NEALT PROCEEDINGS SERIES

VOL. 6

Proceedings of the NODALIDA 2009 workshop

Multimodal Communication – from Human Behaviour to Computational Models

May 14, 2009

Odense, Denmark

Editors

Costanza Navarretta

Patrizia Paggio

Jens Allwood

Elisabeth Alsén

Yasuhiro Katagiri

NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

Proceedings of the NODALIDA 2009 workshop

Multimodal Communication – from Human Behaviour to Computational Models

NEALT Proceedings Series, Vol. 6

ISSN 1736-6305

Published by
Northern European Association for Language
Technology (NEALT)
http://omilia.uio.no/nealt

Electronically published at
Tartu University Library (Estonia)
http://dspace.utlib.ee/dspace/handle/10062/9208

Contents

Preface

Human communication is naturally multimodal, involving the interaction of modalities such as speech, facial expressions, hand gestures and body posture. In order to have a better understanding of human-human communication and to improve human-computer interaction it is essential to identify, describe, formalise and model the interaction of the different modalities in human communication. The past two decades have witnessed numerous initiatives and research efforts to improve the state of the art, including collection and annotation of multimodal corpora, automatic recognition of the different modalities and modelling and generation of multimodal data. However, there are still many questions and problems concerning the annotation of multimodal data and the technology for capturing such data, not to mention the interpretation and reproduction of complex, natural multimodal behaviour.

The main aim of the workshop has been to provide a multidisciplinary forum to present results and discuss issues that concern research on human multimodal communication, its modelling and representation for computational systems. The workshop call mentioned a large range of relevant topics: cognitive aspects of multimodal communication; formal frameworks and descriptions of multimodal communication; representational issues, e.g. definition of annotation units, granularity of descriptions, spatio-temporal models of non-verbal modalities, definition of default values, representation of multimodal meaning and inclusion of world context; interaction of the different modalities; multimodality in intercultural communication; definition of communicative functions in multimodal communication; methodologies and tools to annotate, process and produce multimodal communication; multimodal signal processing and its integration with manual annotation.

Each submitted paper was blindly reviewed by three reviewers. Seven papers were submitted, one paper was withdrawn and five papers were accepted to be presented at the workshop. The accepted papers cover several of the aspects of multimodal communication listed in the workshop call, including annotation, representation, analysis and processing issues.

The invited speaker Nick Campbell in his "Technology for Processing Non-verbal Information in Speech" discusses the importance for researchers of being able to automatically collect and process non-verbal information in speech. This is a prerequisite for using and integrating this information in more and more advanced commercial applications such as machine interpretation, games, robotics and customer services.

In the paper "Gestures that precede and accompany speech – An analysis of their functions and use in the design of virtual agents in different activities", Jens Allwood and Elisabeth Ahlsén discuss the behavioural and functional features of gestures produced before or simultaneously with speech and relate them to two activity types for Embodied Communicative Agents: front-end to database and educational training.

Christopher Habel and Cengiz Acarturk propose, in their paper "Eye-tracking evidence for multimodal language-graphics comprehension: The role of integrated conceptual representations", a modular architecture for the comprehension of textual and graphical input in which the conceptual representations coming from the two modalities' contribution are integrated. Experiments comparing eye-tracking movements caused by graphical input on the one hand and graphs and texts on the other are then described and discussed with respect to the proposed architecture.

Kristiina Jokinen and Minna Vanhasalo in "Stand-up Gestures – Annotation for Communication Management" discuss the form and functions in communication of so called stand-up gestures. The functions of these gestures comprise the regulation and coordination of communication, thus they are quite important in communication management. The authors also propose a way to annotate stand-up gestures via an extension of the MUMIN annotation scheme to include a meta-discursive context level.

"Coarticulation in sign and speech" by Stina Ojala, Tapio Salakoski and Olli Aaltonen presents a

study of coarticulation in Finnish sign language and speech with the main aim of discovering similarities between coarticulation in signing movements and speech. The study's results indicate that the alternations of deceleration and acceleration in signing movements can be compared to deceleration and acceleration patterns which occur at different levels in speech.

The paper "Integration and representation issues in the annotation of multimodal data" by Patrizia Paggio and Costanza Navarretta deals with issues related to the representation of gestures and speech in a multimodal sign in terms of feature structures in a unification-based grammar. The authors also discuss some of the complexities related to the interpretation of the multimodal sign, such as the interaction of gestures and speech at different conceptual levels and their multi-functionality.

The Organizing Committee

Costanza Navarretta, University of Copenhagen

Patrizia Paggio, University of Copenhagen

Jens Allwood, University of Gothenburg

Elisabeth Alsén, University of Gothenburg

Yasuhiro Katagiri, Future University, Hakodate

Acknowledgements

We want to thank the program committee for reviewing the workshop papers:

Nick Campbell, ATR, Osaka

Loredana Cerrato, Acapela Group Sweden

Dirk Heylen, University of Twente

Kristiina Jokinen, University of Helsinki and University of Tartu

Michael Kipp, DFKI Germany

Brian MacWhinney, Carnegie Mellon University

Jean-Claude Martin, CNRS-LIMSI France

Catherine Pelachaud, University of Paris 8

Isabella Poggi, Roma Tre University

Andrei Popescu-Belis, Idiap Research Institute

Rainer Stiefelhagen, Karlsruhe University

Johannes Wagner, University of Southern Denmark

Massimo Zancanaro, Bruno Kessler Foundation, Trento

Technology for Processing Non-verbal Information in Speech

Current speech technology is founded upon text. People don't speak text, so there is often a mismatch between the expectations of the system and the performance of its users. Talk in social interaction *of course* involves the exchange of propositional content (which *can* be expressed through text) but it also involves social networking and the expression of interpersonal relationships, as well as displays of emotion, affect, interest, etc. A computer-based system that processes human speech, whether an information-providing service, a translation device, part of a robot, or entertainment system, must not only be able to process the text of that speech, but must also be able to interpret the underlying intentions, or *acts*, of the speaker who produced it. It is not enough for a machine just to know *what* a person is saying; it must also know *what that person is doing* with each utterance as part of an interactive discourse.

Tone of voice

Previous work carried out in Japan has shown that more than half of interactive speech in everyday conversations takes the form of nonverbal utterances which cannot adequately be transcribed into text. These stylised utterances as well as non-lexical *affective* speech sounds, such as laughs, feedback noises, and grunts, also carry important interpersonal information related to the states, intentions, and beliefs of the discourse participants, and to the progress of the social interaction as a whole. They constitute a small finite set of highly variable sounds in which most of the information is carried by prosody and tone-of-voice. It is this component of speech especially that makes it such a rich and expressive medium for human interaction, but this is an element of the signal that is not yet well modelled, if at all, by machine processing.

A human interlocutor intuitively interprets the nonverbal information in speech and tone-of-voice to aid in the interpretation of each utterance in context. It has been shown, for Japanese, that a machine can be programmed to perform similar interpretation of speech utterances, and currently research is being carried out to generalise and further develop these findings using speech data from other languages. While the academic goal of such research is to show that the use of nonverbal utterances in conversation is a characteristic of human speech *in general* and not limited to only one particular culture or language, the technical goal of the work is to produce devices that are *specifically adapted to interactive or conversational speech* that will enable a friendlier and more efficient speech interface for public services and entertainment.

Recognising that *social actions* are the essential component of intercourse, and that actions, rather than words are the prime units to be processed in a discourse, future speech research must specifically address the question of how new technologies can be produced which are capable of processing not only the lexical content of an utterance, but also its underlying intentions. This might be done by processing prosody & tone-of-voice.

To further the development of such speech technology, it is therefore essential to collect a representative corpus of spoken interactions wherein participants display the *full range of their daily speech strategies* and to use that material to train new modules for interactive speech processing (whether for synthesis or recognition) that can make use of such higher-level information. However, such a corpus requires the prior development of recording techniques that are unobtrusive, and environments which are felicitous.

Discourse dynamics

There is growing international interest in multimodal interaction processing (see e.g., UC *(Universal Communication)* in Japan, AMI *(Augmented Multimodal Interaction)* in Europe, and CHIL *(Computers in the Human Interaction Loop)* in the US) and in the collection of multimodal conversational speech data, which was identified as a principal future task at the LREC (Language Resources and Evaluation Conference) last year.

Whereas traditional approaches to spoken interaction and dialogue systems have tended to assume a “ping-pong” or “push-to-talk” model, wherein either the system or the interlocuting human is active at any given time, it is becoming increasingly apparent that the dynamics of spoken interaction is an important element in itself for speech information processing, and that the typical flow of speech is fragmented and multi-faceted, rather than forming a single uninterrupted stream. This is supported by many recent findings in conversation and discourse analysis, where the definition of a “speech-turn”, or even an “utterance” is proving to be very complex.

People apparently don’t “take turns” to talk in a typical conversational interaction; rather they each contribute actively *and interactively* to the joint emergence of a “common understanding”. The apparent "no gap no overlap" alternation of spoken utterances is actually emergent from a background of continuous behavioural coordination at different levels of behavioural organization. This *interaction synchrony* is a feature yet to be incorporated in modular speech processing technology and might prove to be an important element for dialogue interface design. It should therefore be taken into consideration as a key component of corpus design.

Corpus control

Speech data will continue to be collected from a variety of sources using a variety of capture devices. Techniques will be developed to deal robustly with impoverished or "less-than-perfect" materials, and a corresponding robustness will be reflected in the technology produced as a result. Conversely, in order to derive useful and reliable components for speech information processing, we should ensure that the corpora we collect are representative of the styles and mannerisms of interactive conversational speech, so that future users of this technology will be presented with interface designs that match their (unconscious) expectations and that are able to process the full range of information that is carried by inflections of the voice and from the characteristics of timing and turn-taking.

Conclusion

As we envisage the incorporation of speech processing modules in more and more sophisticated commercial applications, including machine interpretation, robotics, games, and customer-services, a key element of the research will be to develop methods that enable the efficient collection of conversational and interactive speech data without the need for extensive or invasive recordings. Privacy considerations may prevent the use of naturally-occuring samples, so this work may require the development of both capture devices (cameras and recorders) and capture environments (equivalent to a recording studio) that encourage participants ro relax informally and maximise their range of speaking styles and formats.

Nick Campbell Trinity College Dublin, February 2009

Gestures that precede and accompany speech – An analysis of functions and applicability for virtual agents in different activities

Jens Allwood
University of Gothenburg
Gothenburg, Sweden
jens@ling.gu.se

Elisabeth Ahlsén
University of Gothenburg
Gothenburg, Sweden
eliza@ling.gu.se

Abstract

This paper contains an analysis of features of gesture types that are produced before or simultaneously with speech (mainly. nouns and verbs) and in relation to own communication management (choice and change). The types of gestures discussed are arm-hand gestures, head movements and gaze. The analysis is then discussed in relation to two selected social activities, where virtual agents (ECAs) are or can be used. Gesture types and features with different functions are briefly suggested for each of the two activities and also more in general. The analysis is meant provide information about naturally occurring gestures that can serve as a basis for assigning gestural functions to ECAs.

1 Introduction/Background

Human face-to-face interaction is very much characterized by being multimodal. The analysis of spoken interaction and gesture, taking into account also typical interactive features, not traditionally analyzed for written language has been pursued for a couple of decades and has given us more insight into how human-human interactions works on-line. There are, however, still many phenomena related to interactive functions that are not sufficiently studied and understood. Such phenomena include 1) Interactive Communication Management (ICM), i.e. turn taking, feedback and sequences and 2) Own Communication Management (OCM), i.e. choice and change features in speech related to planning and production processes, for example hesitations and self-repeats (Allwood, 2002).

Turning to gestures in spoken interaction and here focusing on arm-hand gestures, head movements and facial expressions, six main types of content of communication have been suggested. A list of what can be conveyed by gestures in face-to-face interaction (Allwood 2007) is the following:

1. Identity: who a communicating person is biologically (e.g. sex and age), psychologically (e.g. character traits such as introvert or extrovert) or socioculturally (e.g. ethnic/cultural background, social class, education, region or role in an activity).

2. Physiological states: e.g., fatigue, illness, fitness etc

3. Emotions and attitudes: expressed continuously with respect to topic, persons etc.

4. Own communication management: gaining time to reflect, plan or concentrate, having difficulties finding a word (cf. Ahlsén, 1985; Ahlsén, 1991) or needing to change what we have said (cf. also Allwood, Nivre & Ahlsén, 1990).

5. Interactive communication management: to regulate turntaking (cf. Duncan & Fiske, 1977; and Sacks, Schegloff & Jefferson, 1975), feedback to show whether we want to continue, whether we have perceived and understood and how we react to the message (cf. Allwood, 1987; and Allwood, Nivre & Ahlsén, 1992).

6. Factual information: especially words and illustrating or emblematic gestures

The role of gestures as enhancing perception and memorization of verbal messages has been

demonstrated by Beattie (2005, 2007). Temporally, gestures can be either preceding, simultaneous with or succeeding the corresponding speech. The effects demonstrated by Beattie can, in principle, be achieved by either of this temporal relations. In making communication more efficient, however, gestures preceding speech or accompanying speech are of special interests. A special case is when a gesture replaces a word or a phrase, which is not spoken.

The relation of the content and function of what is spoken and what is conveyed through gestures can be of different types. Analyzing semantic and semiotic features of what speech and gestures convey can shed more light on how speech and gesture co-contribute to the message. The time relation between gesture and speech with respect to the same target message can also provide clues about the unfolding of the production process.

