Finding Recurrent Features of Image Schema Gestures: the FIGURE corpus

Andy Lücking, Alexander Mehler, Désirée Walther, Marcel Mauri, Dennis Kurfürst
Goethe University Frankfurt
Text Technology Lab
luecking@em.uni-frankfurt.de, mehler@em.uni-frankfurt.de, dwalther@informatik.uni-frankfurt.de,
mauri@informatik.uni-frankfurt.de, kurfuers@informatik.uni-frankfurt.de

Abstract

The **F**rankfurt **I**mage **G**est**URE** corpus (FIGURE) is introduced. The corpus data is collected in an experimental setting where 50 naïve participants spontaneously produced gestures in response to five to six terms from a total of 27 stimulus terms. The stimulus terms have been compiled mainly from image schemata from psycholinguistics, since such schemata provide a panoply of abstract contents derived from natural language use. The gestures have been annotated for kinetic features. FIGURE aims at finding (sets of) stable kinetic feature configurations associated with the stimulus terms. Given such configurations, they can be used for designing HCI gestures that go beyond pre-defined gesture vocabularies or touchpad gestures. It is found, for instance, that movement trajectories are far more informative than handshapes, speaking against purely handshape-based HCI vocabularies. Furthermore, the mean temporal duration of hand and arm movements associated vary with the stimulus terms, indicating a dynamic dimension not covered by vocabulary-based approaches. Descriptive results are presented and related to findings from gesture studies and natural language dialogue.

Keywords: gestures, human-computer interaction, kinetic features

1. Introduction

Gestures for *Human Computer Interaction* (HCI) are usually confined to navigation commands or task-oriented hand and arm movements. Thus, HCI gestures either come as *manipulators* or as part of a specific gesture lexicon (*semaphores*) (Quek et al., 2002) (cf. the touchpad gestures that became popular with the proliferation of touchscreens). While manipulator gestures are highly effective due to providing direct feedback (Brennan, 1998), they are strongly bound to imperative, system-controlling commands. Semaphores can in principle be designed as an input for presumably any kind of operation. However, such semaphores has to be learned in advance, hence they impose a high memorising demand on side of the user. In applications where more abstract declarative interaction is called for (e.g., in the context of education or museums (Mehler and Lücking, 2012)), rich and intuitive interaction means are needed. In order to meet this requirement, Mehler et al. (2014) proposed a conceptual approach to HCI gestures based on the notion of *image schema* propagated in cognitive sciences (Lakoff, 1987). This approach assumes that predicational information can be enacted by an image schema-triggering gesture whose "meaning" is derived from the underlying schema, giving rise to a kind of sign language called "gestural writing" (Mehler et al., 2014). Note that this approach is conceived for touchless HCI, as accomplished by controllers like Kinect or LeapMotion.[1] Gestural writing can be an efficient means for HCI only if there is an association between certain hand and arm movements and concepts (Grandhi et al., 2011) beyond idiosyncrasies in gesture production (Bergmann and Kopp, 2010).

With regard to the coverbal use of gestures in natural language dialogues, an association has been observed between the function of a *kind* of gesture and certain of the *form features* of its *tokens* (Ladewig, 2007; Müller, 2004). These "morphological", i.e. feature-based, invariants induce so called *gesture families* or *gesture fields* (Kendon, 2004; Fricke et al., 2014). In this paper, we adopt a quantitative approach to detect correspondences between kinetic features and image-schematic expressions (Bressem, 2007). In contrast to related field-working studies, we experiment with a controlled vocabulary of image schema-related descriptor terms that are required in more advanced contexts of HCI and employ both a fine-grained feature-based annotation schema and a coarse-grained classification to detect gestures as candidate manifestations of the descriptor terms. This is done to detect "median" gestures that are commonly associated by users with certain descriptors: the stronger the association, the more reliably they are used in HCI, the lower the burden of users to learn an appropriate gesture lexicon. Since it is still unknown to which degree this association holds, our approach is a first effort in detecting them by means of an experimental, corpus-based approach.

