Annotation Schemes for Verbal and Non-verbal Communication: Some General Issues

Niels Ole Bernsen and Laila Dybkjær

NISLab, Denmark
nob@nis.sdu.dk, laila@nis.sdu.dk

Abstract

During the past 5-10 years, increasing efforts have been put into annotation of verbal and non-verbal human-human and human-machine communication in order to better understand the complexities of multimodal communication and model them in computers. This has helped highlight the huge challenges which still confront annotators in this field, from conceptual confusion through lacking or immature coding schemes to inadequate coding tools. We discuss what is an annotation scheme, briefly review previous work on annotation schemes and tools, describe current trends, and discuss challenges ahead.

Introduction

Few, if any, of us actually code many different aspects of verbal and non-verbal human-human or human-machine communication on a daily basis. Rather, we tend to be occupied for long stretches of time annotating a single aspect of a single modality, such as when doing spoken dialogue transcription, or, increasingly, annotating a single aspect, such as emotion expression, across a range of modalities. Data coding tends to be hard work, and difficult, too. One often has to first design and create the data resource to be used before having something appropriate to code, possibly after having spent considerable time looking for re-usable data without finding any. As existing coding schemes often turn out to be inappropriate for the purpose at hand, coding scheme creation might follow, which is often hard theoretical work and for which, moreover, a single data resource is rarely sufficient for creating a new consolidated coding scheme. And coding tools constitute a world of their own, with learning-how-to-use difficulties, programming challenges and sometimes tool inadequacy for what one wants to do. It is tempting to think that things are easier for coders of other types of verbal and non-verbal communication phenomena than one's own and that their world is far more well-organised conceptually. Only an attempt to take a global look can contribute to balancing the picture and provide a common view of what it is that we are all involved in as explorers of the only partially charted land of verbal and non-verbal communication.

In this paper we look at previous and current work on annotation and provide a glimpse of what lies ahead. Section 2 seeks to establish common ground by describing what is a coding scheme and defining the notions of general and consolidated coding schemes. Section 3 briefly refers back to previous work on creating surveys of data, coding schemes, and coding tools for natural interactive communication, and Section 4 addresses current trends in the field. Section 5 discusses future challenges and concludes the paper.

2 What Is A Coding Scheme?

In the context of coding verbal and non-verbal communication, a coding (annotation, markup) scheme is basically a theory of the members (types) of a class of phenomena (tokens) to be found in the data. The data itself may be represented in acoustic – speech and other – files, video files, logfiles, hand-written notes or otherwise. A coding scheme may be based on, or has to be able to support the annotation of, one or several data sets, data resources, or corpora. Within the wealth of information represented in the data, a coding scheme focuses on a single generic kind of information, such as the facial expressions of the participant(s), the parts-of-speech they produce, or the behavioural cues to their emotions whether expressed in speech, facially, in gesture or otherwise. In fact, these three examples, although perfectly legitimate, are far too neat to adequately convey what a coding scheme might be targeting, so let's also include examples, such as nose scratchings, looking carefully around to see if anybody is watching, or increasing heart rate because of sensing danger. You might object that these behaviours, although non-verbal all right, do not constitute communication, but see Section 5. The point we wish to make is that the generic kind of information targeted by a coding scheme solely reflects the scheme's underlying coding purpose, which is why such generic kinds of information are unlimited in number. Quite simply, there is an unlimited number of coding purposes one might have when coding a particular data resource.

To be useful, a coding scheme should include three kinds of information which we might call theory, semantics, and meta-data, respectively. These are discussed in the following Sections 2.1, 2.2 and 2.3.

2.1 Theory and Completeness

The first kind of information a coding scheme should include is a theory of the number and nature of the types of phenomena, relevant to the coding purpose, to be found in the data. If that theory is wrong, so that the data includes more or other types of relevant phenomena than those acknowledged by the coding scheme, more types will have to be added to the scheme. This is a perfectly normal situation for the originator or co-developer of an emerging coding scheme: you approach the data with a theory of the number and nature of the phenomena it includes, discover that there are more, other, or even sometimes fewer types than hypothesised, and revise the coding scheme accordingly. By the same token, however, the coding scheme represents a theory under development and the coding scheme is not yet, at least, a consolidated one.

