Originally published in: Korpora Deutsch als Fremdsprache vol. 3 (2023) no. 2, pp. 225-229.
DOI: *https://doi.org/10.48694/kordaf.3841*

Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 2 (Dezember 2023)
ISSN 2749-4411

THE NOTTDEUYTSCH CORPUS: A corpus of German-language YouTube comments

Louis Cotgrove
Leibniz-Institut für Deutsche Sprache Mannheim

Abstract

In diesem Beitrag wird das Nottinghamer Korpus deutscher YouTube-Sprache (das NottDeuYTSch-Korpus) vorgestellt. Das Korpus hat eine Größe von über 33 Millionen Wörtern, die aus etwa 3 Millionen YouTube-Kommentaren gesammelt wurden. Die Kommentare wurden zwischen 2008 und 2018 veröffentlicht und wurden von einer Gruppe von überwiegend jungen Deutschsprachigen geschrieben. Das NottDeuYTSch-Korpus bietet einen authentischen und repräsentativen sprachlichen Schnappschuss junger Deutschsprachiger und ermöglicht umfangreiche Forschungsmöglichkeiten in verschiedenen linguistischen Bereichen wie Lexik, Morphologie, Syntax, Orthografie, Multilingualismus, sowie Gesprächs- und Diskursanalyse.

Keywords: Korpuslinguistik; digitale Kommunikation; Deutsch; Multilingualismus; Jugendsprache

Abstract

This paper introduces the Nottinghamer Korpus deutscher YouTube-Sprache ('The Nottingham German YouTube Language Corpus' - or NottDeuYTSch corpus). The corpus comprises over 33 million words, taken from roughly 3 million YouTube comments published between 2008 and 2018, written by a young, German-speaking demographic. The *NottDeuYTSch* corpus provides an authentic and representative linguistic snapshot of young German speakers and offers significant opportunities for in-depth research in several linguistic fields, such as lexis, morphology, syntax, orthography, multilingualism, and conversational and discursive analysis.

Keywords: corpus linguistics; YouTube; CMC; online language; German; multilingualism; youth language

1. The importance of researching digital youth language

YouTube is a valuable source of authentic linguistic data written by young German-speakers, yet corpus linguistic scholarship in this field has been limited, despite the widespread use of YouTube by this demographic, with over 85% of young Germans regularly accessing the platform (cf. Bahlo et al. 2019: 80; Statista 2020). While there has been a steady increase in other German-language corpora comprised of digitally-mediated communication (DMC), none of them have explicitly focussed on youth language written on YouTube.[1]

The NottDeuYTSch corpus comprises over 33 million words from YouTube comments under the videos of 112 German-language channels targeted at young people and provides a unique opportunity for exploratory study of colloquial Digitally Mediated Communication (DMC) among young German-speakers. The corpus covers the period of 2008-2018, a crucial decade in the transition from PC to mobile-based communication for many young people. By investigating the linguistic features used by young German-speakers in digital spaces, the NottDeuYTSch corpus can potentially reveal any linguistic changes that may have accompanied technological changes during this period. The

[1] Existing corpora include those using data from websites and forums (e.g. IBK und Social Media-Korpora, cf. Lüngen / Kupietz 2020), Facebook (DiDi Korpus, cf. Glaznieks / Frey 2020), and WhatsApp messages (MoCoDa2 corpus, cf. Beißwenger et al. 2020).

corpus complements and extends the research potential of existing corpora of DMC, and the communicative differences between the corpora demonstrate "unparalleled and rapidly evolving diversity in terms of speakers and settings" in DMC (Barbaresi 2019: 29). To further advance our understanding of the diverse and changing nature of online language, the creation of more specialised corpora of online language, such as the *NottDeuYTSch* corpus, is required, as they can provide valuable information specific written text types and genres.