The corpus of gestures that is build following the above-given rationale is called the **F**rankfurt **I**mage **G**est**URE** corpus (FIGURE). The corpus is derived from a user study, in which users were asked to spontaneously manifest image schema-related terms by means of hand and arm movements. We describe the study, the annotation of the gestures (Section 2.) and provide quantitative results describing them (Section 3.). The annotation manual and the annotation data will be made available *via* `hucompute.org` under the *creative commons share alike* license (CC BY SA).

FIGURE explores the not fully conventionalized space between spontaneous gesticulation and conventionalized languages. According to *Kendon's continuum* (McNeill, 1992), hand and arm movements can be ordered according to their increasing degree of conventionality along the following lines: *gesticulation* → *language-like gestures* → *pantomime* → *emblems* → *sign language*. HCI interaction drawing on manipulators and semaphores is mainly

[1] See `https://dev.windows.com/en-us/kinect` and `https://www.leapmotion.com/`, respectively.

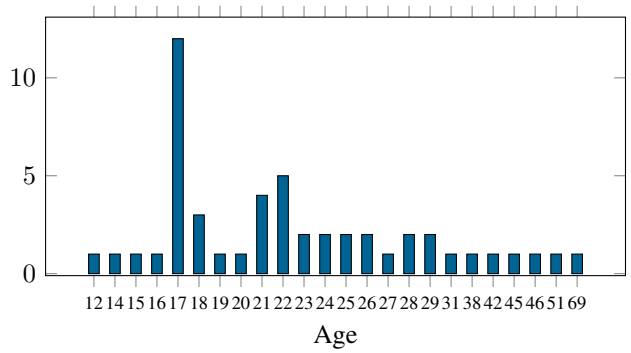

Figure 1: Age range of participants.

based on highly conventionalized emblems according to Kendon's continuum (sign language processing being a special topic in its own (Sáfár and Marshall, 2002)). Hence, FIGURE aspires to increase the expressive power of HCI gestures by exploring the rather unconventionalized realm of gesticulations and pantomime. However, the results reported below are also related to comparable findings from gesture studies in Section 4., connecting FIGURE to multimodal interaction research. But to start with, in Section 2., data gathering, data annotation and reliability assessment is described. Exploratory and descriptive facts and figures are subsequently presented in Section 3..

2. Data and Annotation

2.1. Data Collection

The study proceeds in an interview-like manner, where each subject is asked to depict a number of terms by gestures (i.e., hand and arm movements). We started with creating a catalog of 27 terms, consisting of the image-schema terms collected in Mehler et al. (2014) (e.g., *Equilibrium*, *Contact*), basic navigational terms (like *Left* and *Right*), and evaluative expressions (*Bad*/*Good*).[2] This list of terms has been subdivided into five groups, each comprises five to six words. 50 subjects took part in the study. Participants are from Germany, England, Russia, South Korea, the Kurdish area, China, Indonesia, Italy, Spain, Poland and the Netherlands. The predominant mother tongue was German with 66 %, the remaining 33 % are roughly equally distributed between languages of the above-given countries. Their age ranges from 12 to 69 years (see Figure 1 for a details). Right-handedness predominates in the study; only two subjects are left-handed. That is, the proportion of left-handed people in the study is 4 % and thereby at the lower bound of, but still in accordance with, what would be expected according to geographical variation (Llaurens et al., 2009, p. 882). Each subject depicts one of the five sets of stimulus terms, so that each term is depicted about 10 times.

