RiCL Research in Corpus Linguistics

Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription

Marie-Louise Brunner ‑ Stefan Diemer
Trier University of Applied Sciences / Germany

Abstract – The article discusses how to integrate annotation for nonverbal elements (NVE) from multimodal raw data as part of a standardized corpus transcription. We argue that it is essential to include multimodal elements when investigating conversational data, and that in order to integrate these elements, a structured approach to complex multimodal data is needed. We discuss how to formulate a structured corpus-suitable standard syntax and taxonomy for nonverbal features such as gesture, facial expressions, and physical stance, and how to integrate it in a corpus. Using corpus examples, the article describes the development of a robust annotation system for spoken language in the corpus of *Video-mediated English as a Lingua Franca Conversations* (ViMELF 2018) and illustrates how the system can be used for the study of spoken discourse. The system takes into account previous research on multimodality, transcribes salient nonverbal features in a concise manner, and uses a standard syntax. While such an approach introduces a degree of subjectivity through the criteria of salience and conciseness, the system also offers considerable advantages: it is versatile and adaptable, flexible enough to work with a wide range of multimodal data, and it allows both quantitative and qualitative research on the pragmatics of interaction.

Keywords – corpus annotation; corpus transcription; multimodality; nonverbal elements; spoken discourse; video-mediated communication; gestures

1. Multimodality as part of rich data: The transcriber's dilemma

Complex or 'rich' data poses specific problems in terms of corpus integration. Paralinguistic elements, such as prosody, overlap, laughter in audio data, or nonverbal elements such as gaze, gestures and background interaction in video data, are introducing a level of complexity that is difficult to integrate as part of a replicable and structured transcription system. The question of how to handle such rich data has become increasingly urgent, as more and more datasets have become available through online sources or multimodal compilation projects (cf. Brunner *et al*. 2017). Research acknowledges multimodality as an integral part of the meaning-making process, and studies on multimodal discourse, such as Kress (2011) and Scollon and LeVine (2004),

have established a comprehensive view on language in use as "always and inevitably constructed across multiple modes of communication, including speech and gesture […]." (Scollon and LeVine 2004: 1f.). The realization that the lexical level is only one of many modes and thus only a partial means of meaning making (cf. Kress 2011: 46) creates a problem for corpus researchers. Bezemer and Jewitt (2010: 194) caution that "[m]ultimodality is an eclectic approach" and argue that researchers are faced with a dilemma:

> Too much attention to many different modes may take away from understanding the meanings of a particular mode; too much attention to one single mode and one runs the risk of 'tying things down' to just one of the many ways in which people make meaning. (Bezemer and Jewitt 2010: 194)

A possible solution is to ensure that both corpus data and corpus architecture allow the integration of additional modes, creating the possibility to study various features either independently or in correlation. Multimodality is, of course, a very broad term and, while we consider key features such as paralanguage (e.g. laughter) as part of the general multimodal setting (and as necessary component of spoken corpus transcription), in this article we will focus on the representation of what we term 'nonverbal elements' (NVE), comprising gestures, facial expressions, gaze, and physical stance, as well as camera shifts and background events. This use of NVE constitutes a slight expansion of Adolphs and Carter's term 'nonverbal features' for "gestures which exist in and complement spoken discourse" (Adolphs and Carter 2013: 145), as we also include some affordances of the medium that serve a similar purpose, such as camera shifts and visual and auditory background events (e.g. a person being visible in the background, intruding, or talking to one of the speakers). As part of the transcription and annotation process, corpus compilers and annotators have to achieve the balancing act between preserving and documenting nonverbal features as far as possible in the transcriptions and focusing on those features that are salient in the discourse context. Salience here refers to NVE contributing to or supporting meaning making, as well as NVE that are referred to on a verbal level or that refer to something that is discussed in the conversation (see also Section 3.1). This is, of course, difficult, and researchers have to choose how much of the rich information in a multimodal dataset can and should be included in the finished corpus, either as part of the transcription or as one of various corpus components. This problem of choice is compounded by the complex nature of the data. This rich data creates a second dilemma:

finding a standard way of transcribing it. As DuBois (1991: 73) points out, “there is not, nor ever can be, a single standard way of putting spoken word to paper.”

The question we will explore in this article is how to integrate annotation for NVE elements from multimodal raw data as part of a standard lexical transcription corpus. We argue that it is essential to include multimodal elements when investigating conversational data wherever possible, and that in order to integrate these elements, a structured approach to the complex, unstructured multimodal data is needed. Artificial Intelligence-supported automated gesture recognition does not (yet) provide a satisfactory solution here, and the complex nature of multimodality makes manual annotation necessary in order to obtain gold standard corpus data. We thus need a standard syntax and taxonomy for manually annotating nonverbal features.

We present our approach to creating and implementing such a standard system in the corpus of *Video-mediated English as a Lingua Franca Conversations* (ViMELF 2018). Using examples from the corpus, we describe the bottom-up development of a manual annotation system for spoken language that takes into account previous research on multimodal features, focuses on salience and simplification, and uses a standard syntax. We will also illustrate its potential use in discourse research. Our aim is the creation of a concise and robust transcription system which can be used with a large variety of search tools by researchers from various disciplines who do not need any previous knowledge in gesture research in order to read and understand the data. We see possible uses in varied fields, such as corpus-based multimodal discourse analysis, corpus linguistics, conversation analysis, interactional (socio)linguistics, World Englishes, English as a Lingua franca, and language acquisition.

2. Existing approaches to the description and transcription of nonverbal elements

Multimodal features, and in particular nonverbal elements such as gestures, pose a considerable problem for corpus compilation. The crucial role of gestures in interaction is frequently underlined (e.g. Kendon 2004; Goodwin and Goodwin 2000), and gestures have been studied extensively in multiple branches of linguistics, such as conversation analysis, language acquisition, cognitive linguistics, psycholinguistics, forensic linguistics, multimodal discourse analysis, linguistic anthropology, as well as in psychology. There have also been repeated calls to integrate multimodal data as part of

corpus data (e.g. Adolphs and Carter 2013). However, there is, to our knowledge, no generally recognized and practical transcription system that manages to capture this complex dynamic interaction between gesture, context, and talk. The main problem to overcome in developing an annotation system for nonverbal elements is their complex nature in terms of contribution to discourse, which gesture research has variously commented upon. In our analysis of the various approaches, we will distinguish 'describing' from 'transcribing' nonverbal elements.

