Originally published in: Wildfeuer, Janina/Pflaeging, Jana/Bateman, John/Seizov, Ognyan/Tseng, Chiao-I (eds.): Multimodality. Disciplinary thoughts and the challenge of diversity.-Berlin [u.a.]: de Gruyter Mouton, 2019. Pp. 115-143.
DOI: *https://doi.org/10.1515/9783110608694-005*

Axel Schmidt and Konstanze Marx

Multimodality as Challenge: YouTube Data in Linguistic Corpora

Abstract: A large database is a desirable basis for multimodal analysis. The development of more elaborate methods, data banks, and tools for a stronger empirical grounding of multimodal analysis is a prevailing topic within multimodality. Prerequisite for this are corpora for multimodal data. Our contribution aims at developing a proposal for gathering and building multimodal corpora of audio-visual social media data, predominantly YouTube data.

Our contribution has two parts: First we outline a participation framework which is able to represent the complexity of YouTube communication. To this end we 'dissect' the different communicative and multimodal layers YouTube consists of. Besides the video performance YouTube also integrates comments, social media operators, commercials, and announcements for further YouTube videos. The data consists of various media and modes and is interactively engaged in various discourses. Hence, it is rather difficult to decide what can be considered as a basic communicative unit (or a 'turn') and how it can be mapped. Another decision to be made is which elements are of higher priority than others, thus have to be integrated in an adequate transcription format. We illustrate our conceptual considerations on the example of so-called *Let's Plays*, which are supposed to present and comment computer gaming processes.

The second part is devoted to corpus building. Most previous studies either worked with ad hoc data samples or outlined data mining and data sampling strategies. Our main aim is to delineate in a systematic way and based on the conceptual outline in the first part necessary elements which should be part of a YouTube corpus. To this end we describe in a first step which components (e.g., the video itself, the comments, the metadata, etc.) should be captured. In a second step we outline why and which relations (e.g., screen appearances, hypertextual structures, etc.) are worth to get part of the corpus. In sum, our contribution aims at outlining a proposal for gathering and systematizing multimodal data, specifically audio-visual social media data, in a corpus derived from a conceptual modeling of important communicative processes of the research object itself.

Keywords: multimodality, YouTube, multimodal corpora, Let's Plays, social media

1 Introduction

For multimodal analyses, it is desirable to have a large database (in our case of YouTube data). Within multimodality, the development of more elaborate methods, data banks, and tools to allow a stronger empirical grounding of multimodal analysis is currently an important topic (cf. Bateman et al., 2017, 152–155). Corpora of multimodal data are a prerequisite for this. Our contribution aims at developing a proposal for gathering and building multimodal corpora of audio-visual social media data, predominantly YouTube data. One of the main challenges in constructing corpora is to make them useful for pursuing specific questions that are relevant within a given scientific approach. Within social semiotics, semiotic products like websites, clips on YouTube or comments are viewed as constituting communicative acts (Kress & van Leeuwen, 2006 [1996]; Kress, 2010). Our aim in this contribution is to ground the process of data gathering and corpus building in assumptions about the research object itself. The questions we are asking are: What does a corpus of YouTube data have to look like to allow researchers to tackle questions that are relevant for a study of communicative acts? What components should such a corpus contain?

Although we do think that multimodality is becoming more important because communication has become more diverse and multimodal, we are skeptical about transferring multimodality into a discipline in its own right. This is mainly because it is difficult to identify a unique research object, which would be crucial for a new discipline. The meaning-making functions of different modal resources and their specific relations in communicative processes are not just a topic of multimodality, but of all disciplines that are concerned with reconstructing social meaning (cf. Habermas, 1967). Moreover, a focus on media communication within multimodality, which could be considered a specific focus, seems not to offer a systematic grounding because so-called non-mediated communication (such as, for example, face-to-face-interaction) is also organized multimodally. In sum, it is not clear to us how multimodality could be differentiated from other disciplines, especially sociology, media and communication studies, and semiotics.

Our contribution has two parts: First, we outline a participation framework[1] which can represent the complexity of YouTube communication, drawing mainly

1 If semiotic material on websites is understood as communication, i.e., as something people are doing with another, they establish in and through their communication something Goffman has called a *participation framework* (Goffman, 1981, 137). The notion *participation framework* describes the relations which participants accomplish in and through their communication. Participation frameworks are on the one hand constrained and enabled by situational and technological parameters (for instance, by whether a communication is face-to-face or technically mediated,

on suggestions by Adami (2009b), Dynel (2014), and Eisenlauer (2014). To this end we 'dissect' the different communicative and multimodal layers that YouTube consists of. Besides the video component, YouTube also integrates comments, social media operators, commercials, and suggestions for further YouTube videos. The data consists of various media and modes and is interactively engaged in various discourses. Hence, it is difficult to decide what can be considered the basic communicative unit (or 'turn'). We illustrate our conceptual considerations with an example, the so-called 'Let's Plays'. In this genre, which has become very popular in a very short period of time (Hale, 2013, 3), gamers document their gaming in films and present it to a (potential) mass audience via upload on video hosting websites like YouTube.

The second part of this chapter is devoted to corpus-building. Most previous studies of YouTube and similar media either work with ad hoc data samples or outline data mining and data sampling strategies (for references see Section 3). Our main aim is to identify necessary elements that should be part of a YouTube corpus in a systematic way based on the conceptual outline in the first part. To this end we initially describe which components should be captured (e.g., the video itself, the comments, the metadata, and so on). In a second step we outline which relations ought to be part of the corpus and why (e.g., screen appearances, hypertextual structures, etc.). Another decision to be made is which elements are of higher priority than others and, thus, have to be integrated in an adequate transcription format (Beißwenger, 2009; Recktenwald, 2017; Marx & Schmidt, forthcoming).

In sum, our contribution aims to outline a proposal for gathering multimodal data and making it accessible in a systematic way, specifically audio-visual social media data, via building a corpus that is derived from the conceptual modeling of important communicative processes of the research object itself. What is important, thus, results from a description of the communicative structures and the participation framework on YouTube.

---

whether it is written or oral, etc.; cf. Meyrowitz 1990). On the other hand, participation frameworks are indexed by the ongoing activities of the participants, and therefore are in constant flux (cf. Goodwin, 1986; Goodwin & Goodwin, 2004; Arminen et al., 2016). Goffman has termed the contribution of a single utterance (or parts of it) to the reflexive accomplishment of participation frameworks, *footing* (Goffman, 1981). In our case (communication on YouTube) we are interested in how technological parameters 'afford' the possibilities of accomplishing participation frameworks on YouTube.

2 Modeling Communication on YouTube

The social media service YouTube is described in many different ways, for example as post-television (Tolson, 2010), as creating a distinct aesthetic, often referred to as 'YouTubeness' (Burgess & Greenberg, 2014) or as an alternative business model in comparison to traditional media like television (Vonderau, 2016).

However, our main interest is to consider YouTube as a *specific* medium facilitating a *specific* form of communication which, in turn, enables specific kinds of participation. Forms of communication specify situational and technological conditions (like written/oral/audio-visual; one-way/reciprocal; public/private, etc.) without determining communicative uses (cf. Holly, 1997; Habscheid, 2000; Dürscheid, 2005; Schmitz, 2015). There is a growing body of studies concerned with the specific form of communication made possible by YouTube. Situated somewhere between mediated interpersonal communication and so-called mass media communication like television (Dynel & Chovanec, 2015), the communicative form YouTube enables is described as polylog (Bou-Franch et al., 2012), as video interaction (Adami, 2009a,b, 2015; Schmidt, 2011), as enabling viewer involvement (Frobenius, 2013, 2014), or as dialogical exchanges (Jones & Schieffelin, 2009).

Our theoretical and methodological background is interactional pragmatics (cf. D'hondt et al., 2009). Instead of focusing on language structures or systems, we are interested in how language and other modal resources are used to constitute activities. The starting point, therefore, is the purposeful *doing* of participants and the establishment of a *participation framework* in and through communicative exchanges. We start with the observation that YouTube establishes mediated interactional exchanges, similar to mediated interpersonal communication (Konijn et al., 2008). Although YouTube-communication is asynchronous, physically distant and involves an indeterminate viewership, it holds the possibility of producing 'turns', e.g., posting a video, and reacting to 'turns', e.g., by writing a comment or by posting an answering video (cf. Adami, 2009a; Schmidt, 2011). Since every interaction is situated, exchanges on YouTube can be described as distant situations spliced together via media technology (cf. Thompson, 1995). Those "synthetic situations" (title of a paper by Knorr-Cetina 2009) create a "response presence, without needing to be in one another's physical presence" (Knorr-Cetina, 2009, 69). Thus, Social Media and YouTube generate a specific kind of participation framework modifying familiar forms of interactional participation and involving different communicative levels (Frobenius et al., 2014).

