Originalveröffentlichung in: Kupietz, Marc / Schmidt, Thomas (Hrsg.):
Neue Entwicklungen in der Korpuslandschaft der Germanistik. Beiträge zur IDS-Methodenmesse 2022.
– Tübingen: Narr, 2023. S. 101-114. (Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 11)
DOI: *https://doi.org/10.24053/9783823396024*

LOUIS COTGROVE

# New opportunities for researching digital youth language: The *NottDeuYTSch* corpus

## Abstract

This article details the process of creating the ***Nott****inghamer Korpus* ***deu****tscher* ***Y****ou-****T****ube-****S****pra****ch****e* ('The Nottingham German *YouTube* Language Corpus' – or *NottDeuYTSch* corpus) and outlines potential research opportunities. The corpus was compiled to analyse the online language produced by young German-speakers and offers significant opportunity for in-depth research across several linguistic fields including lexis, morphology, syntax, orthography, and conversational and discursive analysis. The *NottDeuYTSch* corpus contains over 33 million words taken from approximately 3 million *YouTube* comments from videos published between 2008 to 2018 targeted at a young, German-speaking demographic and represent an authentic language snapshot of young German speakers. The corpus was proportionally sampled based on video category[1] and year from a database of 112 popular German-speaking *YouTube* channels in the DACH region for optimal representativeness and balance and contains a considerable amount of associated metadata for each comment that enable further longitudinal cross-sectional analyses. The *NottDeuYTSch* corpus is available for analysis as part of the *German Reference Corpus* (*DeReKo*).

**Keywords:** youth language, CMC, DMC, YouTube, German, digital communication, corpus linguistics

## 1. The need for the *NottDeuYTSch* corpus

*YouTube* is a significant source of authentic linguistic data created by young people. However, there are significant gaps in corpus linguistic scholarship within the field. The linguistic features used by young people in *YouTube* comments have rarely been analysed in studies of either Digitally Mediated Communication (DMC) or youth language, despite *YouTube* becoming one of the most-used online sites of communication in this demographic (Saferinternet.at 2018), with 86% of 12-19-year-olds reporting that they regular watched *YouTube* videos in 2018 (Bahlo et al. 2019, p. 80). To address this underdeveloped field of scholarship, I have constructed the *NottDeuYTSch* specifically to enable the investigation of the language of young German-speakers in digital spaces.

[1] For Germany and Austria, the complete list of 31 categories (translated into English) are as follows: film & animation, autos & vehicles, music, pets & animals, sports, short movies, travel & events, gaming, videoblogging, people & blogs, comedy, entertainment, news & politics, howto & style, education, science & technology, nonprofits & activism, movies, anime/animation, action/adventure, classics, comedy, documentary, drama, family, foreign, horror, sci-fi/fantasy, thriller, shorts, shows, and trailers.

The *NottDeuYTSch* corpus is a collection of over 33 million words written between 2008 and 2018 taken from the comment sections of 112 mainstream German-language *YouTube* channels that produce content targeted at young people. While other corpora of digital German language have been constructed, they have focused on other sources of data, e. g. websites and online forums (the *DECOW* corpus, Schäfer 2015; the *DWDS WebXL Korpus*, Geyken et al. 2017; Barbaresi/Geyken 2020), South Tyrolean *Facebook* texts (the *DiDi Korpus*, Glaznieks/Frey 2020), Internet Relay Chat (IRC) messages from students (the *Dortmunder ChatKorpus*, Beißwenger et al. 2015), *WhatsApp* messages (the *MoCoDa2* corpus, Beißwenger et al. 2020), and SMS, e-mail, IRC, Twitter, and Wikipedia article and discussion pages (the *IBK und Social Media-Korpora*, Lüngen/Kupietz 2020).

Some of the corpora aim to capture a wide range of DMC text types, but the majority are highly specialised: either focusing on one method of communication or on one target group. The range of specialised corpora demonstrates the "unparalleled and rapidly evolving diversity in terms of speakers and settings" in DMC (Barbaresi 2019, p. 29), although none of the above-mentioned corpora have exclusively focused on the language of young people. Indeed, Barbaresi (ibid., p. 30) advocates for the creation of more specialised corpora of online language, "to complement existing collections, as they allow for better coverage of specific written text types and genres, especially the language evolution seen through the lens of user-generated content, which gives access to a number of variants, socio- and idiolects". Androutsopoulos/Tereick (2016, pp. 366 f.) also advocate specifically for more linguistic research using *YouTube*, highlighting "comment interaction, remix and multimodality, discourse participation, performance and stylization of linguistic variability" as potential areas of study.