2.1. Describing nonverbal elements

Adam Kendon, one of the foremost gesture researchers, describes gestures as utterances that contribute to human understanding like vocal elements, as visual behavior with a communicative and not only informative or expressive function (cf. Kendon 2004). From a psycholinguistic perspective, David McNeill and Duncan call gestures dimensions of social interaction that "open a 'window' onto thinking" (McNeill and Duncan 2000: 143). In his own gesture research, Jürgen Streeck foregrounds their complexity as "largely improvised, heterogeneous, partly conventional, partly idiosyncratic, partly culture-specific, partly universal practice to produce situated understandings" (Streeck 2009: 5). For Charles Goodwin, the acknowledged expert on embodied talk in interaction, all interaction is embodied interaction, movement requires talk and talk requires gestures, and all three create a whole that is different from and greater than the individual parts, as "each individual sign is partial and incomplete" (Goodwin 2007: 199). These descriptions set the scene for the various entailed research perspectives.

Gestures can be described structurally, that is, which body parts are involved, the positioning of these body parts, and movement phases, and how this correlates with prosody or speech in general (e.g. Kendon 1980; McNeill 1992). Another way of describing gestures is by describing the semiotic and semantic content of the gesture in combination with underlying cognitive processes, for example, iconic, metaphorical, indexical, or beat gestures (e.g. Kendon 2004; McNeill 2008; Calbris 2011). A holistic approach to describing nonverbal behavior incorporates gesture as part of a broader concept of embodied action, showing nonverbal elements, the body and its positioning with respect to others and the environment, objects and the surroundings, as well as activities that are being carried out in addition to and in interaction with the verbal level

(e.g. Goodwin 2000; Mondada 2014). These general perspectives on describing gestures are variously employed and adapted by the respective linguistic disciplines.

2.2. Transcribing nonverbal elements

Conversation analysis (CA) researchers routinely include prosodic, paralinguistic, and nonverbal elements in their transcriptions, and there are established annotation systems for "the delivery of talk and other bodily conduct" (Hepburn and Bolden 2013: 57) going back to the Jeffersonian annotation scheme (e.g. Jefferson 1973; Sacks *et al*. 1978). The CA scheme is "a shared, standard system for rendering talk-in-interaction" (Hepburn and Bolden 2013: 75) which is insightful, detailed, and highly relevant for studying spoken discourse. Its main shortcoming from the perspective of corpus linguistics is its limited suitability for quantitative research. CA transcription has been characterized as somewhat unsystematic in its representation of selected features (e.g. DuBois 1991). It is also highly individualized depending on the transcribers' research focus and does not provide a general framework for multimodality, but rather allows the inclusion of selected multimodal features as needed when transcribing the data for a particular purpose of analysis. Researchers in the fields of interactional sociolinguistics, semiotics, and pragmatics have also been studying and transcribing nonverbal behavior in discourse. Gestures are transcribed variously as part of a multi-layered score (Kendon 2004), as aligned descriptions with accompanying illustrations (Streeck 2009), as series of images illustrating stages and aligned with the text (Mondada 2014), as dynamic comic-like transcript inserts, or as a combination of all of the above (McNeill 2008, 2017).

For various reasons, these transcription schemes are not ideal for a corpus context: approaches that strive to be descriptive tend to become increasingly elaborate and difficult to understand. Examples are McNeill's verbal transcriptions (McNeill 2008, 2017) or the complex 'Linguistic Annotation System for Gestures' (LASG) developed by Bressem *et al.* (2013). Approaches that classify gestures through additional visual elements (e.g. Mondada 2014) are difficult to analyze quantitatively, and approaches that focus on interaction dynamics (e.g. Goodwin 2007) introduce a considerable degree of interpretation. With the rapid development of Artificial Intelligence (AI) supported automatic gesture recognition since 2015, attempts to automatically map and systematize gestures as part of multimodal construction grammar are under way (e.g. Joo *et al*. 2017). Though this research direction looks promising with further advances in image

recognition, results so far are limited to a basic physical and very detailed taxonomy of hand gestures and body orientation in TV news data based on gold standard, manually transcribed corpora.

When compiling ViMELF, the approaches described above were considered unsuitable for the purpose of providing a manual annotation system for nonverbal elements that is sufficiently systematized, yet robust, and accounts for all salient features in an interactional context. This prompted the development of the transcription system which is described in Sections 3 and 4.

3. TRANSCRIBING VIMELF

3.1. Data and general transcription guidelines

ViMELF (2018) is a small corpus of 20 dyadic video-mediated conversations in an informal setting between previously unacquainted participants from Germany, Spain, Italy, Finland and Bulgaria, using English as a Lingua Franca. The corpus comprises 113,677 words in the plain text version and 154,472 tokens including annotation (NVE, paralanguage, and affordances of the medium). The gestural annotation is integrated as part of the general transcript rather than creating a separate layer for nonverbal elements (see also Section 4.2). There are 7,449 NVEs in total, which are distributed over 6,463 instances of transcribed nonverbal behavior. The full corpus length amounts to 744.5 minutes (ca. 12.5 hours) of recorded conversation with an average conversation length of 37.23 minutes. The corpus was published in 2018 by the research group of the *Corpus of Academic Spoken English* (CASE) at Trier University of Applied Sciences (Germany), where the corpus is also hosted.[1] It is freely available for research, including the anonymized audio and video recordings. The transcripts provide timestamps every 30 seconds as a simple alignment feature in order to facilitate retrieving the corresponding audio or video sequences for a more comprehensive analysis.[2]

ViMELF was transcribed and annotated manually by a team of more than 60 transcribers on the basis of Dressler and Kreuz's (2000) synthetic transcription conventions which were then extended for the particular conditions of spoken computer-mediated communication (CMC) in an international context. In developing an annotation

[1] For further information on ViMELF (2018), see the project website at http://umwelt-campus.de/case
[2] Timestamps were omitted in the transcribed examples used in this paper to facilitate reading.

system for this particular setting, the transcription team followed, as much as possible, DuBois' (1991) and Edwards' (1993) guidelines for spoken discourse transcription, which still constitute best practice in the field. DuBois' (1991) maxims for transcription are: (i) a clear definition of categories, (ii) accessibility, including the use of notations that maximizes access and are easily and intuitively readable, (iii) robustness, (iv) economy, and (v) adaptability. Edwards (1993) requires the established categories to be (i) discriminable, (ii) exhaustive, and (iii) contrastive, with the aim of creating a systematic and predictable scheme that allows multiple transcribers to work on the data while ensuring consistency and retrievability.