Our interest is to describe YouTube's participation framework as a multimodal form of communication involving different levels of participation. This description can then be used as a basis to identify which elements should be part of a YouTube

corpus. Background to this study is a larger project of developing a standard for multimodal corpora of audiovisual YouTube data. Corpora, in general, are compiled to provide an empirical basis for research. A very important aspect of corpus construction, and in particular of the development of a standard for future corpus formation, is the question of which data should be integrated into a corpus (standard) and what form this data should take. Simply gathering data in a corpus without a thought-through design is not a promising strategy. For this reason, we first aim to develop a basic understanding of our research object in order to specify which data are needed and what format the data should take to allow research within the framework of the approach outlined above. A crucial aspect of this basic understanding is the participation framework and its levels, which are established in communication on YouTube.

Before explicating these levels, we introduce an example for illustrative purposes from the data we are currently analyzing. Our data consists of so-called 'Let's Plays' (in the following referred to as LPs). In the literature LPs are defined as "playing videogames for the internet" (Hale, 2013, 3). In the simplest case, a user records his/her playing of a video game and his/her simultaneously produced comments and uploads the result to video hosting websites like YouTube (cf. Ackermann, 2016; Stephan, 2014; Marx & Schmidt, 2018, forthcoming). In addition, the player usually appears in a facecam (see Figure 1).[2] Single player LPs are videos between 30 mins and 2 hours that are accessible via video hosting websites. LPs may be watched in an embedded mode (see Figure 2) or a full screen mode (see Figure 1). Usually (as Figure 1 shows) they consist of the game play that fills almost the entire screen with the player giving comments and appearing in a facecam.

There are variants: In addition to single player LPs as described above, LPs can also be produced by several players; then they are called *Let's Play Together* or *Multiplayer Let's Plays* (for a more detailed description of such *Multiplayer Let's Plays* see below). In addition, LPs can be recorded as described above, or they can be viewed as a live stream as on platforms like Twitch[3] (cf. Recktenwald, 2017). The first Let's Play appeared in 2006 on the website '*something awful*'[4]. Nowadays, Let's Play videos on YouTube as well as the channels of so-called Let's Players achieve high click rates. The German Let's Player Gronkh, for instance, has about 4.8 million YouTube channel subscribers as of August 2018.

2 Let's Players usually use a so-called facecam, which conveys a visual image of the player's face.

3 Twitch is a live streaming video platform that specializes in live streams of Let's Plays.

4 See https://www.somethingawful.com/, last accessed: 29 August 2019.

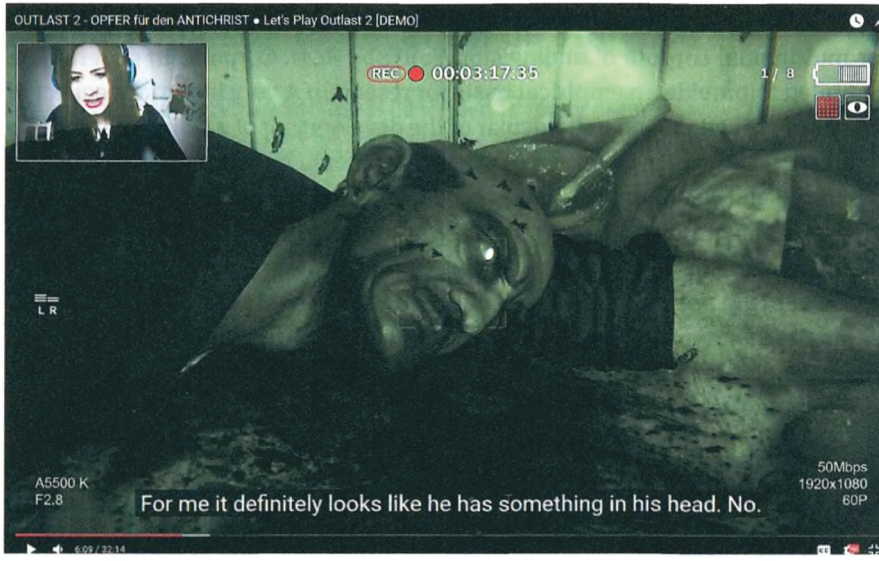

Fig. 1: A screenshot of a Single Player Let's Play in an embedded view.

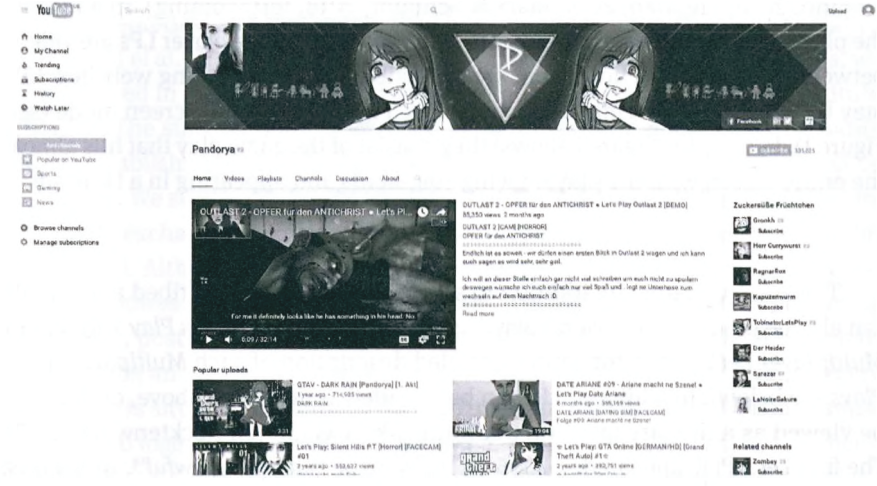

Fig. 2: A screenshot of a Single Player Let's Play in a full screen view.

Our example data consists of an extract lasting 1:25 minutes of a *Let's Play Together* (lasting a total of 23 mins) where four people (Gronkh, Curry, Tobi, and Pan) in spatial distance from each other play a video game together while at the

same time talking to each other (via Teamspeak[5]) and commenting their play moves. The game they are playing together, *Dead by Daylight*[6] (Starbreeze Studios, 2017), is shown on the respective screens of each player. Three of the four screens show an integrated facecam capturing the face of the respective player (see Figure 3).

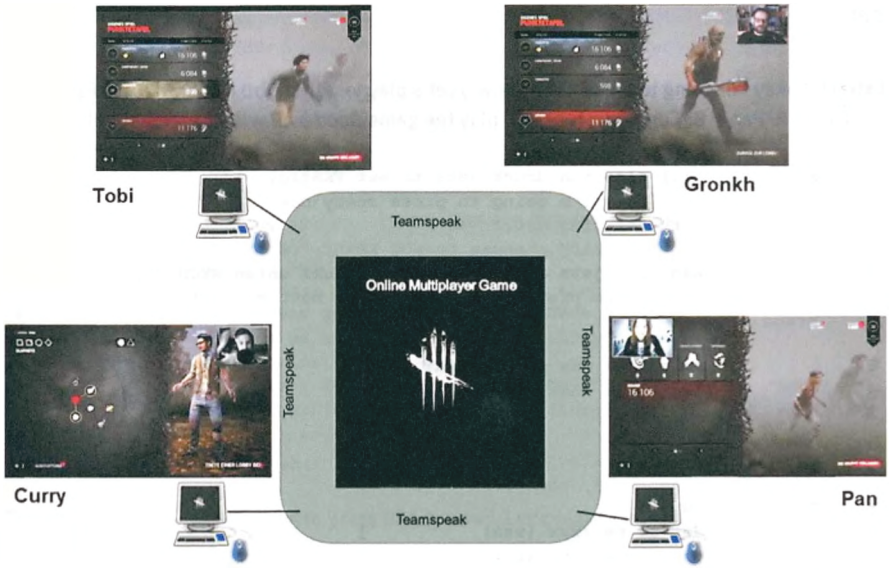

Fig. 3: A schematic diagram of a Let's Play Together.

Each play represented on the respective screen is individually recorded and afterwards uploaded on YouTube. In the following, we rely mainly on the version produced by *Gronkh* (Gronkh, 2016). Extract 1 in Figure 4 is from the beginning of the game session; the four players have just entered the lounge and are now trying to get the game started. In the transcript, we focus on the verbal comments by the participants (in black font), on status displays (appearing after the abbreviation *StaD*), game events (appearing after the abbreviation *GE*) and on game sounds (after *GS*). Talk is assigned to speakers by capital letters (A = Gronkh, B = Pan, C =

5 Teamspeak is an application for audio communication between users on a chat channel, similar to a telephone conference call.

6 *Dead by Daylight* is an asymmetric survival horror game which is played exclusively as a one versus four online multiplayer game.

Curry, D = Tobi). The participants speak German (original language is marked in bold). An indicative English translation is given in the line below. The non-verbal events are aligned with talk and pauses via special characters like %, &, etc. (cf. Mondada, 2014). Letters after abbreviations (a, b, c, d) indicate on which screen the events appear. Standalone small letters (a, b, c, d) indicate activities conveyed by the facecams. Stills are not used as they are not necessary for our argumentation here.