The *NottDeuYTSch* corpus answers these calls, providing an unparalleled opportunity for exploratory study of colloquial DMC of and between young people. The period covered by the corpus, 2008–2018, sits within the internet epoch referred to as Web 2.0 (O'Reilly 2005), an era of online and digital communication that began in the mid-2000s characterised by "social interaction and user-generated content", rather than information repositories (Herring 2013, p. 1). This decade was also an important period of technological transition from PC to mobile-based communication for many young people, who experienced the "digitalisation [of their] everyday lives" (Döring 2010, p. 161), acquiring personal access to the internet through smartphones, rather than being restricted to family or school computers or internet cafes. The corpus therefore can potentially capture any linguistic changes in digital youth language that may have accompanied the technological changes.

The article is divided into four sections. Section 2 presents the methodology behind selecting the data for the *NottDeuYTSch* corpus, including the guiding principles of building the corpus and identifying the *YouTube* channels from which the comments

were collected. Section 3 outlines the processes of constructing the *NottDeuYTSch* corpus, examining methodological concerns, such as corpus balance and size, and explains the sampling procedures used. Section 4 provides an overview of the *NottDeuYTSch* corpus and contains a breakdown of the key statistical features. Finally, section 5 outlines the potential applications of the corpus within future linguistic research.

## 2. Selecting the Data in the *NottDeuYTSch* Corpus

This section presents the methodological processes and principles behind selecting the data for the *NottDeuYTSch* corpus. Section 2.1 presents the aims and objectives of the corpus, and the typical content of the videos selected to provide comments for the corpus. Section 2.2 presents the case for treating the comments collected to construct the corpus can be considered as authentically produced by young people, and the ethical considerations surrounding the data. Finally, section 2.3 details the processes to identify the channels and videos to be included in the pre-corpus database, in preparation for sampling to create the *NottDeuYTSch* corpus.

### 2.1 Principles of building the *NottDeuYTSch* corpus

Five main factors governed the construction of the corpus, which ensure that it is balanced, representative, and able to be used in a wide range of future research:

1) The *NottDeuYTSch* corpus should represent, as best as possible, the language used by young German-speakers online. It is impossible to achieve perfect representativeness, but every effort has been made to ensure that the data were selected according to a strict methodology.
2) The data must be able to be analysed longitudinally.
3) The *NottDeuYTSch* corpus must be able to be used in comparison with other German-language corpora.
4) Only videos with over 100 comments were selected. My previous research on *YouTube* suggests that the average comment contains just over 10 tokens, so selecting videos with over 100 comments, should ensure that every video contributes (on average) over 1.000 words. A 1.000-word minimum sample size helps "to reliably represent the distributions of linguistic features" (Biber 1993, p. 252).
5) Videos must be published between July 2008 and October 2018. This ensured that all videos and comments were created after *YouTube* launched the localised version of the website for Germany on 8th November 2007, which had the effect of promoting German-language content to German speakers.

## 2.2 The identity of the commenters

The *NottDeuYTSch* corpus is intended to be a collection of authentic language created by German-speaking young people. However, verifying the age of the commenters presents a methodological challenge for the construction of the corpus, as this knowledge is not publicly available and is often not disclosed within a comment. The corpus was constructed following approaches suggested by Döring (2010, p. 164) that describe how an online user may present their digital identity to infer that the comments are generally written by young people. These include direct and indirect self-presentation of identifying information, such as statements about oneself and viewing habits, although language use was not considered as this was the focus of study. A more in-depth explanation of the application of Döring's principles to the data selection of the *NottDeuYTSch* corpus can be found in Cotgrove (2022, pp. 62–64). In summary, the videos were specifically selected for the corpus (as detailed in section 2.1.3 below) because they were produced to target a young German-speaking demographic and contain many instances of self-disclosure of relevant age. Therefore, we can assume the corpus reflects German-language youth culture. While there may be commenters who would not be counted as young people, the small size of this group, roughly 5% based on the self-disclosure statistics, would not significantly statistically affect the analyses.

