. However, its' usability for Corpus Linguistic research studies is extremely limited, which could be improved by interdisciplinary research projects for the development of tools and programs for language-based data-processing between Computational Science, Corpus Linguistics, and Social as well as Political Science including a vocabulary-based training of programs and tools. One of the most valuable methodological findings of this study for Linguistics is, however, that an AntConc analysis combined with an analysis with SentiStrength is useful for gaining valid first insights concerning the semantics of a particular digital discourse. This initiative outcome and underlying methods are

useful for further data exploration in a larger corpus. Although there is an abundance of methods such as these, this method enables a quick yet concise examination of the lexicon and the sentiment of digital discourses without requiring specialised knowledge in Computational Linguistics and Computational Science, which makes science more inclusive by reducing methodological complexity. In addition, the methods illustrated in this chapter guided us to the verification of the relevance of the discourse in German on disability and inclusion on Twitter (X) for individuals with a disability and their self-representations. The Sentiment Analysis of the vocabulary demonstrates the significance of computer-mediated communication for an inclusive transformation in a digital society by disability agenda setting, and by vital community organization among individuals with a disability. As there is little knowledge with regards to the communication of individuals with a disability, more language-focused research on disability and inclusion is essential. In summary, research on computer-mediated communication and human-centered data science can be used for gaining insights concerning digital activism for diversity, equity, and inclusion as well as, in general, into digital societies.

References

Abel, Andrea, Aivars Glaznieks, Carolin Müller-Spitzer, Angelika Storrer (eds.), 2020. Themenheft „Textqualität im digitalen Zeitalter“. *Deutsche Sprache* 48 (2). https://doi.org/10.37307/j.1868-775X.2020.02.

Anthony, Laurence. 2023. *AntConc (Version 4.2.2).* Tokyo Waseda University. Available from https://www.laurenceanthony.net/software (last accessed 14 February 2025).

Aragon, Cecilia, Shion Guha, Marina Koga, Michael Muller, Gina Neff. 2022. Human-centred data science. An introduction. Cambridge, MA, USA: MIT Press.

Barbaresi, Adrien. 2019. The vast and the focused: On the need for thematic web and blog corpora. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen & Caroline Iliadi (eds.), *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, 29–32. Mannheim: Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/ids-pub-9025.

Beißwenger, Michael. 2017. *Empirische Erforschung internetbasierter Kommunikation.* Berlin & Boston: De Gruyter. https://doi.org/10.1515/zrs-2018-0027.

Brantner, Cornelia & Jürgen Pfeffer. 2018. Content analysis of Twitter – Big data, big studies. In *The Routledge handbook of developments in digital journalism studies*. 79–92. Abingdon: Taylor & Francis. https://doi.org/10.4324/9781315270449.

Brookes, Gavin & Tony McEnery. 2020. Correlation, collocation and cohesion: A corpus-based scritical analysis of violent jihadist discourse. *Discourse and Society* 31, 4. 351–373. https://doi.org/10.1177/0957926520903528.

Bubenhofer, Noah. 2017. Kollokationen, n-Gramme, Mehrworteinheiten. In Kersten Sven Roth, Martin Wengeler, Alexander Ziem (eds.): *Handbuch Sprache in Politik und Gesellschaft, Sprachwissen*. Berlin & Boston: De Gruyter. 69–93. https://doi.org/10.1515/9783110296310.

Clausen, Yulia & Scheffler, Tatjana. 2020. A corpus-based analysis of meaning variations in German tag questions: Evidence from spoken and written conversational corpora. Corpus Linguistics and Linguistic Theory. *Corpus Linguistics and Linguistic Theory*. 18 (1), 1-31. https://doi.org/10.1515/cllt-2019-0060.

Crystal, David. 2006. *Language and the internet*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139164771.

Dai, Xianfeng, Marwan Bikdash & Bradley Meyer. 2017. From social media to public health surveillance: Word embedding based clustering method for twitter classification. In *SoutheastCon 2017*, 1–7. Concord: IEEE. https://doi.org/10.1109/SECON.2017.7925400.

Dang-Anh, Mark, Jessica Einspänner & Caja Thimm. 2013. Mediatisierung und Medialität in Social Media: Das Diskurssystem „Twitter". In Konstanze Marx & Monika Schwarz-Friesel (eds.): *Sprache und Kommunikation im technischen Zeitalter. Wieviel Internet (v)erträgt unsere Gesellschaft?*, 68–91. Berlin & Boston: De Gruyter. https://doi.org/10.1515/9783110282184.68.