2.2. Annotation

In order to get at kinetic features describing gestures, we developed an annotation scheme based on established gesture annotation formats. Basically, a gesture is represented as a feature vector between a starting position and an end position (Gibbon et al., 2003). The start- and endpoint of every gesture is annotated according to its location in gesture space (McNeill, 1992) and its handshape (ASL handshapes, extended by "thumbs up"). The trajectory of a dynamic gesture is captured in terms of its path (*line*, *arc*, *zigzag*, etc. – cf. (Bressem, 2007; Lausberg and Sloetjes, 2009)) – and its orientation (*away from body*, *up*, *left*, etc. (Lücking et al., 2010)). In order to account for pointing gestures, the value *pointing* has been added as a path.
Furthermore, the relations between the participants' hands are annotated as well as the extension of a gesture in terms of the hands' distance to the body. Eventual contact between fingers or fingers and arms both at the start and at the end of a gesture is also considered. The temporal relationship between movements of the left and the right hand can be explicitly annotated. Additionally, repetitions of movements are captured by counting. Finally, by annotating sequences of gestures we capture that people combine a number of consecutive gestures in order to depict a term.
Note that the kinetic feature-vector annotation is largely independent from the actual timing of a gesture. For this reason, the temporal features of a gesture are explicated on a separate annotation level called "phases" with respect to the video signal. Here, the canonic three-fold partition of a gesture into a *preparation*, a "stroke" and a *retraction* phase is annotated (Kendon, 1980; McNeill, 1992). Additionally, we delimited the time span from the presentation of a stimulus term to the onset of gestural movement in the preparation phase as *presentation phase*. The presentation phase gives a temporal cue to the processing difficulties with regard to the term in question.
Annotation has been carried out using ELAN.[3]

2.3. Reliability

In order to assess the reliability of the annotation scheme, three annotators independently annotated 12 experimental videos, containing 66 gestures in total. For each annotation level, a matching of annotation values is assessed in terms of both raw percentage agreement and chance-adjusted generalized Kappa coefficient (Fleiss, 1971). All calculations were carried out with the the R environment for statistical computing (R Core Team, 2013). The results are detailed in Table 1. In particular, the annotation of start and end locations of gestures within the gesture space (`Start.Pos`, `End.Pos`) turned out to be difficult, with percentage agreement values ranging from about 60 % to 80 %. This seems to be partly due to the underlying being designed for sitting subjects, while in our study participants were standing so that distance reference information like knee position is not applicable. However, the remaining annotations proved to be highly consistent, with percentage agreements ranging from over 80 % to about 98 %. Averaging over all 33 annotation layers, the three annotators reach a pairwise percentage agreement of 90.81 %, 91.64 % and 86.50 %. The mean Kappa coefficient for all raters on all layers is 0.84, meaning that the reliability of the annotation can be regarded as substantial (Krippendorff, 1980) or even "almost perfect" (Rietveld and van Hout, 1993).

[2] See Mehler et al. (2014) for further literature on image schemata.

[3] `https://tla.mpi.nl/tools/tla-tools/elan/`

Table 1: Overview of agreement results.

	%-Agree	Kappa
Gesture.Type	100.00	1.00
R.Start.HS	92.42	0.93
R.Start.Pos	43.94	0.57
R.Start.Palm	86.36	0.89
R.Start.BoH	86.36	0.88
R.Start.Dist	80.30	0.75
R.End.HS	90.91	0.92
R.End.Pos	53.03	0.64
R.End.Palm	81.82	0.84
R.End.BoH	87.88	0.90
R.End.Dist	81.82	0.74
R.Path	75.76	0.77
R.Orient	81.82	0.85
R.Move	98.48	0.96
R.Repeat	84.78	0.72
L.Start.HS	90.91	0.90
L.Start.Pos	66.67	0.69
L.Start.Palm	87.88	0.88
L.Start.BoH	87.88	0.89
L.Start.Dist	86.36	0.85
L.End.HS	86.36	0.86
L.End.Pos	66.67	0.69
L.End.Palm	86.36	0.88
L.End.BoH	90.91	0.92
L.End.Dist	86.36	0.84
L.Path	86.36	0.82
L.Orient	90.91	0.90
L.Move	98.48	0.98
L.Repeat	91.49	0.80
BH.Sym	50.00	0.54
BH.Start.Contact	69.70	0.42
BH.End.Contact	68.18	0.43
BH.Temp.Rel	98.48	0.98

Note that assessing agreement of the temporal markings of the gesture phase segmentation follows a quite different rationale than percentage agreement or Kappa statistics (Lücking et al., 2012). Since the segmentation agreement of the three phases involved (viz., preparation, stroke and retraction) has been evaluated by Lücking et al. (2013), we can safely expect the phase annotation to be "substantially better than what would be expected by accidental coincidences of segmentations" (p. 11).