Because multimodal corpora are impossible to transcribe fully, both DuBois and Edwards recommend that transcribers have to be selective and select a finite number of features for transcription. The key criterion for choosing which features to transcribe is salience, in particular in relation to multimodal elements supplementing the lexical level. Our use of salience here refers to "a property of a linguistic item or feature that makes it in some way perceptually and cognitively prominent" (Kerswill and Williams 2002: 81), that is, contributing in some way to meaning making. As Norris (2002: 118) points out, "salience derives from the interaction," which means that multimodal elements can enhance the verbal level or even acquire their own salience independently of the lexical level (e.g. pointing). DuBois' robustness and economy maxims also reflect the need to establish salient categories, while Edwards' demand for an exhaustive set of categories is more difficult to maintain in this respect. Conversely, not all multimodal elements are salient; they can also be incidental or redundant. To determine salience with a maximum degree of objectivity, transcribers need to compare their respective perceptions during transcription and also consider the baseline of the respective dataset in order to produce a consistent corpus transcription. A speaker may have a certain base speaking speed which can be either slow or fast, so slow speaking in itself may not be salient. Deviating from the base speed may, however, foreground particular items and thus establish salience. The same is true of habitual gestures such as scratching one's nose, in comparison to one-time gestures that may convey meaning in this particular context. Scratching one's nose while pausing and saying *Ummmm* can, for example, convey skepticism. The criterion of salience introduces a certain unavoidable degree of subjective interpretation, but transcription would be impossibly detailed without it. Salience is the only way to satisfy Edwards' maxim for an exhaustive set of categories in a multimodal

dataset, and we thus consider salience to be the most important maxim when transcribing multimodal data.

In terms of transcription procedure, DuBois advocates a transcriber-centered system that allows the transcribers' increasing experience during the transcription process to filter back into the system:

> The system should be convenient and comfortable to use, reasonably easy to learn, and through its implicit categories it should promote insightful perception and classification of discourse phenomena, which in the end may feed back into advances in the system itself. [...] Through the experience of transcribing the transcriber is constantly learning about discourse. (DuBois 1991: 75)

DuBois also cautions that the system needs to be flexible:

> It is the transcriber, immersed in the recorded speech event and grounded in discourse theory, who is in a position to [...] advance the potential of the transcription system and its theoretical framework. (DuBois 1991: 75)

After describing the data and the general guidelines that were followed during transcription and annotation of ViMELF, the transcription process itself is presented in Section 3.2.

3.2. Transcription process

The ViMELF transcription process was designed to ensure that the guidelines presented in Section 3.1 were observed, and that transcribers had opportunities for feedback during the transcription process. The process consisted of a pilot transcription phase, followed by three consecutive main transcription phases: pilot transcription phrase, first transcription phase, and second transcription phase.

In the pilot transcription phase, senior project transcribers transcribed the same randomly selected conversations to identify key issues and potential inconsistencies. The transcripts were then compared, and the guidelines formulated and refined in several transcription rounds until inter-transcriber reliability was above 95 percent.

During the first transcription phase, 50 student transcribers were employed in three consecutive rounds as data became available. The student transcribers were trained in specific transcription tutorials that included a parallel transcription of corpus data by all

transcribers and a joint analysis and discussion of inconsistencies. After training, each student transcriber then transcribed at least 30 minutes of conversation; some conversations were transcribed by multiple student transcribers to check for remaining inconsistencies. Transcription was done with the help of the transcription software *F4trankript*, which facilitates close analysis of the audio and video data through features such as repetition looping, timestamping, and low-speed playback. Transcribers were free to either integrate verbal, nonverbal and paralinguistic features at the same time or to work on each feature (e.g. lexis, pauses, laughter, nonverbal elements) consecutively, as both techniques yielded data of comparable quality. Student transcribers were regularly polled on transcription issues. Based on the results of the transcriber polls, average duration for transcribing one minute of audiovisual data is around two hours for a novice and one hour for a senior transcriber. Not surprisingly, the areas where the most significant issues and inconsistencies were reported were the identification and transcription of nonverbal elements, the transcription of paralinguistic elements, in particular laughter, and the identification of intonation units. This prompted the development and further refinement of separate guidelines for the transcription of nonverbal elements as presented in this article. Separate guidelines were also created for the treatment of paralanguage, in particular laughter (for a discussion of ViMELF transcription guidelines for laughter see Brunner *et al.* 2017).

The second phase of the main transcription consisted of a thorough second transcription and correction by six senior project transcribers. The senior transcribers compared transcripts in regular meetings and discussed inconsistencies and general issues. Remaining inconsistencies were consolidated, and the guidelines were further elaborated if needed and then fed back into the next round of transcriptions. Inter-transcriber reliability at the end of this phase was evaluated at 98 percent.

The third phase consisted of a final correction by four project coordinators to ensure consistency of the final dataset. Project coordinators and senior project transcribers met regularly to discuss differences in transcription, issues of salience, and problematic features.

In sum, regular team meetings at all transcription stages ensured transcriber input on desirable adjustments in the transcription system, contributing to a data-driven, bottom-up formulation of guidelines. The resulting transcription guidelines for ViMELF contain provisions for:

(i) lexical transcription,

(ii) spoken language features (cut-offs, overlap, liaisons, latching),

(ii) prosody (intonation, pitch, volume, speed, pauses) and paralinguistic features (laughter, coughing, sighing, loud breathing),

(iv) nonverbal elements (gestures, facial expressions, gaze, physical stance, camera shifts, background events).

In addition, some specific ELF and video-mediated features are also transcribed, such as code-switching, non-standard pronunciations, and technical issues such as echo.[3] While the guidelines represent the result of an elaborate process, the availability of the anonymized raw data as part of the corpus specifically allows further development and inclusion of additional features at need, depending on the interest of future researchers and transcribers. In the context of this paper, we will focus on just one of the most challenging features to illustrate the design and compilation of transcription guidelines: the transcription of nonverbal behavior.

4. Developing a transcription system for nonverbal elements

4.1. Nonverbal elements in interaction: Examples from ViMELF

In his seminal 1991 article on transcribing spoken discourse data, DuBois does not provide for a multimodal transcription system, but already indicates the need for further research in that direction:

> There are several dimensions along which further development can be hoped for in the coming years ─for example [...] nonverbal cues like eye gaze, body orientation, and so on. (DuBois 1991: 87)

The nature of ViMELF data makes the need for such a development evident. Examples (1) and (2) with Figures 1 and 2 from the ViMELF recordings illustrate the role nonverbal elements can play. In example (1), the German participant SB27 and her Italian conversation partner FL25 talk about the books they own.

(1) Books (03SB27FL25)

SB27:	I have so much books here that I .. bought,
	but .. I can't read them. ((hehe))
FL25:	look, {shifts camera to show bookshelf} {points to bookshelf}

[3] The guidelines are available on the ViMELF homepage at Trier University of Applied Sciences, Germany (ViMELF 2017a).