Embodied Communicative Agents (ECAs) are increasingly being introduced in a number of ICT applications, such as front-ends to databases providing various types of information, pedagogical tools and simulation tools for training. The ECAs used in these applications are often very simple with few communicative functions and very limited variation in means of expressions (e.g. three facial expressions), but there are also advanced ECAs designed, at least partially, for purposes of Artificial Intelligence, i.e. in order to simulate and thereby better understand human interaction and/or to make the ECA appear as human-like as possible, by using a number of salient features from human interaction. Some examples of this are experimenting with the generation of eye gaze and smoothness of gestures (Kipp & Gebhard, 2008, Neff et al. 2008), using gesture dictionaries (Poggi et al., 2005), designing production models for iconic gestures (Kopp, Bergmann et al., 2008), providing models for feedback giving (Kopp, Allwood et al, 2008), studying reactions to behaviors increasing intersubjectivity (Cassell and Tartaro, 2007) and to social versus task only interaction style (Bickmore & Cassell, 2005), comparing direction giving by ECAs and humans (Cassell et al., 2007), evaluating culturally dependent features and intercultural communication (Allwood & Ahlsén, forthc). and creating affective behavior (e.g. Strauss & Kipp, 2008).

This paper is an attempt to, by analyzing human-human communication, focus on a number of features of multimodal communication and discuss gestures with different features in relation ECAs in general and for two different activity types. The paper focuses on two of the main categories of what can be conveyed by gestures, factual content (FC) and own communication management OCM).

2 Method

The analysis was based on a sample of 100 occurrences of gestures preceding or accompanying words, mainly nouns and verbs, in videorecorded spoken face-to-face interaction dyads. 60 of the gestures were primarily identified as illustrating factual content of nouns and verbs, whereas 40 gestures were primarily identified as occurring with own communication management (OCM), i.e. choice and change behavior.

The gestures were coded according to the following features:

- Time: beginning and end

- Target: target word, target word category (mainly for the 60 factual information gestures)

- Contributions: preceding contribution, speech, gesture

- Timing of gesture stroke in relation to spoken contribution/target word: before, same, after (for FC gestures in relation to target word, for OCM gestures in relation to vocal-verbal OCM and target word where a target word can be identified)

- Representational features:

  Description of preparation, prestroke hold, stroke, poststroke hold, retraction

  Gesture form: body part, direction of movement, hand shape

  Complexity: two hands, finger movements, change of hand shape other than fist or open hand shape

- Semantic features of gesture: shape, location in relation to body, functional arm and finger movement, functional hand shape, movement of an object, illustrating action

- Information of gesture in relation to speech: same, added (earlier, more content)

- Gaze direction

- Choice or change function (for OCM gestures)

The features applicable to each of the gestures were coded and used as a basis for the analysis, together with the video.

Two activity types typical for ECA:s were then selected for discussion of types and features of gestures:

- Front-end to database
- Education-training

3 Results

The target word types for factual information gestures are presented in table 1.

	Target word type	
	Noun	**Verb**
Factual information gestures	42%	58%

Table 1. Target word type

There are a more factual information gestures accompanying verbs than nouns in natural spoken interaction.

	Timing	
	Preceding word	**Simultaneous with word**
Factual information gestures	30%	70%

Table 2. Timing of gestures (temporal relation between gesture and target word)

Most of the factual information gestures are produced simultaneously with the target word, as can be seen in table 2, but still a substantial part of them start and have the peak of their strokes before the target word is produced. This is of special interest with respect to the planning and production process as well as for the perception and comprehension process in the interlocutor.

Table 3 presents how much different body parts are used in gestures in the data.

	Body parts		
	Hands		**Head**
	1	2	
Factual information gestures	50%	48%	2%
OCM gestures	0%	88%	12%

Table 3:Body parts used in gestures

Comparing gestures that are mainly arm-hand-finger movements of one hand, two hands or movement of the head, differences are found in distribution between gestures used mainly with factual information and gestures used mainly for own communication management. Gestures used for factual information are fairly evenly distributed between the use of both hands and the use of only one hand, with only very few head movements, Gestures for own communication management, on the other hand are almost always made with one hand only, practically never with two hands, but not infrequently with head movement or gaze.

This indicates that the gestures used with own communication management are most often of a different type than illustrating gestures used with factual information. Iconic gestures which use only one hand are sometimes considered as less complex than if two hands are used. In the case of OCM gestures, this can be one interpretation, while they might also be more fundamentally different, in the typical cases. There is, however, also a considerable overlap and possibly a continuous scale between more representational gestures occurring with nouns and verbs and OCM gestures occurring with communication management. Verbal-vocal OCM as well as OCM gestures often also occur in the context of verb and noun production, when searching for and trying to produce the right noun or verb.
In table 4, a further clue to the planning and production process, i.e. the gaze direction of the speaker during the production of gesture is shown.

	Gaze	
	At interlocutor	**Up/Down/Out**
Factual information gestures	90%	10%
Own Comm.. Management gestures	50%	50%

Table 4. Gaze direction during gesture

Also for gaze direction during gesture, there is a substantial difference between the two types of gestures. During factual information gestures, gaze is almost always directed towards the interlocutor, whereas with own communication management gestures, there is an even distribution of gaze between looking at the interlocutor and "looking away" (up, down, in front of you or at an object or one's own hands).

This also points to illustrating gestures used with nouns and verbs perhaps being used more deliberately in order to enhance the listener's comprehension, by showing/specifying form, size, action, location etc. This does not seem to take so much effort in planning that the gaze has to be averted. With OCM gestures, on the other hand, there is generally a problem of choice or change of verbal-vocal production, which calls for effort and more often requires gaze aversion. In this case, both gaze aversion and gesture indicate planning problems.

For OCM gestures, choice and change functions are distributed as follows (table 5).

	OCM function	
	Choice	**Change**
Own Comm.. Management gestures	82%	18%

Table 5. Own communication management: choice vs. change function

Choice OCM is much more common, both in speech and gesture, than change OCM. This could, however, vary with both individual speaker type and activity type. It is also the case that about 40% of all speech based choice related OCM involves gestures, whereas only 15% of speech based change related OCM is accompanied by gestures (Allwood et al., 2002).

What is, then, the content and function of the factual information gestures? In table 6, the semantic features are presented ranked according to frequency of occurrence.

Semantic features of factual information gestures	
Illustrating action	53%
Illustrating shape	33%
Illustrating location	10%
Functional hand movement	3%
Functional hand shape	1%

Table 6. Semantic features of gestures used for factual information

In the 60 gestures used for factual information with nouns and verbs, 72 semantic features were coded and among these 72 features the distribution was, as shown in table 6, that illustrating an action was the most common feature, followed by shape and location (on the body or in the room). This is consistent with more gestures occurring with verbs than with nouns, although the difference between gestures illustrating actions and gestures illustrating shape is somewhat greater than that between gestures with verbs and gestures with nouns. Gestures illustrating action are also sometimes used to illustrate the meaning of nouns and gestures illustrating shape can be used also with verbs.

Turning to OCM gestures, about 20-25%, according to Allwood & Ahlsén (2002), are illustrating content in a similar way to that of factual content gestures. The rest have more general functions having to do also with self-activation and interaction regulation, especially turn keeping.

4 Discussion

A summary of findings is that
- factual information gestures are used more with verbs than with nouns
- they are most often simultaneous with the noun or verb, but in 30% of the cases precede the target word
- there is an even distribution between use of one and two hands in factual information gestures, but only use of one hand gestures and to some extent head movements (10%) in OCM gestures
- gaze is practically always directed at the interlocutor when factual information gestures are produced, but evenly distributed between gaze at the interlocutor and gaze directed elsewhere with

OCM gestures (gaze aversion could in itself also be considered an OCM gesture).
- more than 80% of the OCM gestures have choice function, rather than change function
- the most frequent semantic feature of factual information gestures is illustration of action, followed by illustrating of shape and location.

This overview is based on a limited sample of data, but can be compared also to earlier studies (e.g. Allwood, Ahlsén et al. 2002) and it gives a general idea of how the two types of gestures are used.

Turning to web-ECAs, the use of both types of gesture (FC and OCM) can be useful in many contexts and the findings reported in this study can be applied more or less directly in the design of ECAs. It is, however, no trivial task to implement gestures of these types in an ECA, so that they will (i) occur with the right context and timing, (ii) be chosen correctly, and (iii) be produced in a way that looks more natural than disturbing. Especially the factual content gestures produced with verbs and nouns are most often quite specific and require either (i) that the ECA has a fixed repertoire of verbal-vocal and gestural output for a specific task, modeled on what a human produces in the same task, i.e. a form of copying of sequences and combinations in context or (ii) that extensive dictionaries of gesture-word correspondences are available and are culturally adapted and can be linked to specific words in production. Considering the OCM gestures, the same is true to some extent, since they often occur when a person has problems producing the right words and then also in a fairly specific way illustrate the content of the intended word or phrase. There are, however, a number of other typical features of OCM gestures, such as discrete pointing to oneself when referring to oneself, to the interlocutor when referring to the interlocutor, pointing to one's head or mouth when referring to own memory or production problems, moving hand in a certain direction in relation to movement verbs and also more metaphorically in relation to more abstract words (forward for words indicating traveling, walking, running, biking etc as well as progress and reference to future; to the side for throwing away, canceling etc; to the back for past time, leaving behind etc.). These types of gestures contain factual content and can also be found with nouns and verbs.

Looking at our two exemplifying activity contexts for an ECA, both types of gesture can be useful in both activity types.

1. The more general OCM gesture types can be used when there is unclarity, "hesitation", time needed for processing, need for change and problems of understanding.
2. The more specific types of gestures illustrating the content of verbs and nouns (most often actions and shapes or locations) can be exploited in enhancing the salience and clarity of spoken (or written) output.

In both cases, timing is essential, as well as gestures that appear fairly natural in the context.

For an ECA as front end to a database, gestures illustrating the content of frequently used words could be included linked to the words, e.g. illustrating the shape of a paper, form, ticket etc., index finger tracing line for reading, writing movement for writing, typing movements for entering data via the computer, hand-to-ear movement for phoning, driving movement (holding driving wheel) for driving, stop sign for stopping etc. Pointing to clock for opening hours, gestures illustrating packing, sending, picking up etc. as well as many kinds of directive pointing can also be useful.

An ECA in an education interface should have specific gestures adapted to what type of education it is used for. Here, pre-prepared sequences of actions for specific procedures can be used, that are specifically designed and related to the words (e.g. nouns and verbs) included, e.g. for learning to make something (practical-procedural education). For more theoretical education, pointing, showing and giving directions by gesture in combination with reference to pictures could be used, also here possibly with gesture-word links from a dictionary.

These are just exemplifications of how the types of gestures in this study can be used in ECAs. IN general, it can be concluded that most of the gestures are fairly specifically linked to specific content words and fairly hard to implement in a natural way, except for pre-prepared and human-based "scenarios" or "sequences" of an ECA. The use of gesture dictionaries is cumbersome and still needs considerable work. For the more general types of OCM gestures, it should, however be much easier to implement them in ECAs in general and they could potentially add to naturalness in the appearance and interaction of ECAs, especially in problematic sequences.

Since gestures are known to enhance perception and memory processes by adding redundancy but also by specifying semantic features, multimodal presentation is a worthwhile enterprise, even though it is fairly complex, as for most of the gestures of this study. The study has presented some of the features to be considered for gestures used for factual content and own communication management.

References

Ahlsén, E. 1985. Discourse Patterns in Aphasia. *Gothenburg Monographs in Linguistics*, 5. University of Gothenburg, Department of Linguistics.

Ahlsén, E. 1991. Body Communication and Speech in a Wemicke's Aphasic - A Longitudinal Study. *Journal of Communication Disorders*, 24:1-12.

Allwood, J. 2002. Bodily communication – dimensions of expression and content. B. Granström, D. House & I. Karlsson (eds) *Multimodality in Language and Speech Systems.* Kluwer, Dordrecht.

Allwood, J. & Ahlsén, E. Multimodal intercultural Information and Communication Technology – A conceptual framework for designing and evaluating Multimodal Intercultural Communicators. (Fortchoming in M. Kipp, J.-C. Martin, P. Paggio & D. Heylen (eds) *Multimodal Corpora*. Springer Verlag.)

Allwood, J., Ahlsén, E. ; Lund, J. et al. 2007. Multimodality in own communication management.. *Current Trends in Research on Spoken Language in the Nordic Countries*. II, 10-19.

Allwood, J., Nivre, J. & Ahlsén, E. 1990. Speech Management: On the Non-Written Life of Speech. *Nordic Journal of Linguistics*, 13:3-48.

Allwood, J., Nivre, J. & Ahlsén, E. 1992. On the Semantics and Pragmatics of Linguistic Feedback. *The Journal of Semantics*, 9.1.

Beattie, G. 2005. Why the spontaneous images created by the hands during talk can help making TV advertismements more effective. *British Journal of Psychology*, 97:21-37.

Beattie, G. 2007. The role of iconic gestures in semantic communication and its theoretical and practical applications. In Duncan, S., Cassell, J. &Levy, E. (eds.) *Gesture and the Dynamic Dimensions of Language*, pp. 221-241.

Bickmore, T., Cassell, J. .2005. Social Dialogue with Embodied Conversational Agents. In van Kuppevelt, J., Dybkjaer, L. & Bernsen, N. (eds.), *Advances in Natural, Multimodal Dialogue Systems.* New York: Kluwer Academic.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K. & Striegnitz, K . 2007. Trading Spaces: How Humans and Humanoids use Speech and Gesture to Give Directions. In Nishida, T. (ed) *Conversational Informatics.* New York: John Wiley & Sons, pp. 133-160

Cassell, J., Thórisson, K.: (1999). The power of a nod and a glance. Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13:519-538.