We use the word "theory" above but "theories" may, in fact, be of two different kinds. The first kind is a scientific theory or hypothesis which aims to categorise all possible types of phenomena of a particular kind as determined by the coding purpose, such as all phonemes in a particular language. We call coding schemes based on a scientific theory general coding schemes, whether consolidated or not. The second kind of theory is a pragmatic theory or hypothesis which merely aims to be complete in the sense of capturing all phenomena that happen to be relevant for a given coding purpose. Since a coding purpose may be nearly anything, such as the speech acts people generally use to agree on meeting dates and times [Alexandersson et al. 1998], or the speech and pointing gesture combinations used to manipulate 2D geometrical shapes [Landragin 2006], the theory underlying coding purposes such as these might not stake any claim to scientific generality or depth of justification – at least not unless or until backed by deeper theory which might explain why these and only these types of phenomena could be used for some purpose. Admittedly, the distinction between scientific and pragmatic theory is thin in some cases. For instance, no existing scientific theory probably explains why English has exactly the set of phonemes it has. But at least our knowledge about English phonemes constitutes a stable scientific generalisation which can be applied in many different contexts. However, no matter which kind of theory is involved, coding aims at completeness relative to coding purpose.

2.2 Coding Scheme Semantics, Criteria

The second kind of information which must be included in a coding scheme is a set of criteria according to which each phenomenon (or each token) in the data can be determined to belong to a particular type among the types of phenomena acknowledged by the coding scheme. These criteria should be made perfectly explicit, clear, and unambiguous as part of the coding scheme representation. This is done by describing criteria for deciding to which type any token belongs and providing useful examples of tokens of each type. Otherwise, coding scheme users will have difficulty applying the coding scheme consistently and in the same way across coders because they will be missing guidance on how to classify the phenomena observed in the data. Coding scheme semantics development is hard work and cannot be done too well.

2.3 Meta-data

The third kind of coding scheme information is meta-data information on the scheme itself. There is no general standard for such meta-data although various initiatives are working towards standardisation, such as the Dublin Core Metadata Initiative (http://-dublincore.org) and the Open Language Archives Community (OLAC) (http://www.-language-archives.org/OLAC/metadata.html). However, it is easy to illustrate the kinds of meta-data that are normally required as well as which additional kinds might be needed in a particular case: What is the coding purpose? Is that a rather unique purpose or could the coding scheme be used more generally, for which other purposes, for instance? Who created the scheme? When? Using which corpora? How well-

tested is it, i.e., on how many and/or which corpora has it been applied and with which results? How reliable is it, has inter-coder agreement been measured and with which results? How difficult is the coding scheme to use, are there any specific problems that should be mentioned, how is it applied in coding practice, how much training/domain experience does it require, are codings from two independent coders needed for obtaining reasonably reliable results? Has annotation based on the coding scheme been automated and with which results compared to human coders? Is the coding scheme underpinned by scientific theory, which theory? How (well) is the scheme documented? How can it be accessed, i.e., at which Internet site, by emailing who, is it for-free, are there any conditions on its use? Whom to contact with questions about the coding scheme? Are there any coding tools that could be used? Are coded corpora available, how, under which conditions? Etc.

2.4 Consolidated Coding Schemes

We can now define a consolidated coding scheme. A consolidated coding scheme is one which has been proved reliable for coding a representative variety of corpora under reasonably achievable conditions to be stated, such as coder experience and training, coding procedure, generic kind of corpora, etc. A consolidated coding scheme may or may not be underpinned by deep scientific theory. It may also have problems, such a inherent difficulties in classifying tokens of particular types, as long as these are well described in the coding manual. In other words, we cannot require, at this stage of coding verbal and non-verbal communication, that coding schemes termed 'consolidated' are perfect in all respects.