Extract: "okay i'm going to press REAdy now"/Let's play together DbD #2/10.6.2016. Four LPers (Gronkh = A; Pan = B; Curry = C; Tobi = D) play the game *Dead by Daylight* together online.

```
1   A       *(.) also ich drück jetz ma auf FERtig,
                 okay I'm going to press ready now
    GS      >>threatening music--->
    StaDa   *button READY changes to NOT READY; red tick above C1
2   A       wenn ihr jetz auch auf fertig drückt unten RECH TS,
            if you too press ready now at the bottom right
    StaDa                                                  *tick above C3,
            line on CS
3           ma gucken was pas *[SIERT,]
4           let's see what happens
5   D                         [ja,   ]
                               yes
    StaDa                     *tick above C4, line on CS, button NOT READY
            disappears, countdown 0.06 starts
6   D       (0.37)
7           das is ne (xxx [xxx)          ]
            that is a (xxx xxx)
8   A                     [SPIELbeginn;  ]
                           start of play
9   D       (.) [oh es geht        ][LOS;     ]
                 oh it's starting
10  C           [okay das geht auch][SO;      ]
                 okay it also works this way
11  B                               [oh;      ]
                                     oh
12  B       (.) ja ja oh   [GEIL; ]
                yes yes oh great
13  A                     [ja    ][nice; ]
                           yes nice
14  D                           [ach du] [SCHEISse; ]
                                  oh my gosh
15  A                                  [halt halt] halt) *YE[A: ]:H;
                                        wait wait wait yeah
16  B                                                     [go-]
                                                          [go-]
    GE/STADa                                             *countdown:
            0.00, lettering disappears, DS turns black
17  B       [dann m         ]oderier ma AN;
             then start the moderation
```

```
18  C       [<<lachend>  eHE,>]
            [<<laughing> eHE,>]
19  A       -(.) #[ihr seid TOT;          ]
                   you are dead
20  C             [((räuspert sich    ))]
                  [((clears his throat))]
    GEc     +black DS and FC appear
    GEd           #black DS appears
21          (0.32)
22  A       $äh JA;
             äh YES;
    GEb     $black DS appears
23  A       HALlo °un herzlich willKOmmen
            HELlo and a warm welcome
    GEa-d         %game screen appears
24  A       § bei °dead ^^by °DAY--light;
              to dead by daylight
    STADb-d                    %game symbols appear
    GEb     §FC appears
    b        °waves, smiles, looks into FC
    c              ~looks into FC, raises his eyebrows
25  A       °°hh un wir gehn_n die nächste RUNde,
            °°hh and we start with the next LAP,
    STADb,d %text fields appear
26  A       heute mit & (0.27) °weiterem beSUCH-   &
            today with& (0.27) °some more VIsitors-&
    GS                &rustling plonk sound--->>
    STADa                    °game symbol appears
```

Fig. 4: Extract 1: "okay i'm going to press REAdy now"/Let's play together DbD #2/10.6.2016.

At the beginning of the transcript (lines 1–16), the four players are concerned with the technical aspects of getting the multiplayer game started. Their talk, although already transmitted to the public, is obviously designed to achieve a joint start of the game. So A's announcement in line 1 `okay I'm going to press ready now` is subsequently expanded to an encouragement addressed to his co-players to do the same, with `if you too press ready now at the bottom right` (line 2). What follows is a successive start of the game which is accompanied by comments of the four players. The comments in this phase merely serve the purpose of mutual coordination. Only at line 17, after having established and announced the start of the game (most obviously by A's exclamation `start of play` in line 8), an intro moderation is requested (with B's `then start the moderation` in line 17), which is subsequently mainly delivered by A (lines 19–26).

In this extract, three kinds of communication are at work at the same time: First, we (as the observers) watch the events on screen (how the videogame gets started) and listen to the accompanying verbal talk between the players. Second, we may feel addressed by the players welcoming the viewers to their joint game event (for instance, by saying `hello and a warm welcome` in line 23). Third, we

may divert our attention from watching the video to reading the comments below the video slot on the screen (not part of the transcript), which may give us an impression of how other audience members perceive the video.

Accordingly, Dynel (2014) differentiates three basic levels in the case of communication on YouTube: (1) the video interaction, (2) the sender-recipient interaction, and (3) the comments. Figure 5 summarizes these three possible forms of communication with respect to the example introduced above. Dark grey is used for level one, black for level two and light grey for level three.

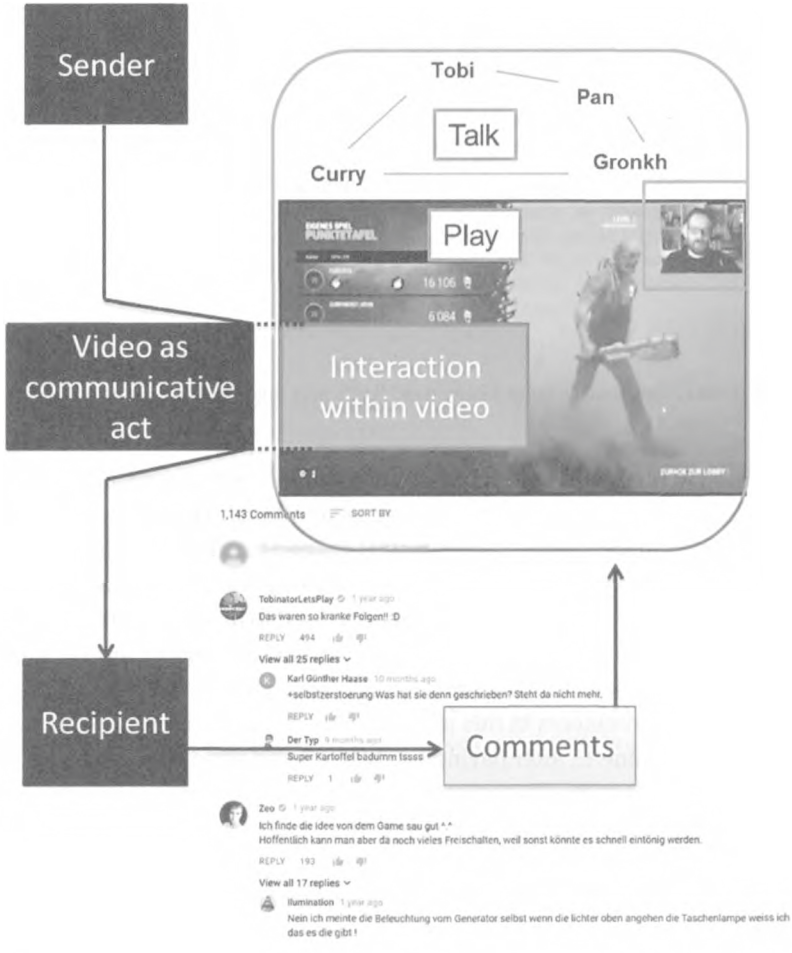

Fig. 5: Basic forms of communication on YouTube.

In the following, the three levels of communicationon YouTube are discussed in more detail. Let us turn to the first level, the video interaction level.

2.1 Video Interaction Level

Videos can show people talking and interacting. Depending on the content of the video, recipients witness forms of participation they are familiar with either from face-to-face interaction or from mass media communication. Videos can show face-to-face interactions (like in talk shows) or pure verbal interactions (like in our case between the four players). In this case the basic participation framework with speaker and hearer categories as outlined first by Goffman's paper on footing (1981; cf. Levinson, 1988) is adaptable, namely the distinctions between ratified and non-ratified and addressed and not addressed participants.[7] In other words: first-level interactions depicted by video resemble face-to-face-interactions with respect to possible participation roles. This, of course, holds only for depicted interactions within the video (and not for other kinds of communication on the platform). In our extract above, for example, we can perceive the content of the video as a focused interaction between four people made possible by technology.

In addition, the idea of different roles within a participation framework is adaptable in a next step to mass media communication (cf. O'Keeffe, 2007; Scannell, 1991). In this sense, videos on YouTube can either show para-social interaction, in which speakers within the video address an audience directly (cf. Horton & Wohl, 1956; Vorderer, 1996), or a video may indirectly target its viewers, thereby creating what has widely been understood as an 'overhearing audience' (cf. Heritage, 1985; Hutchby, 2006; Clayman, 2006).