Duncan, S. and Fiske, D. 1977. *Face-to-Face Interaction.* Lawrence Erlbaum, Hillsdale. N.J.

Kipp, M. & Gebhard, P. 2008. IGaze: Studying reactive gaze behavior in semi-immersive human-avatar interactions. In *Proceedings of the 8th International Conference on Intelligent Virtual Agents (IVA-08), LNAI 5208*, Springer, pp. 191-199.

Kopp, S., Allwood, J., Ahlsén, E. et al. 2008. Modeling Embodied Feedback with Virtual Humans. In: Springer series Lecture Notes in Computer Science (LNBCS) subseries Lecture Notes in Artificial Intelligence (LNAI). I. Wachsmuth & G. Knoblich (Eds.) *Modeling Communication with Robots and Virtual Humans*, LNAI 4930. p. 18-37, Springer, Berlin.

Kopp,S., Bergmann, K. & Wachsmuth, I. 2008. Multimodal communication from multimodal thinking - Towards an integrated model of speech and gesture production. *Int. Journal Semantic Computing* 2(1):115-136.

Neff, M., Kipp, M., Albrecht, I. and Seidel, H.-P. 2008. Gesture Modeling and Animation Based on a Probabilistic Recreation of Speaker Style. In *ACM Transactions on Graphics* 27 (1), ACM Press, pp. 1-24.

Poggi, I., Pelachaud, C., de Rosis, F., Carofiglio, V., De Carolis, N.: 2005. GRETA. A Believable Embodied Conversational Agent. In Stock, O., Zancarano, M. (eds.) *Multimodal Intelligent Information Presentation* Kluwer, Dordrecht.

Sacks, H-, Schegloff, E. & Jeffersson, G. 1974. A Simplest Systematics for the Organization of Turntaking in Conversation. *Language*, 50:696-735.

Strauss, M. & Kipp, M. 2008. ERIC: A Generic Rule-based Framework for an Affective Embodied Commentary Agent. In *Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*

Eye-tracking evidence for multimodal language-graphics comprehension: The role of integrated conceptual representations

Christopher Habel
University of Hamburg
Hamburg, Germany
habel@informatik.uni-hamburg.de

Cengiz Acarturk
University of Hamburg
Hamburg, Germany
acarturk@informatik.uni-hamburg.de

Abstract

In this paper, we propose a computational architecture for multimodal comprehension of text and graphics. A theoretical account of the integrated conceptual structures induced by linguistic and graphical entities is presented. We exemplify these structures with the analysis of an excerpt from a report published by Point Reyes Bird Observatory (PRBO). Experimental evidence, based on the analyses of subject's eye movement recordings was evaluated under the framework of the architecture.

1 Introduction

Multimodal communication combining language and graphics is a successful means to convey information: it includes *persistent* documents, such as newspaper articles, educational material and scientific papers in print media or in electronic media as well as *transient* oral presentations using power point or chalk-and-blackboard lectures.[1] Humans seem to integrate information provided by different modalities—as language and graphics—almost always based on unconscious cognitive processes. Whereas researchers from different disciplines investigated multimodal documents of different types in different domains, research on cognitive mechanisms underlying multimodal integration is currently in a less mature state and detailed computation models of language-graphic comprehension are rare.

The focus of the present study is multimodal comprehension of expository text accompanied by graphics of a specific type, namely line graphs of functions with *time*-arguments and numbers as values. Figure 1 shows an excerpt from a waterbird census report[2], which contained verbal information about the number of birds (1).

(1) Bolinas Lagoon Population Trends

> From a peak of about 60 wintering birds in 1976, numbers have declined to about 20 birds currently.

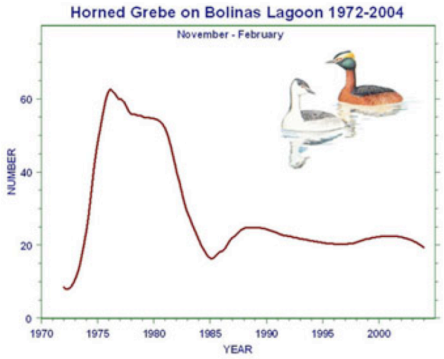

Figure 1. Trend graph depicting the number of wintering birds.

From a linguistic point of view, the process of *referring*, which is constituted by a *referential expression*, as 'peak of about 60', that refers to an *entity* of the domain of discourse, that can contain also abstract entities, as *numbers*, is the core of comprehension. Based on this, *co-reference*, the backbone of text coherence has to be established by speaker and hearer employing internal conceptual representations, which mediate between language and the domain of discourse. In processing text-graphics documents, in which both modalities contribute to a common conceptual representation, additional types of reference and co-reference relations have to be distinguished. Foremost, there exist correspond-

[1] In this paper, we use the term 'modality' as shorthand for 'representational modality'.

[2] "Waterbird Census at Bolinas Lagoon, Marin County, CA" by Wetlands Ecology Division, Point Reyes Bird Observatory (PRBO) Conservation Science: (http://www.prbo.org/cms/index.php, retrieved on 14 April 2009).

ing referential relations (reference links) between graphical entities and entities in the domain of discourse. Furthermore, there exist referential links between linguistic and graphical entities. To sum up, a layer of common conceptual representations is the place where *co-reference links* among conceptual entities introduced by various modalities are constructed where *inter-* and *intra-representational coherence* is established (Seufert, 2003).

A systematic investigation of multimodal comprehension of graph-text documents needs specification of referential link constructions between different representational formats, namely language and graphics. A graph-text document, either in printed or electronic media, is an external representation that includes graphical entities and textual entities.

The purpose of the present paper is to propose a computational model of integrated comprehension of language and graphics based on conceptual representations, which play the crucial role in interfacing between modalities (Jackendoff, 2007). The model is supported by experimental studies using eye-tracking methodology.

2 Integrated Comprehension of Language and Graphics

2.1 Comprehending Language and Comprehending Graphics

Language comprehension, in its most basic form, includes a set of processes that transforms *external* linguistic representations, such as words, phrases, sentences, into internal mental representations, in particular into conceptual structures and spatial representations (Jackendoff, 1996). Comprehension includes phonological, syntactic and semantic processes, which are governed by a set of rules and constraints, often called grammar, and processes of memory retrieval and reasoning to employ knowledge about the world. Furthermore, during the last two decades psycholinguistics has intensively investigated the interaction of—in particular, spoken—language comprehension and visual perception (Ferreira & Tanenhaus, 2007) giving clear evidence that concurrent perception can affect the interpretation of discourse. The 'language module' depicted in Figure 2 is based on Tschander et al. (2003); their approach focuses on 'verbally instructed navigation', i.e., on a language comprehension task, in which processing of spatial language and spatial knowledge is essential (details of the conceptual representation language presented in this paper is discussed in section 2.3). Therefore specific components to process spatial concepts and to match spatial representations with (idealized) visual percepts are foregrounded in their approach.

Comprehension of graphs, in a similar way to language comprehension, can be seen as a set of processes that transform *external* representations, namely graphics, consisting of axes, tick marks, graph lines, etc., into internal conceptual and spatial representations. Graphs, unlike pictorial representations and iconic diagrams, have grammatical structures. Thus graph comprehension involves—particularly in comprehension of statistical information graphics such as line graphs—perceptual, syntactic and semantic processes (Kosslyn, 1989).

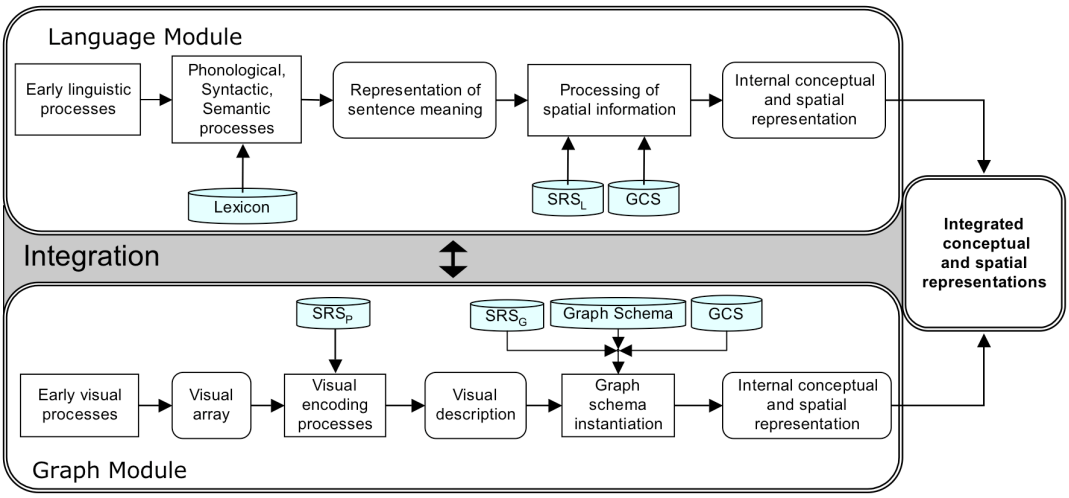

Figure 2. The three basic components of the information flow architecture.

The 'graph module' depicted in Figure 2 is an adaptation of Pinker's (1990) graph comprehension architecture. It transforms the information induced by external graphical representations, such as shape and position of graph line segments, into *visual array* and then into *visual description* by employing visual encoding processes (c.f. *visual routines*, Ullman, 1984). Visual description represents information about relative spatial positions of graphical entities (e.g., horizontal and vertical lines as well as segmented graph lines) and textual entities (e.g., axis labels, value labels). Visual description is then transformed into internal conceptual and spatial representations via instantiation of *graph schemata*. The graph schema is a long-term memory structure that includes information for specifications of gestalt atoms in graphs. For example, for a line graph, these gestalt atoms are the diagonal lines ' / ' and ' \ ' leading to INCREASE and DECREASE concepts (see section 2.3). It is the graph schema that makes possible to process perceptual information provided by the lines on paper or on screen as entities belonging to a line graph. Whereas visual encoding corresponds to the phonological, morphological and syntactic stages of language comprehension, graph schema instantiation corresponds to the semantic and pragmatic stages.

2.2 Multimodal comprehension: Integration

Multimodal comprehension of a text-graphics document requires the integration of information contributed by both representational modalities, namely language and graphs, or in other words, the interaction between the language comprehension module(s) and the graph comprehension module(s) (cf. Schnotz, 2005, and Holsanova 2008, for kindred approaches). As discussed by Habel and Acarturk (2007), in processing text-graphics documents humans construct different types of reference and co-reference relations (cf. section 1). The underlying idea of the present study is that *integrated conceptual representations* mediate between language, graphics and domain entities in multimodal comprehension of language and graphics.

Figure 2 depicts the information flow between the modality specific modules and the integration processes as proposed in this paper. Since humans do language comprehension as well as graph comprehension incrementally—as empirical research in psychology and neuroscience convincingly argues for—the core research questions concerning the internal structure of the integration module are: (a) *which level of incremental entities are involved in integration?*, (b) *which types of representations are constructed by the modality specific modules to be transferred to integration?*, (c) *how are these representations constructed by modality specific modules be processed?*, and (d) *how do integrated representations influence modality specific comprehension.?*[3] In the present paper we focus on questions (b) – (d), in particular on the construction of referential and co-referential links.

2.3 The role of conceptual representations in integration

In a first step, we exemplify the construction of conceptual structures by the language module with example sentence (1).[4] The lexical information of 'decline' provides a conceptual representation containing a process concept

$$DECREASE\_OF\_VALUE(\__{TEMP}, \__{VALUE}, \ldots).$$

We focus here only on two arguments of this process, namely a temporal argument, which can be filled by an interval, and value argument, that can be filled by an entity of an ordered structure, which functions as the domain of the value, here the NUMBER-domain. By using such abstract representations, which generalize over different value domains, it is possible to catch the common properties 'decline of number', 'loss of weight', and others. The temporal argument, which is necessary for all process and event concepts, stands for the 'temporal interval during which the whole process is occurring'; in sentence (1) the beginning of the interval is explicitly specified. Putting this together, the process concept *DECREASE_OF_VALUE* stands for a specification of a mapping from the temporal domain in the value domain, or—using the terminology of topology—for a 'path' in the value space. Moreover, the lexical information of 'decline' provides *SOURCE* and *GOAL* arguments to be filled optionally. Sentence (1) supplies 'peak of about 60' [via a *from-PP*] and 'about 20'[via a *to-PP*].

The task of the second phase of line graph comprehension (as depicted in figure 2), the con-

[3] Figure 2 is undetermined with respect to the internal structure of the integration module as well as to the details of the interaction processes since these question are only partially answered up to now.

[4] The system of conceptual and spatial representations we use is a computation-oriented extension of Jackendoff's conceptual semantics (see, Jackendoff 2007) described in Eschenbach et al. (2000) Tschander et al. (2003).

struction of structured visual descriptions, in particular contains this: descriptions of relevant parts of the graph line, their geometrical properties and spatial relations between these parts. In this step, the system of spatial representations plays the role of a descriptional inventory, which is accessed by visual routines. We exemplify this with salient parts of the trend graph for Horned Grebes (cf. figure 1). Visual segmentation of the line graph leads to—inter alia—a line, which overall direction is vertical and which possesses one local maximum of curvature. Figure 3.a depicts the correspondence between an idealized shape of this type and its structured description by a spatial representation. Figure 3.b depicts the correspondence between a complex part of the line graph, namely a sequence of line segments, which has an overall horizontal orientation, and a abbreviated spatial description of that graphical constellation.[5]

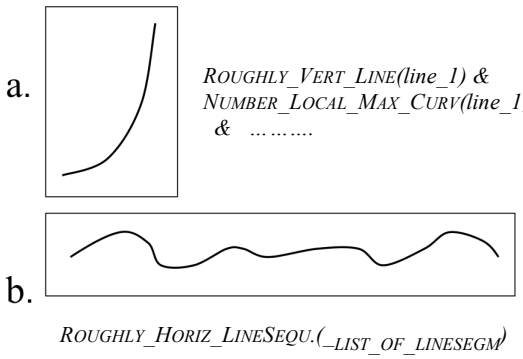

Figure 3. A sample set of integral conceptual representation.

