# Annotating *candy speech* in German YouTube comments

**Yulia Clausen and Tatjana Scheffler**
CRC 1567 Virtual Lifeworlds
Ruhr-University Bochum, Germany
{yulia.clausen|tatjana.scheffler}@rub.de

## Abstract

We describe the phenomenon of *candy speech* – positive emotional speech in online communication – and introduce a categorization of its various types based on the theoretical framework of social interaction by Goffman (1967). We provide a dataset of 46,286 German YouTube comments manually annotated with candy speech types; 14,580 comments in this data contain a total of 21,785 candy speech expressions. We discuss issues in the annotation and evaluation of such higher-level semantic properties of text.

## 1 Introduction

The theoretical framework of social interaction introduced by Goffman (1967) is centered around *face-work*, where *face* represents a 'positive social value a person effectively claims for [themselves] [...] an image of self delineated in terms of approved social attributes' (p. 5). In this approach, social interactions involve emotionally charged linguistic utterances which directly influence a person's image or face. Goffman (1967) assumes various states and processes related to face: An individual is said to be 'in face' when they feel confident and assured, hence one strives to 'maintain one's face', i.e., to sustain a positive image of oneself. At the same time, one fears to 'lose face', which could result in a damage to one's image. In cooperative discourse, mutual face support is desired and even expected, and, if heeded, ensures that faces are maintained. Furthermore, 'face-saving' and 'face-giving' strategies can be applied when face is lost. The former allows an individual to sustain an impression that they have not lost their face, while the latter refers to the process by which others help an individual to 'gain face'.

In linguistics, face-work plays a central role, as it provides insight into how language functions not only as a medium for conveying information, but also as a means to manage social relationships, shape interpersonal dynamics, and construct identities in interactions. Nonetheless, very few studies have addressed positive interactions in social media from a corpus-based perspective via annotation of significant amounts of realistic data or using computational approaches. Annotation efforts have so far centered on *negative* online interactions, and linguistic expressions that negatively influence another person's or group's public image have been extensively studied. The area of negative communication practices has been delineated in great detail, with distinctions between hate speech, offensive language, toxicity, and many other subtypes (see Poletto et al., 2021, for a survey, and references therein). In contrast, little empirical work has been done on the positive side, despite the fact that (as we believe) positive face-work is similarly complex, and despite the fact that positive social engagement leads many users to strongly associate with certain virtual communities and spend large amounts of time interacting online. The lack of empirical research on positive face-work means that we know very little on how it looks and how to identify it in online data. Studying the types of phenomena that make up positive interactions in digital media may enable us to automatically find and possibly enhance positive face-work, and may help us understand how virtual communities and identities are constructed through language.

In this study, we focus on *candy speech* – a term we use for positive face-work in online discourse that provides face support for others. We develop a classification of candy speech types that allows for a differentiated view of face-supporting strategies. Some previous work has already documented the prevalence of (certain types of) positive speech in social media (e.g., Chakravarthi and Muralidaran 2021; Jiménez-Zafra et al. 2023 on 'hope speech' or Njoo et al. 2023 on 'empowerment language'). Face-work, in particular positive face-work, has

however rarely been directly addressed in corpus or computational linguistic studies (but see Dutt et al., 2020; Klüwer, 2011; Klüwer, 2015; Virtanen, 2022). Specifically, Klüwer's (2011; 2015) work on small talk in task-oriented dialogs, which she frames in face-work terms, is relevant for our study. Klüwer (2011; 2015) develops a taxonomy of dialog acts for non-task-oriented passages in virtual reality dialogs based on the notion that these interactions typically serve social purposes: to either request support for one's own face, or to provide face support for the interlocutor. In our classification of candy speech, we build on and extend Klüwer's face supporting dialog acts based on social media interactions between real humans.

Our main contributions are the following:

- We develop a definition and subcategorization of candy speech in social media comments.
- We annotate a subset of a German YouTube corpus and discuss first observations regarding the distribution of candy speech expressions.
- We present an evaluation method for comparing span-based candy speech annotations and apply it to our corpus data.

## 2 Dataset

We work with the data from the NottDeuYTSch corpus (Cotgrove, 2018), which contains over 33 million words taken from approximately 3 million YouTube comments published between 2008 and 2018 by a young German-speaking audience. Comments posted on social media platforms often represent emotional discourse. In addition, it is known that YouTube comments in particular contain many positive social interactions, for example within fan groups and other communities (Cotgrove, 2025), thus being suitable for our purposes.