In the above example, both of these roles are present: When Gronkh is welcoming the viewers with `hello and a warm welcome` in line 23, he is addressing the

7 Due to lack of space the basic categories of the participation framework by Goffman (1981) are only briefly reviewed in this footnote: Goffman distinguishes on the side of the 'hearers' between ratified and unratified participants. The former build what he calls a focused interaction, that is a certain number of people sharing a common attentional focus for a certain amount of time. The latter, the unratified participants, are all others who are in response presence but not officially part of the interaction. Ratified participants are further divided into addressees, who are addressed by a particular utterance, and side participants, who are part of the focused interaction but are not addressed. Unratified participants are understood as bystanders and subdivided into overhearers, who witness ongoing talk accidentally, and eavesdroppers, who purposefully eavesdrop a conversation. Goffman's model is the basis for most approaches dealing with participation. It has been widely taken up and extended, reformulated, and adapted (cf. Levinson, 1988; Goodwin, 1986; Dynel, 2014).

audience directly. In contrast, when he is playing and talking with other players (for instance in lines 1–16), he is engaged in an interaction which is both carried out for the purpose of coordination with his co-players and produced for viewers as he is talking in a public space. Therefore, this kind of interaction targets the audience indirectly. In both cases we are dealing with podium or platform formats (Goffman, 1981, 1983) which involve unequally distributed possibilities to participate.

Accordingly, YouTube viewers can focus on the content of a video (e.g., a TV series, parts of a movie or, as in our case, a commented video game), or they can treat it as a release of a YouTube sender who communicates something with it. This is often recognizable in the comments, which either deal with the content or with the sender (cf. Dynel, 2014, 44). This brings us to the second communicative level, which is the sender-recipient level.

2.2 Sender-Recipient Level

Beginning with the *reception end*, the second level is constituted by the fact that videos are uploaded for viewers, who are, in any case, ratified participants. This holds even if recipients are unregistered, not logged in or remain passive, as each of these participation statuses is possible and legitimate on YouTube. In all cases recipients are external to the interaction shown in the video at the first level because they are not able to intervene and, thus, unable to alter the symbolic flow of the video itself. In other words: Interaction depicted *within* the video is not contingent on the activities of the recipients. Thus, *depicted* participation frameworks within the video remain unaffected by recipients' activities. In our extract above, for example, recipients are able to listen to the conversation of the four players and they may comment on it *afterwards*, but they are not able to participate or intervene in the interaction depicted within the video. This is due to the recorded character of the video.

In addition, recipients are distant and distributed. As long as they do not post anything, it is not possible to determine who is watching. The "response presence" (Knorr-Cetina, 2009, 69) of a (potentially) mass audience only becomes evident via website metrics such as number of views, subscribers, and so on. In our case, we know from the metrics shown below the video that the video has had over one million views and over one-thousand comments. Therefore, YouTube 'senders'[8] are to a large extent always oriented towards imagined (mass) audiences (Androutsopoulos, 2014).

8 The term *sender* is chosen in accordance with (Dynel, 2014, 42–45). The term should indicate the mass media-like aspect of communication on YouTube.

As opposed to the reception end, there is a video author and/or releaser at the *production end* (technically termed as sender). While the term *author* indicates an involvement in the production process of a video, the term *releaser* merely refers to the distribution of a video (in this case the process of uploading).[9]

Dynel (2014, 43–45) differentiates between three production formats: first, *re-publications* of mass media content; second, *modified versions* of existing videos; third, self-made videos termed as *vlogging*. Depending on the production format, the authorial status of the sender varies. In re-publications, a collective professional sender is embedded[10]; the releaser does not gain authorial status. Modified videos have an original author (the 'first sender') and the releaser (the 'new sender') becomes a 'top layer author' of the new version. Finally, *vlogging* videos are authored by individuals, who are usually the owner of an account or a channel on YouTube.

Furthermore, videos are not presented in isolation. They are embedded within a website containing manifold cues for understanding, partly authored by the releaser, like the title of the video or its short description. In our case, for example, the video is entitled "*GRONKH macht GYROS! – DEAD BY DAYLIGHT #002 – Gronkh*", which contains information about the sender (communicating under the pseudonym *Gronkh*), the game that is played (*Dead by Daylight*), the episode (*#002*) and a satirical description of the content (*GRONKH macht GYROS!*, to be translated as *GRONKH makes GYROS!*, which alludes to the fact that the game belongs to the horror/slasher genre). Independent of the video's content and its production format, the releaser and thus the owner of the account is seen as the (last) sender (but not necessarily as its author).In our case, the Let's Player *Gronkh* is simultaneously the author and the releaser of a self-made video uploaded on his eponymous YouTube channel.

The roles of speaker, sender, author, and account-owner coincide particularly in those cases where vloggers speak to the camera directly (as in our case *Gronkh*). In such cases, it is more likely that the video author/releaser is personally addressed by recipients increasing the likelihood of a verbal exchange between sender and recipients.

---

9 We adopt both terms—*author* and *releaser*—from (Dynel, 2014, 42–45). *Author*, in addition, is a term used by Goffman to indicate who is responsible for constructing an interactional move (who has chosen the words, the images, etc.).

10 *Embedded* means that original content—for example a Hollywood movie or parts of it—are re-published by a YouTube user (a 'new sender', so to speak) who was not part of the original production or distribution context. The original sender, in this case the film company, is thus embedded as a first sender within the re-publication on YouTube by a new sender.

Communication on the sender-recipient-level is established through uploading and watching videos. It does not require any comments to be sufficiently established. However, comments are an additional option for recipients to engage further. This brings us to the third level, the level of comments.

2.3 Level of Comments

In contrast to mass media, YouTube allows an alternation of sender and recipient roles. Everyone can release videos and comment on videos. In this way, YouTubers can interact with one another publicly but, of course, without being able to change the content of released videos on the first level. Though meant as a video sharing website originally (Vonderau, 2016), YouTube enables interactional exchanges of several different kinds: only between video releasers through video responses (Adami, 2009b), between releasers and commenters, or only between commenters.

Therefore, comments are often specifically addressed either by technical options (such as a reply-button), or by using specific signs (like the @-symbol). However, as YouTube communication is persistent, everybody can witness the whole interaction and join in at any time (indexed by the time stamps).[11] In this way, communication on YouTube creates an endless 'open state of talk' (Goffman, 1981) leading to asynchronous communication typical for social media platforms. In our case, *Gronkh*'s video was released on June 10, 2016 and has received 1,143 comments as of August 2018. Most of them, in turn, received answers by other commentators. The oldest comment is from June 2016, while the newest one (at the time of writing this chapter) was posted in July 2018. Thus, the communication that was generated by this video of *Gronkh* currently bridges a time span of approximately two years (and is still ongoing).

The model from Dynel (2014) that we have discussed so far should be extended by a fourth level, which we call website-user interaction.

2.4 Level of Website-User Interaction

Besides interaction between a video sender and recipients, there is interaction between the website as a communicator, in this case the platform YouTube as a part of Google, and a YouTuber as a user, whether as a producer and/or as a recipient (Eisenlauer, 2014). YouTube provides a designed space, including for

11 The persistence of YouTube comments is limited, however, because video releasers have the ability to moderate comment-level interaction by deleting comments and blocking users.

example templates, basic functionalities, indexes, and metrics as well as a basic broadcast structure like channels and multichannel networks (Vonderau, 2016). The platform's basic structures can be seen as affordances (Gibson, 1979; Arminen et al., 2016) creating exchange in the form of a human-machine interaction on a separate level. This means that, while watching videos and posting comments, users also interact with software designed for specific purposes. This level also has to be taken into consideration when building a corpus.

In our case, for example, watching the video of *Gronkh* happens in a predefined template provided by the website. At the same time, watching the video triggers the display of several automatically generated features like indexes, playlists, and ads. Both are 'communicative acts' which are not attributable to the video author/releaser but to the hosting/distributing website.

In addition, on this fourth level the question of ratified/unratified participation reappears as the difference between (un-)registered and (not) logged-in participants as it is the platform which regulates the formal dimension of participation via technical implementations (Boyd, 2014).

2.5 Interim Conclusion: Participation on YouTube

The basic structure of communication on YouTube looks like this: A (an active user) is doing C (posting a video or posting a comment) for B (indeterminate group of viewers), who, in turn, have the chance to answer either by writing a comment or by releasing a video response. The result of this is an interlacing structure that includes an interaction within the video to be seen and commented on by the audience, creating a second (sender-receiver), a third (comments/response video), and a fourth (website-user) level of interaction.

In this model, communicative acts (or turns, or moves) can occur at all levels in any modal form (verbal or non-verbal, spoken or written). Every logged-in person is able to either just watch or to get engaged at any level (e.g., video producing and uploading, writing comments, and so on). Comments may or may not address video releaser(s)/author(s), or other commenter(s). 'Participation' means any activity at the production or reception end. However, identifying participants drawing only on online data are limited to its active users, as passive ones are not visible on the surface of the website.[12]

The basic participation framework underlying communication on YouTube outlined above is summarized in Figure 6.

12 Although passive users are not interacting, they, as imagined audiences, are still part of the participation framework.