In contrast to the more general *visual encoding processes* the following sub-module, *graph schema instantiation* in Pinker's (1990) terminology, has the task to interpret elements as parts of graphs. In the Horned Grebe graph, for example, the vertically extreme *POINT_OF_MAXIMAL_CURVATURE*, which is characterized as connection point between two roughly vertically oriented lines, will be determined as a *PEAK* of graph line. Since the x-axis of the trend graph in question refers to the temporal domain, the 'natural order' of time, leads to an inherent orientation of the line segments: Thus the most left part of the trend graph has to be interpreted as an *INCREASE*, the following—after the *PEAK*—as a *DECREASE*.

As the Horned Grebe example shows, both modality specific comprehension via contribute via conceptual representations based on a common conceptual inventory and the referential links build up during comprehension to an integrated and—hopefully—coherent interpretation of the text graphics document: The verb 'decline' provides *DECREASE* conceptualizations, as well as the application of graph schemata. The linguistically mentioned 'peak' is source of two referential links, one the one hand, to a domain-entity, namely an approximate number of birds, on the other hand, to a graphical entity.

In the second part of this paper, we present empirical evidence for the process descriptions presented in this section.

3 Eye Movements in Multimodal Comprehension

The investigation of eye movement parameters has been a widely used research method for the investigation of online comprehension processes in psycholinguistics research (Staub & Rayner, 2007), graph comprehension research (Shah & Vekiri, 2005), as well as in multimodal discourse analysis (Holsanova, 2008). But, as of our knowledge, there is no systematic analysis of eye movement behavior on graph-text documents.

3.1 The Experiment

We conducted two experiments, both based on the material exemplified in Section 1. In Experiment 1, ninety-one graduate or undergraduate students were presented 42 graph lines in rectangular frame, without any labels or numbers. The graphs were redrawn based on the original source (see fn. 2). The subjects were informed that they would see a set of graphs on the screen, each for three seconds; and they were expected to inspect the graphs as they change automatically.

In Experiment 2, text-graph constellations were presented to 36 graduate or undergraduate students. Each subject was presented twelve text-graph documents, similar to the one in Figure 4. The figure also shows resulting eye movement patterns on the presented stimuli.

The stimuli were based on the texts and graphs in the original source, after redrawing of graphs and modifications of text for systematic investigation. There were two factors in Experiment 2: the shape of the graph line (with three condi-

[5] A detailed description of these steps is beyond the scope of this paper. De Winter & Wagemans (2006) give a thorough overview about segmentation processes in perceiving line drawings.

tions) and the number of graph-related sentences in the text (with four conditions). We call these *target sentences*. The text in each stimulus consisted of three parts in the mentioned order: (1) several sentences before the target sentences (namely, *pre-target sentences*). These were not related to the graph, presenting information such as breeding, migration etc. (2) The *target sentences*. (3) Several sentences after the target sentences (*post-target sentences*). These were not related to the graph. The subjects were informed that they would see inventory information about wintering birds; they were expected to investigate the presented information and then to answer some questions. A 50 Hz eye tracker recorded eye movements of the subjects in both experiments.

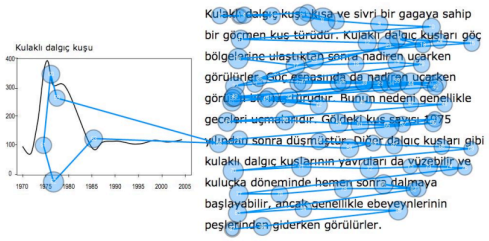

Figure 4. Sample eye movement protocol.

3.2 Results

In this section, a partial summary of the results of the experiments is presented. First, we discuss the results concerning the characteristics of eye movement behavior in a qualitative manner. Based on the eye movement recordings, fixation maps can be drawn, as exemplified in Figure 5. This fixation map is based on the fixation counts of all the subjects on one of the graphs presented in Experiment 1. In this figure, the red, yellow and green regions show fixation distribution in decreasing order.

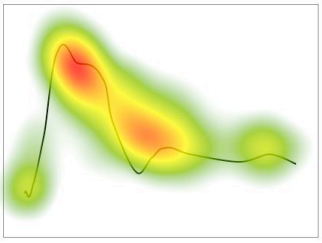

Figure 5. Sample fixation map.

Since the graphs were not accompanied by text in Experiment 1, the resulting fixation maps reflect the visually salient regions. In other words, these patterns were *not-linguistically-guided* fixation patterns. In Experiment 2, part of the stimuli of the first experiment was presented with accompanying text. We have divided the fixations on the graph region in Experiment 2 into three groups for analysis: (1) the fixations before the target sentences were read (namely, *pre-target phase* fixations). These occurred generally at the beginning of the reading of the text. (2) The fixations immediately after reading the target sentences (*target-phase* fixations). (3) The fixations after the target sentences were read. (*post-target phase* fixations). These occurred generally at the end of the reading phase. In this study, we focus on the first two types of fixations.

The fixations on the graph region were transcribed based on their location and total gaze time. A qualitative comparison of the fixation patterns, based on the exemplified stimuli in Figure 4 revealed that in Experiment 2, the *pre-target phase* fixation patterns were different than the *target-phase* fixation pattern*s*. Furthermore, the *target phase* fixation patterns of Experiment 2 were different than the *not-linguistically-guided* fixation patterns obtained in Experiment 1. In other words, different fixation maps were obtained in *linguistically-guided* and *not-linguistically-guided* inspection of the same graph.

A further analysis of the *target-phase fixations* in Experiment 2 was performed by quantitative comparisons of the fixation counts and gaze times on the graph proper (the fixations on the numbers and labels were excluded) that occurred after the two target sentences of the accompanying text: "*The number of birds declined after 1975*" and "*The number of birds remained stable around 100 after 1985*".[6] The results showed that after the 'decline' target sentence, the mean fixation count was higher on the decline-line than the mean fixation count on the remain-line of the graph, $t(16) = 4.76$, $p < .01$. On the other hand, after the 'remain' target sentence, the mean fixation count was higher on the remain-line than the mean fixation count on the decline-line of the graph, $t(11) = -5.70$, $p < .01$. On the other hand, in Experiment 1, there was no significant difference between the mean fixation count on the decline-line and the one on the remain-line, $t(90) = 1.86$, $p = .07$. Parallel results were found for gaze times. In summary, quantitative analyses revealed partial evidence that the linguistic repre-

[6] The number of target sentences had four conditions from one sentence to four sentences. For the purpose of this study, we compare the fixations only after the first and second target sentences.

sentations with different conceptual representations, in our case the '*decline*' and the '*remain*' sentences resulted in significant differences between mean fixation counts and gaze times.

4 Discussion

In the present paper we proposed a computational architecture for multimodal comprehension of text-graphics documents. We analyzed comprehension processes in terms of the interaction between the information induced by graphical and linguistic entities at conceptual level. We presented experimental support for the architecture by the analysis of eye movement patterns and parameters. First, we presented evidence for the difference between linguistically-guided and not-linguistically-guided inspection of graphs. Second, the findings of Experiment 2 revealed a difference between the fixations that followed the 'decline' sentence and the 'remain' sentence.

5 Conclusion

The interaction between language and graphs, as the two representational modalities, is not a well-investigated domain compared to research on multimodal comprehension of pictorial or diagrammatical illustrations. Methodologically, compared to research on eye movement control in reading, the studies that investigate eye movement behavior in multimodal documents have a relatively premature state due to abundant types of visual representations. This study contributes on both theoretical and experimental aspects of research on multimodal graph-text comprehension.

Acknowledgments
The research reported in this paper has been partially supported by the DFG (German Science Foundation) in ITRG 1247 'Cross-modal Interaction in Natural and Artificial Cognitive Systems' (CINACS). We thank the HCI Lab at the Middle East Technical University and the two anonymous reviewers for their helpful comments.

References

De Winter, J. & Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale integrative study. *Cognition, 99*. 275–325.

Eschenbach, C., Tschander, L., Habel, C., & Kulik, L. (2000). Lexical specifications of paths. In C. Freksa, W. Brauer, C. Habel & K. F. Wender (Eds.), *Spatial Cognition II* (pp. 127-144). Berlin: Springer.

Ferreira, F. & Tanenhaus, M.K. (2007). Introduction to special issue on language-vision interactions. *Journal of Memory and Language, 57*. 455-459.

Habel, C., & Acarturk, C. (2007). On reciprocal improvement in multimodal generation: Co-reference by text and information graphics. In I. van der Sluis, M. Theune, E. Reiter & E. Krahmer (Eds.), *Proceedings of the Workshop on Multimodal Output Generation: MOG 2007* (pp. 69-80). United Kingdom: University of Aberdeen.

Holsanova, J. (2008). *Discourse, vision, and cognition. Human Cognitive Processes 23*. John Benjamins Publishing Company: Amsterdam / Philadelphia.

Jackendoff, R. (1996). The architecture of the linguistic-spatial interface. In P. Bloom; M. A. Peterson; L. Nadel & M. F. Garrett (eds.), *Language and Space*. (pp. 1-30). Cambridge, MA: The MIT Press.

Jackendoff, R. (2007). Linguistics in cognitive science: The state of the art. *The Linguistic Review 24*. 347-401.

Kosslyn, S.M. (1989). Understanding Charts and Graphs. *Applied Cognitive Psychology, 3*. 185-226.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73-126). Hillsdale, NJ: Erlbaum.

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction*, *13*, 227-237.

Schnotz, W. (2005). An Integrated Model of Text and Picture Comprehension. In R.E. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning*. (pp. 49-69). Cambridge: Cambridge University Press.

Shah, P., Freedman, E., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge Handbook of Visuospatial Thinking* (pp. 426-476). New York: Cambridge University Press.

Staub, A., & Rayner, K. (2007). Eye movements and on-line comprehension processes. In G. Gaskell (Ed.), *The Oxford* handbook *of psycholinguistics* (pp. 327–342). New York: Oxford University Press.

Tschander, L., Schmidtke, H., Habel, C., Eschenbach, C., & Kulik, L. (2003). A geometric agent following route instructions. In C. Freksa, W. Brauer, C. Habel & K. F. Wender (Eds.), *Spatial Cognition III*. (pp. 89–111). Berlin: Springer.

Ullman, S. (1984). Visual routines. *Cognition, 18*, 97-106.

Stand-up Gestures – Annotation for Communication Management

Kristiina Jokinen
Department of Speech Sciences
University of Helsinki
Kristiina.Jokinen@helsinki.fi

Minna Vanhasalo
Department of Finnish Language
University of Tampere
Minna.Vanhasalo@uta.fi

Abstract

This paper deals with the analysis and annotation of gestures which we call stand-up gestures. These gestures distinguish themselves from the flow of verbal information exchange by their regulating and coordinating function in communication, independent from the spoken content. The name also bears reference to stand-up comedy where these gestures occur as part of the normal repertoire of successful performance. Besides analysing the functions of stand-up gestures, the paper also discusses their annotation using the MUMIN annotation scheme and proposes extensions to the scheme in terms of a meta-discursive context level.

1 Introduction

In order to maintain smooth communication, dialogue participants need to pay attention to subtle gesturing by the partner. Gestures seem to have several important functions in communication, ranging from the actual content level contributions (iconic gestures) to the coordination of communication (own communication management and interaction management, see Allwood, 2002; Allwood et al., 2007), and to the giving of rhythm to spoken utterances (McNeill, 2005). Gesture studies have thus been important in sociolinguistics, intercultural communication and behavioral studies, so as to have a better understanding of how human communication takes place. For instance in second language learning, it is important to understand how gestures are used in communication: the students need to learn to observe the relevant communicative signals and to produce suitable gestures themselves. Gestures are also important for computer animations and interaction technology in order to allow more natural interactions with a computer. Besides ECAs (Cassel et al. 2003), recently also robotic companions have developed so that they can recognize gestures and thus become engaged with multimodal communication (Bennewitz et al., 2007). New application areas are also various game and educational toys that would allow especially autistic or disabled children to enjoy and be empowered by the new technology.

This paper deals with the analysis and annotation of certain kinds of gestures which have a regulating and coordinating function in dialogues. They distinguish themselves from the flow of verbal information exchange in that they not only accompany or complement the spoken content but rather function as independent means for communication management. They are related to interactive gestures (Bavelas & Chovil, 2000) and gestures on meta-discursive levels (Kendon, 2004). We call them stand-up gestures, as they typically single out one word or phrase from the utterance as important making the expression to stand up from the flow of speech, and since they are typical to the normal repertoire of successful stand-up comedy performance.

The paper is structured as follows. Section 2 presents the data and provides two examples of stand-up gestures. Section 3 discusses the MUMIN annotation scheme and its suitability for annotating stand-up gestures. Section 4 discusses multifunctionality of gestures, and Section 5 provides an extension to annotation schemes in terms of extended contexts. Section 6 draws conclusions and points to further research topics.