3. Some Results

In the following subsections, descriptive figures of FIGURE with regard to HCI are reported. Statistical calculations and data plots (except for pie charts) were carried out with the the R environment for statistical computing (R Core Team, 2013) and the *Grammar of Graphics* library (Wickham, 2009).

3.1. Frequent Form Features

The subjects' gestures mainly describe four kinds of paths: we found 60.2 % straight line trajectories, 13.1 % pointing gestures, 11.9 % arc-shaped paths, and 10.3 % full curves – see Figure 2, where other paths include *Zigzag* or *U*-shaped trajectories.

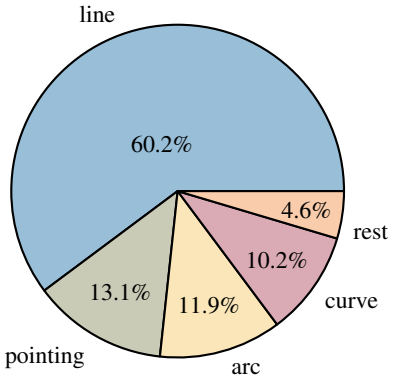

Figure 2: Most commonly used paths (percentages).

With regard to handshapes we get similar results. The mostly used handshapes at the beginning of a gestures were (using ASL notation) *5* (56.4 %), *d*, *c* (each 8.9 %) and the newly introduced *10* (*thumbs up*, 6.2 %).
The same handshapes are also the most frequent handshapes at the end of gestures, with the exception of a higher usage of a terminal *b* handshape. The most frequent "end handshapes" are: *5* (54.2 %), *d* (12.8 %), *c* (7.3 %), *10* (6.6 %) and *b* (4.6 %).
All in all, there is a trend to perform simple, geometrical paths and plain handshapes, which is manifested in the high usage of the path *line* and the open hand (*5*) as handshape.
The linear complexity of the gestures also tends to be low: sequences were an exception. Only 1.5 % of the gestures were *right-handed sequences* and 4.2 % were *complex sequences* (there is no gesture sequence performed by only the left hand (though there are single gestures)). The lion's share of FIGURE gesture (94.2 %) is made up of *simplex* gestures.[4]

3.2. Simplex Gesture Types

Each gesture from a non-sequence was classified according to the mutual relation of the hands. If only one hand was used, the gesture is said to be of type *simplex*, with a suffix indicating which hand was involved (i.e., *-lh* for *left hand* and *-rh* for right hand). In case of two-handed gestures, two possibilities obtain: firstly, both hands interact, often in a symmetric way, in order to depict a single entity; or secondly, each hand depicts a separate entity individually, which are related in a more complex scene depiction. The first case is called *complex-sym*, the second is called *complex-ind*. The most common gesture type was *complex-sym*, which comprised about half of all gestures. The proportions of gesture types are displayed in Figure 4.

3.3. Recurrent Forms

It turned out that most subjects tend to perform similar gestures as response for some stimulus terms. The most striking example are gesture produced to depict *Good/Like* and *Bad/Dislike*. 80 % of the subjects performed a *thumbs up* gesture (handshape 10) for *Good/Like* and even 100 % of the subjects used *thumbs down* for *Bad/Dislike*. These gestures are found to occur in one of two variations, distin-

[4]The "missing" 0.01 % is due to rounding of numbers.

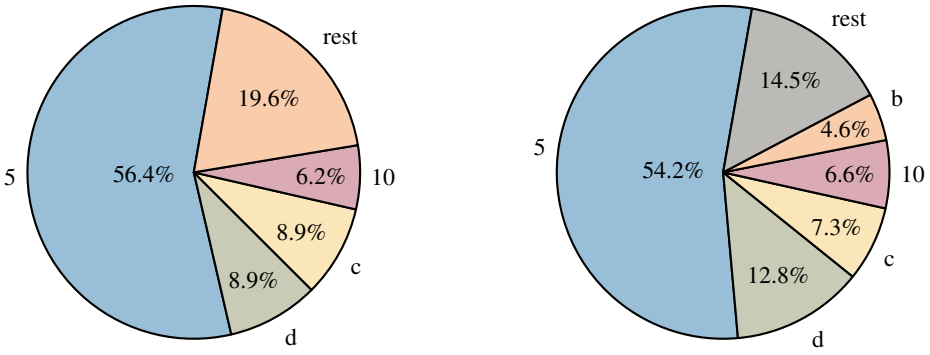

Figure 3: Most commonly used ASL handshapes (percentages): start position (left) and end position (right).