## 2.3 Identifying relevant *YouTube* channels

In order to select the comments that comprise the *NottDeuYTSch* corpus, a database of channels was created. The process of identifying the channels was initially informed by my previous exposure to German-language *YouTube* culture. The channels identified had either received considerable media attention due to their *YouTube* popularity, such as *BibisBeautyPalace*, or were owned by media companies specifically targeted at young people, such as the *YouTube* channel of the radio station *1Live* (the youth station of *WDR*). Background information collected on Bibi from *BibisBeautyPalace* revealed that she often appeared on the front cover of *BRAVO*, the teen magazine. Due to the magazine's prominent role in German-speaking youth culture and regular news items involving German-language *YouTubers*, 63 of the 112 *YouTube* channels in the database featured in *BRAVO* cover stories and home page articles.

Additionally, music channels were added to the database by analysing the German music charts for successful German-speaking artists over the past 10 years and German music *YouTube* channels, such as *AggroTV*. Eight artist or music channels were selected who had the highest chart success, largest *YouTube* presence, and highest number of appearances in youth media (including *BRAVO*).

Five successful youth/online media platforms that have a high number of views and subscriptions on *YouTube*, such as *PromiFlash*, the leading *YouTube*-based news service aimed at young German speakers, were also included. Each of the five media outlet channels included in the corpus (*1Live*, *AGGRO.TV*, *Promiflash*, *RTL*, *World Wide Wohnzimmer*) has at least 50m views and has uploaded 500 videos, with *PromiFlash* leading the way with almost 2bn views and 1.2m comments.

I used a *YouTube* social aggregation website *SocialBlade* which lists the 250 channels in each of Germany, Austria, and Switzerland with the most subscribers. This achieved two goals. Firstly, it verified whether the channels had a large enough number of subscribers to be eligible for inclusion in the corpus. Secondly, using the Internet Archive[2] to view the page at various times since 2014, I was able to identify *YouTube* channels aimed at the demographic that were popular in the past and include them in the database. This was crucial to ensuring that the *NottDeuYTSch* corpus is as representative as possible of all years encompassed by the corpus, not just at time of its construction. This process added 18 further channels to the database, such as *Coldmirror*,[3] famous for Harry Potter parody videos.

The final process in expanding the database was to explore the 'Related channels' section on the 'About' page (as in fig. 3.3) from the 101 channels in the database identified up to this point. To do so, I used the 'YouTube Tools Channel Network Module',[4] which produces a list of channels that are similar to, or recommended by, the list of channels inputted. Combined with manual checks of the 'Related channels' sections, I added eleven more channels to the database. A breakdown of the sources of the channels for the *YouTube* corpus is presented in table 1.

| Channel identification process | Number of channels identified |
|---|---|
| Existing knowledge | 7 |
| BRAVO magazine covers and website | 63 |
| Music channels | 8 |
| Youth media channels | 5 |
| SocialBlade.com | 18 |
| Related channels | 11 |
| Total | 112 |

Table 1: Breakdown of sources used to identify channels included in the NottDeuYTSch Corpus

[2] https://archive.org/web (last accessed: 17-11-2022).

[3] www.youtube.com/user/coldmirror (last accessed: 17-11-2022).

[4] https://tools.digitalmethods.net/netvizz/youtube/mod_channels_net.php (last accessed: 17-11-2022).

## 3. Constructing the *NottDeuYTSch* Corpus

This section explains the methods taken to construct the *NottDeuYTSch* corpus. Section 3.1 outlines the process of extracting and cleaning of the data. Section 3.2 outlines the steps taken to ensure the corpus is as balanced and representative as can be. Section 3.3 explains how the corpus can be considered an appropriate size for a wide range of future linguistic analyses.

### 3.1 Extracting and cleaning the data

Using the statistical software, R (R Core Team 2021), custom code was written to interact with the *YouTube* Application Programming Interface (API) to import data on the channels in the database. This meant that the number of videos and comments could be established, and how they were distributed across video category and year for further sampling. The initial size of the pre-corpus database was 102.115 videos, and approximately 3.000 videos were removed as they did not have any comments that could be extracted, because the uploader either had disabled comments for that video or had streamed the video live through *YouTube*. This brought the total number of videos to 99.334. Whilst comments under a live-streamed video can be extracted using other methods, I chose not to include them as the interaction between commenters and the nature of their participation in a 'live' environment creates a different communicative environment: for example, comments simply express that a user is virtually present, rather than interacting with the content of the video or other users (Stenson 2020, p. 233).