2 Stand-up Gestures

2.1 Pointing in repairing

In the first example there are four people playing a North-Finnish game called *tuppi* (similar to bridge). Players play as partners, one pair against the other pair, and the main rule is that each player must, if possible, play a card of the suit led. A player with no card of the suit led may play any card, which is called *sakata* (in the example there is a past tense 1. person form of *sakata – sakkasin*). When a player has to *sakata*

he usually just plays the least useful card he has. However, at times a player can put forth a really good card and use *sakata* as an opportunity to give a signal to the partner of a desired suit to be played next, this is called *merkkisakkuu.* The players are aware that the meaning of the choice of a card in certain circumstances can be either a neutral *sakata* or a marked *merkkisakkuu*, but they are strictly not allowed to express explicitly which one of the two moves they make when playing the other suit. Most typically the *sakata* situations occur when one's partner, the co-player, is forced to play a different suit on one's card. These situations are also easiest for the players to notice and interpret correctly, because they represent highly conventionalized practice. However, the players also follow closely what cards the other pair puts forth when they have to *sakata*, because these can be – although very rarely are – a *merkkisakkuu.* The better the players can read the game and distinguish between *sakata* and *merkkisakkuu*, the better players they are.

Figure 1 shows the relevant stand-up gesture that occurs when the players discuss the game they have just finished. M and his partner T have severely lost, and M had given an explanation for their losing (line 1): he had misread a neutral *sakata* as a marked *merkkisakkuu*.. This is misunderstood by A who asks for clarification – the other initiated other repair – on line 3 with the question *which spade.* Soon after asking this A makes a self initiated self repair with the turn *oh the one that I:: did sakata* with an accompanying pointing gesture, Index Finger Extended. The gesture is on the elongated pronoun *I::*, i.e. it points out the most important word of the sentence, and of the repair sequence.

What happens in the dialogue is that A first misinterprets M's reference of *merkkisakkuu* to be some spade played by M's co-player, but then understands that M actually means the spade she herself had *sakata*. A's understanding is evident when looking at her self correction together with the gesture. The stand-up gesture points out the most important word from the utterance, i.e. the correction of misunderstanding, and is made towards M. The reason of the original misunderstanding can be spelled out explicitly as follows: "Which spade played by your co-player? Oh you mean the spade that I (and not your co-player) have to *sakata*." The misunderstanding and its solution is conveyed by the accompanying gesture which is synchronized with the relevant word of the repair indicating which part of the misunderstanding is the repairable part.

Figure 1. "oh the one that I *did sakata"*

```
1 M: mä luulin et se pata olis ollu
2    me(h)rkkisakkuuh hh
     I thought that the spade would have been a merkkisakkuu
3 T: hehe
4 A: $m(h)ikä patah$ .hhhh ai se
5    minkä [mä:] sakkaasin
     $which spade$ .hhh oh the one that [I::] did sakata
```

[LH Extended index finger flicks to M hand rotating from palm lateral position to palm up position. Arm rests on the table whole time.]

Furthermore, with her repair, A also shows that she trusts that all the players have common ground and general knowledge about playing the game: she does not explicitly explain why she had the kind of misunderstanding she had. Her repair *the one* I *did sakata* is interpreted correctly by all the parties with (and in) a flick of a hand.

2.2 Pointing in Managing Information

Our other example is from a situation where two young women chat over a lunch. One of them is telling stories about a janitor and snow plowing, and the relevant stand-up gesture occurs between two story-telling episodes. The first story goes as follows: "Our janitor has started plowing snow again, and as you remember from last winter, he used to plow snow at an inconvenient time at 5 am., the snow tractor made utterly annoying noise, and the job took 2 hours to finish even if the yard to be plowed is really small". The second story concerns the janitor doing snow plowing again this year. However, to justify the story about the same janitor doing the same task, the narrator gives a piece of new information that explains why the follow-up story is new and its telling worthwhile. One of the complaints of the snow plowing last year was that the janitor started to work too early. The new information is that the janitor now starts later, and this fact had allowed the narrator to observe the janitor's working in a more detailed way: she now knows why the snow plowing takes so much time. The new information is accompanied by a pointing gesture

that singles out the newsworthy content (Figure 2), and in the follow-up story the narrator gives an account for the lengthy snow plowing.

Figure 2. "started a bit later *than before"*

1 N: .hhh me $j(h)ust niinku-$ nyt se on
2 alottanu vähä [**myöhem**min] siitä niinku,
$We just like-$ Now he has started a bit ***[later]*** *than before like,*
[**RH index finger stretched, slightly crooked, palm towards oneself, points quickly straight to the interlocutor twice**]

When the narrator marks the word *later* with the gesture, not only does she mark the word as important in relation to the content of the story already told, but it also refers to the parts to come: the new detail in the shared information expounds the premises for understanding the conditions of the new story to come. The narrator has been able to watch the snow plowing exactly because it has been done at a reasonable time in the morning when she has been awake. This is not explicitly said in words, but conveyed together with the gesture in the given context.

3 Gesture Annotations

For the applications and research mentioned in Section 1, it is important that the collection and analysis of large multimodal corpora are available and accompanied with rich annotations comprising of verbal and non-verbal phenomena. For instance the AMI corpus (Carletta, 2006) is a large video corpus of meetings and spontaneous interactions and it is accompanied with annotations that also deal with multimodal aspects of communication. Several other video corpora and annotation schemes have been developed as part of projects or individual efforts, see e.g. Martin et al. (2007), and the examples in this paper.

As part of our analysis of stand-up gestures, we have used the MUMIN annotation scheme (Allwood et al., 2007) which is intended as a general instrument for the study of hand gestures, facial displays and body posture in interpersonal communication. The annotation scheme contains categories to describe the form and dynamics of communicative elements as well as their function in managing feedback, turn-taking, and sequencing. The distinctive feature in the scheme is the use of semiotic categories to encode elements as semiotic signs: *Indexical Deictic, Indexical Non-deictic*, *Iconic,* and *Symbolic*.

Considering the analysis of stand-up gestures, their description is distributed among the categories for interaction and communication management. The MUMIN scheme provides annotation categories for their form (hand shape, orientation, location, direction of the movement) and functioning in the information structure (opening, continuing or closing of topics; emphasis), turn management (opening, holding, yielding, etc.), and sequencing (opening, continuing or closing speech act sequences). This is useful when interpreting communicative signs via a dynamic process where the combination of characteristic features determines the sign's interpretation, i.e. gesture signs are not fixed categories but form a continuum along the defined features. By defining elementary features and modelling their combinations it is possible to construct a flexible framework in which similarities and interpretations of various communicative gestures can be compared and studied. From the computational view-point, this supports modelling and experimentation with various classification and clustering algorithms.

Since gestures are multifunctional and multidimensional, this many-to-many nature needs to be incorporated in the annotation. However, the analysis of stand-up gestures also seems to require understanding of the linguistic, pragmatic and social contexts in which they occur, and how the different contexts affect the layering of more than one meaning and function on a gesture. We will return to the different contexts in Section 5, but will first look at the example gestures and how their form, meaning and functions are motivated by the lexical affiliate, parts of speech and common ground between the participants.

4 Multifunctional gestures

4.1 Local meaning and function

Kendon (2004) has identified different gesture families, e.g. Open Hand Prone and Open Hand Supine families. Based on his observations he suggests that each gesture family has its own *semantic theme*. Gestures in Open Hand Prone family express in general stopping or halting of an action (own or other), whereas those in Open Hand Supine family express general offering and

giving of ideas and concepts. According to Kendon the Index Finger Extended is yet another gesture family which, however, has not been thoroughly identified nor classified. The main semantic theme of the Index Finger Extended family seems to be the same as that of the Open Hand families with one distinction: the gestures in this family are precise and explicit. Our analyses of the two stand-up gestures support this distinction. The gestures explicitly single out the important word of the utterance, i.e. the one that refers to what has been repaired in the previous misunderstanding or is the relevant new content in the story telling episode. The exact hand shape, index finger extended is motivated by the communicative needs on the utterance level (to point out a particular expression from the speech), while the orientation of the palm is motivated by the needs of communication management (halt conversation, offer information).

4.2 Communication management

In example 2.1, A's index finger is oriented horizontally and the hand rotates from a palm down position to a palm up position, thus offering the repair to the interlocutor. This resembles the Palm Open Supine family's semantic theme of giving and offering. In example 2.2 the index finger is obliquely horizontal, but the palm is facing the speaker (i.e. back of the hand is towards the listener) and the finger points straight to the interlocutor. The narrator halted telling the second story for a moment in order to give some new information with respect to the given information (i.e. the annoying snow plowing starts later in the mornings this year). This resembles the semantic theme of stopping and halting of the Open Hand Prone (vertical) family.

Allwood (2002) talks about own communication management and interaction management, referring to the aspects of communication that concern meta-level control of the interaction and can include such functions as repairs, initiations of topics, direction of the focus of attention, etc. In example 2.1 the gesture in conjunction with the repair of one's own speech belongs to the own communication management plane. Furthermore, the orientation of the gesture, palm up, is sensitive to the local negotiation of context. The speaker knows that she has made a correct repair of the person who used the spade for *sakata*, and with the orientation of her gesture she signifies her understanding and hands the understanding of a successful repair over to the interlocutor. The gesture in example 2.2, however, manages the structuring of information. The palm orientation of the pointing gesture shows that the speaker is temporarily halting the flow of storytelling but not halting it altogether. She is not merely offering a new piece of information but rather stopping the storytelling in order to give the particular piece of information that motivates the later story. The palm orientation away from the listener cuts the interlocutor's opportunity to take the floor during the stop.

Gestures are often related directly to the information flow of the dialogue. However, stand-up gestures require that the speaker is aware of the means to coordinate the conversational situation and to focus the partner's mind on some particular aspect in the exchanged information or to prepare the partner to have the right stance in order to interpret the message in the intended way. Stand-up gestures often occur in everyday contexts (as in our examples) where the speaker controls a story telling situation and indicates the start of a new topic, a repair, or otherwise important new information. They also often indicate the speaker's dominance over the floor, since the speaker can thus coordinate the flow of information, turn-taking, and interpretation of the presented ideas. The speaker's role as the initiator of a topic also allows her to control the topic management, to continue or close the chosen topic. This kind of control can be especially seen if the speaker has a dominant role in the activity (e.g. chairing a meeting), and in storytelling situations and stand-up comedies where the gestures are frequently used to manage the flow of information and lead the story towards its punch-line.

Instead of getting their meaning from the content of the verbal flow of information, stand-up gestures indicate to the partner non-verbally how the conversation is to be understood and divided into communicatively important segments. They are distinguished from the normal flow of information so as to catch the partner's attention, and by so doing they also control the dialogue flow.

4.3 Social Interaction

With the notion of catchment McNeill (2005) refers to the social-interactive nature of all gestures: gestures have an active role in creating, shifting and updating the common ground between the interlocutors. Catchment is used only in the context of cohesives, i.e. similar kind of gestures that keep recurring in the dialogue. We propose, however, that not only repetitive cohesives, but single gestures (the stand-up gestures) can create and indicate the common ground be-

tween interlocutors. In other words, catchments can be seen as part of the constant ongoing negotiation of context in conversations – yet another level of context in addition to the local utterance and communication management levels.

Pointing straight to the interlocutor is usually considered insulting unless the social relationship is such that this is acceptable, in which case pointing can act as a bonding strategy. For instance, in example 2.2 the narrator recognizes, that the previously given information (the first story) is already part of their shared knowledge, so she starts the storytelling with summoning *as you remember from last winter*. By making the pointing gesture straight to the interlocutor, the narrator also seems to want to gain heightened attention of the interlocutor: it is at this point that the truly new information begins which the interlocutor has not heard before. The narrator has thus taken their social relationship into account and acknowledged their long shared history of similar discussions: the gesture points out that the follow-up story will update their shared knowledge of the janitor and snowplowing, and therefore asks for intensive attention.

The last interesting observation is that the gesture in 2.1 is made in the periphery, whereas the gesture in 2.2 is made in a more central place. This can be accounted for with the help of the notion of common ground. Bavelas & Gewing (2004) showed that interlocutors use less explicit, smaller and peripheral gestures when reference is made to the common ground, and when the reference is not to the common ground, the gestures become larger, more central and explicit. In example 2.1 the speaker is handing over information that is self explanatory for all the parties because it is based on their common ground: shared knowledge of the game conventions. The gesture is thus rather small and peripheral. In 2.2, however, the narrator updates the common ground as she is about to move from the first story (given information) to the follow-up story (brand new information), and the gesture is consequently larger and more central. The place of the stand-up gesture can thus be said to be motivated by the social interactive level where the notion of common ground explains the choice between the periphery and central place.

5 Stand-up Gestures and Context

As shown above, interpretation of the gesture is related to the context in which the gesture occurs. The context influences the form and function of the gesture, and depending on the closeness of the interlocutors' relationship, also the gesture's acceptability and interpretation. Concerning interactive situations, we especially like to emphasise the communicative context in terms of activity types and the speakers' roles (cf. Levinson, 1992; Allwood, 2002). Activity types impose constraints on acceptable contributions in a given communicative context, and roles set up strong expectations on the appropriate behaviour and how contributions should be interpreted.

Often, however, gestures have different relations to their context, or the relation of the gesture to its context is not explicitly spelled out: the gestures are multi-contextual. Figure 3 depicts the five different context levels that we consider important when analyzing gestures.