I mean .. we have dictionaries,

yes, dictionaries,

but ther- but there are also books there somewhere, {makes brushing-away gesture; arm still extended to back}

Figure 1: Pointing gesture in Books (03SB27FL25). Click on image to see the full video sequence

Even just focusing on the visual level and disregarding, for the moment, paralinguistic features, several interesting features can be shown. FL25 shifts her stance by leaning back and out of the screen so the bookshelf in the background is no longer obscured, indicating awareness of her conversation partner's field of view. This shift of orientation is accompanied by the invitation *look* while FL25 moves her laptop computer so that the camera points to the bookshelf, forcing a shift of perspective also for SB27, who responds with a backchanneling smile and nodding, signaling understanding and marking agreement and engagement, all nonverbally. This is then immediately followed by FL25 extending her right hand to point at the bookshelf and the books in it.

In example (2), the German participant SB73 explains Bavarian traditional male dress code but does not recall the word for braces.

(2) Braces (06SB73ST14)

SB73: ... an:d uhm: they have, {lifts head & rolls eyes}
(1.1) how do you call it uhm,
(1.3) uhm .t, [((ehh))] {imitates braces with both hands}
th- it's uhm .t, {imitates braces with both hands}
... uh like a rubber band,
it goes .. [on your trousers],

ST14: [two things],= {imitates braces with both hands}

SB73: =yeah, {imitates braces with both hands}

ST14: right okay, {smiles & nods}

SB73: yeah [I didn’t], {points at herself with both hands}
ST14: [>I don’t know what is-<]
... >I don’t know what it's-< what's the name for it right.
yeah I know what you mean, {closes eyes} ((hehe))
SB73: yeah, {nods}
it hol- holds the trousers .h? {looks down & lifts arms} ((snuffles))

Figure 2: Imitating gesture in Braces (06SB73ST14) with QR video link. Click on image to see the full video sequence

SB73 uses imitative gestures to convey her meaning; the gesture is then mirrored by her Spanish conversation partner ST14. Shared understanding is negotiated and achieved through nonverbal elements without using, at any point, the lexical item that denotes the referent.

4.2. Transcription guidelines for nonverbal elements: Basic guidelines

The complex nature of gesture and other nonverbal elements that we already commented upon raises the question of how to proceed when developing a systematic annotation system. Whether we consider NVE and speech to be overlapping and complementary, or to represent a single system may depend on the type of gesture we analyze. What is true for all interpretations is that NVE represent an essential part of meaning-making in interaction and cannot be ignored when analyzing multimodal data. The approach taken by the ViMELF transcription team focused on four basic guidelines that tie in with the general transcription guidelines discussed above:

(i) the system should take into account the function of the NVE in interaction,
(ii) it should be as systematic as possible and use a regular and predictable syntax that allows quantitative research,
(iii) it should be as descriptive, but also as simple as possible, and
(iv) it should remain adaptable.

In the development of the ViMELF annotation system the compilers refrained, as far as possible, from interpreting NVE during annotation in order to make the transcripts as objective as possible, leaving it to the researchers to draw their own conclusions. At the same time, Hepburn and Bolden's observation on the complex nature of visible behavior is of particular relevance:

> Although the transcription of both talk and visible behavior is necessarily selective, the transcription of visible behavior may be even more so due to the substantial number of parameters. Moreover, visible behavior involving facial expressions, body posture, gestures and gaze can occur in overlap with each other and with talk. (Hepburn and Bolden 2013: 70)

The video component of ViMELF remains available as an integral part of the corpus, so that researchers can return to the raw data in order to supplement the transcript in a context of a more exhaustive multimodal analysis. The annotation system is specifically left open for additions —if salient gestures are observed that are not yet codified, transcribers can easily expand the taxonomy following general guidelines and mark-up syntax.

The ViMELF project team decided to integrate the gestural annotation into the general transcript rather than creating a separate layer for nonverbal elements. In line with keeping the general transcript syntax as simple and readable as possible, nonverbal elements are universally marked with curly brackets, thus: {shrugs}. As mentioned in Section 1, our definition of nonverbal elements includes not only gestures, but also other embodied talk, such as salient head movements and facial expressions, gaze, physical stance shifts, camera shifts, and interaction happening in interlocutors' surroundings, as these can all be considered salient nonverbal contributions to meaning-making.

4.3. Transcription guidelines for nonverbal elements: Development

Similar to the general transcription process, the development of an annotation scheme for nonverbal elements was a data-driven, bottom-up process that integrated continuous feedback by transcribers. Its aim was the classification of salient nonverbal elements in the form of a clear taxonomy. The development process can be divided into four phases:

(i) a survey of existing transcription practices for gestural research,

(ii) a survey of salient nonverbal elements marked by transcribers in pilot transcriptions,

(iii) the formulation of general guidelines for NVE transcription, and

(iv) an inventory of NVE documented in the data as the basis for transcription.

The phases are briefly illustrated below.

(i) Survey of existing transcription practices. The aim of this phase was to establish whether there is a best-practice approach for the annotation of nonverbal elements that can be used or adapted to the multimodal data. While our approach is informed by the more general transcription practices for NVE in CA and interactional sociolinguistics, none of these schemes, as already discussed, fulfils the specified requirements. There is a number of corpora that integrate gestures in their annotation, mostly aligned and integrated into multi-layer display tools. The *Augmented Multi-party Interaction Corpus* (AMI; cf. Carletta *et al.* 2006) and the *SmartKom Multimodal Corpus* (Schiel *et al.* 2002) both use experimental annotation systems that focus on distinguishing conversational (or interactional) and nonconversational gestures with the aim of enhancing machine gesture recognition and are, due to this narrow focus, not suitable for adaptation with ViMELF data. Several corpora use the MUMIN multimodal coding scheme (Allwood *et al*. 2007), which was developed to experiment with annotation of multimodal communication in television data, for example the *Multimodal Human-Computer Interaction Technologies Corpus* (MM HuComTech; cf. Pápay *et al*. 2011). MUMIN focuses on the interpretation of the communicative function of NVE and proposes mutually exclusive categories, "since the focus of the annotation scheme is on the explicit communicative function of the phenomenon under analysis" (Allwood *et. al*. 2007: 278). In other words, "the annotator is asked to select the most noticeable communicative function" (Allwood *et. al*. 2007: 278). This focus on interpretation and the absence of multifunctionality are the key reasons why the system was not considered for use with ViMELF, though the systems share several features, in particular in the basic differentiation of behavior attributes (such as hand shape or head movement).

(ii) Survey of salient NVE marked by transcribers. In a separate pilot phase for the transcription of nonverbal elements, the six project transcribers were asked to transcribe salient NVE in six sample transcripts, constituting about a quarter of ViMELF corpus data, and to describe them in a concise manner. The NVE identified after this round were systematized, resulting in a list of salient NVE and another list of potential inconsistencies regarding descriptive syntax, the concept of salience and the issue of gesture overlap. The transcription team then compared their transcripts and discussed these inconsistencies, contributing to the formulation of general guidelines for transcription.