We selected 16 videos authored by seven creators, together with all their comments. To reflect the topic distribution in the original corpus, the creators/videos were selected randomly; however, we made sure that the creators represent different sectors (e.g., music, tutorials) so that the commenting communities can be expected to differ in the frequency and types of candy speech expressions. The annotated dataset consists of a total of 46,286 comments, grouped into 16 'documents' according to the video they relate to.[1]

[1]The dataset and annotation guidelines are available via the OSF platform: https://osf.io/r9uek/.

## 3 Candy speech

### 3.1 Definition

Following Goffman's (1967) theory, we define candy speech as face-support that aims to help others maintain and restore their positive (self-)image. Candy speech thus is constituted by expressions of positive attitudes and feelings on social media towards individuals (e.g., content creators or commenters) and their posts (videos, comments, etc.). The purpose of candy speech is to encourage, cheer up, support or empower others. Candy speech can be viewed as the counterpart to hate speech, as it likewise aims to influence the self-image of the target person or group, but in a positive way. In the following section, we describe our classification of candy speech expressions against the backdrop of face-work strategies.

### 3.2 Classification

Our classification includes 10 annotation categories: eight distinct types of candy speech and two additional categories. An overview of all candy speech types is given in Table 1. The additional categories are *implicit* and *ambiguous*. The annotation *implicit* is used for indirect expressions of one of the eight explicit types. The label *ambiguous* applies to cases in which the lack of context prevents an expression from being clearly classified as candy speech or not.

The candy speech types realize face-supporting strategies directed at others, which we broadly divide into two classes: those conveying positive disposition toward individuals and those claiming shared common ground (Stalnaker, 2002) with an individual or a group. Positive disposition is realized by the types *affection declaration*, *compliment*, *encouragement*, *gratitude*, *positive feedback* and *sympathy*. It can also be expressed implicitly. Claiming of common ground is done via using markers of *group membership* or signaling *agreement*.

Additionally, we label each comment containing candy speech as *initiative* or *reactive*, which allows us to differentiate between spontaneous acts of face support (initiative) and replies to other comments (reactive). Reactive comments can represent face-supporting or face-saving acts, depending on whether they refer to candy speech expressions (e.g., agreement) or aim at counteracting face threats initiated by others (e.g., compliments on positive achievements of the target person).

| Type | Short definition | Example |
|---|---|---|
| affection declaration | admiration, love and affection towards others | *I like you XD* |
| compliment | acknowledgment of skills, personal characteristics or achievements of others | *You create really great videos !* |
| encouragement | comments that aim to encourage others | *Keep at it !* |
| gratitude | sincere gratitude expressed unprompted | *Thanks for motivating me !* |
| group membership | markers of group membership, e.g., belonging to a fan community | *I am a #lochinator* |
| positive feedback | positive attitude toward a post, video, comment etc. | *The song is mega mega cool .* |
| sympathy | words of compassion and understanding | *the new ones are worth a chance, too !* |
| agreement | agreement with an opinion or statement that represents candy speech | *Yeaaah so amazing* |
| implicit | indirect expression of candy speech | *Why don't you go to Supertalent ?* |
| ambiguous | unclear whether candy speech or not | *OMG* |

Table 1: Types of candy speech expressions (examples are translated from German).

## 4 Annotation

### 4.1 Procedure

The annotations were performed with the annotation tool Inception (Klie et al., 2018). Each comment was checked for the presence of candy speech, and the identified candy speech expressions were annotated on the exact span level with one of the predefined types. Note that one comment can contain several candy speech expressions, and such expressions can also overlap. For each expression, we aimed at labeling the shortest possible span, e.g., instead of annotating several consecutive expressions of the same type as one span, each clause was annotated separately. Furthermore, our annotation scheme allows for overlapping spans in order to preserve the grammaticality of each annotated expression. E.g., *Ihr seit soooooo süss und eure Parodien der Hammer* ('You are soooooo sweet and your parodies are awesome') was labeled both as *affection declaration* and *positive feedback*.

The annotations were conducted by two annotators – an author of this paper (annotator 1) and a graduate student with linguistic background (annotator 2). At the beginning of the annotation process, the annotation guidelines with the definition of candy speech and a number of predefined candy speech types were compiled and shared with annotator 2. In the annotation training period, both annotators annotated the same portion of the data and discussed the results. Annotator 2 proceeded with the annotation, while regularly discussing the results with annotator 1. When new cases/types emerged, the annotation guidelines were updated and previous annotations were adapted accordingly.