### 3.2 Corpus representativeness and balance

One of the most important principles for the construction of the *NottDeuYTSch* corpus, is that it is 'representative' of the language used by young German-speakers in comments under mainstream *YouTube* videos, i. e. the findings in the corpus can be generalised to the wider population from which the data were sampled (Biber 1993, p. 243). The database contains information on the upload date and video category for every video uploaded by the 112 channels, as well as the timestamp for every comment written under the videos. The upload year and category of the video were selected to be the two parameters used to ensure the representativeness of the *NottDeuYTSch* corpus using stratified random sampling, which is an optimal method to ensure corpus 'balance' where a corpus contains "a wide range of text categories" (McEnery/Xiao/Tono 2006, p. 16), as is the case here.

## 3.3 Determining the size of the *NottDeuYTSch* corpus

As researchers on corpus linguistics have observed (e. g. Baker 2010), the appropriate size of a corpus varies depending on the features that are to be analysed. From a purely statistical standpoint, a chi-square test requires an expected value of at least five occurrences of a linguistic feature to successfully run the test. If the frequency of this feature occurs once every 10-000 tokens, then the corpus must contain at least 50.000 tokens.[5] The *NottDeuYTSch* corpus is intended to be large enough to analyse lexical, orthographical, morphological, and syntactic features, the last of which requires a corpus size of at least one million tokens, according to Baker (2010, pp. 95 f.). This should also be large enough to provide a suitable number of features for grammatical and morphosyntactic analysis, as well as offer the opportunity for longitudinal examination over the ten-year period covered by the corpus.

The total number of comments under the 99.334 videos in the database was over 150 million, which equates to roughly 1,5bn tokens. This amount of data would take too long to process and analyse within the scope the project, so, as noted above, I used stratified random sampling of the pre-corpus database based on the proportions of videos under each video category and year. The smallest acceptable size for the corpus was based on the number of videos that would contribute at least 1.000 comments in the smallest category (in this case 'pets & animals'), which would also provide a minimum of 10.000 tokens per category. This number of tokens was sufficient for the analyses planned for the project, as well as enabling possible future inter-categorical research, i. e. genre analyses. I therefore scaled the corpus down to find the number of comments needed per video category and year when the total number of comments in the 'pets & animals' category was equal to 1,000 comments. Based on this figure, the *NottDeuYTSch* corpus should therefore have 4.8 million comments with an approximate token count of 50 million.

The final proportions of the database were adjusted so that every intersection of video category and upload year contained at least one video, and each set of comments extracted from under a video contained complete conversational threads. This ensured that there was complete data for longitudinal, genre, and conversational analyses. Within each intersection, I devised a programmatic method to select videos with the closest number of comments to the proportion required, ensuring that a wide range of channels were selected to provide videos, as well as videos with a wide spread of the number of comments. For an in-depth explanation of these methods see Cotgrove (2022, pp. 78 f.). The final number of comments extracted in every intersection is provided in the appendix (Table 3).

[5] However, see Kilgarriff (2005) and Koplenig (2017) on the pitfalls of statistical significance testing in corpus linguistics.

## 4. Statistical overview of the *NottDeuYTSch* corpus

A statistical overview of the *NottDeuYTSch* corpus is presented in table 2 outlining the token count, total number of comments, and key averages of the corpus. The mean number of tokens per comment (10,72) correlates with the average found in my previous research on the language of young German-speakers on *YouTube* (Cotgrove 2017). The type-token ratio of the *NottDeuYTSch* corpus (0,017) is slightly lower than that of the *DWDS-Kernkorpus* (0,021) (Geyken 2010, p. 1), which indicates less lexical diversity, i.e. commenters use the same words more often (Kettunen 2014, p. 223), but the closeness of the figures implies that young people's vocabulary in *YouTube* comments is almost as broad as that found in general written communication by adults.