Interestingly, the fact that a coding scheme can be underpinned by scientific theory does not, by itself, guarantee that the coding scheme is a consolidated one. Data coding may constitute a hard test of the theory underlying the scheme. Scientific theories themselves need justification and they sometimes compete in accounting for phenomena in a particular field. Attempts at data coding based on each of them may contribute to selecting the theory which best accounts for the data. We saw that ourselves some years ago when we developed a coding scheme for communication problems in spoken dialogue. Having done that, we compared the results with Grice's theory of conversational implicature and its typology of cooperativity issues that may arise in spoken dialogue [Grice 1975]. In the literature at the time, the scope of Grice's theory had been subject to various proposed reductions but none of the critics had raised serious doubt with respect to the theory's validity for human-human shared-goal dialogue, i.e., dialogue in which the interlocutors try to cooperatively solve a problem. Nonetheless, we found that Grice's theory had to be extended in order to account for the types of phenomena which we found in our data corpora from human-computer shared-goal dialogue [Bernsen et al. 1996].

Despite the possible imperfections of consolidated coding schemes, it is a great advantage for the coder to use a consolidated coding scheme which comes with the three kinds of information described above. The advantage is that you can simply follow the coding manual and code the data in the expectation that that's it. The alternative of using an unconsolidated coding scheme may carry a range of implications depending on what's in the data. At the very least, the coding task becomes the double one

of (i) coding the data and (ii) testing the coding scheme. If the test turns out reasonably well, you will have accomplished two things, i.e., coded your data and contributed, however slightly, to making the coding scheme a consolidated one, possibly contributing useful observations for its coding manual as well. But if the test fails, for instance because a large fraction of the phenomena in your corpus cannot be coded using the scheme, you are left with no coded data and the choice of whether to (iii) look for an alternative coding scheme that might work better, (iv) become a coding scheme co-developer who tries to extend the scheme to cover your corpus, (v) try to develop an alternative coding scheme from scratch, or give up coding the data, which may not be an option because other work depends on the planned annotation.

However, in order to use an existing coding scheme – consolidated or not – you need to find it first, which may not be easy since there are no catalogues available.

3 Previous Work

Some years ago, we were involved in carrying out global surveys of natural interactivity data, coding schemes and coding tools in EU-projects MATE, NITE and ISLE. MATE made a survey of annotation schemes for aspects of spoken dialogue, e.g., prosody and dialogue acts [Klein et al. 1998]. NITE described a number of gesture, facial expression and cross-modality schemes [Serenari et al. 2002], drawing heavily on ISLE which had reviewed 21 different coding schemes of which 7 concerned facial expression possibly combined with speech, and 14 concerned gesture possibly accompanied by speech [Knudsen et al. 2002a]. In two other reports, ISLE reviewed multimodal data resources [Knudsen et al. 2002b] and coding tools [Dybkjær et al. 2001].

In the period since around the turn of the century, others have looked at verbal and non-verbal communication coding schemes and tools as well. Some did this as part of comparing their own coding scheme to the state of the art or related schemes, e.g., [Martell 2005], or while looking for a tool to use, e.g., [Garg et al. 2004]. Others did it as part of surveying multimodality and natural interaction without specifically focusing on annotation [Gibbon et al. 2000]. Other examples are the following. Until around 2002 the Linguistic Data Consortium (LDC) maintained a web page with brief descriptions of linguistic annotation schemes and tools (http://www.ldc.upenn.edu/-annotation/). Michael Kipp, the developer of the Anvil multimodal annotation tool, maintains a page (http://www.dfki.de/~kipp/anvil/users.html) listing users of Anvil. This list mentions various coding schemes which are being applied using Anvil.

To our knowledge, however, there has not been any large-scale initiative in surveying multimodal and natural interaction annotation schemes since ISLE. Maybe the task has simply grown too complex as will be discussed in the next section.

4 Current Trends

While MATE looked at aspects of spoken dialogue annotation, ISLE focused on gesture-only annotation, gesture combined with speech, facial expression-only and

facial expression combined with speech. Multimodal annotation, more generally, is a vast area. An interest in any combination of two or more communication modalities requires a multimodal annotation scheme or some cross-modal annotation to see the interactions between the modalities. This adds up to very many possible combinations, such as, e.g., speech and hand gesture, head and eye brow movements, lip movements and speech, gaze and speech, speech, body posture and facial expression, to mention but a few, and it would take considerable effort to compile an overview of the annotation schemes that have been proposed in recent years for all possible combinations, especially since activity in the field would seem to continue to increase. We discuss the increasing activity and some project examples in the following where we also briefly mention consolidation and standardisation efforts.

