Project ref. no.	IST-1999-10647
Project title	ISLE Natural Interactivity and Multimodality Working Group

Deliverable status	Public
Contractual date of delivery	30 September 2002
Actual date of delivery	14 February 2003
Deliverable number	D9.2
Deliverable title	Guidelines for the Creation of NIMM Annotation Schemes
Type	Report
Status & version	Final
Number of pages	36
WP contributing to the deliverable	WP9
WP / Task responsible	NISLab
Editors	Laila Dybkjær and Niels Ole Bernsen
Authors	Laila Dybkjær, Niels Ole Bernsen, Malene Wegener Knudsen, Joaquim Llisterri, María Machuca, Jean-Claude Martin, Catherine Pelachaud, Montse Riera, Peter Wittenburg
EC Project Officer	Philippe Gelin
Keywords	Natural interactivity, multimodality, annotation schemes, creation guidelines
Abstract (for dissemination)	This ISLE Report 9.2 from the ISLE Natural Interactivity and Multimodality (NIMM) Working Group reviews the current state of the art in NIMM coding schemes, summarising the conclusions in ISLE Report D9.1 which surveys NIMM annotation schemes and best practice, and discusses work on standardisation by describing a series of activities world-wide which share the aim of influencing NIMM coding scheme-related standardisation. Finally, the report discusses and presents recommendations and guidelines for NIMM coding scheme creation.

ISLE Natural Interactivity and Multimodality Working Group Report D9.2

Guidelines for the Creation of NIMM Annotation Schemes

February 2003

Authors

Laila Dybkjær[1], Niels Ole Bernsen[1], Malene Wegener Knudsen[1], Joaquim Llisterri[2], María Machuca[2], Jean-Claude Martin[3], Catherine Pelachaud[4], Montse Riera[2], Peter Wittenburg[5]

1: NISLab, University of Southern Denmark. 2. DFE:, Barcelona, Spain. 3: LIMSI-CNRS, Orsay, France.

4: DIS, University of Rome, Italy. 5: MPI, Nijmegen, The Netherlands.

Contents

1. Introduction

This document, ISLE NIMM (Natural Interactivity and Multimodality) Working Group Report D9.2, is a successor to the ISLE NIMM Working Group Report D9.1 [Knudsen et al. 2002a] which surveyed Natural Interactivity and Multimodality Annotation Schemes and Best Practice. Like all other ISLE NIMM Working Group reports, ISLE Report D9.1 can be downloaded from http://isle.nis.sdu.dk. As we will often refer to NIMM coding tools below because the existence of an appropriate coding tool makes coding so much more efficient, we would also like to refer the reader to the ISLE NIMM Working Group's Survey Report D11.1 of the state of the art in NIMM coding tools [Dybkjær et al. 2001].

Based on the comprehensive survey in D9.1, the present report aims to contribute to future standards by providing guidelines for the creation of NIMM (Natural Interactivity and Multimodality) annotation schemes. We are happy to note that D9.1 is already being quoted in the literature, see e.g. [Pirker and Krenn 2002], [Piwek et al. 2002].

It may be helpful to explain our use of terms and concepts relating to standardisation as these are understood in the present report. By a NIMM annotation scheme *standard* we understand a standard which has been approved by a recognised standards organisation, such as ISO. A standard is a guidelines documentation that reflects agreements on products, practices, or operations by nationally or internationally recognized industrial, professional, trade associations or governmental bodies. By a *de facto standard* we understand a standard that is widely accepted and used, but lacks formal approval by a recognised standards organisation. Sets of *guidelines,* such as those presented below, may be the precursors of de facto standards and ultimately of standards. Guidelines are not formally approved by a recognised standards organisation and they are not necessarily widely accepted yet but they propose a best practice - often based on experience and current practice – which may have the potential for eventually becoming a (de facto or otherwise) standards.

Proposals for general standards, de facto standards, and guidelines in the NIMM annotation scheme area could concern:

- how to create NIMM coding schemes;
- how to document NIMM coding schemes;
- how to represent NIMM coding schemes and annotations in a computer readable format;
- how to locate and select an appropriate existing coding scheme;
- how to adapt an appropriate existing coding scheme.