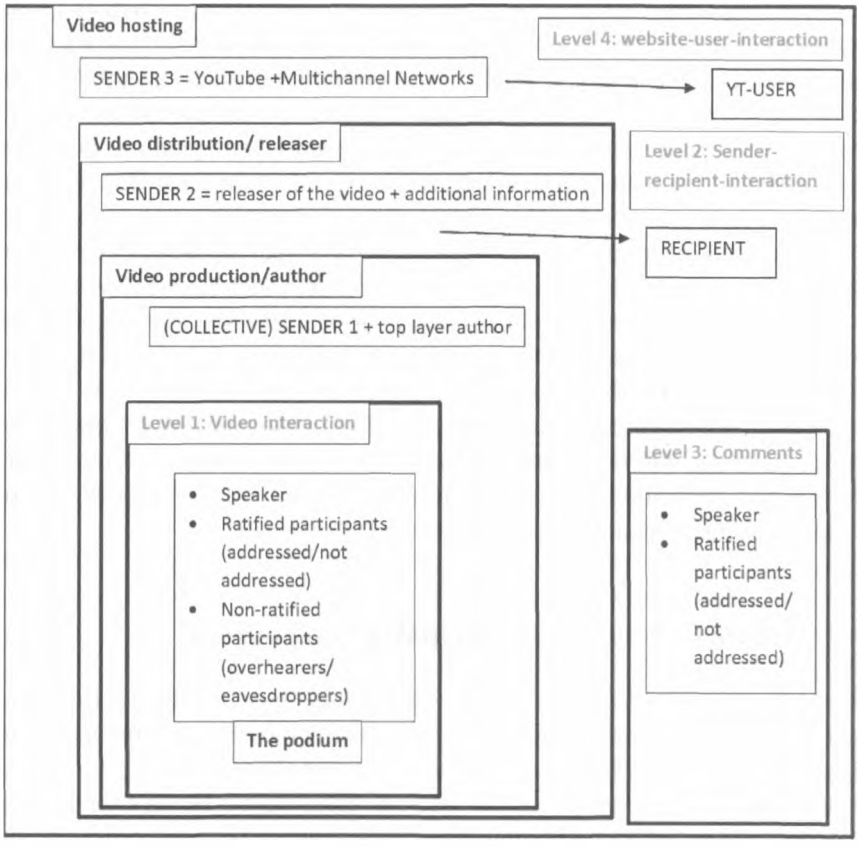

Fig. 6: Basic participation framework of communication on YouTube.

3 YouTube Corpora

In the second part of this chapter, we use the model developed above to formulate requirements for a multimodal corpus of YouTube data. As a multimodal form of communication, YouTube draws heavily on multimodal resources like audio-visuality, images, and design. In order to understand communication on YouTube, it is necessary to include everything that is relevant for meaning making. The decision what is relevant follows from the conceptual schema we introduced above, which is, in turn, rooted in our theoretical framework of interactional pragmatics we outlined above. In other words: we delineate which elements of YouTube's communication should be part of a corpus in order to be able to do research in the

framework of our approach. Thus, this delineation has to be as broad as possible. The corpus structure should be designed in such a way that it allows the study of many different research questions.

Recent studies using YouTube corpora are concerned mainly with data sampling and data mining strategies (cf. Abidi et al., 2017; Androutsopoulos, 2014; Bou-Franch & Garcés-Conejos Blitvich, 2014; Frobenius, 2013, 2014; Ivkovic, 2013; Maybury, 2012; Siersdorfer et al., 2014; Tereick, 2013, 2016; Uryupina et al., 2014; Zhang & Kramarae, 2014). In contrast, our main concern is not how to gather or store data but to outline in a systematic way which components should be included in a YouTube corpus. This chapter represents a first step towards these goals by defining such requirements, which remain purely conceptual at this stage.[13]

We distinguish between components and relations. *Components* are single elements which can be identified visually on the webpage by recipients using a browser (for example the videos or the comment section). Components should be captured individually. *Relations* refer to the overall structure of the webpage, in particular relations between components resulting in a 'designed space' of the webpage (for example, how a video is embedded within a webpage). Those relations have to be captured as a whole, preserving the semiotic structure. Components and relations are different forms of capturing communication on YouTube and, of course, overlap in its representation within the corpus.

The next sections deal first with the components and then with the relations. As the construction of a corpus cannot start with everything at once, we outline a stepwise strategy, starting with the most important parts.

3.1 Components

With respect to *components* we differentiate between *video data*, *interactional data*, and *metadata*.

Video data are the core element as all communication on YouTube is about single clips: They are searchable, titled, ranked, and commented on, and they are unambiguously locatable by a URL. Therefore, the basic communicative unit on YouTube is the single video accessible by its URL.

13 That is, we are not dealing with the question of how and with which tools the data are gathered and stored at this stage. This also means that we are not dealing with questions of implementation of a corpus and what a corpus should look like exactly on a description level in this paper. These aspects will be addressed in a next step, after we have specified what kind of data are needed to do research according to our approach.

Videos should be transcribed to allow a search for linguistic and multimodal elements (like words, phrases, gazes, visual elements, etc.). As transcription is a complex and time-consuming process, it is impossible to transcribe all aspects of a video right at the beginning. For practical research reasons, the transcription should therefore be successively refined, depending on the research question at hand. In a first step, a rather rough transcription should record the verbal exchange according to the transcription conventions of GAT2 (cf. Selting et al., 2011). Verbal transcriptions allow capturing the entire verbal exchange and thus have a clear cut outcome. Multimodal transcriptions, on the other hand, need to be neatly adjusted to the research question at hand. Not everything that is visually accessible can be part of the transcript. Multimodal transcriptions are, thus, in contrast to verbal transcriptions, highly selective (cf. Mondada, 2018; Stukenbrock, 2009).[14] In addition, multimodal transcriptionsare even more complex than verbal transcripts and should only be included if they are relevant for the research question at hand.

Therefore, in further steps, only selected sections are transcribed using a multimodal extended GAT2-system (Mondada, 2014, 2018). Multimodal transcripts of this kind represent different modal resources, e.g., talk, embodied conduct like gaze, gestures, posture shifts, etc., or game events on a screen in their temporal unfolding interplay. The transcripts can also include screenshots whose location is indicated within the transcript (Stukenbrock, 2009). The transcripts are only auxiliary means to produce a working document and to represent the exact temporal relations between different resources which are otherwise not accessible. However, audio and video data remains an indispensable basis and should not be replaced by the transcript at any stage in the process of analyzing. In our example above, as Figure 7 illustrates, besides the spoken material (here after the capital A), we also note game events (GE), game sounds (GS), status displays (StaD), and physical activities (FC) in case there is a visual representation of the gamer(s), a so-called facecam (FC).

As one can see in the transcript, when player A (*Gronkh*) is announcing that he is going to press a button for starting the game (`okay i'm going to press ready now`), the action of pressing has already been conducted as the 'ready-button' on his screen changes before A actually announces his action. This is conveyed by the status display line abbreviated with StaD. In addition, we see that in this stage, other modalities are either not yet available (like the facecam which is inserted later) or are in auto-play mode (like the music and the movements of the avatar). This

14 One reason for this is that in contrast to the verbal mode , the visual mode is not necessarily based on action units. Thus, with respect to the visual mode it is often unclear whether events are accountable actions or not.

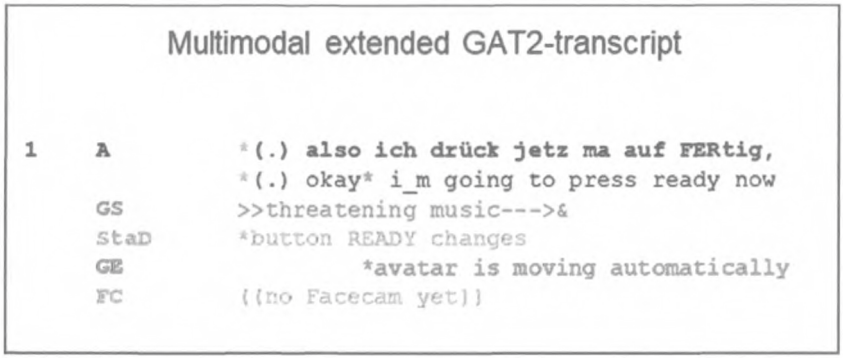

Fig. 7: Example of a multimodal transcript.

observation can, additionally, be represented by a screenshot. Having access to the temporal unfolding of both talk and the effects of controlling actions represented on the screen allows us to see that A actually produces a retrospective comment in an announcement-like form.[15] This sheds light on how gamers are dealing with time in coordinating different temporalities (of talking, gaming, and presenting). To add a relation to absolute time, the transcript may be extended by a timeline or time stamps (cf. Stukenbrock, 2009). Multimodal extended transcripts are an essential basis to pursue questions of temporal relations.[16]

In addition, YouTube videos are posted by a *sender* who communicates under a name, usually a pseudonym. The sender represents, together with the video, the basic pragmatic structure of YouTube communication. In our case, a sender, calling himself *Gronkh*, is releasing a video accessible via a unique address, i.e., a URL. The structure of YouTube suggests an understanding of the video as a communicative act of its releaser.