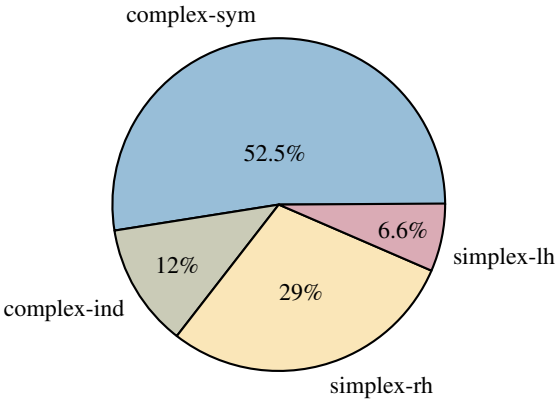

Figure 4: Proportion of gesture types for simplex gestures.

guished by using one hand or both hands. This stable result is presumable due to the emblematic nature of the *thumbs up/down* gestures.

In response to the keyword *Equilibrium*, 70 % of the subjects stretch out their arms horizontally by slightly swinging them. This depiction obviously is influenced by the image of an artist balancing on a high wire.

To our surprise, we found a common frequent gesture for *Source-Path-Goal*. This was insofar surprising as this term has been rated as one of the most difficult concepts in a pre-study evaluation of the keyword catalog among four raters. However, the subjects tend to move an invisible object through the space. There were two variants, a one-hand version and two-hand parallel one. The gesture presumably mimics the "drag and drop" action known from graphical user interfaces of operating systems.

As regards the gestures for the stimulus terms *Left* and *Right* we find that 60 % of the *Left*-gestures are performed with the left hand only, while 70 % of the *Right*-gestures were produced using the right hand. Since only 4 % of our subjects are left-handed, this result indicates a lateral bias of such "egocentric" terms.

3.4. Stroke Duration

The length of the stroke phase is a hint for the complexity of the gestural movement. Given the inclination to use simplex gestures when possible (see the findings reported in Section 3.2.), participants can also be assumed to prefer short gesture. In this line of thinking, longer stroke phases point to a greater extent of depiction demands on side of the participant. A respective visual data inspection, displayed in Figure 5, shows that *Part* is the most demanding notion in our set of stimulus terms, followed by *Rotation* and *Texture* (*Collection* and *Blockage* show larger standard deviations, but smaller means).

Whether such findings have to be explained in terms of intrinsic structure of notions or in terms of possible recency effects of presentation of the notions cannot be decided on the basis of such exploratory data. However, the data facilitates to generate a hypotheses concerning depiction strategies which can be tested in specifically designed experiments.

4. Discussion

We provided a data set consisting of "morphological" annotations of gestures, which in turn have been produced by naïve participants in response to a controlled set of stimulus terms mainly drawn from the cognitive paradigm of image schemata. Our results show that the depiction of stimulus terms does not rest on the handshape parameter in a crucial way. This an interesting difference to co-speech gestures from free speech, where it is found to be highly manifold (Bressem, 2007). Previous findings (Bressem, 2007) demonstrating the significance of paths have been corroborated. These results provide a number of conclusions for HCI, since they suggest that gestural interaction going beyond manipulator gestures relies mainly on the recognition of trajectories rather than on exact hand recognition. HCI gesture vocabularies tend to be oriented at static hand postures. Such static vocabularies should be extended by dynamic vocabularies formulated in terms of temporal movement pattern. The dynamics of gestures is also highlighted by differences in the mean durations of movements per stroke phase.