4.1 Increasing Coding Activity

Since natural interactivity and multimodality gained popularity and became buzzwords in the late 1990s, many initiatives have addressed the construction of increasingly sophisticated systems incorporating various aspects of human communication. This typically requires data resources and annotation of phenomena which in many cases have not been studied in great detail before, implying a strong need for new coding schemes with a heavy emphasis on multimodal or cross-modal markup.

4.2 Project Examples

In recent years, several large-scale projects have been launched in focused areas of natural interactivity and multimodality, such as emotion or multi-party interaction. We will look at multimodal corpus annotation work done in a couple of these projects and stress that several other projects could have been mentioned instead.

The European HUMAINE Network addresses emotion and human-machine interaction (http://emotion-research.net/). Researchers in the network have proposed EARL (http://emotion-research.net/earl, the HUMAINE Emotion Annotation and Representation Language), an XML-based language for representing and annotating emotions. The language is aimed for use in corpus annotation as well as for recognising and generating emotions. Figure 1 shows an example of audio-visual annotation from the EARL website. The annotation can be done using, e.g., Anvil (Section 3).

```
<emotion category="pleasure" probability="0.4" start="0.5"
end="1.02"/>
<emotion modality="voice" category="pleasure" probability="0.9"
start="0.5" end="1.02"/>
<emotion modality="face" category="neutral" probability="0.5"
start="0" end="2"/>
<emotion modality="text" probability="0.4" start="0.5" end="1.02"
arousal="-0.5" valence="0.1"/>
```

Figure 1. EARL markup.

Face-to-face communication is multimodal and may include emotions in one or several participants. Magno Caldognetto et al. [2004] use – within the framework of three different projects - the Multimodal Score annotation scheme implemented in Anvil to synchronously mark up speech, prosody, gesture, facial (mouth, gaze, eyes, eyebrows), and head and body posture in order to facilitate analysis of cross-modal interactions. The investigation aims at better understanding the elements of human communication. Figure 2 only shows part of this enormous coding representation which, in fact, represents several dozens of coding schemes at various stages of development combined into a single coding representation. The top tier shows the common timeline followed by three tiers presenting the speech signal, the words spoken and their segmentation. Then follows the pitch and intensity aspects of prosody (5 tiers each). Since the right hand does nothing, this tier is greyed out whereas the left hand's behaviour is described in 7 tiers, the last of which relates the gesture to what is being spoken at the same time. The gesture type (Tier 2) is labelled "other" which is typical of coding schemes under development which still lack complete semantics. The codings in Figure 2 provide a glimpse of the huge complexity of future codings of human-human and human-machine communication.