---

15 This, of course, only holds for the audio-visual presentation of the Let's Play, i.e., what viewers of the video get to see. Often there is a slight delay between facecam/multiplayer audio communication and the video feed. It would be interesting to integrate such questions of delay caused by processes of technical transmission. In principle, this would require an approach that also includes the integration of production data. At the current time, however, we focus only on the product.

16 There are, of course, other proposals for transcribing multimodal data (cf. Baldry & Thibault, 2006; Norris, 2004; Flewitt et al., 2009). Our approach proposes to use talk and pauses as a scaffold for the temporal alignment of other events. Transcripts of this kind focus on interaction and are useful to investigate dynamic, talk-based processes (Mondada, 2018). For other kinds of material, like static websites, they are less useful.

Another crucial element to be integrated in a YouTube corpus is *interactional data*. This concerns first of all the *comment section*, which we categorize as 'first stage interactional data' since comments are responsive actions either to the posted video or to other comments. The comments should be, at first, gathered as a coherent block. They need to be annotated following TEI standards[17] for computer-mediated communication, especially regarding so-called interaction signs such as emojis, interaction words, and addressing terms (as proposed by Beißwenger et al., 2012).

A YouTube corpus should also integrate *metadata*. We distinguish three types of metadata: First, *automatically generated metadata* consisting mainly of the release date and the URL, which also contains the video ID. Secondly, *semi-automatically generated metadata*, which is generated by clicks and involves metrics like the number of views, likes and dislikes, or comments. Finally, *self-generated metadata*refers, for example, to the name of the releaser's channel, the selected pseudonym or the short content description below the video.

However, the components outlined above—videos, interactional data, and metadata—are not isolated elements. Rather, they are embedded within the given structure of a webpage. This means that not only components but also their *relations* need to be considered when generating corpora of multimodal website data. The next section is devoted to this aspect of relations between elements.

3.2 Relations

Besides the elements listed und discussed above, their relations also have to be taken into account. Consequently, the visible structure of the webpage is important, as predefined slots have a meaning-making function for the text elements.[18] Each webpage accessed on YouTube has a similar structure, thus entailing a certain recognition value. This means that before we discover the specific content of particular sections and its semiotic features, we already know the basic frame of meaning and its relation to other sections. This partly stems from the specific

17 TEI (Text Encoding Initiative) is a non-profit membership consortium. The goal of TEI is to develop a 'set of high-quality guidelines for the encoding of humanities texts, and to support their use by a wide community of projects, institutions, and individuals', which is based on ongoing research due to the dynamic textual domains that are still being explored (see http://www.tei-c.org for further information).

18 We are not going into analysis or interpretation at this stage. How relations between elements on a website are to be interpreted, and what particular meaning they may have, is not within the scope of this paper. The only thing which is relevant here is that they have a meaning-making function and that they have to be captured for this reason.

arrangement of elements. Figure 8 illustrates the typical structure of YouTube sites, their sections, and typical relations.[19]

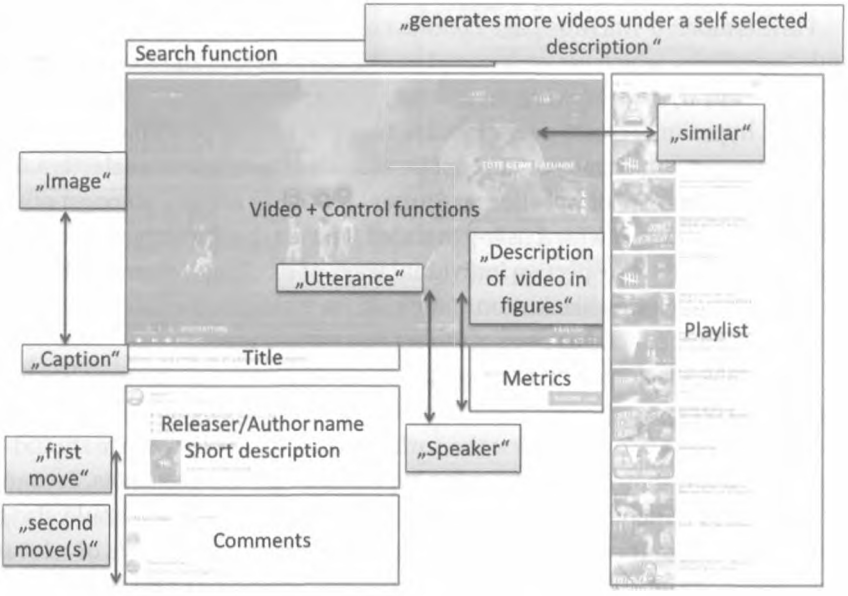

Fig. 8: Typical structure of YouTube pages, their sections, and typical relations.

As the diagram shows, all YouTube pages are composed of the same sections (indicated by the white boxes). Thus, we know the basic meaning of corresponding content, like 'the title' of the video, which is always placed below the video, 'the comments', which are placed at the bottom end of the webpage and may require scrolling and/or clicking to view all of them. In addition, sections are not only prefiguring a certain meaning, but also implicate meaningful relations between certain sections (indicated by the grey boxes and corresponding arrows). For example, the 'video section' and 'the title section' below are understood as an 'image' and its 'caption'. Similarly, the metrics are taken as 'describing the present video in figures', the 'play list section' as 'video suggestions similar to the present one' and the 'search section' as a function to generate more video search results by using

19 The nature of those relations (logical, rhetorical, etc.) is of no relevance for our argumentation at this stage and only becomes relevant in the analysis. The examples only intend to illustrate that relations of this kind exist and that they should be represented in a corpus.

own search queries. The name in the 'releaser section' is read as 'the speaker' (the one to whom the video is ultimately attributed to as a communicative act) and the 'comment section' in relation to 'the speaker section' and its 'video utterance' as a 'second move' in relation to the video releaser's 'first move'.

Furthermore, a YouTube page contains several *active elements*, most importantly hyperlinks, which are understood as enabling specific kinds of actions, i.e., selecting and, by that, moving on to new content (Huber, 2002; Storrer, 2000). All in all, hyperlinks create a relational network of further potential contents in the background, which may or may not be activated by recipient's selection. The hypertextual structure of websites, as Storrer (2000) has argued, amongst others, mean to be confronted with a user-generated 'text' as the linearity of traditional texts are replaced by respective individual ways of reading. Moreover, there are further active elements like buttons (such as, for instance, a subscribe-button), which can be understood as prompts to act in a certain way.

With respect to corpus building purposes, the hypertextual structure has to be preserved in order to represent basic interactional affordances of the website. This should be done in two ways. First the surface structure of the website should be captured in a way that allows identifying hyperlinks, e.g., as screenshots. Secondly, the covert deep structure of interconnected contents via hyperlinks should also be represented. Thus, for corpus generating purposes both screenshots of relevant websites and the underlying hypertextual structure should be integrated. How the latter can be represented is a technical problem which is not in the scope of this paper.

Finally, websites are not only hypertextual but furthermore they change due to individual reception (as for instance the viewing modes of the player) and due to user interfaces and devices (the website looks different on a computer compared to a smartphone, see Zichel, 2016). In addition, webpages are adapted to individual user habits and search histories based on website-specific algorithms. This variation is difficult to represent in a corpus. The only feasible solution is to define a default standard for representation (and, if relevant, possible variations).

Figure 9 provides a final overview of which elements should be included in a YouTube corpus.

4 Conclusion and Further Problems

What we have presented so far is work in progress. The background of our chapter is a larger project based at the Institute for German Language and the University of

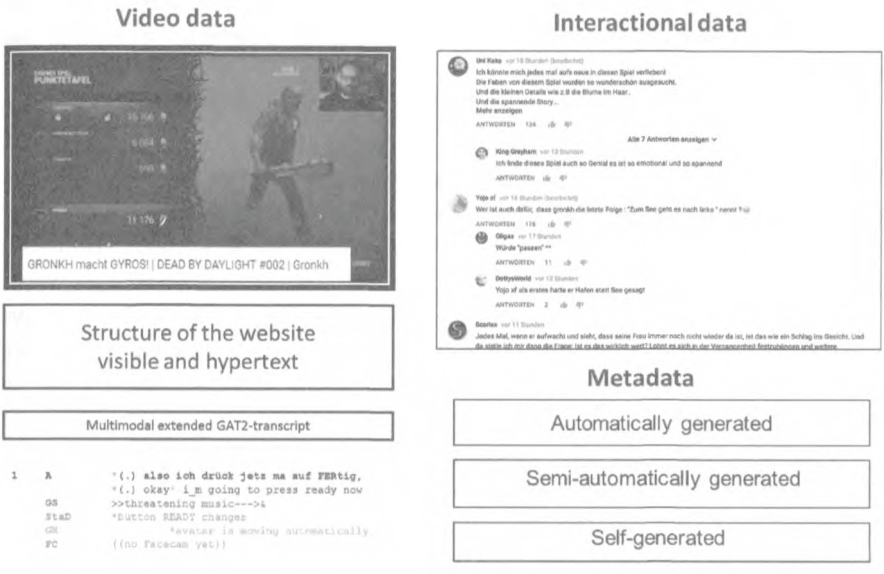

Fig. 9: Elements of a YouTube corpus.