Proposals for guidelines concerning these issues are presented in Chapter 4. It should be noted that general proposals for the *creation* and *documentation* of NIMM coding schemes are closely related because of the fact that documentation guidelines, de facto standards, or standards provide a wealth of information on how to create NIMM coding schemes. In addition, annotation scheme generality may attach to NIMM general-purpose coding tools but software coding tools, being potential or actual products on the market, are not subject to „real" standardisation but at most to de facto standardisation.

In the present context, *natural interactivity* means natural interactive communication, i.e. communication, or exchange of information, which uses the full set of means for conveying information used by humans in situated information exchange in which humans communicate about anything whilst sharing time, location, and perceivable physical context. These means (or modalities, see below) include not only speech but also gaze, facial expression, gesture, body posture, use of referenced objects and artefacts during communication, interpersonal (physical) distance, etc. Natural interactive communication is, by nature, *multimodal.* For instance, if speech is considered a single modality in natural interaction, then gaze may be considered a second, different modality, facial expression a third, etc. Moreover, when communicating with machines, humans may be required to use modalities which have no correspondence in natural interactive human-human communication,

such as mouse click codes. When communicating with other humans, people use touch only modestly in order to exchange information, and they do not use anything like touch codes for this purpose, except for exchanging information with the severely disabled. For this reason, multimodality constitutes a wider area of modalities for information representation and exchange than does natural interactivity [Bernsen 2002].

Particular annotation scheme standards or de facto standards exist for certain NIMM sub-areas, such as speech transcription and facial expression, cf. Chapter 3. Standardisation efforts are ongoing in other NIMM sub-areas and there are now even efforts to span across the entire NIMM area in order to propose general guidelines for NIMM coding scheme internal representation, creation and documentation, cf. Sections 3.7 on ATLAS and 3.8 on NITE. However, it is worth noting that existing (de facto) standards only cover areas for which a common tag set could be agreed upon, sometimes including extensibility of the tag set according to certain rules as in the case of TEI, see Chapter 3. There are not yet any general or cross-NIMM (de facto or official) standards for annotation schemes.

Today's researchers and industrial developers in human(s)-machine and human(s)-machine-human(s) communication are beginning to venture far beyond spoken dialogue systems into more or less full natural interactive systems development. Like the maturing field of speech applications, the wider field of natural interactive systems development needs annotated data for many different purposes. Annotated data requires annotation schemes for use in annotating the data, and preferably also annotation tools support. Researchers and developers in this massively expanding field come from many different areas of research and development, such as spoken language dialogue systems, machine vision, computer graphics, and telecommunications, and they do not necessarily bring with them a baggage of knowledge about data annotation and annotation schemes in the spoken dialogue field or elsewhere. This report may be useful to these colleagues by outlining the state-of-the-art, pointing to ongoing and past standardisation efforts, and providing guidelines on how to create annotation schemes which hopefully will be more easy to use and re-use than most NIMM coding schemes are today.

With the growing need for annotated data resources there is also an increasing need for tools to support annotation and annotation analysis. Coding tools are typically rather expensive to develop and they are only easy to re-use across coding schemes and annotated data for purposes, such as data statistics, if the data have the same format and if the coding schemes can be made to fit into a single framework so that existing tool features can be exploited during markup using a new coding scheme.

Once the initial conceptual work has been done it often may be easy to jot down one's own ideas about an annotation scheme consisting of a simple tag set and little else. However, well-documented, repeatedly evaluated, high-quality coding schemes require much more time and effort to create but also return a much higher benefit in terms of reliability of the annotated data and usability and re-usability of the coding scheme.

Standards and guidelines for creating NIMM annotation schemes is becoming a vital factor in ensuring usability and re-usability not only of the annotation schemes themselves but also of the tools which support the use of the annotation schemes. Let us briefly mention some of the reasons.

Annotation scheme retrieval for possible re-use. Most annotation schemes today are home-made and tailored to a particular coding purpose. They tend to be poorly documented and hard to find and use by others than their creators, cf. ISLE Report D9.1 and Chapter 2. Relatively few NIMM coding schemes are widely known, well-documented, and supported by tools. All coding schemes reviewed in ISLE D9.1 are described on the basis of the same information template. However, in many cases even the simple template we designed for the purpose could only be filled by contacting the creators and getting their input on a bilateral basis. The information simply was not available elsewhere. This demonstrates, we believe, the need for documentation standardisation in the field since we may save considerable effort by finding a suitable existing coding scheme, possibly even supported by a tool, instead of having to re-invent the wheel ourselves.