De Decker, Benny & Reinhild Vandekerckhove. 2017. Global features of online communication in local Flemish: Social and medium-related determinants. *Folia Linguistica* 51 (1). 253–281. https://doi.org/10.1515/flin-2017-0007.

Fábián, Annamária. 2020. Verblose Sätze und kommunikative Praktiken in den Sozialen Medien am Beispiel der #MeToo-Bewegung. In Anne-Laure Daux & Anne Larory (eds.): *Kurze Formen in der Sprache / Formes brèves de la langue. Syntaktische, semantische und textuelle Aspekte / aspects syntaxiques, sémantiques et textuels*, 215-227. Tübingen: Stauffenburg.

Fábián, Annamária, Igor Trost, Kevin Altmann & Mara Schwind (2024). The analysis of "inclusion" and "accessibility" in Computer-Mediated-Communication for an inclusive transformation in digital societies. In Céline Poudat, Matilda Guernut. *Proceedings of the 11th Conference on CMC and Social Media Corpora for the Humanities*. 11th Conference on CMC and Social Media Corpora for the Humanities (CMC 2024), CORLI; Université Côte d'Azur, 2024. 20-26. https://shs.hal.science/halshs-04673776 (last accessed 14 February 2025).

Gnau, Birte C. & Eva L. Wyss. 2019. Der #MeToo-Protest. Diskurswandel durch alternative Öffentlichkeit. In Stefan Hauser, Roman Opiłowski & Eva L. Wyss (eds.), *Alternative Öffentlichkeiten. Soziale Medien zwischen Partizipation, Sharing und Vergemeinschaftung*, 131–165. Bielefeld: transcript. https://doi.org/10.14361/9783839436127-006.

Grieve, Jack & Helena Woodfield. 2023. *The language of fake news*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781009349161.

Grue, Jan. 2014. *Disability and Discourse Analysis*. London: Routledge. https://doi.org/10.4324/9781315577302.

Harvey, Kevin. 2012. Disclosures of depression: Using corpus linguistics methods to interrogate young people's online health concerns. *International Journal of Corpus Linguistics* 17 (3). 349–379. https://doi.org/10.1075/ijcl.17.3.03har.

Heaton, Dan, Jeremie Clos, Elena Nichele & Joel Fischer. 2023. Critical reflections on three popular computational linguistic approaches to examine Twitter discourses. *PeerJ Computer Science* 9: e1211. https://doi.org/10.7717/peerj-cs.1211.

Heritage, Frazer & Paul Baker. 2022. Crime or culture? Representations of chemsex in the British press and magazines aimed at GBTQ+ men. *Critical Discourse Studies* 19 (4). 435–453. https://doi.org/10.1080/17405904.2021.1910052.

Herrera, Lucia Castro & Terje Gjøsæter. 2022. Community segmentation and inclusive social media listening. In Rob Grace & Hossein Baharmand (eds.), *ISCRAM 2022 Conference Proceedings – 19th International Conference on Information Systems for Crisis Response and Management*, 1012–1023. Tarbes, France. https://idl.iscram.org/files/luciacastroherrera/2022/2467_LuciaCastroHerrera+TerjeGjosaeter2022.pdf (last accessed 14 February 2025).

Jaborooty, Maryam Paknahad & Paul Baker. 2016. Resisting silence: Moments of empowerment in Iranian women's blogs. *Gender and Language* 11 (1). 77–99. https://doi.org/10.1558/genl.22212

Kiritchenko, Svetlana, Xiaodan Zhu & Saif M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research* 50. 723–762. https://doi.org/10.1613/jair.4272.

Knuchel, Daniel & Noah Bubenhofer. 2023. Machine Learning und Korpuspragmatik. Word Embeddings als Beispiel für einen kreativen Umgang mit NLP-Tools. In Simon Meier-Vieracker, Lars Bülow, Konstanze Marx & Robert Mroczynski (eds.), *Digitale Pragmatik.* Digitale Linguistik 1, 213–235. Berlin & Heidelberg: Springer. https://doi.org/10.1007/978-3-662-65373-9_10.

Lui, Marco & Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, 17–25. Gothenburg, Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1303.

Marx, Konstanze & Georg Weidacher. 2020. *Internetlinguistik. Ein Lehr - und Arbeitsbuch*, 2nd edition. Tübingen: Narr.

Oussalah, Mourad Chabane, B. Escallier & D. Daher. 2016. An automated system for grammatical analysis of Twitter messages. A learning task application. *Knowledge-Based Systems* 101. 31–47. https://doi.org/10.1016/j.knosys.2016.02.015.