(iii) Formulation of general guidelines for NVE transcription. In order to make sure transcribers in the main transcription phase would mark gestures consistently, several general principles for transcribing NVE were formulated on the basis of the survey in phase (ii). The general principles formulated during this process are presented in Table 1.

1. Salience	Only salient NVE are transcribed. Salience here refers to NVE contributing to or supporting meaning making, as well as NVE that are referred to on a verbal level or that refer to something that is discussed in the conversation.
2. Markup	All transcribed salient NVE are marked by curly brackets, creating self-contained, searchable markup units that can easily be converted to other data formats such as Extended Markup Language (XML), while being part of an easy to read, lexical transcript.
3. Conciseness and syntax	The transcription syntax of NVE is verb-based and concise. The verb should be in the third person present tense, not the present participle, that is, {nods}, not {nodding}. If there are commonly used verbs that already encompass the NVE they should be used instead of disassembling the NVE into single verb components, such as {makes peace sign} instead of {lifts hand; palm outward} {spreads two fingers}. It is understood that this will always include some level of abstraction and/or interpretation; the aim is not to provide a semiotically precise representation, but an easy to read, concise description.
4. Treatment of consecutive and co-occurring NVE	Consecutive NVE can be transcribed consecutively if no concise transcription exists: {smiles} {nods} {makes thumbs-up gesture}. If several NVE co-occur, they are transcribed in one bracket and connected with &: {smiles & nods}. If the NVE consists of separate stages that could be part of the same NVE, it should not be disassembled into phases but transcribed concisely, e.g. {imitates breathalyzer by blowing into top end of pen}, not {imitates breathalyzer by lifting pen to mouth and blowing into top end of pen and setting it down}.
5. Position and alignment	NVE transcription should follow the intonation unit containing the most salient use, or, if limited to smaller units (e.g. words), follow those units. It is not aligned and not marked for duration, intensity, or speed.
6. Modification	The following modifiers can be added (if co-occurring, in this order): a) direct object, if the main verb does not already contain the object sense (e.g. *nods* includes the object *head* and does not need to be repeated), e.g. *lifts arm*, *lifts hand*. If a NVE is already conventionally named, it may be used as object of *make*, e.g. {makes throwaway gesture} {makes peace sign} {makes air quotes} b) temporal adverb (*three times*, *repeatedly*, ...) c) directional adverb(ial) (e.g. *up*, *down*, *behind ear*), if necessary. Directions are not separately denoted if the specification does not have an influence on the meaning of a NVE —thus, left and right are usually not distinguished. Directions are always given from the speaker's perspective (*outward*, *inward*, *front*, *back* etc.). d) *to* ..., if a target/level needs to be added (*to chair*, *to eye*, etc.) e) *with* ..., indicating for example the hand(s), body parts or objects used in the NVE (e.g. *with left hand*, *with left index finger*), if salient f) *by* [...]-*ing*, if a further modification is needed (e.g. in the case of imitating, as in *imitates breathalyzer by blowing into top end of pen*) g) other additions, for example, if further specification is needed, separated with semicolon {lifts hands; palms outward} {lifts hand; palm up}. This modification should be used sparingly. The sequence should be as short as possible, following the conciseness principle.

Table 1: General principles for the transcription of nonverbal elements in ViMELF (adapted from ViMELF 2017b)

In general, the taxonomy makes use of names of conventionalized Western European NVE as descriptions to reduce complexity in the annotation (e.g. {shrugs}, {nods}). Additional explanations are added in the taxonomy, clarifying NVE that might be culturally specific (e.g. 'peace sign'). The annotation aims to be as descriptive as possible, but in some cases, the terminology used may imply certain meanings or interpret NVE to a certain degree to clarify the context (e.g. 'fist pump'). The actual interpretation of the individual functions of an NVE should remain with the researcher, taking into account the conversational setting, particularly as speakers' different cultural backgrounds increase the probability of diverging functions for similar NVE.

(iv) Inventory of NVE documented in the data. The NVE transcribed during phase (ii) were collected in order to serve as salient examples of NVE that could occur in the transcription guidelines, and to create a data-driven, bottom-up taxonomy for nonverbal elements in ViMELF. The taxonomy was specifically left open so that new salient instances of NVE could be added following the guidelines. This resulted in a taxonomy of salient nonverbal features as presented in Table 2, current as of March 2020.

The taxonomy currently comprises 55 NVE, of which nine are associated with facial expressions, four with the head, including gaze features, three with physical stance and two with the speakers' background. 39 features are associated with hand or body movement; these do not only include movement, but also actions such as standing up, walking, or camera movements that force a shift of perspective. Both guidelines and taxonomy were integrated into the general transcription guidelines and included in the training sessions for student and project transcribers. The current version of guidelines and taxonomy is available online.[4]

[4] Transcription of nonverbal elements (ViMELF 2017b)

Head, including gaze
- Looks (up, down, to side, to upper corner ...)
- Nods (head moves up and down)
- Shakes head (head turns left and right)
- Tilts head (repeatedly)
Facial expressions
- Frowns
- Grimaces (implying negative connotation)
- Purses lips
- Raises eyebrow(s)
- Rolls eyes
- Smiles
- Squints
- Winks
- Yawns
Hands and body
- Claps
- Clasps hands (in front of chest if not otherwise specified)
- Drinks from …
- Drums fingers (rapid movements with fingers)
- Eats …
- Folds arms
- Hits/thumps with … on …
- Holds … to …
- Holds up … (two fingers, glass of wine, etc.)
- Imitates … (drinking, breathalyzer, braces, shape of …, size of …, etc., by ...)
- Lifts hand (to…, or lifts hand; palm up)
- Makes ...
- air quotes (imitates quotation marks with index and middle fingers)
- beat gesture (up and down hand movement during speech; cf. McNeill 1992)
- box gesture (raises hands and moves them, palms vertical, cf. Cassell 1998)
- fist pump (makes fist with one hand, moves fist quickly downwards)
- fist(s)
- okay sign (index finger and thumb together, other fingers extended)
- peace sign (makes a 'V' with index and middle finger, palm outward)
- swiping gesture (moves hand sidewards quickly)
- throwing-away gesture (downward hand movement, palm down, cf. Bressem *et al.* 2013)
- brushing-away gesture (upward hand movement, palm down, cf. Bressem *et al.* 2013)
- thumbs-up gesture, thumbs-down gesture (fist with thumb extended)
- Moves hand to … (to mouth, to forehead, etc.)
- Moves hands … (in circle, outwards, up, etc.)
- Opens hand(s) (outwards movement, palm upwards)
- Points to … (with …) (with index finger/hand, etc.)
- Puts … on …
- Rubs … against ... (rubs thumb against index and middle fingers)
- Scratches (head) (if salient, e.g. in combination with hesitation, thinking, etc.)
- Shifts camera to show …
- Shows … (moves object in front of camera/closer to screen to focus attention)
- Shrugs
- Stands up / sits down