Annotator 1 annotated one document; annotator 2 annotated 13 documents. Annotations performed by annotator 2 were reviewed by annotator 1 and any disagreements were discussed until a consensus was reached and corrected if necessary. Two additional documents were annotated separately by each annotator; these results were not discussed and used to calculate the inter-annotator agreement.

### 4.2 Inter-annotator agreement

The basic inter-annotator agreement (IAA) was measured on the comment level in binary form, i.e., whether a given comment contains candy speech or not. The results based on percentage agreement and Cohen's $\kappa$ (Cohen, 1960) are given in Table 2. The annotators show good agreement of $\kappa \geq 0.7$ on the detection of whether comments contain candy speech. Note that most comments are quite short, with an average of 16.5 tokens per comment.

Evaluating agreement for span annotations such as candy speech expressions is not a trivial task.

| Document | # comments | % | $\kappa$ |
|---|---|---|---|
| Doc1 | 204 | 85.2 | .70 |
| Doc2 | 242 | 89.6 | .76 |

Table 2: Binary IAA on the comment level.

There are generally two options: First, classical chance-corrected inter-annotator agreement (Artstein and Poesio, 2008) could be applied if the task is seen as a classification task, assigning items to classes. However, in this case we should choose a suitable method which allows for multiple classes to be assigned to the same token. In addition, the most likely item choice (for practical reasons) for evaluation would be word tokens – and this does not take into account that several words often belong together to make up one candy speech expression (see Table 1). Thus, missing one candy speech expression should not count for different numbers of mismatches depending on the length of the phrase. Similar issues arise for other span-based annotations, such as named entity recognition (NER). A second option for evaluating span-based annotations comes from the NER literature and is based on matching markables (labeled spans) between a candidate and a reference annotation. Since all standardly available NER scorers however share the assumption that spans cannot overlap (Nakayama, 2018; Batista and Upson, 2020; Palen-Michel et al., 2021; Lignos et al., 2023), we implemented our own span-based F-score to compare two candy speech annotations. We calculate precision (P), recall (R) and F1 scores by counting whether the type and character span of each annotated candy speech expression matches between the two annotators (strict agreement) as well as whether both annotators identified the same type(s) of candy speech in a given comment (type agreement only; disregarding spans). The results show good agreement at the type level, and moderate agreement in the (very strict) fine-grained evaluation (see Table 3).

| | | Strict | | | Type | | |
|---|---|---|---|---|---|---|---|
| Doc | # | P | R | F1 | P | R | F1 |
| Doc1 | 204 | .66 | .51 | .58 | .79 | .61 | .69 |
| Doc2 | 242 | .55 | .48 | .51 | .84 | .73 | .78 |

Table 3: IAA on the fine-grained annotation.

### 4.3 Statistics on the annotated data

14,580 (31.5%) of the comments contain at least one candy speech expression.[2] In total, 21,785 expressions of candy speech were found. Table 4 shows the distribution per type.

| Type | Count | % |
|---|---|---|
| affection declaration | 3,933 | 18.1 |
| compliment | 3,504 | 16.1 |
| encouragement | 1,009 | 4.6 |
| gratitude | 474 | 2.2 |
| group membership | 558 | 2.6 |
| positive feedback | 11,403 | 52.3 |
| sympathy | 101 | 0.5 |
| agreement | 269 | 1.2 |
| implicit | 255 | 1.2 |
| ambiguous | 279 | 1.3 |
| Total | 21,785 | 100 |

Table 4: Distribution of candy speech types.

*Positive feedback* is the most frequent type and covers over 50% of all annotated expressions. It represents a more 'general' type of candy speech that occurs with all kinds of videos. *Affection declaration* and *compliment* are also frequent, with a proportion of 18% and 16%, respectively. The other types were found in less than 5% of all candy speech expressions, which can be explained by the fact that they are more specific and often closely linked to the video theme. For example, *sympathy* occurred mainly in the comments to a video about a natural disaster, while *gratitude* was most frequently found in the comments to a fitness tutorial.

Emojis/emoticons occurring without accompanying text, but with a clear positive meaning, were counted as *positive feedback* (275 instances; 2.4%). Beißwenger and Pappert (2019) have previously noted the significance of emojis for face-work of this kind. Other single emojis were counted as *group membership* (if they were clearly interpretable as the creator's symbol; see Scheffler 2024) or as *ambiguous* (if both negative and positive interpretations could in principle be possible; Scheffler and Nenchev 2024). These were less frequent, however (3 and 29 instances, respectively).