| Statistic | Value |
|---|---|
| Number of tokens (including emoji and emoticons) | 33.760.494 |
| Number of tokens (only lexemes) | 32.549.462 |
| Number of types | 567.086 |
| Type-token ratio (TTR) | 0,017 |
| Number of comments | 3.149.457 |
| Number of videos | 296 |
| YouTube channels represented | 63 |
| Mean tokens per comment | 10,72 |
| Median tokens per comment | 5 |
| Mean comments per video | 1.914 |

Table 2: Statistical overview of the *NottDeuYTSch* corpus

The extracted numbers of comments for each intersection were consistently lower than the target, as shown in table 2 above. This was a trend for most intersections, and it was discovered that the reported number of comments by the *YouTube* Application Programming Interface was different to the number of comments that it was possible to extract. Some of the differences can be explained by the videos selected having fewer comments than the target number. However, 1,7 million comments of the predicted 4,8 million were not available to download using the *YouTube* API. The main reason for this shortfall is that the comments had been removed from *YouTube* but were still counted by the *YouTube* API.[6] Despite this shortfall, the *NottDeuYTSch*

[6] Comments can be removed by the commenter, the channel owner, or by *YouTube* themselves, if the comment violates their community guidelines.

corpus, with a total of 3,1 million comments from 296 covering 10 years of data from 2008 to 2018, is still a suitable size to answer the research areas covered above.

Furthermore, the targeted proportions for the distribution of comments per video category and year were generally met, although the comment timestamps in the *NottDeuYTSch* corpus are slightly more weighted towards later years. For most videos, the bulk of the comments are posted within the first two months of the upload date. However, commenters revisit older *YouTube* videos and leave comments, such as "Who is still watching this in 2017?" under a video uploaded in 2008, which is the major contributing factor to the slight difference between the targeted and achieved proportions for the distribution of comments per video category and year. This does not pose any thorny methodological problems, as the comments are timestamped.

## 5. Applications of the *NottDeuYTSch* corpus

The *NottDeuYTSch* corpus is one of the first large corpora of linguistic data containing language written specifically by young German-speakers in *YouTube* comments, an important and popular site of youth culture and discourse. The corpus is thus a significant contribution to corpora of online data, complementing existing corpora mentioned in section 1, which focus on other areas of online language, such as the *MoCoDa2* corpus of WhatsApp messages (Beißwenger et al. 2020), the *DiDi* corpus of Facebook texts (Glaznieks/Frey 2020), and the *IBK* corpus of multiple online sources, e. g. emails, IRC chats, and blogs (Lüngen/Kupietz 2020). The *NottDeuYTSch* corpus offers a wide range of new possibilities for study and is now available as part of the *German Reference Corpus* (*DeReKo*, Leibniz-Institut für Deutsche Sprache 2022). The structured sampling of the data over the time frame of the corpus enables a wide range of longitudinal studies for lexical, orthographical, and morphosyntactic features, as shown in examples 1 to 3. Videos and comments contain a wealth of metadata, which can facilitate a wide range of future research, e. g., analyses of video genres, time frames, users, or *YouTubers*. The metadata also allow interactional and discourse analyses of interactions between commenters as it preserves the comment structure on a page, i. e., parent comments and replies.

**Example 1 (2015)**
Hey [YOUTUBER] Es wär oberMEGAsuperHammerGeilo wenn ich dabei sein könnte 😁
('Hey [YOUTUBER] It would be "above-MEGA-super-hammercool-o" if I could be there 😁')

**Example 2 (2012)**
Ich möchte gewinnen weil wegen is so ;D
('I would like to win because cos of it is like that ;D')

**Example 3 (2018)**
Hey ich liebe 😍3 Uhr nachts wiedios 🐴🐴🐴🐴👧
('hey I love 😍3am videos 🐴🐴🐴🐴👧')

The large size of the *NottDeuYTSch* corpus allows for considerable quantitative research, including the investigation of features that do not occur frequently, such as some syntactic constructions, as well as linguistic features specific to Digitally-Mediated Communication, such as emoji and hashtags, where a large amount of data is required for linguistic study beyond qualitative analysis. The comments in the *NottDeuYTSch* corpus are predominantly written in German (including dialect use), but there is also a significant presence of other languages, such as English, Turkish, and Russian, including linguistic elements from multiple languages within the same comment, and the corpus can also be used for potential quantitative and qualitative analyses of multilingualism. For example, in Cotgrove (2022), three linguistic case studies were presented, which each focus on a different area of linguistics; lexis, morphosyntax, and orthography, demonstrating the wide applicability of the *NottDeuYTSch* corpus to analyse the digital writing of young people. It is hoped that further research in this vein can be produced with the *NottDeuYTSch* corpus.