Mannheim.[20] One aim of the project is, in a first step, to consider what a corpus of audio-visual, multimodal social media data should look like. Further steps concern questions around technical implementation, issues of data protection and privacy, issues of hosting and retrieval, and finally the actual construction of such a corpus. The present paper is a contribution to the first step: What kind of data are required in the corpus given our theoretical background? Our outline of a possible corpus construction was rooted in conceptual considerations about how communication and participation on YouTube are organized. In Part 1 of our chapter, we delineated levels of interaction: level of video interaction, sender-recipient-level, level of comments, and level of website-user interaction. As we are interested in interaction processes, corpus construction should entail data which are suitable for investigating these processes. Accordingly, in Part 2 we sketched out which elements and relations should make up such a corpus.

As YouTube communication is multimodal in nature, it should be represented as such on the data level. This means, as we outlined above, that all potentially relevant meaning-making processes should be captured both as single components

20 The project envisages setting up a center of multimodality bringing together different researchers dealing with questions of multimodality and/or multimodal data.

(e.g., the video) and in their relatedness to surrounding components (e.g., the video embedded in a webpage). This is the only way to create a basis for reconstructing communication processes exhaustively.

There are still remaining problems. One of the main and most notorious problems in the area of corpus compilation is the dynamic nature of social media websites like YouTube. First, the *videos* as time-based media are dynamic. This problem can be partially solved by not only archiving the videos as films but by transcribing the videos including screenshots as mentioned above. In doing so, parts of the videos' content can be provided in a fixed and thus searchable way.

Secondly, *websites* change over time. Data collections are, therefore, always only snapshots of a certain moment in time. This can only be solved by multiple data acquisitions representing different stages in the 'life' of a website and its content.

Finally, websites are often *customized versions* adapted to individual users. However, since customization usually does not concern the core elements of the platform like the video or the comments, customized elements like individually adapted playlists can be disregarded initially. Moreover, they only appear in the screenshot versions of the website.

Further central problems are questions of *data protection* and, related to this, procedures of *anonymization*, as well as the *technical implementation* of data gathering and data archiving. This applies in particular to technical solutions for hosting, retrieving, mining, and processing YouTube data.

Appendix: Transcriptions Conventions

Speaker's signs/Display Screens/Facecams

A/a **Gronkh**
B/b **Pan**(dorya)
C/c (Herr) **Curry**(wurst)
D/d **Tobi**(nator)

Abbreviations for other events

GE = Game Event
GS = Game Sound
StaD = Status Display

Conventions for the notation of physical activities (cf. Mondada, 2014)
Nonlinguistic events and activities

- appear after the abbreviations GE, GS, StaD and FC
- in lines following pauses or conversation activities (without own number)
- are aligned with conversation/pauses with the help of special characters (like §, + etc.) indicating the beginning (simple sign) and the end (double sign) of events
- are assigned to the players (A/B/C/D; e.g. 'GEa' means 'game event on the screen of Gronkh', etc.)

Further conventions for the notation of physical movement

—>	movement continues
—>$	movement continues after the line until reaching
$	the same sign
—> >	continues after transcript ends
> >	starts before transcript

Special conventions for the notation of screen events
Abbreviations

CS	Counting Sign
C(1-4)	Characters/Avatars
FC	Facecam

Bibliography

Abidi, Karima, Mohamed-Amine Menacer & Kamel Smaili. 2017. CALYOU: A Comparable Spoken Algerian Corpus Harvested from YouTube. In *18th Annual Conference of the International Communication Association (Interspeech)*, 3742–3746.

Ackermann, Judith. 2016. *Phänomen Let's play-Video: Entstehung, Ästhetik, Aneignung und Faszination aufgezeichneten Computerhandelns*. Wiesbaden: VS Verlag.

Adami, Elisabetta. 2009a. Do You Tube? When Communication Turns into Video e-Interaction. In Domenico Torretta, Marina Dossena & Annamaria Sportelli (eds.), *Migration of Forms, Forms of Migration: Atti de XXIII Convegno Nazionale AIA*, Bari: Progedit.

Adami, Elisabetta. 2009b. 'We/YouTube': Exploring Sign-Making in Video-Interaction. *Visual Communication* 8. 379–399.

Adami, Elisabetta. 2015. What I Can (Re)Make out of it: Incoherence, Non-Cohesion, and Re-Interpretation of YouTube Video Responses. In Marta Dynel & Jan Chovanec (eds.), *Participation in Public and Social Media Interactions*, 233–257. Philadelphia: John Benjamins.

Androutsopoulos, Jannis. 2014. Languaging when Contexts Collapse: Audience Design in Social Networking. *Discourse, Context & Media* 49(4). 290–309.

Arminen, Ilkka, Christian Licoppe & Anna Spagnolli. 2016. Respecifying Mediated Interaction. *Research on Language and Social Interaction* 49(4). 290–309.

Baldry, Anthony & Paul J. Thibault. 2006. *Multimodal Transcription and Text Analysis: A Multimedia Toolkit and Coursebook with Associated On-Line Course*. London and New York: Equinox.

Bateman, John A., Janina Wildfeuer & Tuomo Hiippala. 2017. *Multimodality – Foundations, Research and Analysis. A Problem-Oriented Introduction*. Berlin: Mouton de Gruyter.

Beißwenger, Michael. 2009. Multimodale Analyse von Chat-Kommunikation. In Karin Birkner & Anja Stukenbrock (eds.), *Die Arbeit mit Transkripten in Fortbildung, Lehre und Forschung*, 117–143. Mannheim: Verlag für Gesprächsforschung.

Beißwenger, Michael, Maria Ermakowa, Alexander Geyken, Lothar Lemnitzer & Angelika Storrer. 2012. A TEI Schema for the Representation of Computer-Mediated Communication. *Journal of the Text Encoding Initiative* 3. 1–36.

Bou-Franch, Patricia & Pilar Garcés-Conejos Blitvich. 2014. Conflict Management in Massive Polylogues: A Case Study from YouTube. *Journal of Pragmatics* 73. 19–36.

Bou-Franch, Patricia, Nuria Lorenzo-Dus & Pilar Garcés-Conejos Blitvich. 2012. Social Interaction in YouTube Text-Based Polylogues: A Study of Coherence. *Journal of Computer-Mediated Communication* 17. 501–521.

Boyd, Michael S. 2014. (New) Participatory Framework on YouTube? Commenter Interaction in US Political Speeches. *Journal of Pragmatics* 72. 46–58.

Burgess, Jean & Joshua R. Greenberg. 2014. *YouTube: Online Video and Participatory Culture*. Cambridge: Polity.

Clayman, Steven E. 2006. Understanding News Media: The Relevance of Interaction. In Paul Drew, Geoffrey Raymond & Geoffrey Darin (eds.), *Talk and Interaction in Social Research Methods*, 135–154. London: Sage.

D'hondt, Sigurd, Jan-Ola Östman & Jef Verschueren (eds.). 2009. *The Pragmatics of Interaction. Handbook of Pragmatics*. Philadelphia: John Benjamins.

Dürscheid, Christa. 2005. Medien, Kommunikationsformen, kommunikative Gattungen. *Linguistik online* 22(1). 3–16.

Dynel, Marta. 2014. Participation Framework underlying YouTube Interaction. *Journal of Pragmatics* 73. 37–52.

Dynel, Marta & Jan Chovanec. 2015. *Participation in Public and Social Media Interactions*. Philadelphia: Benjamins.

Eisenlauer, Volker. 2014. Facebook as a Third Author – (Semi-)Automated Participation Framework in Social Network Sites. *Journal of Pragmatics* 72. 73–85.

Flewitt, Rosie, Regine Hampel, Mirjam Hauck & Lesley Lancaster. 2009. What are Multimodal Data and Transcription? In Carey Jewitt (ed.), *The Routledge Handbook of Multimodal Analysis*, 40–53. London: Routledge.

Frobenius, Maximiliane. 2013. Pointing Gestures in Video Blogs. *Text & Talk* 33. 1–23.

Frobenius, Maximiliane. 2014. Audience Design in Monologues: How Vloggers Involve their Viewers. *Journal of Pragmatics* 72. 59–72.

Frobenius, Maximiliane, Volker Eisenlauer & Cornelia Gerhardt. 2014. Special Edition: Participation Framework Revisited: (New) Media and their Audiences/Users. *Journal of Pragmatics* 72. 1–90.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Goffman, Erving. 1981. Footing. In Erving Goffman (ed.), *Forms of Talk*, 124–159. Philadelphia: University of Pennsylvania Press.