Initiative comments prevail over the reactive ones (92% vs. 8%, respectively). All types of

[2] For the documents annotated by both annotators, we consider the version of annotator 1.

candy speech occurred in both modes, except for *agreement*, which is only possible in responses.

## 5 Conclusion and discussion

This study contributes to the identification and promotion of positive online discourse. We have defined the phenomenon of candy speech as positive face-work in online communication and provided a detailed annotation scheme for its different types. Further, we discussed challenges related to the annotation and evaluation of this type of span-based semantic properties.

Our work facilitates a deeper understanding of positive face-work in online settings by showing that candy speech varies across several dimensions: its 'target' (e.g., an individual or their output), the domain/topic of the creator/video (e.g., expressions of *gratitude* are most common with videos offering practical advice), and the level of intensity (e.g., *affection declaration* may reflect stronger emotions than *compliments* or *positive feedback*). Empirical research into candy speech and its linguistic realizations can yield insights into how virtual communities constitute themselves and support each other. The dataset we provide can be used to train computational models to detect (and potentially generate) various types of candy speech, and positive language more broadly, e.g., for mitigating face threats.

As the next step, we plan to look into a finer-grained differentiation of our majority class *positive feedback* as well as of the reactive comments with respect to face-supporting and face-saving acts.

## Acknowledgments

We are grateful to Dennis Reisloh for providing a large part of the annotations. We would like to thank the anonymous reviewers for their helpful comments. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), CRC 1567, Project ID 470106373.

## References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

David Batista and Matthew Antony Upson. 2020. nervaluate.

Michael Beißwenger and Steffen Pappert. 2019. How to be polite with emojis: a pragmatic analysis of face work strategies in an online learning environment. *European Journal for Applied Linguistics*, 7(2):225–253.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Louis A. Cotgrove. 2018. Nottinghamer Korpus Deutscher YouTube-Sprache (The NottDeuYTSch Corpus). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Louis A. Cotgrove. 2025. *Abogeil! The language of German teens on YouTube*. Number 63 in amades - Arbeiten und Materialien zur deutschen Sprache. IDS-Verlag, Mannheim.

Ritam Dutt, Rishabh Joshi, and Carolyn Rose. 2020. Keeping up appearances: Computational modeling of face acts in persuasion oriented discussions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7473–7485, Online. Association for Computational Linguistics.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Behavior*. Pantheon Books, New York.

Salud María Jiménez-Zafra, Miguel Ángel Garcia-Cumbreras, Daniel García-Baena, José Antonio Garcia-Díaz, Bharathi Raja Chakravarthi, Rafael Valencia-García, and Luis Alfonso Ureña-López. 2023. Overview of HOPE at IberLEF 2023: Multilingual hope speech detection. In *Procesamiento del Lenguaje Natural, Revista*, 71, pages 371–381.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Tina Klüwer. 2011. "I like your shirt" - dialogue acts for enabling social talk in conversational agents. In *Intelligent Virtual Agents*, pages 14–27, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tina Klüwer. 2015. *Social talk capabilities for dialogue systems*. Ph.D. thesis, Saarland University, Saarbrücken.

Constantine Lignos, Maya Kruse, and Andrew Rueda. 2023. Improving NER research workflows with SeqScore. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 147–152, Singapore. Association for Computational Linguistics.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Lucille Njoo, Chan Park, Octavia Stappart, Marvin Thielk, Yi Chu, and Yulia Tsvetkov. 2023. TalkUp: Paving the way for understanding empowering language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9334–9354, Singapore. Association for Computational Linguistics.

Chester Palen-Michel, Nolan Holley, and Constantine Lignos. 2021. SeqScore: Addressing barriers to reproducible named entity recognition evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 40–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection. *Language Resources and Evaluation*, 55(2):477–523.

Tatjana Scheffler. 2024. Emojis und Gruppenidentität auf Twitter. In Simon Meier-Vieracker, editor, *Reingegrätscht: Eine kleine Linguistik des Fußballs*, pages 73–81. Narr, Tübingen.

Tatjana Scheffler and Ivan Nenchev. 2024. Affective, semantic, frequency, and descriptive norms for 107 face emojis. *Behavior Research Methods*, 56(8):8159–8180.

Robert Stalnaker. 2002. Common Ground. *Linguistics and Philosophy*, 25(5):701–721.

Tuija Virtanen. 2022. Virtual performatives as facework practices on Twitter: Relying on self-reference and humour. *Journal of Pragmatics*, 189:134–146.