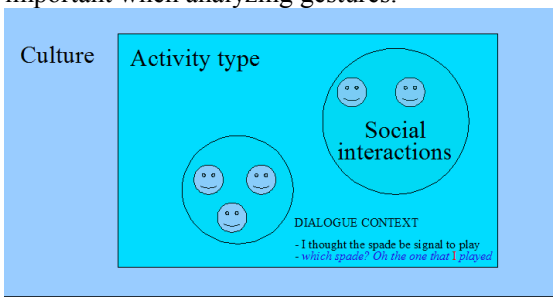

Figure 3 Contexts that influence the form, meaning and function of a gesture.

The most local context is the representational context of the gesture, which consists of a lexical affiliate. For instance, a beat can give emphasis on a word highlighting it, and an iconic gesture can express semantic features of the referent by similarity or homomorphism. A stand-up gesture also singles out the most important word of the utterance and thus resembles beats, but rather than being repetitive and rhythmical as beats, a stand-up gesture is a single "stand-alone" gesture. The next level context is the dialogue context. Gestures operating on this level deal with the relationship between speech segments, sequencing and structuring of information, managing contributions and turn taking ("what I said previously", "the next point", new vs. given information, repairs). The third level context deals with social interactions. Gestures on this level denote the relationship between interlocutors and the common ground between them. The fourth context level concerns the activity type that the speakers are engaged in (ranging from everyday chatting to task-oriented discussions, from informal events to formal performances). For instance, pointing a finger to the listener of a story or to the audience of a comedy act asks for heightened attention to the shift in focus: the ges-

ture indicates that there is a transition from an old story to a new one, or that the punch line is coming. The largest context is the cultural context that concerns social norms and relationships, i.e. culturally conditioned behaviour patterns that limit the appropriateness and interpretation of gestures. Emblems are typical examples of gestures on this level.

The five contexts interact with each other and the larger contexts usually affect the more specific ones. Each context also influences the form and function of the gestures in various degrees. We propose that the contexts be taken into account in the MUMIN annotation scheme. They can be included as a special annotation feature "Context" with five values (lexical, segmental, social, activity, culture) or via a more sophisticated linking system based on the gesture's multifunctionality and multidimensionality.

As discussed in Chapter 3, the hierarchical feature-based annotation seems reasonable compared with a simple gesture categorisation, especially when thinking of the continuum that different gestures make with respect to their form and function in general. However, as always with annotations, an important yet open issue is how much detail will be sufficient in the annotation scheme without getting too deep into the micro-analysis of gestures and lose useful generalisations. On one hand we have views about highly organised interactions where no phenomenon is too small to be considered meaningful (cf. Goodwin, 1984). On the other hand, there are practical goals and needs for developing models for interactive systems for which a certain level of generality, frequency, and categorisation is desirable and necessary. Gesture families as suggested by Kendon (2004) seem useful in this respect.

6 Conclusions

We have discussed the form and function of stand-up gestures on the basis of corpus examples. The gestures are important in coordinating interaction on meta-discursive level: constructing common ground and regulating information flow so that the verbal activity is not disrupted. The speakers need to learn how to distinguish communicatively meaningful gestures from those that do not matter, and also to provide a correct interpretation for them. It is through this kind of gestural communication that the speakers construct mutual knowledge and create social bonds.

We have also proposed five contextual levels in which the gestures can be interpreted: linguistic, dialogue, social interaction, activity type and cultural context. For various applications and further modelling (e.g. gesture lexicons for ECAs and in human communication studies), the contexts need to be included in the annotation scheme, so as to be able to describe gestures on a meta-discursive level where they can be related to the whole dialogue and the dialogue partners.

References

Allwood, J. 2002. Bodily Communication - Dimensions of Expression and Content. In Granström, B., House, D., Karlsson, I. (Eds.) *Multimodality in Language and Speech Systems.* Kluwer. pp. 7-26.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In Martin, J. et al. (Eds.), pp. 273–287

Bavelas, J., Chovil, N., 2000 Visible acts of meaning. An Integrated Message Model of Language in Face-to-Face Dialogue. *Journal of Language and Social Psychology* 19(2), 163–194.

Bavelas, J., Gerwing, J. 2004 Linguistic influences on gesture's form. *Gesture* 4(2), 157–195.

Bennewitz, M., Faber, F., Joho, D., Behnke, S. 2007. Fritz - A Humanoid Communication Robot. Procs of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).

Carletta, J. (2006) Announcing the AMI Meeting Corpus. *The ELRA Newsletter* 11(1), 3–5.

Cassell, J., Sullivan, J., Prevost, S., Churchill, E. (Eds.) 2003. *Embodied Conversational Agents*. MIT Press, Cambridge, MA

Goodwin, C. 1984 Story structure and organization of participation. In Atkison, J.M., Heritage, J. (Eds), *Structures of Social Action. Studies in Conversation analysis. Studies in Social Emotion and Interaction.* Cambridge University Press. pp. 225–246.

Kendon, A. 2004. *Gesture: Visible action as utterance*. Cambridge.

Levinson S. 1992 Activity types and language. In Drew P., Heritage J. (Eds.) *Talk at work*. Cambridge University Press. pp. 67–107.

Martin, J.C., Paggio, P., Kuenlein, P., Stiefelhagen, R., Pianesi, F. (Eds), Multimodal corpora for modelling human multimodal behaviour. *Special issue of the International Journal of Language Resources and Evaluation,* 41(3–4).

McNeill, D. 2005. *Gesture and Thought.* University of Chicago Press, Chicago and London.

Coarticulation in sign and speech

Stina Ojala
Department of Information Technology, University of Turku, Finland
stina.ojala@utu.fi

Salakoski, Tapio
Department of Information Technology, University of Turku, Finland
tapio.salakoski@utu.fi

Aaltonen, Olli
Department of Speech Sciences, University of Helsinki, Finland
olli.aaltonen@helsinki.fi

Abstract

Different manifestations of coarticulation have been within focus of speech sciences for quite some time now. In sign research the focus has recently covered also coarticulation research through latest techniques. Studying coarticulation is easier in sign because all the articulators are visible all the time, which makes it different from speech where articulators are mostly hidden. In speech research the study of coarticulation is thus concentrated on the manifestations of coarticulatory phenomena in acoustic signal.

1 Introduction

In speech the sounds are not discrete phenomena but speech is a continuous string of articulatory movements, where previous and following sounds affect each other. This on-going movement pattern from one sound to another is called coarticulation. Coarticulation is a way to make transitions from one sound to another easier. In this way we are acting according to the *ease of articulation* –principle (e.g. Shariatmadari, 2000). Simultaneously, we also tend to use all the capacity available if needed (Lindblom, 1981). That interplay of coordinating movements makes speech easier and faster.

When recording speech usually coarticulatory phenomena are controlled by using carrier words and sentences such that different sounds occur in similar positions coarticulatorily. This is a way to control what is said in order to make subjects' production comparable to each other. Coarticulation provides us knowledge on how different sounds are represented in different contexts. Coarticulation studies are made through acoustics in speech because the articulators are not visible. This is different from studies of sign language since in sign the articulatory patterns are visible and thus it is more straightforward to study coarticulation in sign.

Coarticulation has the same function in sign as in speech, it functions to make the message smoother and more compact. In sign the compactness aspect is even more important since hands are slower as articulators than speech organs. Speech rate usually ranges 90-160 words per minute while within this study signing rate is between 20-30 signs per minute. In signing both hands participate in coarticulation, so there are two levels of coarticulatory patterns – each hand separately and then the interarticulation – both hands together. Coarticulation is also present in facial expressions and gestures but those are left out of scope in this study. There is a distinction between manual and non-manual coarticulation and this study concentrates on manual coarticulation.

The controlling of coarticulation in sign language studies is through task design. This is the only way to have control on what is signed because there is no written form of any signed languages. In this study task design was based on an imaginary floor plan and a map task along with spontaneous signing.

Studies on coarticulation in sign first concentrated on fingerspelling research. Fingerspelling is converting text into a visible form by means of manual alphabet. It has a very limited amount of coarticulation: in most inventories of manual alphabets only one hand participates and the hand has a very limited movement patterns. There is

though an exception to this: British Manual Alphabet, which uses both hands. But still the effects of coarticulation are great also within the scope of fingerspelling (Wilcox, 1992). Recently the scientists have made first investigations on coarticulation of signing: the effects are similar to those on speech – coarticulation makes articulation more fluent and its effects can be seen both on handshapes and places of articulation (Mauk, 2003; Ann, 1996). Thus also signing obeys the ease of articulation –principle.

Especially anatomy and physiology of the hand and fingers affect articulation of handshapes. The economy of articulation is for its major parts depended on dimensions and movement ranges of individual fingers. According to physiological research results state that thumb has the widest range of movement patterns and that the ring finger is the most restricted in movement patterns (Ann, 1996).

2 Material and methods

The aim of this study was to investigate whether there are similarities in coarticulation between sign and speech, since they both are means of human communication and in the previous studies various researchers have stated analogous findings between speech and sign, e.g. in perception studies (on handshape perception see e.g. Ojala & Aaltonen, 2007).

The data have been gathered as a part of larger study, which concentrates on gathering data from the production and perception of Finnish Sign Language. Data consists of signed answers to a set of questions and tasks. The questions ranged from given tasks to spontaneous signing. The study design was equivalent and analogous to speech research sound samples, where tasks include different carrier sentences (read speech) and spontaneous speech (informants are asked to tell how they spent their last holidays or they had the option of telling a 2-3 minute story of whatever subject). The productions were gathered into a digital video which were then further processed and analysed with video software.

The data of this study consists of one informant's production HERE APARTMENT *Here is an apartment, which has a rectangle shape*.[1] The signed sentence consists of 3 individual signs and 6 rhythm units. The rhythm units were defined visually from the alternation of accelerations and decelerations. Furthermore the coarticulation analysis was made frame by frame, that is every 42 milliseconds. In each frame the coarticulation points were measured and that analysis served as the basis of the study of movements in time for both hands. There were 10 measurement points in each hand in order to gather precise material on how different handshapes manifestate in a continuous sign flow.

In this preliminary study we have measured 6 of those 10 measurement points from each hand. Other 4 measurement points were used to specify the orientation of the hand when it was possible.

Figure 1. The coarticulation measurement points in the whole study (dark and light diamonds) and in this data (light diamonds). These points are measured in each hand.

The orientation of the hand within this study translates as the orientation of the palm of the hand in relation to the body of the signer. The pixel coordinates of the coarticulation measurement points were inserted in a matrix. The matrix was the input for the 3D image of movements of coarticulation points in time. This served also as an input of 2D images of the speed and acceleration as changes of each coarticulation point frame by frame. The matrix was also the basis according to which the median of the speed was calculated. All figures and calculations were accomplished by MatLab scripts.

3 Results

Coarticulation in sign can be studied within two different scopes: a broader one with focus on the interarticulation of both hands and a more precise one considering coarticulation within the

[1] There are no standardized ways of transcriptions, so within this article glossing of signs is used. Glosses consist of capitalized transcriptions of signs and translations of signs in English with italics.

movement patterns of one hand only. Interarticulation seems to effect in such way that the hands move faster when they are more apart than when they are closer to each other.

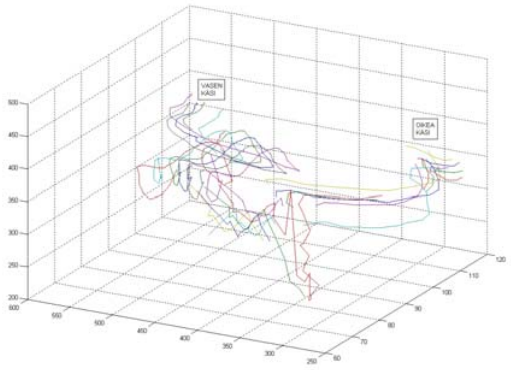

Figure 2. 3D graph of movement envelopes of individual fingers in space. The discontinuities in graphs are due to coarticulation measurement points not being visible to camera.

Most of the time all fingers move simultaneously and with the same speed, but both index fingers have broader and faster movement envelopes when compared to other fingers. When comparing between the index fingers the right index finger has slightly broader and faster movements. A similar handedness effect can be noted in thumb movement patterns but not on other fingers. The movement envelopes become smaller when going from the index finger to little finger, however, little finger has a broader movement envelope than ring finger. In other words, the ring fingers in both hands have the most compact movement patterns. The thumb has a broader movement envelope than index finger, but the movements are slower. The overall movement patterns in both hands are quite similar, both in timing and in broadness.

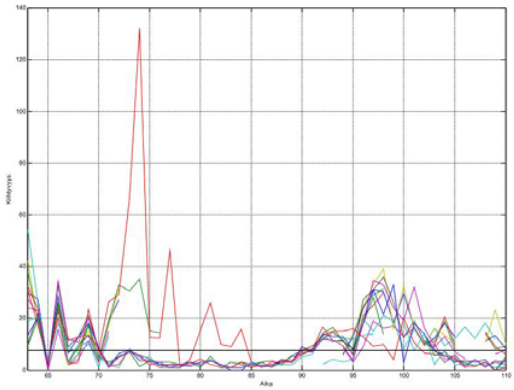

Figure 3. Changes in speed of individual fingers in time.

Changes in movement were investigated frame by frame. The graph shows both the speed of individual fingers and the relations between different finger movement patterns. Most of the time within this material the momentary/instantaneous differences between individual fingers' speeds are so minute that the graph would be sufficient with just one finger's movement description, but there are exceptions too. According to preliminary results it seems so, that individual vertical movements are faster than movements along the horizontal plane, but this observation needs further investigations. The graph also shows the tendency to keep movements as slow as possible but the scarce data might distort the results.