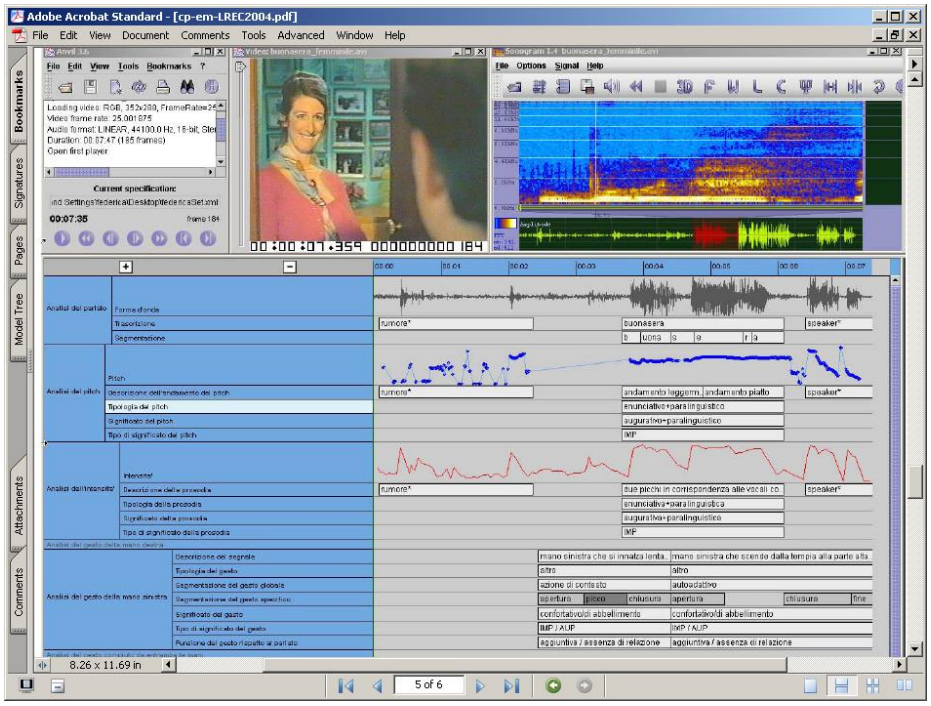

Figure 2. Multimodal Score annotation in Anvil.

The European AMI (Augmented Multiparty Interaction) project (http://www.amiproject.org) is one among several large projects in the area of multi-party meeting interaction. One project result is the AMI video Meeting Corpus which consists of 100

hours of meeting recordings. The corpus has been orthographically transcribed and annotated with dialogue acts, topic segmentation, extractive and abstractive summaries, named entities, the types of head gesture, hand gesture, and gaze direction that are most related to communicative intention, movement around the room, emotional state, and where heads are located in the video frames. Markup has been done using the NITE XML toolkit (http://www.ltg.ed.ac.uk/NITE/). Figure 3 shows a screenshot of the coding representation.

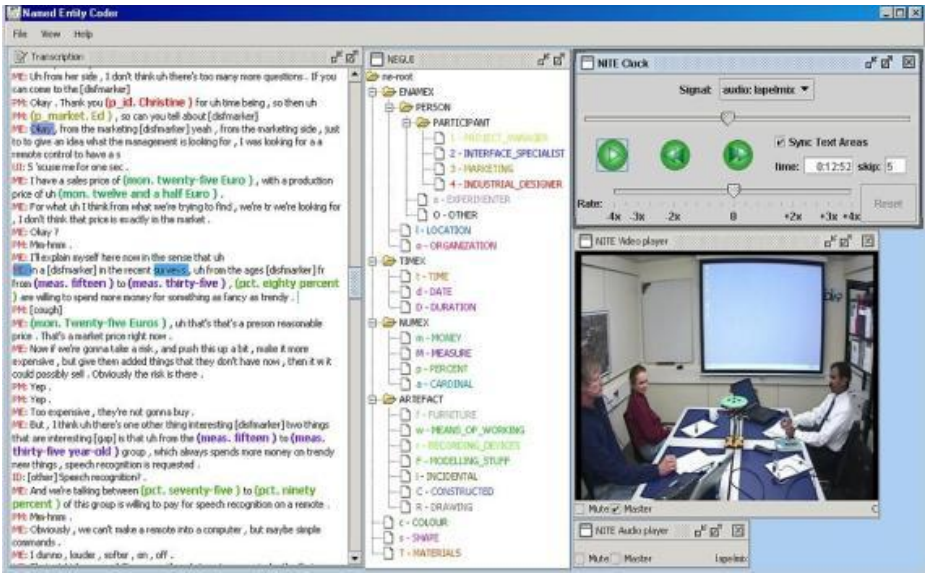

Figure 3. AMI corpus spoken dialogue annotation with NITE XML toolkit.

