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

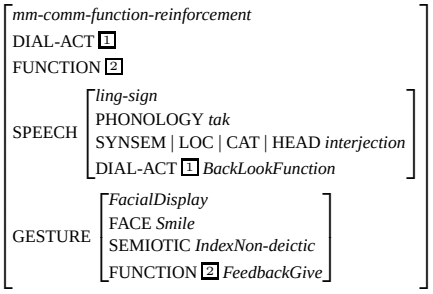

Figure 1: Feature structure representation of a feedback multimodal sign

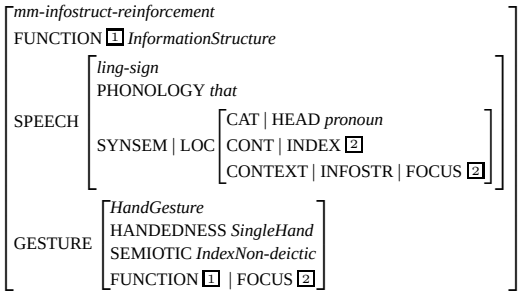

Figure 2: Feature structure representation of focus in a multimodal sign

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.

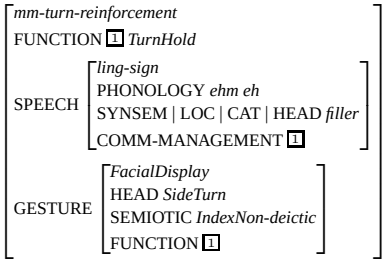

Figure 3: Feature structure representation of a turn holding multimodal sign

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).

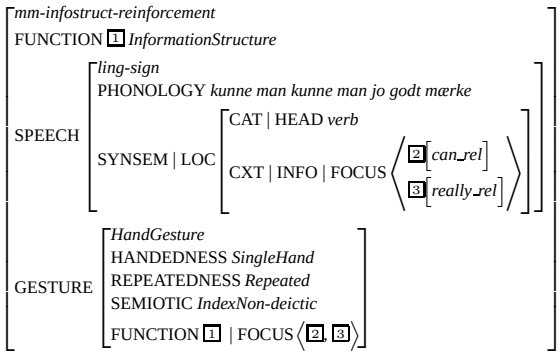

Figure 4: Multiple focus in a multimodal sign

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.

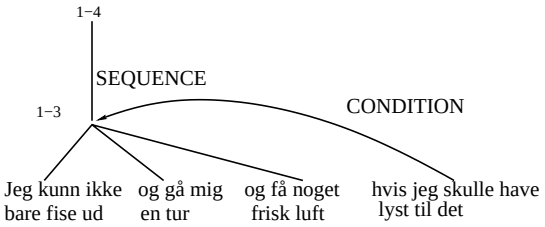

Figure 5: **RST** diagram

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.