Annotation scheme creation. If there do not seem to be any NIMM annotation schemes out there which fits one's needs, the only option seems to be to develop a new one. In this case, it appears that the annotation scheme developer's first inclination is to simply develop the annotation scheme –

maybe just in terms of a tag set - and then use it, ignoring all but the simplest documentation. Normally, there is no tools support available for coding data in compliance with the new coding scheme unless the coding scheme developer takes the additional major step of developing a coding tool specifically for the this scheme. Sometimes it also turns out that the annotated data are stored in a format which makes them difficult to reuse in other contexts. Standardisation of the representation of coding schemes and coded data, and of coding scheme documentation would facilitate reuse of both tools, annotated data, and annotation schemes.

Difficulty. Once the initial idea has been fleshed out conceptually, it is not difficult to create a coding scheme according to a well-defined set of guidelines. The problem rather is that we don't have such a set of agreed-upon guidelines yet.

This report does not offer a complete guide to NIMM annotation scheme creation. The field is far from being mature enough for full-scale standardisation. However, there are now several interesting initiatives going on as regards improved annotation scheme documentation and as regards representation formats. We hope that the presentation and discussion below will provide the reader with a fairly clear picture of the current situation, encourage him/her to use the presented guidelines, and thus help pushing coding scheme best practice in the right direction in terms of improved usability and re-usability of both annotation schemes and tools.

In the following, Chapter 2 provides a summary of the conclusions in ISLE Report D9.1 which surveys NIMM annotation schemes and best practice, and discusses work on standardisation. Chapter 3 presents a series of activities world-wide which share the aim of influencing NIMM coding scheme-related standardisation. Chapter 4 concludes the report and discusses recommendations and guidelines for NIMM coding scheme creation.

2. State-of-the-art in coding schemes

Based on the review of natural interactivity and multimodal annotation schemes in ISLE NIMM Report D9.1, [Knudsen et al. 2002a] (see http://isle.nis.sdu.dk -> Reports), conclusions were made on state-of-the-art and best practice in the field. In the following we summarise the contents and conclusions of the report.

The survey in D9.1 comprises a total of 21 coding scheme descriptions, including four general descriptions which do not detail particular coding schemes. Some of the coding schemes serve to mark up a single modality while others support markup of modality combinations. Seven descriptions present coding schemes for the coding of facial expression, including, e.g., eyes, eye brows and lips. Jointly, these schemes cover faces of adults as well as faces of babies and children, and they cover human faces as well as cartoon faces. The remaining fourteen descriptions present coding schemes with a main emphasis on the coding of gesture (hand, arm, other), hand manipulation of objects, body movement, or any of the preceding as accompanied by speech. Nearly all the reviewed coding schemes are aimed at markup of video, sometimes including audio. A couple of facial expression coding schemes address static image markup. "Pure" spoken language annotation schemes were not reviewed in ISLE NIMM D9.1 as such schemes had already been reviewed in [Klein et al. 1998].

There probably exist far more coding schemes than those described in D9.1. Most of them are very likely to be tailored to some particular purpose, such as one of developing a particular application, and to be used solely by their creators or at the creators' site. Such coding schemes tend not to be very well described and they tend to be hard to find. The D9.1 survey includes a number of such coding schemes, many of which were created by ISLE NIMM participants or by people known to the ISLE NIMM participants, which is the main reason why we were aware of them. D9.1 also describes a few coding schemes which are in frequent use, or even considered standards in their field. Figure 2.1 provides an overview of the coding schemes reviewed, including the annotation purpose for which they were created. Figure 2.1 does not include the four general descriptions mentioned above which do not detail particular coding schemes.

D9.1 mainly focuses on coding schemes for NIMM video recorded data. It should be noted that coding schemes for specifying output behaviours of embodied conversational agents may also be of interest to NIMM behaviour coders more generally, even if their final goal is different, namely that of specifying the behaviour of a virtual character, see, e.g., [Pirker and Krenn 2002], [VHML 2001], [HF2002], [AAMAS 2002], and [PRICAI 2002]. Similarly, efforts towards standardisation of multimodal input should be followed, cf. [W3C 2002].