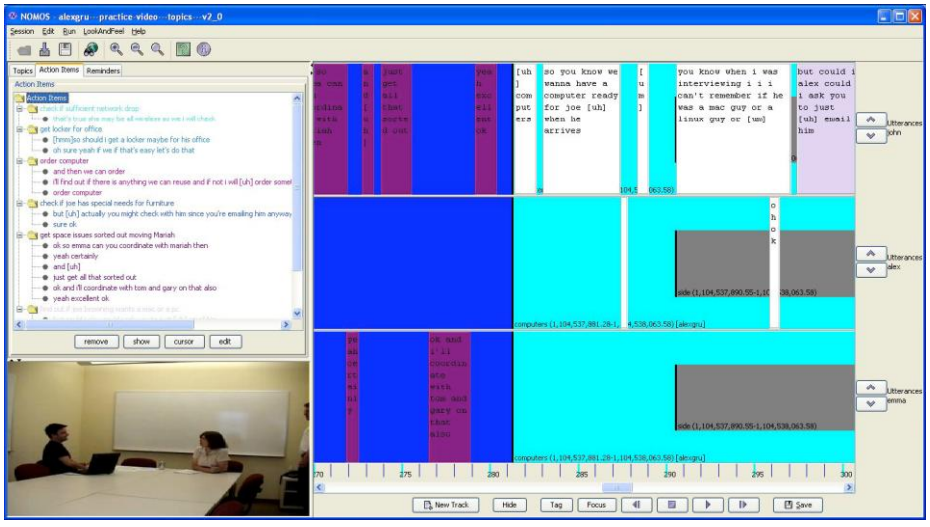

Figure 4. Meeting topic segmentation annotation with Nomos.

A similar corpus is the ICSI Meeting Corpus collected at International Computer Science Institute at Berkeley, CA, which contains 75 meeting recordings (audio and video). The audio has been transcribed at word level and the transcription is distributed along with the corpus (http://www.idiap.ch/mmm/corpora/icsi). The corpus has been used by several people who have annotated various phenomena, such as hierarchical topic segmentation and action items, see, e.g., [Gruenstein et al. 2005] who have used the Nomos annotation software (Figure 4) [Niekrasz 2006].

What this small list of projects illustrates is that (i) corpus annotation groundwork is going on in order to better understand multimodal and cross-modal aspects of human communication; (ii) annotation tools are highly desirable for supporting the annotation process; and (iii) annotation schemes for verbal and non-verbal communication are still often at an exploratory stage although steps are being taken towards standardisation, common formats, and consolidation as briefly discussed next.

4.3 Towards Consolidation and Standards

For most multimodal and natural interactive areas there are no standards and few consolidated annotation schemes. At the same time it is acknowledged that consolidated coding schemes, standardisation, and common formats could significantly facilitate analysis and data reuse. The problem is that it is not an easy task to consolidate annotation schemes in the areas we are talking about but there are ongoing attempts in this direction. An example is the W3C incubator group on emotions (Emotion XG) (http://www.w3.org/2005/Incubator/emotion/) proposed by the HUMAINE Network (Section 4.2). As there is no standard annotation scheme or markup language for emotions, the purpose of the Emotion XG is to "discuss and propose scientifically valid representations of those aspects of emotional states that appear to be relevant for a number of use cases. The group will condense these considerations into a formal draft specification" for an emotion annotation and representation language. Clearly, the scope of the planned result will very much depend of the collective representativity of emotional behaviour in general of the use cases selected.

Another example is the International Standards Organisation (ISO) TC37/SC4 group on Language Resources Management (http://www.tc37sc4.org). Focus is on language resources and aspects of their standardisation. To this end, the Linguistic Annotation Framework (LAF) has been established [Ide and Romary 2007]. It aims to provide a standard infrastructure for representing language resources and their annotation. The underlying abstract data model builds on a clear separation of structure and contents. The goal is to achieve an internationally accepted standard that will enable far more flexible use, reuse, comparison, and evaluation of language resources than is the case today.

It is worth noting that, in both cases just mentioned, the aim is a theoretically well-founded, consolidated or even standardised representation and annotation language rather than a particular coding scheme with a fixed set of tags. We agree that, in many cases, this is the right level of abstraction to aim for at this stage given that (i) theoretically complete coding schemes are still a long way off in many areas of multimodal annotation of verbal and non-verbal communication, and (ii) in some cases completeness is not even theoretically feasible because of the open-ended nature of what is

being coded, such as human action or iconic gesture. Common formats will facilitate the construction and use of common tools and the reuse/further use of existing data resources never mind the theoretical completeness of the coding schemes supported.

5 Future Challenges

We have discussed the notion of an annotation scheme, briefly presented previous work on annotation schemes and tools, and discussed current trends in coding verbal and non-verbal communication. The work described suggests that we are to a great extent exploring new land where general and/or consolidated coding schemes often do not exist. However, as we have said far too little about what lies ahead we will try to add some more glimpses in the following.