2. Constructing the NottDeuYTSch corpus

The NottDeuYTSch corpus was built using five guiding principles to ensure balance and representativeness, as well as future application to a wide range of linguistic research:

1. The corpus is representative of the language used by young German-speakers in YouTube comments.
2. The corpus can be analysed longitudinally.
3. The corpus can be analysed comparatively with other corpora.
4. Every video must have comments amounting to a 1,000-word minimum sample size "to reliably represent the distributions of linguistic features" (Biber 1993: 252).
5. Every video must be published between July 2008 and October 2018 to ensure all comments were written after YouTube launched the localised German version of the website.[2]

2.1 Data selection

The NottDeuYTSch corpus was created by selecting comments from YouTube channels in German-speaking countries. The channels were identified based on my own previous exposure to German-language YouTube culture, appearances in youth media, such as *BRAVO* magazine, and ownership by media companies targeting young people, e.g. *1Live* from *WDR*. To explore popular YouTube channels in the start of the corpus 2008, I used a combination of the Internet Archive,[3] which allows a user to view websites at particular points in time with *SocialBlade*,[4] a website which lists the 250 channels in each of Germany, Austria, and Switzerland with the most subscribers.

It was important to ensure that the corpus was representative of the language used by young German-speakers online. To achieve this, custom R code was used to extract data using the YouTube API, which was then used to create a database of comments from all the videos of the selected channels, except comments under live-streamed videos, as this would not create a consistent communicative environment. This would have resulted in a corpus of over 1.5 billion tokens, which was too large for the scope of the project. The database was then sampled down to roughly 3 million comments using stratified random sampling of the year and video category to maintain "a wide range of text categories" for optimal balance (McEnery / Xiao / Tono 2006: 16). For further information on the construction and sampling process, see Cotgrove (2022: 59). Table 1 provides a statistical overview of the NottDeuYTSch corpus.

[2] Please note, the videos or transcripts of the videos are not included in the corpus.
[3] https://web.archive.org/ (14.04.2023).
[4] https://socialblade.com/youtube/top/country/de/mostsubscribed (14.04.2023).

Statistic	Value
Tokens (inc. emoji and emoticons)	33,760,494
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	10.720
Median Tokens per Comment	5
Mean Comments per Video	1,914

Table 1
Statistical overview of the *NottDeuYTSch* corpus, adapted from Cotgrove (2022: 343)

3. The NottDeuYTSch corpus as a resource for linguistic research

The NottDeuYTSch corpus is a valuable resource for linguistic research, as it provides a large, representative sample of the language used by young German-speakers online. The NottDeuYTSch corpus is suitable for many different kinds of quantitative and qualitative projects, lexical, orthographical, morphosyntactic, and syntactic studies, interactional and discourse analyses, as well as investigations into multilingualism. Furthermore, the metadata enables longitudinal studies of language changes, as well as text and genre studies. For example, Figure 1 shows a comparison between the change in frequencies in three intensifiers, *geil*, *cool*, and *mega*:

RegEx Queries: [/g+e+i+l/gi] [/c+o{2,}l/gi] [/m+e+g+a/gi]