Intended for markup of modalities:	**Name**	**Purpose of creation**
Gaze	The alphabet of eyes	Analyse any single item of gaze in videotaped data.
Facial expression	FACS (facial action coding system)	Encode facial expressions by breaking them down into component movements of individual facial muscles (Action Units). Suitable for video or static images.
	BABYFACS	Based on FACS but tailored to infants.
	MAX (Maximally Discriminative Facial Movement Coding System)	Measure emotion signals in the facial behaviours of infants and young children. Suitable for video or static images.
	MPEG-4	Define a set of parameters for defining and controlling facial models.

	ToonFace	Code facial expression with limited detail. Developed for easy creation of 2D synthetic agents.
Gesture	HamNoSys	Designed as a transcription scheme for various sign languages.
	SWML (SignWriting Markup Language)	Code utterances in sign languages written in the SignWriting System.
	MPI GesturePhone	Transcription of signs and gestures.
	MPI Movement Phase Coding Scheme	Coding of co-speech gestures and signs.
Speech and gesture	DIME (Multimodal extension of DAMSL)	Code multimodal behaviour (speech and mouse-based) observed in simulated sessions in order to specify a multimodal information system.
	HIAT (Halbinterpretative Arbeitstranskriptionen)	Describe and annotate parallel tracks of verbal and non-verbal (e.g. gesture) communication in a simple way.
	TYCOON	Annotation of available referable objects and references to such objects in each modality.
Text and gesture	TUSNELDA	Annotation of text -and-image-sequences, e.g. from comic strips.
Speech, gesture, gaze	LIMSI Coding Scheme for Multimodal Dialogues between Car Driver and Co-pilot	Annotation of a resource which contains multimodal dialogues between drivers and co-pilots during real car driving tasks. Speech, hand gesture, head gesture, gaze.
Speech, gesture and body movement	MPML (A Multimodal Presentation Markup Language with Character Agent Control Functions)	Allow users to encode the voice and animation of an agent guiding a website visitor through a website.
Speech, gesture, facial expression	SmartKom Coding scheme	Provide information about the intentional information contained in a gesture.

Figure 2.1. Overview of the NIMM coding schemes reviewed in ISLE D9.1, including the annotation purpose for which they were created [Knudsen et al. 2002a].

The picture of a proliferation of home-grown coding schemes provided by the D9.1 survey was supported by the 28 questionnaires included in ISLE NIMM Report D8.1 [Knudsen et al. 2002b], asking people at a multimodal interaction workshop, e.g., which coding scheme(s) they had used or planned to use for data markup. Some respondents did not answer the question at all or had not made any decision yet. However, in 15 cases the answers indicated that a custom-made scheme would be, or was being, used. Only a few respondents mentioned more frequently used annotation schemes, such as TEI, BAS, or HamNoSys. Fore more information visit http://isle.nis.sdu.dk ->Reports.

For facial expression coding, MPEG-4, FACS and ToonFace are the most frequently used coding schemes. MPEG-4 is, in fact, among the coding schemes reviewed which may be considered a de facto standard. FACS is used by many colleagues as well but is not considered a standard and is not really well suited for markup of lip movements. ToonFace is being used less widely than MPEG-4 and FACS but is being used by a goodly number of sites. It is useful for 2D caricature coding but not for real facial expression annotation [Knudsen et al. 2002a].

The reviewed facial coding schemes are all language-independent and focus entirely on facial expression. However, within facial expression they cover a multitude of different features including,

e.g., gaze, eye brows, eye lids, wrinkles, and lips. A couple of schemes have baby faces or children's faces as their target but most of the schemes focus on adult faces.

The area of gesture coding schemes was found to be even more diverse than that of facial expression. Whereas facial expression is often the sole focus of the reviewed coding schemes, gesture is often being marked up in combination with other modalities, each modality being coded separately. It was only in the specialised field of sign languages that focus was on gesture alone. Many other gesture coding schemes have been created to study gesture in combination with one or several other modalities with the purpose of supporting the development of a particular multimodal system. When several modalities are involved, it becomes important to be able to handle interrelationships among communicative phenomena expressed in different modalities. Time alignment appears to be basic as the common point of reference for this purpose.

By contrast with facial markup. no real standards for gesture markup were found although well-known methodologies for collecting, annotating and classifying gesture data have been used for several years with VCR [McNeill 1992]. As regards gesture annotation only, HamNoSys seems to be the most frequently used among the schemes we looked at. As regards annotation of gesture in combination with other modalities, there are many schemes around, most of which are being used by few people, and no standardisation in sight.