At first glance, the question of what we annotate when coding verbal and non-verbal communication might appear to have a rather straightforward answer: we code all the different kinds of observable behaviour which humans use to communicate intended meaning to other humans and/or machines, including speech, facial expression and gaze, gesture, head and body posture, and body action as part of the communication. However, this answer is radically incomplete because (i) humans communicate more than deliberately intended meaning and (ii) machines are capable of perceiving information that humans cannot perceive. For instance, (i) our voice may unintentionally reveal our mood, or (ii) bio-sensing is becoming an important source of information for computers during interaction. Moreover, (iii) one-way "communication" is common among humans and is emerging between humans and machines as well, such as in surveillance and friendly observation aimed at learning more about the user. In Figure 5 from [Bernsen and Dybkjær, in press], we replace "communication" by the more inclusive "information presentation and exchange" and propose a taxonomy of the many different types of the latter which annotators may have to deal with.

A second way in which to put into perspective future challenges in annotating verbal and non-verbal information presentation and exchange is to consider the media and modalities involved. Modality theory [Bernsen 2002, Bernsen and Dybkjær, in press] provides an exhaustive taxonomy of the large numbers of possible modalities in the three media of light/vision, sound/hearing or audition, and mechanical impact/touch sensing or haptics. Basically, they are all relevant to annotation and their combinatorics is staggering, as Figure 2 is beginning to illustrate. Bio-sensing is becoming important as well, and even smell (olfaction) and taste (gustation) should be kept in mind even though they are not (yet) being much used by machines and normally don't play any significant role in human-human information exchange.

We need a third typology as well, orthogonal to the one in Figure 5 and to the modality taxonomy, which describes the different possible levels of annotation from low-level, non-semantic, such as phonemes and mouth shapes, through non-semantic structures, such as the phases of some types of gesture, to basic semantics, such as words or smiles, composite semantics, cross-modal semantic combinations, and the semantics of global personal states, such as emotion or cognition, see [Bernsen and Dybkjær, in press] for a proposal. In addition, there is a strong need for standardised concepts and terminology as even basic terms like 'gesture' have no agreed definition.

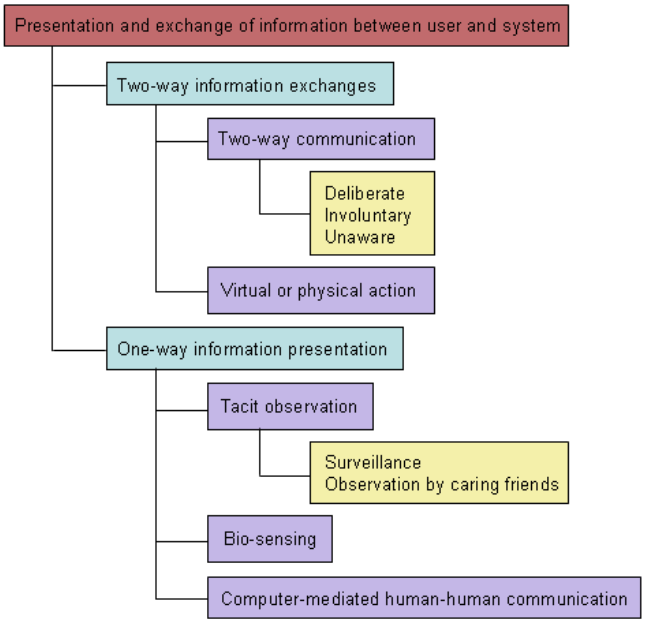

Figure 5. Taxonomy of information representation and exchange.

Although very different, all three typologies just mentioned as well as the fact that even basic terms lack common definitions, suggest the same conclusion. It is that there is a long way to go before we have anything like a comprehensive and systematic grasp of how to annotate full human-human and human-machine presentation and exchange of information in context, and before we have general and consolidated coding schemes for more than a small fraction of what humans do when they communicate and observe one another during communication.

Acknowledgements

This paper was written as part of the collaboration in COST Action 2102 Cross-Modal Analysis of Verbal and Non-verbal Communication (CAVeNC). We gratefully acknowledge the support.

References

Alexandersson, J., Buschbeck-Wolf, B., Fujinami, T., Kipp, M., Koch, S., Maier, E., Reithinger, N., Schmitz, B. and Siegel, M.: Dialogue Acts in VERBMOBIL-2, Second Edition. Report 226, Saarbrücken, Germany, 1998.