The unconventionalized space of non-emblematic movements seems to be at least partly governed by motor programs associated with certain concepts, as diagnosed in embodiment-approaches to gestures (Hostetter and Alibali, 2008). In this regard, the results reported for FIGURE provide a starting point for further hypothesis testing beyond HCI.

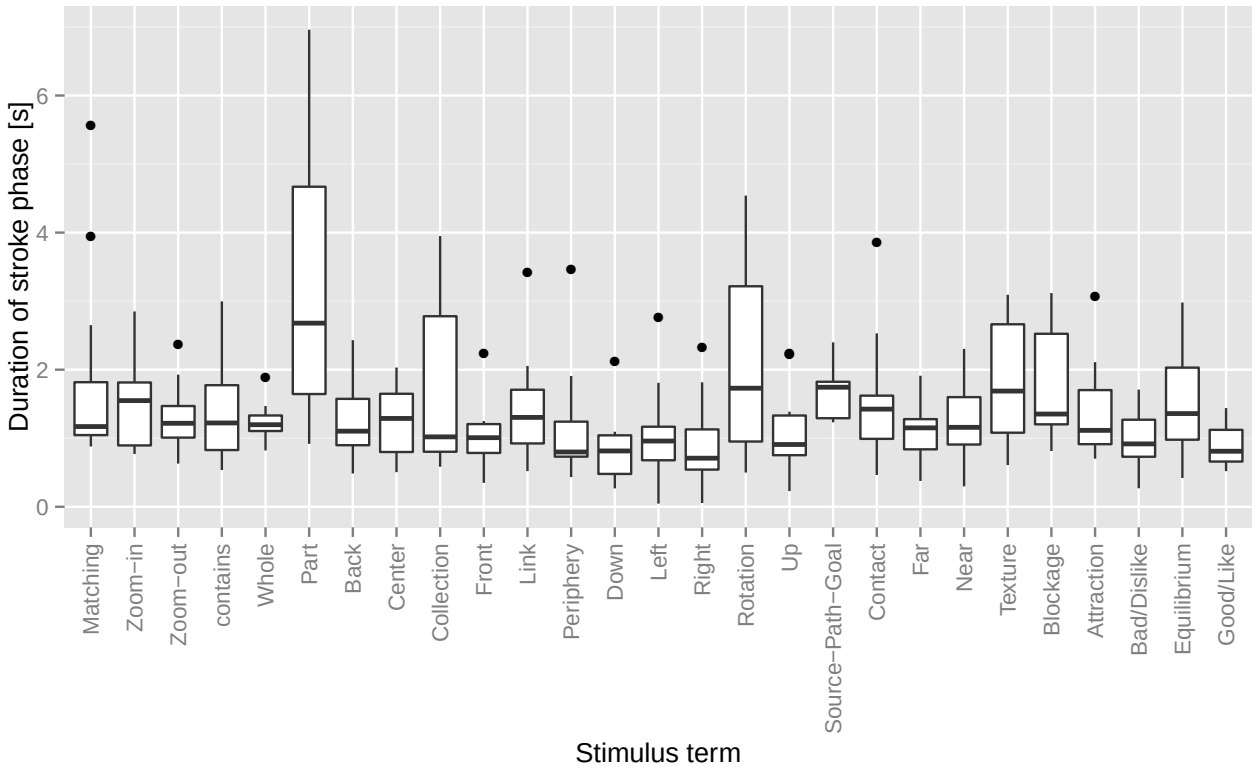

Figure 5: Mean duration of stroke phases for the stimulus terms.

5. Acknowledgements

We gratefully acknowledge the support of the *Frankfurter Goethe-Haus / Freies Deutsches Hochstift* in conducting the study. As representatives, we thank in particular Dr. Petra Maisak and Prof. Dr. Anne Bohnenkamp-Renken.

6. References

Bergmann, K. and Kopp, S. (2010). Systematicity and idiosyncrasy in iconic gesture use: Empirical analysis and computational modeling. In Stefan Kopp et al., editors, *Gesture in Embodied Communication and Human-Computer Interaction*, volume 5934 of *Lecture Notes in Computer Science*, pages 182–194. Springer, Berlin.