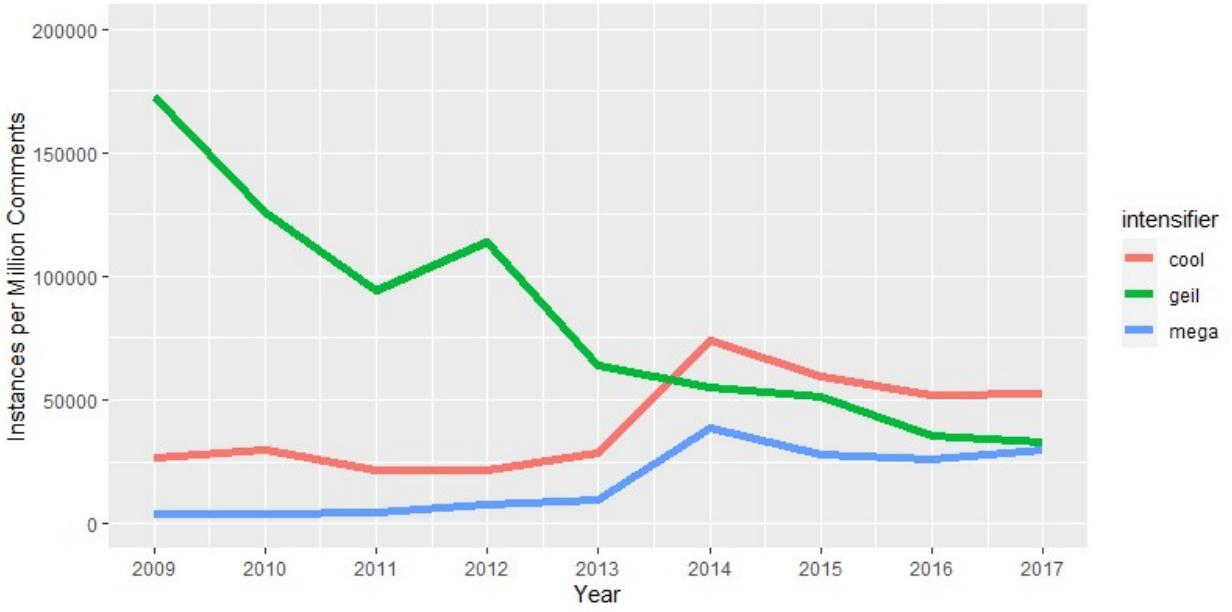

Figure 1
Frequency of comments containing selected intensifiers over time

From the graph, we see dramatic changes in the choice of intensifier used by young people, with *geil* falling out of fashion, *cool* steadily increasing, overtaking *geil*, and *mega* dramatically increasing over the time period of the corpus. This demonstrates lexical change on a microdiachronic scale.

The *NottDeuYTSch* corpus is available for download in many different formats (see Cotgrove 2018), and has been integrated into the German Reference Corpus (DeReKo) (Leibniz-Institut für Deutsche Sprache 2022).

Bibliography

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

Barbaresi, Adrien (2019): The Vast and the Focused: On the Need for Thematic Web and Blog Corpora. In: Bański, Piotr; Barbaresi, Adrien / 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*. Mannheim: Leibniz-Institut für Deutsche Sprache, 29-32.

Beißwenger, Michael et al. (2020): Die Mobile Communication Database 2 (MoCoDa 2). In: Marx, Konstanze / Lobin, Henning / Schmidt, Axel (Hg.): *Deutsch in Sozialen Medien: Interaktiv – Multimodal – Vielfältig*. Berlin: de Gruyter, 349-352.

Biber, Douglas (1993): Representativeness in Corpus Design. In: Literary and Linguistic Computing, 8: 243-257.

Cotgrove, Louis Alexander (2018): Das Nottinghamer Korpus Deutscher YouTube-Sprache (the Nott-DeuYTSch corpus). LINDAT/CLARIAH-CZ. http://hdl.handle.net/11372/LRT-4806.

Cotgrove, Louis Alexander (2022): #GlockeAktiv: A corpus linguistic investigation of German online youth language. Nottingham: University of Nottingham. https://eprints.nottingham.ac.uk/id/eprint/69043 (21.07.2023).

Glaznieks, Aivars / Frey, Jennifer-Carmen (2020): Das DiDi-Korpus: Internetbasierte Kommunikation aus Südtirol. In: Marx, Konstanze / Lobin, Henning / Schmidt, Axel (Hg.): *Deutsch in Sozialen Medien: Interaktiv – Multimodal – Vielfältig*. Berlin: de Gruyter, 353-354.

Leibniz-Institut für Deutsche Sprache (2022): IDS: Korpuslinguistik: Korpusausbau. http://www1.ids-mannheim.de/kl/projekte/korpora.html (14.04.2023).

Lüngen, Harald / Kupietz, Marc (2020): IBK- und Social Media-Korpora am Leibniz-Institut für Deutsche Sprache. In: Marx, Konstanze / Lobin, Henning / Schmidt, Axel (Hg.): *Deutsch in Sozialen Medien: Interaktiv – Multimodal – Vielfältig*. Berlin: de Gruyter, 319-342. https://doi.org/10.1515/9783110679885-016.

McEnery, Tony / Xiao, Richard / Tono, Yukio (2006): *Corpus-Based Language Studies: An Advanced Resource Book*. Abingdon: Routledge.

Statista (2020): Jugendliche - Beliebteste Internetangebote 2020. https://de.statista.com/statistik/daten/studie/419810/umfrage/beliebteste-internetangebote-bei-jugendlichen/ (14.04.2023).

Biographische Notiz: Dr Louis Cotgrove is a researcher in the department of Lexicology at the Leibniz Institute for the German Language (IDS) in Mannheim. His research specialities include corpus linguistic investigation of youth and online language, as well as text technology and analysis in online digital lexicography and empirical lexicology, and developing data infrastructure for online dictionaries.

Contact address:
Dr. Louis Cotgrove
Leibniz-Institut für Deutsche Sprachen
68161, Mannheim
Germany
cotgrove@ids-mannheim.de