Since the publication of ISLE NIMM Report D9.1, new gesture coding schemes have emerged. One example is the FORM coding scheme which was developed in order to capture the kinematic information of gesture from videos of speakers. This is done by annotating the video with geometric descriptions of the positions and movements of the upper and lower arms, and the hands and wrists. FORM uses Annotation Graphs as its logical representation. [Martell et al. 2002]

Several of the analysed gesture coding schemes are meta-schemes in the sense that they are general information containers (or meta-schemes) that can be filled with different specific coding schemes, e.g. for a particular sign language. The gesture meta-schemes are often defined a way in which links gesture with other modalities.

We found a trend towards using XML across the gesture schemes analysed. Even the creators of the recent schemes tended to announce plans or ongoing efforts to convert their coding schemes into XML (HIAT, HamNoSys (in the ViSiCAST project), and the more recent gesture coding schemes (e.g. SWML, TUSNELDA) were designed from the start for being represented in XML.

With respect to coding scheme evaluation, the pattern observed for the two groups of facial and gesture coding schemes, respectively, was, again, rather different. Most of the reviewed facial coding schemes had been evaluated whereas only a couple of the gesture coding schemes had been evaluated and only on a small scale. All the other gesture coding schemes had either not been evaluated or no information on evaluation could be found.

It was only when it came to tool support that no significant difference was found between the groups of facial and gesture coding schemes. Most coding schemes come with some kind of tools support for using the coding scheme or for processing the results of using it. Only in three cases was there no tools support at all.

Based on the survey in ISLE NIMM D9.1, we concluded that there is still a long way to go before we achieve the ideal goal of being able to code natural interactive communication and multimodal information exchange in all their forms, at any relevant level of detail (or abstraction), generally or exhaustively per level, and in all their cross-level and cross-modality forms. In fact, this is already true for the coding of speech at several important levels of abstraction, such as dialogue acts and co-reference, as concluded in the MATE Report D1.1 on multi-level annotation of spoken dialogue [Klein et al. 1998], which is available at the MATE website at http://mate.nis.sdu.dk. When we move to considering facial coding, we do find a number of general and substantially evaluated coding schemes for different aspects of the facial expression of information (eyes, facial muscles), but it seems clear that we still need a number of higher-level coding schemes based on solid science for how the face manages to express cognitive properties, such as emotions, purposes, attitudes and character. In the general field of gesture, the state of the art is even further remote from the ideal described above. General coding schemes, as opposed to schemes designed for the study of particular kinds of task-

dependent gesture, are hard to find at all, except for the field of sign languages, and the state of evaluation of particular schemes is generally poor. Finally, when it comes to cross-modality coding of natural interactive behaviour, no coding scheme of a general nature would seem to exist at all.

As noted above, most existing facial and gesture coding schemes are coupled with a dedicated coding tool which makes the enterprise of applying the coding scheme to substantial amounts of data feasible in the first place. Although the nature of the intellectual effort required is quite different, dedicated coding tools tend to be at least as costly to develop as the coding schemes they support. It would seem to follow that a key to progress in understanding and making practical use of natural interactivity and multimodal data is the availability of *general-purpose* coding tools which enable the coding scheme creator to focus on coding scheme development, data coding, and analysis, rather than always having to face the dual task of coding scheme and supporting tool development in order to get started with data coding and analysis. Such general-purpose NIMM data coding tools do not yet exist, as shown in the ISLE coding tools state of the art survey D11.1 [Dybkjær et al. 2001]. Their existence, however, could herald a breakthrough in the scientific study of how humans express information through intriguingly complex and massively coordinated use of multiple modalities and at multiple levels of abstraction within each of the modalities involved.

3. Important initiatives

In this section we describe initiatives in the area of natural interactivity and multimodal communication which have set standards for NIMM coding schemes or are currently seeking to do so. There are, at present, no general standards for NIMM coding schemes. However, several ongoing initiatives aim to create coding standards for the NIMM area in general, cf. Chapter 1.

The initiatives presented below are all described according to the following template to the extent that the information has been available.

- Brief description.
- Important websites and other information sources.
- Area covered by the initiative.
- Standardisation.
- Evaluation of the initiative.
- Tools support.