Goffman, Erving. 1983. The Interaction Order. *American Sociological Review* 48. 1–17.
Goodwin, Charles. 1986. Audience Diversity, Participation and Interpretation. *Text* 6(3). 283–316.
Goodwin, Charles & Marjorie Harness Goodwin. 2004. Participation. In Alessandro Duranti (ed.), *A Companion to Linguistic Anthropology*, 222–244. Malden, Mass: Blackwell.
Gronkh. 2016. *Multiplayer Let's Play "GRONKH macht GYROS– DEAD BY DAYLIGHT #002 – Gronkh*. YouTube. https://www.youtube.com/watch?v=ikDqiW_DRqc. Last accessed: 17 September 2019.
Habermas, Jürgen. 1967. *Zur Logik der Sozialwissenschaften*. Tübingen: Mohr.
Habscheid, Stephan. 2000. 'Medium' in der Pragmatik. Eine kritische Bestandsaufnahme. *Deutsche Sprache* 28. 126–143.
Hale, Thomas. 2013. From Jackasses to Superstars: A Case for the Study of 'Let's Play'. Last accessed: 17 September 2019. http://www.academia.edu/5260639/From_Jackasses_to_Superstars_A_Case_for_the_Study_of_Let_s_Play_September_2013.
Heritage, John. 1985. Analyzing News Interviews. Aspects of the Production of Talk for an Overhearing Audience. In Teun Adrianus van Dijk (ed.), *Handbook of discourse analysis*, 95–117. London: Academic Press.
Holly, Werner. 1997. Zur Rolle von Sprache in Medien. Semiotische und kommunikationsstrukturelle Grundlagen. *Muttersprache* 19. 215–229.
Horton, Donald & Richard R. Wohl. 1956. Mass Communication and Para-Social Interaction: Observations on Intimacy at a Distance. *Psychiatry* 19. 215–229.
Huber, Oliver. 2002. *Hyper-Text-Linguistik. TAH: Ein textlinguistisches Analysemodell für Hypertexte. Theoretisch und praktisch exemplifiziert am Problemfeld der typisierten Links von Hypertexten im WWW*: Universität München dissertation.
Hutchby, Ian. 2006. *Media Talk. Conversation Analysis and the Study of Broadcasting*. Maidenhead: Open University Press.
Ivkovic, Dejan. 2013. The Eurovision Song Contest on YouTube: A Corpus-Based Analysis of Language Attitudes. *Language & Internet* 10(1). 1–25.
Jones, Graham & Bambi Schieffelin. 2009. Talking Text and Talking Back: 'My BFF Jill' from Boob Tube to YouTube. *Journal of Computer-Mediated Communication* 1050–1079.
Knorr-Cetina, Karin. 2009. The synthetic situation: Interactionism for a global world. *Symbolic Interaction* 32. 61–87.
Konijn, Elly A., Sonja Utz, Martin Tanis & Susan B. Barnes. 2008. *Mediated Interpersonal Communication*. London and New York: Routledge.
Kress, Gunther. 2010. *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London: Routledge.
Kress, Gunther & Theo van Leeuwen. 2006 [1996]. *Reading Images: The Grammar of Visual Design*. London and New York: Routledge.
Levinson, Stephen C. 1988. Putting Linguistics on a Proper Footing: Explorations in Goffman's Concepts of Participation. In Paul Drew & Anthony Wotton (eds.), *Erving Goffman. Exploring the Interactional Order*, 161–227. Cambridge: Polity Press.
Marx, Konstanze & Axel Schmidt. 2018. Let's Play (together) oder schau mal, wie ich spiele – (Interaktive) Praktiken der Attraktionssteigerung auf YouTube. In Konstanze Marx & Axel Schmidt (eds.), *Interaktion und Medien – Interaktionsanalytische Zugänge zu medienvermittelter Kommunikation*, 319–352. Mannheim: OraLingua.
Marx, Konstanze & Axel Schmidt. forthcoming. Making Let's Plays Watchable: An Interactional Approach to Multimodality. In Crispin Thurlow, Christa Dürscheid & Federica Diémoz (eds.), *Visualizing (in) the New Media*, London: John Benjamins.

Maybury, Mark T. 2012. *Multimedia Information Extraction: Advances in Video, Audio any Imagery Analysis for Search, Data Mining, Surveillance, and Authoring*. Hoboken and New Jersey: Wiley.

Meyrowitz, Joshua. 1990. Redefining the Situation: Extending Dramaturgy into a Theory of Social Change and Media Effects. In Stephen H. Riggins (ed.), *Beyond Goffman: Studies on Communication, Institution, and Social Interaction*, Berlin: Mouton De Gruyter.

Mondada, Lorenza. 2014. Conventions for Multimodal Transcription. Last accessed: 17 September 2019. https://franzoesistik.philhist.unibas.ch/fileadmin/user_upload/franzoesistik/mondada_multimodal_conventions.pdf.

Mondada, Lorenza. 2018. Multiple Temporalities of Language and Body in Interaction: Challenges for Transcribing Multimodality. *Research on Language and Society* 51. 85–106.

Norris, Sigrid. 2004. *Analyzing Multimodal Interaction: A Methodological Framework*. London and New York: Routledge.

O'Keeffe, Anne. 2007. *Investigating Media Discourse*. London: Routledge.

Recktenwald, Daniel. 2017. Toward a Transcription and Analysis of Live Streaming on Twitch. *Journal of Pragmatics* 115. 68–81.

Scannell, Paddy. 1991. Introduction: The Relevance of Talk. In Paddy Scannell (ed.), *Broadcast Talk*, London: Sage.

Schmidt, Axel. 2011. How to Deal Methodologically with Entertaining Hatred and Aggressive Humor on the Web (and Television). *Studies in Communication Sciences* 11(2). 133–166.

Schmitz, Ulrich. 2015. *Einführung in die Medienlinguistik*. Darmstadt: WBG.

Selting, Margret, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stuckenbrock & Susanne Uhmann. 2011. A System for Transcribing Talk-in-Interaction: GAT 2, translated and adapted for English by Elizabeth Couper-Kuhlen and Dagmar Barth-Weingarten. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 12. 1–15.

Siersdorfer, Stefan, Sergiu Chelaru, Jose San Pedro, Ismail Altingovde & Wolfgang Nejdl. 2014. Analyzing and Mining Comments and Comment Ratings on the Social Web. *ACM Transactions on the Web (TWEB)* 8. 17.

Starbreeze Studios. 2017. *Dead by Daylight*. Starbreeze Studios.

Stephan, Heike. 2014. *Let's Plays: Kommentierte Spielevideos und ihre Auswirkungen auf die Spielemagazine*. Hamburg: Diplomica-Verlag.

Storrer, Angelika. 2000. Was ist "hyper" am Hypertext? In Werner Kallmeyer (ed.), *Sprache und neue Medien*, 222–249. Berlin and Boston: De Gruyter.

Stukenbrock, Anja. 2009. Herausforderungen der multimodalen Transkription: Methodische und theoretische Überlegungen aus der wissenschaftlichen Praxis. In Karin Birkner & Anja Stukenbrock (eds.), *Die Arbeit mit Transkripten in Fortbildung, Lehre und Forschung*, 145–169. Mannheim: Verlag für Gesprächsforschung.

Tereick, Jana. 2013. Die Klimalüge auf YouTube: Eine korpusgesteuerte Diskursanalyse der Aushandlung subversiver Positionen in der partizipatorischen Kultur. In Claudia Fraas, Stefan Meier & Christian Pentzold (eds.), *Online Diskurse. Theorien und Methoden transmedialer Online-Diskursforschung*, 226–257. Köln: Halem.

Tereick, Jana. 2016. *Klimawandel im Diskurs: Multimodale Diskursanalyse crossmedialer Korpora*. Berlin and Boston: De Gruyter.

Thompson, John. 1995. *The Media and Modernity: A Social Theory of the Media*. Cambridge: Polity Press.

Tolson, Andrew. 2010. A New Authenticity? Communicative Practices on YouTube. *Critical Discourse Studies* 7(4). 277–289.

Uryupina, Olga, Barbara Plank, Aliaksei Severyn, Agata Rotondi & Alessandro Moschitti. 2014. SenTube: A Corpus for Sentiment Analysis on YouTube Social Media. *LREC* 4244–4249.

Vonderau, Patrick. 2016. The Video Bubble: Multichannel Networks and the Transformation of YouTube. *Convergence: The International Journal of Research into New Media Technologies* 22(4). 361–375.

Vorderer, Peter (ed.). 1996. *Fernsehen als 'Beziehungskiste': Parasoziale Beziehungen und Interaktionen mit TV-Personen*. Opladen: Westdeutscher Verlag.

Zhang, Wei & Cheris Kramarae. 2014. "SlutWalk" on Connected Screens: Multiple Framings of a Social Media Discussion. *Journal of Pragmatics* 73. 66–81.

Zichel, Jana. 2016. *Sprache im Layout. Analyse von kohäsionsstiftenden Mitteln auf Websites aus internetlinguistischer Perspektive*. Technische Universität Berlin MA thesis.