| Category | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Autos & Vehicles** | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 3.508 (0,11%) | 2.306 (0,07%) | 704 (0,02%) | 6.518 (0,21%) |
| **Comedy** | 434 (0,01%) | 439 (0,01%) | 2.986 (0,09%) | 1.686 (0,05%) | 5.405 (0,17%) | 9.326 (0,3%) | 23.731 (0,75%) | 10.279 (0,33%) | 25.447 (0,81%) | 36.908 (1,17%) | 13.116 (0,42%) | 129.757 (4,12%) |
| **Education** | 0 (0%) | 29 (<0,01%) | 196 (0,01%) | 300 (0,01%) | 139 (<0,01%) | 19 (<0,01%) | 115 (<0,01%) | 2.470 (0,08%) | 28.046 (0,89%) | 5.422 (0,17%) | 3.914 (0,12%) | 40.650 (1,29%) |
| **Entertainment** | 494 (0,02%) | 2.590 (0,08%) | 7.236 (0,23%) | 13.854 (0,44%) | 22.588 (0,72%) | 70.378 (2,23%) | 39.747 (1,26%) | 81.248 (2,58%) | 340.062 (10,8%) | 348.523 (11,07%) | 203.871 (6,47%) | 1.130.591 (35,9%) |
| **Film & Animation** | 0 (0%) | 86 (<0,01%) | 3.781 (0,12%) | 3.204 (0,1%) | 3.503 (0,11%) | 1.056 (0,03%) | 4.561 (0,14%) | 2.124 (0,07%) | 4.373 (0,14%) | 4.044 (0,13%) | 2.069 (0,07%) | 28.801 (0,91%) |
| **Gaming** | 0 (0%) | 24 (<0,01%) | 118 (<0,01%) | 883 (0,03%) | 49.482 (1,57%) | 89.822 (2,85%) | 78.498 (2,49%) | 97.392 (3,09%) | 124.231 (3,94%) | 70.262 (2,23%) | 30.754 (0,98%) | 541.466 (17,19%) |
| **Howto & Style** | 5 (<0,01%) | 380 (0,01%) | 3.585 (0,11%) | 5.119 (0,16%) | 5.188 (0,16%) | 15.998 (0,51%) | 52.969 (1,68%) | 63.920 (2,03%) | 323.636 (10,28%) | 345.654 (10,98%) | 51.115 (1,62%) | 867.569 (27,55%) |
| **Music** | 0 (0%) | 93 (<0,01%) | 1.219 (0,04%) | 849 (0,03%) | 1.610 (0,05%) | 1.723 (0,05%) | 2.227 (0,07%) | 1.210 (0,04%) | 5.446 (0,17%) | 35.710 (1,13%) | 7.436 (0,24%) | 57.523 (1,83%) |
| **News & Politics** | 0 (0%) | 86 (<0,01%) | 2.699 (0,09%) | 416 (0,01%) | 195 (0,01%) | 195 (0,01%) | 339 (0,01%) | 46 (<0,01%) | 3.390 (0,11%) | 1.071 (0,03%) | 1.654 (0,05%) | 10.091 (0,32%) |
| **Nonprofits & Activism** | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 184 (0,01%) | 18 (<0,01%) | 68 (<0,01%) | 1.571 (0,05%) | 142 (<0,01%) | 34 (<0,01%) | 2.017 (0,06%) |
| **People & Blogs** | 25 (<0,01%) | 818 (0,03%) | 566 (0,02%) | 3.405 (0,11%) | 6.290 (0,2%) | 12.501 (0,4%) | 7.258 (0,23%) | 16.762 (0,53%) | 52.608 (1,67%) | 77.513 (2,46%) | 16.871 (0,54%) | 194.617 (6,18%) |
| **Pets & Animals** | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 0 (0%) | 1.053 (0,03%) | 8 (<0,01%) | 3 (<0,01%) | 1.064 (0,03%) |
| **Science & Technology** | 0 (0%) | 0 (0%) | 32 (<0,01%) | 54 (<0,01%) | 20 (0%) | 6 (<0,01%) | 12 (<0,01%) | 51 (<0,01%) | 986 (0,03%) | 1.101 (0,03%) | 11 (<0,01%) | 2.273 (0,07%) |
| **Shows** | 0 (0%) | 245 (0,01%) | 4.896 (0,16%) | 5.747 (0,18%) | 26.482 (0,84%) | 37.874 (1,2%) | 1.951 (0,06%) | 1.775 (0,06%) | 1.865 (0,06%) | 1.152 (0,04%) | 616 (0,02%) | 82.603 (2,62%) |
| **Sports** | 11 (<0,01%) | 56 (<0,01%) | 170 (0,01%) | 565 (0,02%) | 948 (0,03%) | 1.520 (0,05%) | 1.801 (0,06%) | 479 (0,02%) | 5.801 (0,18%) | 11.819 (0,38%) | 2.368 (0,08%) | 25.538 (0,81%) |
| **Travel & Events** | 0 (0%) | 26 (<0,01%) | 138 (<0,01%) | 198 (0,01%) | 204 (0,01%) | 384 (0,01%) | 2.079 (0,07%) | 4.956 (0,16%) | 12.149 (0,39%) | 6.272 (0,2%) | 1.973 (0,06%) | 28.379 (0,9%) |
| **Total** | 969 (0,03%) | 4.872 (0,15%) | 27.622 (0,88%) | 36.280 (1,15%) | 122.054 (3,88%) | 240.986 (7,65%) | 215.306 (6,84%) | 282.780 (8,98%) | 934.172 (29,66%) | 947.907 (30,1%) | 336.509 (10,68%) | 3.149.457 (100%) |