Table 2: Taxonomy of salient nonverbal elements in ViMELF (2017b)

Hands and body (cont.)
- Touches … (head, ear, shoulder, etc.)
- Types
- Walks to ...
- Waves
Physical stance
- Leans … (forward, backward, towards …)
- Sits up (straighter than before)
- Shifts position
Background
- Movement: Noun + verb in third person (roommate walks past screen, etc.)
- Background sounds: Noun + verb in third person (baby cries etc.), noun (clicking sound, etc.) if source is unclear

Table 2 (continuation)

4.4. Transcription guidelines for nonverbal elements: Implementation

Transcription of ViMELF in its current version 1.0 took place over a period of two years, from April 2016 until April 2018. This was due to the fact that recordings of ViMELF data were still ongoing and that transcription phases needed to coincide with research periods of student transcribers. Both transcription guidelines and taxonomy were updated and expanded repeatedly during the transcription period and transcripts checked repeatedly for conformity with the latest version before publication in May 2018.

4.5. Transcription guidelines for nonverbal elements: Advantages and disadvantages

One potential disadvantage of the ViMELF transcription system for nonverbal elements is the selective perspective introduced by the maxims of salience and conciseness.

The salience maxim was a central component in transcriber training in order to ensure a maximum degree of agreement in the transcripts. While this aim was reached with inter-transcriber reliability at more than nine percent at the end of the second transcription phase, it also had the effect that features that were classified as nonsalient were almost uniformly excluded from the transcripts. The resulting transcript is thus necessarily a selective representation of the interactions. Features that are not transcribed but that are of interest to researchers in another context, for example, idiosyncratic or incidental gestures, will need to be extrapolated from the raw audio and video data which is an integral part of the corpus.

To illustrate the consequences of the focus on conciseness, we will consider the transcription of {imitates}, one of the most intriguing subcategories of NVE documented in the corpus. {Imitates} refers to instances where gestures are used to imitate an action, activity, object, shape, size, etc. while explaining or referring to it verbally, as shown in example (2) and Figure 2. There are 130 instances of imitation in ViMELF ranging from imitations of clothing items, actions, states of mind, or physical distances to cultural traditions. Obviously, these can be very complex sequences that are condensed by the transcription team so as to facilitate comparative research. An example is the imitative action by one of the speakers who explains the word *alcohol* by first imitating a breathalyzer and then the action of drinking, as shown in example (3) and Figure 3.

(3) HE19: oh which one? {leans forward}
SB93: alcohol ((/ˈalkɔ:l)). [was that just in Norway],
HE19: [{shakes head once, leans forward}]
what is it, aikai ((/aikai/))? {leans forward}
SB93: so, alcohol ((/ˈalkɔ:l/)),
<alcohol> ((/ˈalkohəʊl/)). {imitates breathalyzer by blowing in top of pen}
[what you drink].
HE19: [oh: the], {scratches head with left hand} the brand?
SB93: nO, what you drink, {imitates drinking}

Figure 3: Imitating gestures in Alcohol (05SB93HE19). Click on image to see the full video sequence

Example (3) shows an imitation sequence in the interaction between a German and a Finnish conversation partner. SB93 asks her Finnish interlocutor about the price of alcohol in Finland, but HE19 does not understand the word *alcohol*, probably because

SB93 pronounces it very quickly and without aspirating the /h/. Since HE19 still does not understand the intended meaning after repeatedly asking for and receiving a repetition of the problematic word, SB93 finally uses gesture to support her meaning making. She blows in the top of a pen as if to imitate a breathalyzer (see Figure 3, first picture, the circle indicates the top of the pen). In this case, the pen at hand is 'recruited' as imaginary breathalyzer, which is then handled accordingly by blowing into the top. When HE19 still does not understand, confusing SB93's pronunciation with the name of the Finnish national alcohol retailer (*Alko*), SB93 adds the combined verbal/nonverbal explanation *what you drink,* {imitates drinking} (see Figure 3, second picture), using her left hand to illustrate the act of drinking. In the subsequent exchange it becomes clear that the negotiation of meaning is successful.

While the conciseness maxim may lead to a simplification of complex sequences, it is also necessary to ensure that the features can be systematically retrieved, which is an important aspect from a corpus analytical perspective.

On balance, we argue that the selective focus and the integrated standardization is what makes the proposed transcription system feasible for use in a corpus linguistic context. A decisive advantage of the proposed system is the quantification of nonverbal elements which allows a mixed-methods approach. Is the effort involved in such a detailed transcription justified in view of its potential for linguistic research? We would argue that despite the considerable time necessary for transcription of NVE, in ViMELF roughly between one and two hours per minute of recorded data, even a comparatively small corpus such as this, with roughly 150,000 tokens, is large enough to quantify selected features (as will be shown in Section 5), and small enough for a meaningful qualitative analysis of multimodal features. The proposed transcription scheme provides considerable benefits: it helps the researcher to find specific instances for closer analysis, and it provides quantitative observations that can be used to guide the analyst's perspective, and that would not be possible to make by close qualitative analysis of the data alone.

The concise (if necessarily less detailed) multimodal transcript and the possibility to access the original data allow a more complete picture of conversational interaction and open up new perspectives on multimodal conversation.

In order to show the research potential of this resource, two approaches are briefly illustrated in Section 5; for a more extensive study of multimodality in ViMELF see Brunner (2021).

5. Using the transcription system

5.1. Quantification of nonverbal elements

One of the main advantages of having a searchable multimodal corpus is the possibility to use quantitative methods to investigate the role of NVE in interaction. A quantitative analysis of NVE in ViMELF (2018) is easy to carry out and provides first insights into how NVE contribute to meaning-making in interaction, and how they correlate with other discourse features. There are 7,449 salient transcribed NVE in ViMELF, distributed over 6,463 instances of transcribed nonverbal behavior (one instance of non-verbal behavior may contain several parallel NVE). Interestingly, only 35 NVE account for 80.4 percent of all transcribed nonverbal behavior, as illustrated in Figure 3. Of those, the six most frequent NVE ({nods}, {shakes head}, {shrugs}, {raises eyebrows}, {tilts head}, and {smiles} already account for 50.9 percent of the total.