Brennan, S. E. (1998). The grounding problem in conversations with and through computers. In Susan R. Fussell et al., editors, *Social and cognitive psychological approaches to interpersonal communication*, chapter 9, pages 201–225. Lawrence Erlbaum, Hillsdale, NJ.

Bressem, J. (2007). Recurrent form features in coverbal gestures. `www.janabressem.de/Downloads/Bressem-recurrentformfeatures.pdf`.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Fricke, E., Bressem, J., and Müller, C. (2014). Gesture families and gestural fields. In *Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK)*, volume 2, chapter 123, pages 1630–1640. De Gruyter, Berlin and Boston.

Gibbon, D., Gut, U., Hell, B., Looks, K., Thies, A., and Trippel, T. (2003). A computational model of arm gestures in conversation. In *Proceedings of the 8th European Conference on Speech Communication and Technology*, EUROSPEECH 2003, pages 813–816.

Grandhi, S. A., Joue, G., and Mittelberg, I. (2011). Understanding naturalness and intuitiveness in gesture production: insights for touchless gestural interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 821–824.

Hostetter, A. B. and Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, 15(3):495–514.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In Mary Ritchie Key, editor, *The Relationship of Verbal and Nonverbal Communication*, volume 25 of *Contributions to the Sociology of Language*, pages 207–227. Mouton, The Hague.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge University Press.

Krippendorff, K. (1980). *Content Analysis*, volume 5 of *The SAGE KommTexT Series*. SAGE Publications.

Ladewig, S. (2007). The crank gesture – systematic variation of form and context. Talk given at the third congress of the ISGS, Chicago, Northwestern University, USA.

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Lausberg, H. and Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES–ELAN system. *Behavior*

Research Methods, 41(3):841–849.

Llaurens, V., Raymond, M., and Faurie, C. (2009). Why are some people left-handed? An evolutionary perspective. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1519):881–894.

Lücking, A., Ptock, S., and Bergmann, K. (2012). Assessing agreement on segmentations by means of *Staccato*, the *Segmentation Agreement Calculator according to Thomann*. In Eleni Efthimiou, et al., editors, *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, volume 7206 of *Lecture Notes in Artificial Intelligence*, pages 129–138. Springer, Berlin and Heidelberg.

Lücking, A., Bergman, K., Hahn, F., Kopp, S., and Rieser, H. (2013). Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2):5–18.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2010). The Bielefeld Speech and Gesture Alignment Corpus (SaGA). In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010, pages 92–98. 7th International Conference for Language Resources and Evaluation.

McNeill, D. (1992). *Hand and Mind–What Gestures Reveal about Thought*. Chicago University Press.

Mehler, A. and Lücking, A. (2012). WikiNect: Towards a gestural writing system for kinetic museum wikis. In *Proceedings of the International Workshop On User Experience in e-Learning and Augmented Technologies in Education*, UXeLATE 2012, pages 7–12, Nara, Japan.

Mehler, A., Lücking, A., and Abrami, G. (2014). WikiNect: Image schemata as a basis of gestural writing for kinetic museum wikis. *Universal Access in the Information Society*, pages 1–17.

Müller, C. (2004). Forms and uses of the Palm Up Open Hand: A case of a gesture family? In Cornelia Müller et al., editors, *The semantics and pragmatics of everyday gestures*, volume 9 of *Körper – Zeichen – Kultur*, pages 233–256. Weidler. Proceedings of the Berlin conference 1998.

Quek, F. K. H., McNeill, D., Bryll, R. K., Duncan, S., Ma, X.-F., Kirbas, C., McCullough, K. E., and Ansari, R. (2002). Multimodal human discourse: Gesture and speech. *ACM Transactions on Computer-Human Interaction*, 9(3):171–193.

R Core Team, (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rietveld, T. and van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour*. Mouton de Gruyter.

Sáfár, E. and Marshall, I. (2002). Sign language translation via DRT and HPSG. In *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science*, pages 58–68. Springer, Berlin and Heidelberg.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer, New York.