3.1 TEI - Text Encoding Initiative

3.1.1 Brief description

The Text Encoding Initiative (TEI) began in 1987 as a research project within the humanities. It is now since 2000 an international membership consortium. The consortium's goal is to provide support for the encoding of all kinds of text by describing what to encode and how to do it.

The comprehensive TEI guidelines provide specific guidelines for the structured markup of any kind of text and make recommendations on their use. The rules and recommendations made are designed to enable the creation of documents that conform to either the Standard Generalised Markup Language (SGML, defined by ISO 8879 – http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16387.) or the Extensible Markup Language (XML, defined by the World Wide Web Consortium's XML Recommendation – http://www.w3.org/XML/). The support for XML was introduced in TEI P4 [Sperberg-McQueen and Burnard 2002].

3.1.2 Important websites and other information sources

The TEI website: http://www.tei-c.org/

University of Virginia TEI site: http://etext.lib.virginia.edu/TEI.html

Sperberg-McQueen, C. M. and Burnard, L. (Eds.): Guidelines for Electronic Text Encoding and Interchange. TEI P3. Text Encoding Initiative. ACH, ACL, ALLC, Chicago, Oxford, April 1994.

Sperberg-McQueen, C. M. and Burnard, L. (Eds.): TEI P4: Guidelines for Electronic Text Encoding and Interchange. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen, 2002.

3.1.3 Area covered by TEI

The TEI Guidelines have their main emphasis on written text of any kind. However, the Guidelines also include a proposal for the transcription of spoken language [Sperberg-McQueen and Burnard 1994, Chapter 11]. The TEI Guidelines, however, do not go beyond spoken language orthographic transcription. The chapter on transcription guidelines has not been revised to reflect developments in

the field of speech transcription and multimodal language annotation since its first publication in 1994. It is planned that a future revision of the chapter will reflect those developments.

The TEI Guidelines are designed to be applicable across a broad range of applications and disciplines and therefore address not only a vast range of textual phenomena, but are also designed for purposes of maximising generality and flexibility. The TEI standard is deliberately open-ended. The focus has been on defining a common encoding reference and guidelines for adding new (TEI) compliant or conformant DTDs in a way that allows for the interchange of annotated corpora. Generality and extensibility have been achieved, to some extent, at the cost of excessive complexity of the underlying representation. It may therefore be expected that more specialised (less general) and simpler extensions will be specified for different corpus groups, and that support tools will be developed to enable easy and fast coding according to the encoding standard.

3.1.4 Standardisation

The TEI Guidelines called *P3* were published in 1994 and have become the de facto standard for encoding of literary and linguistic texts, corpora, and the like. The TEI Guidelines called *P4* were published in 2002. The major change from P3 to P4 was to make the tag set and documentation compatible with XML.

3.1.5 Evaluation of TEI

An evaluation of the Text Encoding Initiative (TEI) recommendations on symbolic transcription of spoken language can be found at http://www.ilc.pi.cnr.it/EAGLES96/spokentx/node14.html

Arguments for and against rich TEI markup can be found at http://www.iath.virginia.edu/ach-allc.99/proceedings/mah.html

3.1.6 Tools support

A number of supporting tools are listed at the TEI software page: http://www.tei-c.org/Software/

Windows 95/NT software for creating, editing, checking, and doing other interesting things with files marked up according to the principles of the TEI is available at:
http://www.umanitoba.ca/faculties/arts/linguistics/russell/ebenezer.htm

3.2 ToBI - Tones and Break Indices

3.2.1 Brief description

Tones and Break Indices (ToBI) is a framework for developing a standard for transcribing the intonation and prosodic structure of spoken utterances in a variety of languages, see Section 3.2.3. A ToBI framework system for a particular language is grounded in careful research on the intonation system and the relationship between intonation and the prosodic structures of the language, such as tonally marked phrases and any smaller prosodic constituents that are distinctively marked by other phonological means. ToBI is an adaptation of Pierrehumbert' s phonological model of English intonation [Pierrehumbert 1980].

According to Llisterri (1996) “in the domain of prosodic transcription systems to be used in speech research and in speech technology, ToBI (Tone and Break Index Tier) was developed to fulfil the need of a prosodic notation system providing a common core to which different researchers can add additional detail within the format of the system. It focuses on the structure of American English, but transcribes word grouping and prominence, two aspects which are considered to be rather universal [Price 1992]”.