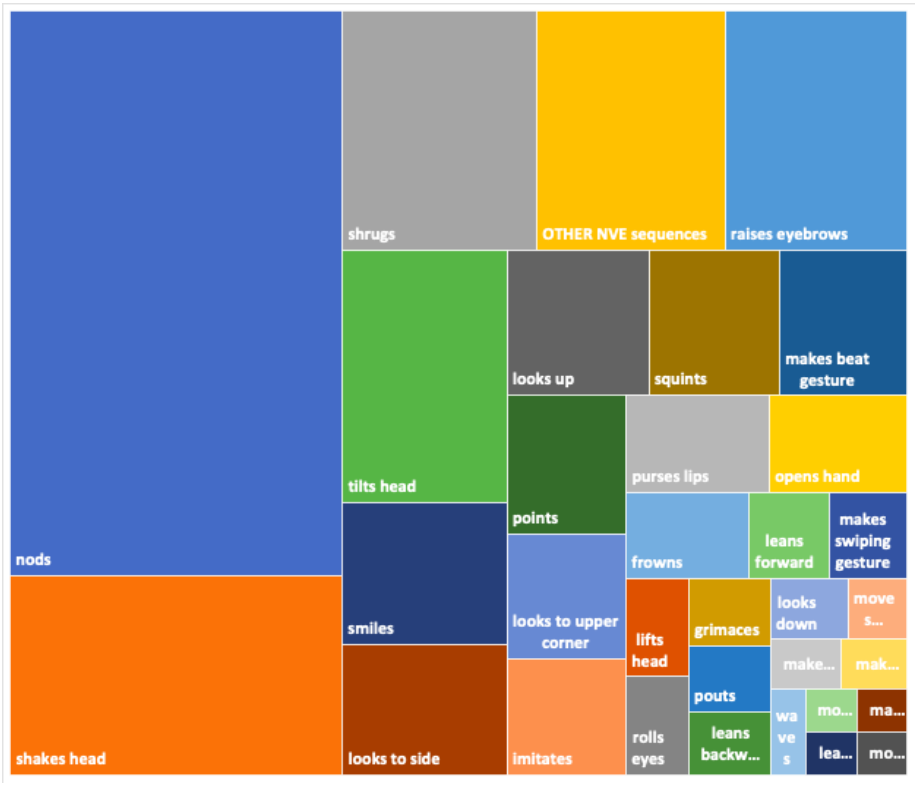

Figure 3: Visualization of relative distribution of nonverbal elements in ViMELF

The descriptive taxonomy advocated in the NVE transcription system means that these instances need to be analyzed qualitatively to determine the function of the respective NVE in the discursive context. If we examine, for example, the 430 instances of {shrugs}, the third frequent NVE, a qualitative analysis reveals a multitude of situational interpretations and functions (see also Brunner *et al*. 2017 for additional examples). {Shrugs} can, for example, express uncertainty (*maybe*. {shrugs}); indicate normalcy and a lack of excitement (*basically the same thing*. {shrugs}); mark a lack of knowledge (*I don't know much about Germany anyway. so.* {shrugs}); mark resignation (*I think it is*, {shrugs} (1.3) *almost impossible*); indicate agreement ({shrugs} *right*); indicate a lack of preferences (*you want to go first or should I?* [...] {shrugs} *go ahead*); signal exasperation (*it doesn't make sense, why is the table female?* {shrugs}); and express disapproval (*and the government* {shrugs} *is not doing anything*).

The example illustrates the complexity of possible interpretations and functions in interaction and shows both the advantages and disadvantages of a mainly descriptive annotation system. On the one hand, it allows the quantification of an additional mode without having to refer to the original data in every case; on the other hand, it will still be necessary to perform a detailed manual analysis of the context. Even in this case, though, relevant instances will be easier to retrieve without going through all of the original recordings.

Another clear advantage lies in the possibility to correlate NVE with other elements and with each other. On a basic level, NVE can correlate with lexical items: the 1,741 instances of {nods}, for example, collocate ($p<0.05$) with *yeah*, *mhm*, *right*, and *okay*, while {shakes head} correlates with *no* and *not*. Both correlations are not surprising. But correlations can become complex very fast: the 386 instances of {tilts head}, for example, correlate with *well*, and *then*, but also with the NVE {nods}, {tilts head}, and {raises eyebrows} as well as the paralinguistic elements ((*ehh*)), ((*heh*)), ((*laughs*)), which are various types of laughter. A correlation analysis like this has the potential to enhance our understanding of meaning making. The gesture {tilts head} clearly is part of a complex negotiation sequence that may include hesitation markers, laughter, and other gestures. It can thus contribute one additional facet to a mixed-method analysis of talk-in-interaction with quantitative and qualitative elements.

5.2. Discursive functions of nonverbal elements

A corpus that is annotated for multimodality also allows researchers to easily extract nonverbal elements and to focus on their broader functions in discourse. One comprehensive recent study uses ViMELF data for the development of a model for multimodal meaning negotiation in video-mediated interactions based on ViMELF data (Brunner 2021). Preliminary results show that although interlocutors are separated by the computer screen and in different environments, they make use of nonverbal elements to complement, replace, nuance, and support their verbal utterances multimodally. Understanding is signaled through both verbal and nonverbal back-channeling. Interlocutors notice aspects of their respective surroundings and can focus attention on them through both verbal means and complementary focusing NVE, for example through pointing, object showings, or camera shifts. Interlocutors also interact with their immediate environment, causing disruptions that have to be negotiated. These first results show the potential for further work with multimodally annotated corpora in order to investigate spoken discourse.

6. CONCLUSION

In our article we propose a concise annotation system for nonverbal elements in spoken discourse and illustrate its application in the context of the ViMELF corpus as a way of integrating unstructured multimodal data into a corpus context. We also show several applications of a corpus annotated with the proposed system for both quantitative and qualitative research on multimodal discourse. The main challenges in creating annotation for multimodal features are (i) the necessity to create systematic criteria for selecting which multimodal features to transcribe and (ii) the need to create an annotation syntax that facilitates systematic quantitative research while preserving a consolidated transcript including lexical, nonverbal, and paralinguistic elements. The resulting taxonomy is based on the two principles of salience and conciseness, and constitutes a systematic, descriptive and comprehensive annotation system. Our aim is not to replace existing approaches, but to provide a robust, easy-to-use tool for the annotation of nonverbal elements as key elements of linguistic meaning-making. In considering both benefits and drawbacks of such a system, we argue that it represents a balanced approach that allows researchers to structure rich, multimodal data and contributes to opening the way for the development of more rich-data corpora and a wide range of applications.

REFERENCES

Adolphs, Svenja and Ronald Carter. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. London: Routledge.

Allwood, Jens, Loredana Cerrato, Kristina Jokinen, Constanza Navarretta and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41: 273–287.