Bernsen, N. O.: Multimodality in Language and Speech Systems - From Theory to Design Support Tool. In B. Granström, D. House and I. Karlsson (Eds.): Multimodality in Language and Speech Systems. Dordrecht: Kluwer Academic Publishers, 2002, 93-148.

Bernsen, N. O., Dybkjær, H. and Dybkjær, L.: Cooperativity in Human-Machine and Human-Human Spoken Dialogue. Discourse Processes, Vol. 21, No. 2, 1996, 213-236.

Bernsen, N. O. and Dybkjær, L.: Multimodal Usability. To appear.

Dybkjær, L., Berman, S., Kipp, M., Olsen, M.W., Pirrelli, V., Reithinger, N. and Soria, C.: Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data. ISLE Deliverable D11.1, January 2001.

Garg, S., Martinovski, B., Robinson, S., Stephan, J., Tetreault, J. and Traum, D.: Evaluation of Transcription and Annotation Tools for a Multi-modal, Multi-party Dialogue Corpus. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC), 2004, 2163-2166.

Gibbon, D., Mertins, I. and Moore, R. (Eds.): Handbook of Multimodal and Spoken Dialogue Systems. Kluwer Academic Publishers, 2000.

Grice, P.: Logic and conversation. In P. Cole and J. L. Morgan (Eds.), Syntax and Semantics Vol. 3: Speech Acts. New York: Academic Press, 1975, 41-58. Reprinted in Grice, P.: Studies in the Way of Words. Cambridge, MA, Harvard University Press, 1989.

Gruenstein, A., Niekrasz J. and Purver, M.: Meeting Structure Annotation: Data and Tools. Proceedings of the Sixth SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal, 2005, 117-127.

Ide, N. and Romary, L.: Towards International Standards for Language Resources. In L. Dybkjær, H. Hemsen and W. Minker (Eds.): Evaluation of Text and Speech Systems, Springer, Text, Speech and Language Technology Series, Vol. 37, 2007, 263-284.

Klein, M., Bernsen, N.O., Davies, S., Dybkjær, L., Garrido, J., Kasch, H., Mengel, A., Pirrelli, V., Poesio, M., Quazza, S. and Soria, C.: Supported Coding Schemes. MATE Deliverable D1.1, July 1998.

Knudsen, M. W., Martin, J.-C., Dybkjær, L., Ayuso, M. J. M, N., Bernsen, N. O., Carletta, J., Kita, S., Heid, U., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van ElsWijk, G. and Wittenburg, P.: Survey of Multimodal Annotation Schemes and Best Practice. ISLE Deliverable D9.1, 2002a.

Knudsen, M. W., Martin, J.-C., Dybkjær, L., Berman, S., Bernsen, N. O., Choukri, K., Heid, U., Mapelli, V., Pelachaud, C., Poggi, I., van ElsWijk, G. and Wittenburg, P.: Survey of NIMM Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources. ISLE Deliverable D8.1, 2002b.

Landragin, F.: Visual Perception, Language and Gesture: A Model for their Understanding in Multimodal Dialogue Systems. Signal Processing 86(12), 2006, 3578-3595.

Magno Caldognetto, E., Poggi, I., Cosi, P., Cavicchio, F. and Merola G.: Multimodal Score: An ANVIL Based Annotation Scheme for Multimodal Audio-Video Analysis. Proceedings of LREC Workshop on Multimodal Corpora, Models of Human Behaviour for the Specification and Evaluation of Multimodal Input and Output Interfaces, Lisbon, Portugal, 2004, 29-33.

Martell, C.: FORM. In van Kuppevelt, J., Dybkjær, L. and Bernsen, N. O. (Eds.): Advances in Natural Multimodal Dialogue Systems. Springer. Series: Text, Speech and Language Technology, Vol. 30, 2005, 79-95.

Niekrasz, J.: NOMOS: A Semantic Web Software Framework for Multimodal Corpus Annotation. Demonstration Session Guide of the 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI), Washington DC, USA, 2006, 11.

Serenari, M., Dybkjær, L., Heid, U., Kipp, M., and Reithinger, N.: Survey of Existing Gesture, Facial Expression, and Cross-Modality Coding Schemes. NITE Deliverable D2.1, September 2002.