3.2.2 Important websites and other information sources

The Ohio State University Department of Linguistics ToBI website: http://www.ling.ohio-state.edu/~tobi/

Guidelines for ToBI Labelling: http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/

The ToBI Annotation Conventions:
http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html

ToBI creators have developed two labelling guides [Beckman & Ayers, 1994; Beckman & Hirschberg, 1994].

Beckman, M. E., Ayers, G. M. - Guidelines for ToBI Labelling. Version 2.0, February 1994.

Beckman, M. E., Hirschberg. J.: The ToBI Annotation Conventions. Appendix A of M. E. Beckman, G. M. Ayers. Guidelines for ToBI Labelling. Version 2.0, February 1994.

Hart, J., Collier, R. and Cohen, A.: A perceptual study of intonation. Cambridge University Press, 1990.

Grice, M., Reyelt, M. Benzmuller, R., Mayer, J., Batliner, A.: Consistency in Transcription and Labelling of German Intonation with GtoBI. Proc. ICSLP 96.

Llisterri, J.: Prosodic encoding survey. Multext (LRE62-050) Report. WP1, 1994.

Llisterri, J.: Preliminary Recommendations on Spoken Texts. EAGLES Documents EAG-TCWG-STP/P, May 1996. http://www.ilc.pi.cnr.it/EAGLES96/spokentx/spokentx.html

Pierrehumbert, J.: *The Phonology and Phonetics of English Intonation*, Bloomington, Indiana University Linguistics Club, 1980.

Pitrelli, J. F., Beckman, M. E., and Hirschberg, J.: Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *ICSLP94*, volume 1, 123-126, 1994.

Price, P.: Summary of the Second Prosodic Transcription Workshop: the TOBI (TOnes and Break Indices) Labelling System. Nynex Science and Technology, Inc., April, 1992. Linguist List vol 3, 761, October 1992.

3.2.3 Area covered by ToBI

ToBI covers intonation and prosodic structure of spoken utterances in a variety of languages. Complete ToBI systems have been developed for English (American, British, Australian), German, Japanese and Korean. A nearly complete system with training materials but no published inter-transcriber consistency tests as of yet, has been developed for Greek. Systems are under development for Serbo-Croatian, Cantonese, English-Glasgow variety, Mandarin and Spanish. Finally, an annotation system with training materials in a similar auto-segmental framework for intonational tunes (but without the parallel annotation of prosodic structure) is available for Dutch.

3.2.4 Standardisation

ToBI has become a standard due to the fact that prior to 1992 (the year ToBI was developed) there were no widely accepted systems for the transcription of prosody that addressed both intonation and phrasing in a integrated way. Furthermore, back in 1992 there was a growing need for computational methods for annotation and ToBI offered a solution [Wightman, Colin W.: ToBI or not ToBI? Can be downloaded from: http://www.lpl.univ-aix.fr/sp2002/pdf/wightman.pdf].

There are also other systems for the transcription of prosody. INTSINT is a coding system of intonation developed at Aix-en-Provence University. More information about the INSTINT system can be found at htttp://www.lpl.univ-aix.fr/~hirst/intsint.html. The coding scheme of IPO has been also considered in the field of intonation research [Hart et al. 1990].

3.2.5 Evaluation of ToBI

ToBI has been used by numerous researchers for annotation work ranging from linguistic research to systems engineering. An evaluation of the performance of ToBI is presented in [Pitrelli et al.1994]. The German version GToBI has been evaluated in [Grice et al. 1996].

3.2.6 Tools support

There is software to support ToBI transcription using Waves and UNIX programmes. [NOB: how do I get it? There does not seem to be a reference to ToBI software in 3.2.2, and the reference should perhaps in any case be repeated here.]

3.3 SAMPA - Speech Assessment Methods Phonetic Alphabet

3.3.1 Brief description

Speech Assessment Methods Phonetic Alphabet (SAMPA) is a machine-readable phonetic alphabet. It was originally developed by an international group of phoneticians in the ESPRIT project SAM (Speech Assessment Methods) in 1987-89, and was first applied to the European Community languages Danish, Dutch, English, French, German, and Italian (1989); later to Norwegian and Swedish (1992); and subsequently to Greek, Portuguese, and Spanish (1993). In the BABEL project, it has now been extended to Bulgarian, Estonian, Hungarian, Polish, and Romanian (1996). Under the aegis of COCOSDA (The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques – http://www.cocosda.org/), it is hoped to extend SAMPA to cover many other languages (and in principle all languages). Recent additions include Arabic, Cantonese, Hebrew, and Thai.