4 Conclusion

The alternations of deceleration and acceleration in signing movements are a similar pattern to speech – also speech has intertwining rhythms in different levels. The bases for these rhythms at least for some parts vary according to the individual language (e.g. O'Dell & Nieminen 2001). In speech the rhythm is achieved with the coordination of articulatory movements and it might be that the alternations between consonant and vowel sounds are one of the very corner stones of human evolution (MacNeilage, 1998). In this preliminary study we have concentrated on the observations of movement patterns on the lower level of coarticulation. Previously the alternations of decelerations and accelerations have been studied by Loomis et al (1983). The movement patterns in sign seem to have an oscillation pattern and cyclical form as do the articulatory movements in speech (See also Lindblom et al., 2006). The basic rhythm on the higher level of hierarchy in signing materialises in the movements and holds within and between individual signs in on-going signing (Liddell & Johnson 1989; Kita et al. 1998).

According to this material the index finger seems to be the determining fact in the amplitude and the rate of signing. Other fingers follow the movement patterns of the index finger, but are more restricted in their patterns. The index finger also has a special task: pointing. Pointing is an important part of both signing and speaking – it is a convenient way to refer to something which is present and visible, let it be an object or a person. (More on pointing, please see Corballis 2002.) The thumb seems to have more independent movement patterns than other fingers. This

demonstrates the tendency of signing to exploit the capabilities in individual fingers' movement patterns as widely as possible whereas the ring finger's more restricted movement patterns demonstrate that signing tends to avoid such patterns that are more difficult to produce. In this way, signing as a form of communication acknowledges the physiological restrictions in the hands and operates accordingly.

References

Ann, Jean. 1996. On the relation between ease of articulation and frequency of occurrence of handshapes in two sign languages. *Lingua* 98: 19-41.

Corballis, Michael C. 2002. *From hand to mouth*. Princeton University Press: Princeton.

Kita, Sotaro, Ingeborg Van Gijn and Harry Van der Hulst 1998. Movement phases in signs and co-speech gestures, and their transcription by human coders. Lecture Notes in Computer Science. *Proceedings of the Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop, Bielefeld, Germany, September 1997.* Springer Verlag: Heidelberg. Volume 1371/1998

Liddell, Scott K., Robert E. Johnson 1989. American Sign Language: The phonological base. *Sign Language Studies* 64: 195-277.

Lindblom, Björn 1981. Economy of speech gestures. In: MacNeilage, Peter (ed.) *The production of speech* pp. 217-246. Springer Verlag: New York.

Lindblom, Björn, Claude Mauk and Seung-Jae Moon 2006. Dynamic specification in the production of speech and sign. In: Divenyi, Pierre L., Steven Greenberg and Georg Meyer (eds.) *Dynamics of Speech Production and Perception.* Volume 374 NATO Science Series: Life and Behavioural Sciences.

Loomis, Jeffrey, Howard Poizner, Ursula Bellugi, Alynn Blakemore and John Hollerbach 1983. Computer graphic modelling of American Sign Language. *Computer Graphics* 17(3): 105-114.

MacNeilage, Peter 1998. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences* 21: 499-511.

Mauk, Claude 2003. *Undershoot in two modalities: evidence from fast speech and fast signing.* PhD Thesis, University of Texas, Austin.

O'Dell, Michael and Tommi Nieminen 2001. Speech rhythms as cyclical activity. In: Ojala, Stina and Jyrki (eds.) *21. Fonetiikan päivät Turku 4.-5.1.2001* Publications of the Department of Finnish and General Linguistics of the University of Turku, 67: 159-168.

Ojala, Stina and Olli Aaltonen 2007. Speech and sign – it's all in the motion. *Proceedings of ICPhS XVI 2007 in Saarbrücken*, 2169-2172.

Shariatmadari, David 2006. Sounds difficult? Why phonological theory needs 'ease of articulation'. – *SOAS Working papers in linguistics* 14: 207-226.

Wilcox, Sherman 1992. *Phonetics of fingerspelling*. Studies in speech pathology and clinical linguistics 4. John Benjamins.

Integration and representation issues in the annotation of multimodal data

Patrizia Paggio
University of Copenhagen
Centre for Language Technology
Copenhagen, Denmark
paggio@hum.ku.dk

Costanza Navarretta
University of Copenhagen
Centre for Language Technology
Copenhagen, Denmark
costanza@hum.ku.dk

Abstract

This paper deals with the issue of how to represent different types of multimodal interaction. We argue that, from a syntactic point of view, it is not possible to characterise the speech segments involved in a multimodal relation in uniform grammatical terms. In addition, the interpretation of the multimodal sign is also complex in that gestures interact with speech at different conceptual levels. We discuss examples of such complexity from empirical Danish data, and give suggestions for how they could be formalised in feature structures and how they could contribute to dialogue and discourse structure.

1 Introduction

Human communication is *situated* in the human body: we cannot avoid using our face, hands and body while we speak, and in face-to-face conversation we clearly react not only to our interlocutor's words but also to their gestures[1]. A possible cognitive explanation of this tight relation between speech and non-verbal behaviour may be that language emerged millions of years ago on top of our ancestors' ability to interpret and replicate gestures, so that speaking and gesturing partly depend on the same neurological mechanisms (Arbib, 2005).

However, speech and gestures are very different in nature, therefore it is difficult to formalise the way in which they interact.

First of all, since gestures are largely non-conventionalised, a fact that in turn depends on their essentially indexical and iconic rather than symbolic nature (Allwood et al., 2008), we cannot apply to them well-established abstract categories similar to phonemes or words. Attempts have been made to categorise hand gestures into meaningful types. Kendon (2004) describes for instance iconic types that share common physical features. However, such typologies are necessarily incomplete due to the very nature of the phenomenon.

Furthermore, gestures interact with the linguistic sign at different levels, from prosody to pragmatics (McNeill, 1992). An account of the different interacton types must therefore cope with segmentation and representation problems. In other words, which segment of speech should a specific gesture be associated with, and what representation should be given to the integrated multimodal contribution? In this study, we give tentative answers to these two questions drawing on examples from annotated video clips in Danish. We start by shortly presenting the annotation scheme and relating it to relevant work in Section 2. In Sections 3 and 4 we discuss examples where gestures accompany single words vs longer speech sequences. We show what the multimodal contributions look like in the XML annotation, and discuss how they could be represented in feature-based formalisms. In Section 5 we discuss how multimodal representations can contribute to discourse or dialogue structure representation. In Section 6 we summarise and indicate issues for future research.

2 Gesture annotation

In this work, multimodal communication is annotated by means of an annotation scheme (Allwood et al., 2007) where each modality is described by means of a list of attributes. The scheme is a general framework for the study of gestures in interpersonal communication that has been applied to multimodal video data in several languages. In order to circumvent the inherent difficulties related to describing the shape of gestures in formal terms, this is done in rather coarse-grained terms. Ex-

[1] We use *gesture* to mean non-verbal behaviour in general, not only hand gestures.

amples of shape annotation are "from down upwards" for a head movement, "away from interlocutor" for an eye movement, or "single-handed" for a hand gesture. The main purpose of the annotation is being able to distinguish different communicative functions rather than providing a precise description of the gestures. This is in line with the emerging standard for a functional markup language that is being developed for the generation of multimodal behaviour in robots and virtual agents (Heylen et al., 2008).

The functional annotation in MUMIN consists in a number of features relating to *feedback*, *turn management*, *sequencing* and *information structuring*. Only gestures that are deemed relevant to one of these phenomena are annotated.

Semiotic categories are also annotated for each gesture following Peirce (1931). The categories are the following: *indexical deictic* used for gestures pointing to some object in the conversation situation, *indexical non-deictic* assigned to gestures based on the result of a causal process, *iconic* assigned to gestures making use of similarity, *symbolic* characterising gestures making use of an arbitrary conventional relation.

For each gesture under consideration, a relation with the corresponding speech expression[2] is annotated following Poggi and Magno Caldognetto (1996), who propose the types *reinforcement*, *addition*, *substitution* and *contradiction*. Similar relations have been described in other proposals, e.g. in Martin (1999), where they are applied to cooperation between multimodal software agents.

The properties of the MUMIN schema and its application to data in several languages with satisfactory intercoder agreement have been described in (Allwood et al., 2007). It has also been shown how the transcribed data can be used to train machine learning algorithms to recognise some of the functions of multimodal behaviour (Jokinen et al., 2008; Jokinen and Ragni, 2007). The present study focuses on the issue of how to integrate the information provided by the gesture – as expressed through the annotation categories used in MUMIN – with the content of the linguistic sign. Understanding how this should be done is relatively straightforward in case a gesture seems clearly associated with a word, but this is by no means the only or even the most typical case. In fact, it doesn't seem possible to characterise the speech segment involved in a multimodal relation in uniform grammatical terms. We suggest, on the contrary, that different grammatical categories and different integration levels are involved.

[2]Here we assume that to correspond to each other, a speech and a gesture expression must overlap temporally.

3 Gestures and single words

In the simplest case, gestures coincide with single words or syllables. This is in general true of batonic gestures, a type of *indexical non-deictic* in the MUMIN scheme. Iconic hand gestures can also coincide with single words. Finally, there are also single gestures combining symbolic and indexical aspects which relate to isolated words. For example in our material, one of the dialogue participants smiles while saying *Tak* (Thanks). The gesture starts before and ends after the brief utterance. It is coded as a *feedback* gesture that reinforces the word it overlaps with. The semiotic type is *indexical non-deictic*.

The following excerpt shows the representation in the XML annotation produced by means of the ANVIL coding tool (Kipp, 2005):

```
<track name="SpeakerA.FacialDisplay" type="primary"\>
   <attribute name="Reinforcement">
   <value-link ref-track="SpeakerA.words" ref-index="0" />
   </attribute>
   <attribute name="FeedbackBasic">
    FeedbackGive
   </attribute>
   <attribute name="Face">
    Smile
   </attribute>
   <attribute name="SemioticType">
    IndexNon-deictic
   </attribute>
<track name="SpeakerA.words" type="primary">
   <el index="0" start="4.84459" end="5.11858">
   <attribute name="token">
    tak
   </attribute>
   </el>
```

A representation of this kind, while serving the intended practical purpose (annotating the actual multimodal interaction), is not the most concise way of modelling the multimodal behaviour. Previous proposals have suggested that feature structures are a convenient and elegant way of representing the unimodal content of each modality as well as their integration for instance for parsing purposes (Johnston et al., 1997; Paggio and Jongejan, 2005). We will then recast the XML code in feature structures terms. Our feature structures partly rely on Head-driven Phrase Structure Theory (HPSG) (Pollard and Sag, 1994) for the representation of the speech utterances, although our discussion is intended in very general terms rather than as a direct contribution to HPSG.

In Figure (1), then, the multimodal contribution is represented as a typed feature structure that

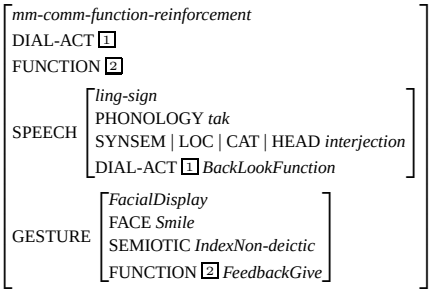

Figure 1: Feature structure representation of a feedback multimodal sign

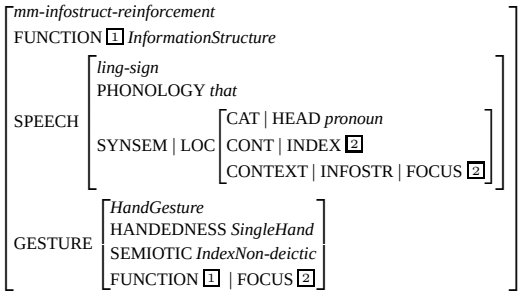

Figure 2: Feature structure representation of focus in a multimodal sign

includes information from both modalities. The attributes associated with the linguistic sign are a subset of those that the word would be given in HPSG. Since the word is also an utterance, we have added a dialogue act feature inspired by the DAMLS annotation system (Allen and Core, 1997). The attributes associated with the gesture are taken from the MUMIN categories. The numerical index means that the FUNCTION attributes of the gesture and the whole multimodal sign share the same value, i.e. *FeedbackGive*. The same is true of the DIAL-ACT feature, which is shared between linguistic and multimodal sign. In this case then, reinforcement should be understood in the sense that the communicative function of the gesture and the dialogue act expressed by the utterance are compatible and reinforce each other. Various reinforcement types can be defined based on the different values that these two attributes can take: in general, *BackwardLookingFunction* values in DAMLS correspond to *FeedbackGive* in MUMIN, and *ForwardLookingFunction* values correspond to *FeedbackElicit*.

While the cases in which a gesture is associated with a single word seem similar from the point of view of segmentation, they differ with respect to the conceptual level at which the multimodal relation applies. For batonic gestures, the level is that of information structure, or perhaps focus. In a constraint-based approach to information structure (Vallduví and Engdahl, 1996; Paggio, 2009), the multimodal relation could be represented in terms of structure sharing between the representation of the gesture and the information packaging features of the linguistic sign. For instance, in an example where a batonic gesture corresponds to the single accented word *det* (*that*), the representation could be as shown in Figure (2). Indices express structure sharing of two different features: the communicative function is still shared between gesture and multimodal sign; furthermore, the FOCUS attribute is structure-shared between the gesture, the semantic index of the linguistic expression and the focus value of its context.

In the case of iconic gestures, structure sharing would occur between the gesture and the content part of the corresponding linguistic expression. This should be done by adding a CONTENT attribute to the representation of the gesture and letting the value of this attribute be structure-shared with elements of the linguistic content. Thus, a different type of reinforcement is involved.