Table 3: Number of comments per video category and year in the *NottDeuYTSch* corpus

## References

Androutsopoulos, Jannis/Tereick, Jana (2016): YouTube: language and discourse practices in participatory culture. In: Georgakopoulou, Alexandra/Spilioti, Tereza (eds.): The Routledge handbook of language and digital communication. (= Routledge Handbooks in Applied Linguistics). Abingdon: Routledge, pp. 354–370.

Bahlo, Nils/Becker, Tabea/Kalkavan-Aydın, Zeynep/Lotze, Netaya/Marx, Konstanze/Schwarz, Christian/Şimşek, Yazgül (2019): Jugendsprache: Eine Einführung. Berlin: Metzler.

Baker, Paul (2010): Sociolinguistics and Corpus Linguistics. (= Edinburgh Sociolinguistics). Edinburgh: Edinburgh University Press.

Barbaresi, Adrien (2019): The vast and the focused: on the need for thematic web and blog corpora. In: Bánski, Piotr/Barbaresi, Adrian/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Iliadi, Caroline (eds.): Proceedings of the workshop on challenges in the management of large corpora (CMLC-7), Cardiff, 22 July 2019. Mannheim: Leibniz-Institut für Deutsche Sprache, pp. 29–32.

Barbaresi, Adrien/Geyken, Alexander (2020): Die Webkorpora im DWDS –Strategien des Korpusaufbaus und Nutzungsmöglichkeiten. In: Marx/Lobin/Schmidt (eds.), pp. 345–348.

Beißwenger, Michael/Ehrhardt, Eric/Horbach, Andrea/Lüngen, Harald/Steffen, Diana/Storrer, Angelika (2015): Adding value to CMC corpora: CLARINification and part-of-speech annotation of the Dortmund Chat Corpus. In: Beißwenger, Michael/Zesch, Torsten (eds.): Proceedings of the 2nd workshop on natural language processing for computer-mediated communication/social media at GSCL2015 (NLP4CMC2015) University of Duisburg-Essen, September 28. German Society for Computational Linguistics & Language Technology, pp. 12–16.

Beißwenger, Michael/Fladrich, Marcel/Imo, Wolfgang/Ziegler, Evelyn (2020): Die Mobile Communication Database 2 (MoCoDa 2). In: Marx/Lobin/Schmidt (eds.), pp. 349–352.

Biber, Douglas (1993): Representativeness in corpus design. In: Literary and Linguistic Computing 8,4, pp. 243–257.

Cotgrove, Louis A. (2017): #GlockeAktiv: gender and ethnicity differences in German-language YouTube comments. Master's Thesis. Nottingham: University of Nottingham.