Palomino, Marco A., Aditya Padmanabhan Varma, Gowriprasad Kuruba Bedala & Aidan Connelly. 2020. Investigating the lack of consensus among sentiment analysis tools. In Zygmunt Vetulani, Patrick Paroubek & Marek Kubis (eds.), *Human Language Technology. Challenges for Computer Science and Linguistics. LTC 2017. Lecture Notes in Computer Science 12598*, 58–72. Cham: Springer. https://doi.org/10.1007/978-3-030-66527-2_5.

Pan, Zhao, Yao-bin Lu & Sumeet Gupta. 2014. How heterogeneous community engage newcomers? The effect of community diversity on newcomers' perception of inclusion: An empirical study in social media service. *Computers in Human Behavior* 39. 100–111. https://doi.org/10.1016/j.chb.2014.05.034.

Ralev, Radoslav & Pfeffer, Jürgen. 2022. Hate speech classification in Bulgarian. In *Proceedings of the 5th International Conference on Computational Linguistics in Bulgaria (CLIB 2022)*, 49–58. Sofia, Bulgaria: Department of Computational Linguistics, IBL – BAS. https://dcl.bas.bg/clib/wp-content/uploads/2022/09/CLIB2022_PROCEEDINGS_v1.0.pdf (last accessed 14 February 2025).

Scheffler, Tatjana, Berfin Aktaş, Debopam Das & Manfred Stede. 2019. Annotating shallow discourse relations in Twitter conversations. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, 50–55. Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2707.

Sinclair, Stephen & Bramley, Glen. 2011. Beyond virtual inclusion – communications inclusion and digital divisions. *Social Policy and Society* 10 (1). 1–11. https://doi.org/10.1017/S1474746410000345.

Sties, Norat. 2013. Diskursive Produktion von Behinderung: Die marginalisierende Funktion von Personengruppenbezeichnungen. In Jörg Meibauer (ed.), *Hassrede/Hate Speech. Interdisziplinäre Beiträge zu einer aktuellen Diskussion*, 194–222. Gießen: Gießener Elektronische Bibliothek. http://geb.uni-giessen.de/geb/volltexte/2013/10121/ (last accessed 14 February 2025).

Strathern, Wienke, Raji Ghawi, Mirco Schönfeld, & Pfeffer, Jürgen. 2022. Identifying lexical change in negative word-of-mouth on social media. *Social Network Analysis and Mining* 12: 59. https://doi.org/10.1007/s13278-022-00881-0.

Thelwall, Mike. 2023. *SentiStrength (Version 2.3).* University of Wolverhampton. Available from http://sentistrength.wlv.ac.uk/ (last accessed 14 February 2025).

Trost, Igor. 2023. Corona als Basis sprachlicher Argumentation an die eigene Nation und andere Nationen – vom Impfnationalismus bis hin zur Public Diplomacy. In Aleksandra Salamurović (ed.), *Konzepte der NATION im europäischen Kontext im 21. Jahrhundert*, 311–327. Berlin & Heidelberg: J.B. Metzler. https://doi.org/10.1007/978-3-662-66332-5_15.

Viola, Lorella & Andreas Musolff. 2019. *Migration and media. Discourses about identities in crisis*. Amsterdam & Philadelphia: Benjamins. https://doi.org/10.1075/dapsac.81.

Viola, Lorella. 2023. On the use of sentiment analysis for linguistics research. Observations on sentiment polarity and the use of the progressive in Italian. *Frontiers in Artificial Intelligenc*, 6, 1101364. https://doi.org/10.3389/frai.2023.1101364.

Wright, David. 2020. The discursive construction of resistance to sex in an online community. *Discourse, Context & Media* 36: 100402. https://doi.org/10.1016/j.dcm.2020.100402.

Yang, Chao & Padmini Srinivasan. 2014. Translating surveys to surveillance on social media: Methodological challenges & solutions. In *Proceedings of the 2014 ACM conference on Web Science*, 4–12. New York: Association for Computing Machinery. https://doi.org/10.1145/2615569.2615696.

Yurchenko, Oleno & Nataliia Ugolnikova. 2021. Linguistic methods in social media marketing. In *Proceedings to the "International Conference on Computational Linguistics and Intelligent Systems"*, 12pp. https://ceur-ws.org/Vol-2870/paper55.pdf (last accessed 20 May 2022).

Zappavigna, Michele (2012). *Discourse of Twitter and social media: How we use language to create affiliation on the web*. London: Bloomsbury.

Zelena, András. 2020. The psychology of inclusion on new media platforms and the online communication. *Acta Universitatis Sapientiae, Communicatio* 7 (1). 54–67. https://doi.org/10.2478/auscom-2020-0005.