A relevant question here is how conventionalised the meaning of different iconic gestures is. We have already mentioned that several attempts, Kendon (2004) among others, have been made to describe classes of iconic gestures that share general characteristics both in terms of shape and meaning. Recently, Kipp et al. (2007) have argued, based on a proposal originally advanced by Schegloff (1984), that the content of iconic gestures can be expressed in terms of pre-defined categories of lexical meaning. The authors' iconic gesture lexicon consists of 35 entries including lexemes such as "cup", "wipe" and "progressive". The lexeme is the content part of the gesture annotation, and it is complemented by features concerning e.g. trajectory and amplitude.

For all three cases discussed so far, the gesture reinforces different parts of the linguistic sign. Gestures can also add meaning, for example by further specifying the meaning of the utterance (*addition*), or contradict what is said (*contradition*). While addition can be expressed in typed feature structures in terms of structure sharing between a type and a more specific subtype, contradiction is not as straightforward. In principle, it implies that the linguistic sign and the gesture

refer to disjoint content values. The last multimodal relation mentioned by Poggi and Magno Caldognetto (op.cit.) is *substitution*, which expresses the fact that the gesture stands alone: this can be modelled by letting the linguistic sign be empty.

4 Gestures and word sequences

Combinations of more complex hand gestures[3] and face displays are often associated with longer linguistic contributions that only rarely correspond to syntactic phrases. For instance, repeated nodding accompanied by intense gazing towards the speaker – again a feedback sign – may start in the middle of the speaker's utterance and continue up to a breathing pause. The speech transcription reads in one of our examples:

> *så vi ses %breath*
> See you then.
> (lit. "so we see(PASS)")

The utterance corresponds here to a sentence, so that a feature structure representation of the multimodal sign would include here the linguistic sign corresponding to the whole sentence, and otherwise be similar to the representation in Figure (1). Phrase structure information is not shown, but the feature structure can be conceived of as the top node of the syntactic tree corresponding to the sentence.

Turn holding gestures, where the speaker maybe slightly turns the head and looks away while finding the right words, are often more difficult to integrate in the linguistic representation, since they typically span over a speech sequence of varying size. The overlapping speech often starts with fillers like *og* (and), *ehm* and contains several word repetitions or self-repairs. From a syntactic point of view, these speech segments are sometimes but not always full syntactic phrases, since they also include chunks like verb groups, adjective lists, or fragments that get interrupted. In fact in some of these cases, the gesture also has a discourse resuming function, i.e. the speaker has made a false start, abandons the current line of discourse and goes on by resuming a preceding discourse segment.

An interesting question that merits further investigation on the basis of a larger corpus, is whether the non-verbal behaviour interacts with prosodic cues to segment the speech signal in utterances that do not necessarily correspond to grammatical units. Jensen (2003) argues that in Danish speech there is reasonable correspondance between syntactic units and prosodic units, although prosodic units often include additional elements such as interjections and discourse markers. This seems also true of the speech units that interact with gesture behaviour, and therefore the representation of multimodal signs should be able to accommodate fragmentary and 'noisy' utterances as well as phrases and sentences.

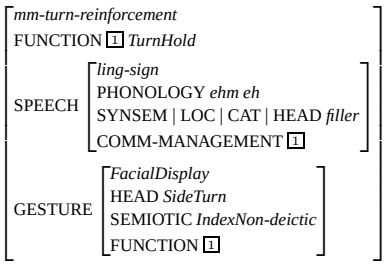

Figure 3: Feature structure representation of a turn holding multimodal sign

If the segmentation problem can be solved by making the definition of a grammatical sign more flexible, how should the turn management information provided by the gesture be expressed in a feature structure representation? The solution we propose here, shown in Figure (3), is to use the attribute FUNCTION to express the information coming from the gesture. Whether this is a reinforcement or an addition depends on whether the speech modality also provides communication management information (as would be the case if fillers like *ehm* or *eh* are used).

The last complex case we want to mention is that of sequences of batonic hand gestures, where several strokes in rapid succession accompany two or three stressed syllables within the same utterance, for example:

> *'kunne man kunne man jo 'godt mærke*
> One could, could ideed really feel.
> (lit. "COULD one could indeed REALLY feel")

The accented words are marked by an accent in the Danish text and written in small caps in the literal gloss. They are accompanied by two strokes of the hand. The utterance here spans over a grammatical sentence the two first words of which are

[3] In the literature also called gesture phrases, i.a (Kendon, 2004; Kipp, 2005).

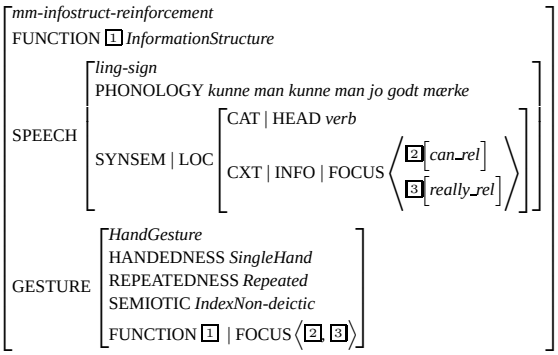

Figure 4: Multiple focus in a multimodal sign

repeated. The intonation clearly marks the sequence as a prosodic unit, and the two strokes come so quickly after each other that it seems reasonable to consider them as one complex gesture. However, the focus that they reinforce falls on two single words and not on the entire sequence. This is expressed in the feature structure in Figure (4) by letting the FOCUS attribute be a list of two indices, which correspond to the contents of the two accented words.

5 The contribution of gestures to discourse and dialogue structures

So far, we have seen how gesture and speech could be represented in an integrated fashion in feature structures that express syntactic, semantic and pragmatic features at the utterance level (from single words to more complex utterances). This could be referred to as the grammar of multimodal signs. However, it is also interesting to discuss how such multimodal signs can contribute to the representation of whole discourses or dialogues. This is of course a very complex issue. We can only hint at some of the relevant issues.

We have seen that feedback or turn managing gestures can be attached to words as well as longer speech sequences. The resulting multimodal sign plays a role at the level of dialogue acts and dialogue structure, i.a. (Traum and Hinkelman, 1992; Allen and Core, 1997). Provided that the feedback functions expressed by gestures are mapped onto the relevant dialogue acts (the specific repertoire depends on the theory one decides to adopt), the dialogue structure can then include multimodal representations on the same level as utterance representations. However, there are also numerous cases were gestures alone signal feedback and turn management. They should be included in the dialogue representation in the same way.

A final type of gesture we would like to discuss are discourse structuring gestures. Their contribution can be modelled in terms of discourse relations that make explicit how coherence between the various discourse parts is achieved. Discourse relations are formalised i.a. in Rhetorical Structure Theory (RST) (Mann and Thompson, 2007). For example, the *list* relation can be expressed by a multimodal sign. The speaker is explaining that there were many things she could not do when she was working at a film in prison:

> *jeg kunne ikke bare fise ud og gå mig en tur og få noget frisk luft hvis jeg skulle have lyst til det*
> I could not just dash out and take a walk and get some fresh air if I felt like it.

At the same time she marks the various items in the list by moving the right arm repeatedly from the center of the body to the right side. The function of the repeated gesture corresponds in MUMIN to a SEQUENCE attribute, and helps establish the corresponding rhetorical relation SEQUENCE in RST terms. The speaker stops moving her arm when the sequence is finished and she utters the hypothetic sentence *hvis jeg skulle have lyst til det* (if I felt like it) as a condition to the preceding list of actions (CONDITION rhetorical relation). The rhetorical structure for the example is in Figure 5.

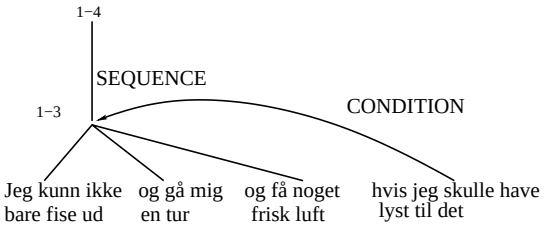

Figure 5: **RST** diagram

Linguistically, the example is quite complex, involving coordination, ellipsis and clausal modification. It can be observed, however, that the beginning of each arm movement in the complex gesture also marks the beginning of a list item. So the most obvious way of formalising the multimodal interaction seems that of binding the gesture to each of the conjuncts. The appropriate type would be *mm-sequence-reinforcement*.

6 Conclusion

We have discussed issues related to the segmentation of speech for multimodal annotation and the representation of the relation of gestures and speech in a multimodal sign. In particular we have shown, for a number of simple cases of interaction of gestures and speech, how this relation can be formalised in terms of feature structures in a unification-based formalism. These formalisations can be thought of the first fragments of a multimodal grammar. In addition, we have also touched on how the representations produced by such a grammar could be included in a discourse or dialogue model.

Although the examples we discuss are natural ones, taken from TV interviews, the empirical coverage of our grammar representations is extremely limited. Much more insight must come from the analysis and formalisation of more empirical data. However, interesting issues have already emerged. We have thus pointed out that gestures and speech can reinforce each other in different ways, and shown how the various reinforcement types can be represented. And we have indicated cases in which the interpretation of the multimodal sign fits well with well-known discourse and dialogue models. Other issues – e.g. how to cope with contradiction, or how to account for the interaction of gestures and prosody for speech segmentation purposes – we have left open.

An additional complexity is the fact that gestures are often multifunctional and can belong to several semiotic categories at the same time. In our data we have a number of examples in which batonic gestures also display iconic properties, or in which feedback gestures also play a role in the turn management system. An issue we want to investigate in future is how to represent such complex cases.

References

James F. Allen and Mark G. Core. 1997. *Draft of DAMSL: Dialog Annotation Markup in Several Layers*. The Multiparty Discourse Group. University of Rochester, Rochester, USA.

Jens Allwood, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In J.C. Martin, et al (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation, 41(3–4), 273–287. Springer.

Jens Allwood 2008. Dimensions of Embodied Communication - towards a typology of embodied communication. In Ipke Wachsmuth, Manuela Lenzen, Gnther Knoblich (eds) *Embodied Communication in Humans and Machines*. Oxford University Press.

Michael A. Arbib 2005. From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics *Behavioral and Brain Sciences*, 28, 105–124. Cambridge University Press

Dirk Heylen, Stefan Kopp, Stacy C. Marsella, Catherine Pelachaud and Hannes Vilhjálmsson. 2008. The Next Step towards a Function Markup Language. In H. Prendinger, J. Lester, and M. Ishizuka (eds.) *IVA 2008, LNAI 5208*, pp. 270-280, 2008. Springer-Verlag, Berlin Heidelberg.

Anne K. Jensen 2003. *Clause Linkage in Spoken Danish* PhD Dissertation. Department of General and Applied Linguistics. University of Copenhagen.

Michael Johnston, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman and Ira Smith 1997. Unification-based Multimodal Integration. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, 1997, pp. 281–288.

Kristiina Jokinen and Anton Ragni. 2007. Clustering experiments on the communicative properties of gaze and gestures. In *Proceeding of the 3rd. Baltic Conference on Human Language Technologies*. Kaunas.

Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. 2008. Distinguishing the communicative functions of gestures. In *Proceedings of the 5th Joint Workshop on Machine Learning and Multimodal Interaction* 8-10 September 2008, Utrecht, The Netherlands. Springer LNCS 5237, pp. 38–49.

Adam Kendon. 2004. *Gesture*. Cambridge.

Michael Kipp. 2005. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Boca Raton, Florida: Dissertation.com

Michael Kipp, Michael Neff and Irene Albrecht. 2007. *An annotation scheme for conversational gestures*. In Martin, J.C. et al (eds) *Multimodal Corpora for Modelling Human Multimodal Behaviour*. Special issue of the International Journal of Language Resources and Evaluation, 41(3–4). Springer.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical Structure Theory: Description and Construction of Text Structures, *in* G. Kempen, ed., 'Natural Language Generation', number 135. In *'NATO ASI'*, Martinus Nijhoff Publishers, pp. 85–95.

Jean-Claude Martin. 1999. *TYCOON: six primitive types of cooperation for observing, evaluating and specifying cooperations.*. In *Proceedings of AAAI Fall 1999 Symposium on Psychological Models of Communication in Collaborative Systems*.

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

Patrizia Paggio. 2009. The information structure of Danish grammar constructions. Nordic Journal of Linguistics (in press).

Patrizia Paggio and Bart Jongejan. 2005 Multimodal Communication in Virtual Environments: Communicating with the Staging virtual farm. In O. Stock and M. Zancanaro (eds) *Multimodal In-telligent Information Presentation*, Kluwer Academic Publishers, pp.27–47. ISBN: 1-4020-3051-7.

Charles S. Peirce. 1931. *Elements of Logic.* Collected Papers of Charles Sanders Peirce. Volume Two. Hartshorne, C. & Weiss, P. editors Cambridge: Harvard University Press.

Isabella Poggi and Emanuela Magno Caldognetto. 1996. A score for the analysis of gestures in multimodal communication. In *Proceedings of the Workshop on the Integration of Gesture and Language in Speech*. Applied Science and Engineering Laboratories. L. Messing, Newark and Wilmington, Del, pp. 235–244.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: The University of Chicago Press.

Emanuel Schegloff. On some gestures' relation to talk. In J. M. Atkinson and J. Heritage (eds.) *Structures of Social Action*, 266–298. Cambridge University Press.

David R. Traum and Elizabeth A. Hinkelman. Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence*, 8:575-599.

Enric Vallduví and Elisabeth Engdahl. 1996. The linguistic realisation of information packaging. *Linguistics*, 34(33), 459–519, de Gruyter, 1996.