Cotgrove, Louis A. (2022): #GlockeAktiv: a corpus linguistic investigation of German online youth language. PhD Thesis. Nottingham: University of Nottingham.

Döring, Nicola (2010): Sozialkontakte online: Identitäten, Beziehungen, Gemeinschaften. In: Schweiger, Wolfgang/Beck, Klaus (eds.): Handbuch Online-Kommunikation. Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 159–83.

Geyken, Alexander (2010): Statistical variations of German support verb constructions in very large corpora. In: A Way with Words, pp. 169–186.

Geyken, Alexander/Barbaresi, Adrien/Didakowski, Jörg/Jurish, Bryan/Wiegand, Franck/Lemnitzer, Lothar (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen

Sprache“ (DWDS). In: Zeitschrift für germanistische Linguistik 45, 2, pp. 327–344. https://doi.org/10.1515/zgl-2017-0017.

Glaznieks, Aivars/Frey, Jennifer-Carmen (2020): Das DiDi-Korpus: Internetbasierte Kommunikation aus Südtirol. In: Marx/Lobin/Schmidt (eds.), pp. 353–354.

Herring, Susan C. (2013): Discourse in web 2.0: familiar, reconfigured, and emergent. In: Tannen, Deborah/Trester, Anne M. (eds.): Discourse 2.0: language and new media. Washington DC: Georgetown University Press, pp. 1–26.

Kettunen, Kimmo (2014): Can type-token ratio be used to show morphological complexity of languages?In: Journal of Quantitative Linguistics 21, 3, pp. 223–45. https://doi.org/10.1080/09296174.2014.911506.

Kilgarriff, Adam (2005): Language is never, ever, ever, random. In: Corpus Linguistics and Linguistic Theory 1, 2, pp. 263–276. https://doi.org/doi:10.1515/cllt.2005.1.2.263.

Koplenig, Alexander (2017): Against statistical significance testing in corpus linguistics. In: Corpus Linguistics and Linguistic Theory 15, 2, pp. 321–346.

Leibniz-Institut für Deutsche Sprache (2022): Deutsches Referenzkorpus/Archiv der Korpora geschriebener Gegenwartssprache 2022-I. (Release vom 08.03.2022). Mannheim: Leibniz-Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (last accessed: 14-11-2022).

Lüngen, Harald/Kupietz, Marc (2020): IBK- und Social Media-Korpora am Leibniz-Institut für Deutsche Sprache. In: Marx/Lobin/Schmidt (eds.), pp. 319–342. https://doi.org/10.1515/9783110679885-016.

Marx, Konstanze/Lobin, Henning/Schmidt, Axel (eds.): Deutsch in Sozialen Medien: Interaktiv – Multimodal – Vielfältig. (= Jahrbuch des Instituts für Deutsche Sprache 2019). Berlin/Boston: De Gruyter.

McEnery, Tony/Xiao, Richard/Tono, Yukio (2006): Corpus-based language studies: an advanced resource book. (= Routledge applied linguistics). Abingdon: Routledge.

O’Reilly, Tim (2005): What is web 2.0?: design patterns and business models for the next generation of software. www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html (last accessed: 14-11-2022).

R Core Team (2021): R: a language and environment for statistical omputing. Vienna, Austria: R Foundation for Statistical Computing. www.R-project.org/ (last accessed: 14-11-2022).

Saferinternet.at. (2018): Jugend-Internet-Monitor 2018. Saferinternet.at. 2018. www.saferinternet.at/presse-detail/jugend-internet-monitor-2018/ (last accessed: 14-11-2022).

Schäfer, Roland (2015):Processing and querying large web corpora with the COW14 architecture. In: Bański, Piotr/Biber, Hanno/Breiteneder, Evelyn/Kupietz, Marc/Lüngen, Harald/Witt, Andreas (eds.): Proceedings of the 3rd workshop on challenges in the management of large corpora (CMLC-3), Lancaster, 20 July 2015. Mannheim: Institut für Deutsche Sprache, pp. 28–34. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/3836 (last accessed: 14-12-2022).

Stenson, Robert (2020): "TUNE IN/JOIN US": mobilising liveness as a promotional strategy in film trailer exhibition. PhD Thesis. Nottingham: University of Nottingham.