Bezemer, Jeff and Carey Jewitt. 2010. *Multimodal Analysis: Key Issues.* London: Continuum.

Bressem, Jana, Silva H. Ladewig and Cornelia Müller. 2013. Linguistic Annotation System for Gestures (LASG). In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf eds. *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction.* Berlin: Walter de Gruyter, 1098–1125.

Brunner, Marie-Louise. 2021. *Understanding Intercultural Communication: Negotiating Meaning and Identities in English as a Lingua Franca Skype Conversations*. Saarbrücken: Saarland University PhD dissertation.

Brunner, Marie-Louise, Stefan Diemer and Selina Schmidt. 2017. "... okay so good luck with that ((laughing))?" - Managing rich data in a corpus of Skype conversations. In Turo Hiltunen, Joe McVeigh and Tanja Säily. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations* [Studies in Variation, Contacts and Change in English 19]. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/19/brunner_diemer_schmidt/ (01 May, 2021.)

Calbris, Geneviève. 2011. *Elements of Meaning in Gesture.* Amsterdam: John Benjamins.

Carletta, Jean, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthatl, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio eds. *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11–13, 2005, Revised Selected Papers* (Lecture Notes in Computer Sciences 3869). Berlin: Springer, 28–39. https://link.springer.com/chapter/10.1007/11677482_3

Cassell, Justine. 1998. A framework for gesture generation and interpretation. In Roberto Cipolla and Alex Pentland eds. *Computer Vision in Human-machine Interaction*. Cambridge: Cambridge University Press, 191–215.

Dressler, Richard A. and Roger J. Kreuz. 2000. Transcribing oral discourse: A survey and a model system. *Discourse Processes* 29/1: 25–36.

Du Bois, John W. 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1/1: 71–106.

Edwards, Jane A. 1993. Principles and contrasting systems of discourse transcription. In Jane A. Edwards and Martin D. Lampert eds. *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale: Lawrence Erlbaum Associates, 3–31.

F4transkript. Dr. Dresing & Pehl GmbH. https://www.audiotranskription.de/f4transkript/ (07 May, 2021.)

Goodwin, Charles. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics* 32/10: 1489–1522.

Goodwin, Charles. 2007. Environmentally coupled gestures. In Charles Goodwin, Susan D. Duncan, Justine Cassell and Elena Levy eds. *Gesture and the Dynamic Dimensions of Language.* Amsterdam: John Benjamins, 195–212.

Goodwin, Marjorie H. and Charles Goodwin. 2000. Emotion within situated activity. In Nancy Budwig, Ina Č. Užgiris and James V. Wertsch eds. *Communication: An Arena of Development*. Stamford: Greenwood Publishing Group, 33–53.

Hepburn, Alexa and Galina Bolden. 2013. The conversation analytic approach to transcription. In Jack Sidnell and Tanya Stivers eds. *The Handbook of Conversation Analysis*. Hoboken: John Wiley & Sons, 57–76.

Jefferson, Gayle. 1973. A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica* 9/1: 47–96.

Joo, Jungseock, Francis F. Steen and Mark Turner. 2017. Red Hen Lab: Dataset and tools for multimodal human communication research. *KI-Künstliche Intelligenz* 31/4: 357–361.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Richie Key ed. *The Relationship of Verbal and Nonverbal Communication*. Berlin: Mouton de Gruyter, 207–228.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

Kerswill, Paul and Ann William. 2002. "Salience" as an explanatory factor in language change: Evidence from dialect levelling in urban England. In Mari C. Jones and Edith Esch eds. *Language Change: The Interplay of Internal, External and Extra-Linguistic Factors.* Berlin: Mouton de Gruyter, 81–110.

Kress, Gunther. 2011. Multimodal discourse analysis. In John P. Gee and Michael Handford eds. *The Routledge handbook of discourse analysis*. London: Routledge, 35–50.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill, David. 2008. *Gesture and Thought.* Chicago: University of Chicago press.

McNeill, David. 2017. *Brief Introduction to Annotation.* http://mcneilllab.uchicago.edu/analyzing-gesture/intro_to_annotation.html (01 May, 2021.)

McNeill, David and Susan Duncan. 2000. Growth points in thinking-for-speaking. In David McNeill ed. *Language and Gesture.* Cambridge: Cambridge University Press, 141–161.

Mondada, Lorenza. 2014. Pointing, talk, and the bodies. In Mandana Seyfeddinipur and Marianne Gullberg eds. *From Gesture in Conversation to Visible Action as Utterance: Essays in Honor of Adam Kendon*. Amsterdam: John Benjamins, 95–124.

Norris, Sigrid. 2002. The implication of visual research for discourse analysis: Transcription beyond language. *Visual Communication* 1/1: 97–121.

Pápay, Kinga, Szilvia Szeghalmy and István Szekrényes. 2011. Hucomtech multimodal corpus annotation. *Argumentum* 7: 330–347.

Sacks, Harvey, Emanuel A. Schegloff and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In Jim Schenkein ed. *Studies in the Organization of Conversational Interaction*. New York: Academic Press, 7–55.

Schiel, Florian, Silke Steininger and Ulrich Türk. 2002. *The SmartKom Multimodal Corpus at BAS.* München: Ludwig-Maximilians Universität München Press.

Scollon, Ron and Philip LeVine. 2004. Multimodal discourse analysis as the confluence of discourse and technology. In Philip LeVine and Ron Scollon (eds.), *Discourse*

and Technology: Multimodal Discourse Analysis. Washington: Georgetown University Press, 1–6.
Streeck, Jürgen. 2009. *Gesturecraft: The Manu-facture of Meaning.* Amsterdam: John Benjamins.
ViMELF. 2017a. *ViMELF Transcription Conventions.* Birkenfeld: Trier University of Applied Sciences. http://umwelt-campus.de/case-conventions (01 May, 2021.)
ViMELF. 2017b. *Transcription of Nonverbal Elements.* Birkenfeld: Trier University of Applied Sciences. https://www.umwelt-campus.de/fileadmin/Umwelt-Campus/SK-Weiterbildung/Dateien/Transcription_of_non-verbal_elements_in_CASE.pdf (01 May, 2021.)
ViMELF. 2018. *Corpus of Video-Mediated English as a Lingua Franca Conversations.* Birkenfeld: Trier University of Applied Sciences. http://umwelt-campus.de/case (01 May, 2021.)

Corresponding author
Marie-Louise Brunner
Trier University of Applied Sciences
Environmental Campus Birkenfeld
P.O. Box 13 80
55761 Birkenfeld
Germany
e-mail: ml.brunner@umwelt-campus.de

received: February 2020
accepted: May 2021