Unlike other proposals for mapping the IPA (International Phonetic Alphabet) onto ASCII, SAMPA is not one single author's scheme, but represents the outcome of collaboration and consultation among speech researchers in many different countries. The SAMPA transcription symbols have been developed by, or in consultation with, native speakers of every language to which they have been applied.

A recent proposal for an extended version of the segmental alphabet, X-SAMPA, would extend the presently agreed SAMPA conventions so as to make provision for every symbol on the Chart of the International Phonetic Association, including all diacritics. In principle, this would make it possible to produce a machine-readable phonetic transcription for every known human language.

Length, stress and tone marks according to SAMPA are shown in Table I

Length, stress and tone marks	ASCII no.	Unicode	Explanation
:	58	colon 02D0, 720	length mark
"	34	vertical stroke 02C8, 712	primary stress
%	37	low vert. str. 02CC, 716	secondary stress
`	96	*(see note in text)*	falling tone
'	39	*(see note in text)*	rising tone

Table I. Length, stress and tone marks according to SAMPA.

Note: The SAMPA tone mark recommendations were based on the IPA as it was up to 1989-90. Since then, however, the IPA has changed its symbols for falling and rising tones. The SAMPA tone marks may now be considered obsolete, having in practice been superseded by the SAMPROSA proposals (http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm).

3.3.2 Important websites and other information sources

The SAMPA website: http://www.phon.ucl.ac.uk/home/sampa/home.htm

SAMPROSA : http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm

Definitions and information: http://coral.lili.uni-bielefeld.de/Documents/sampa.html

3.3.3 Area covered by SAMPA

In its basic or initial form, SAMPA was seen as essentially catering for segmental transcription, particularly of a traditional phonemic or near-phonemic kind. Prosodic notation was not adequately developed. This shortcoming has now been remedied by a proposed parallel system of prosodic notation, SAMPROSA. It is important that prosodic and segmental transcriptions be kept distinct from one another, on separate representational tiers, because certain symbols have different meanings in SAMPROSA from their meaning in SAMPA. For instance, *H* denotes a labial-palatal semivowel in SAMPA, but High tone in SAMPROSA.

3.3.4 Standardisation

SAMPA has been applied not only by the SAM partners collaborating on EUROM 1 (http://www.icp.inpg.fr/Relator/multiling/eurom1.html), but also in other speech research projects (e.g. BABEL, Onomastica) and by Oxford University Press. SAMPA is included among the resources listed by the Linguistic Data Consortium (http://www.ldc.upenn.edu/annotation/). It has been applied to more than twenty languages .

3.3.5 Evaluation of SAMPA

Information about evaluation of SAMPA is not available but modifications and extensions have been suggested for reasons which have arisen from practical use in speech technology and spoken language lexicography.

3.3.6 Tools support

SAMPA does not need tools support since it consists of a mapping of symbols of the International Phonetic Alphabet onto ASCII codes.

3.4 ISO sub-committee TC37/SC4

3.4.1 Brief description

ISO/TC 37/SC4 is a recent initiative which was formally established at the LREC 2002 conference. The ISO/Technical Committee 37 is dealing with all aspects of terminology and other language resources. By 2002, ISO/TC 37 already included three subcommittees: (1) SC1 works on Principles and Methods; (2) SC2 works on Terminology and Lexicography, and (3) works on Computer Applications for Terminology. Sub-committee SC4 addresses all aspects of Language Resource Management.

The objective of ISO/TC37/SC4 is to prepare various standards by specifying principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes. Standards produced by ISO/TC37/SC4 will particularly address the needs of industry and international trade as well as the global economy regarding multilingual information retrieval and cross-cultural technical communication and information management. The sub-committee's technical work will promote international standards through technical reports that cover language resource management principles and methods, as well as various aspects of computer-assisted lexicography and language engineering, including their implementation in a broad array of applications.

To achieve its goals, ISO/TC37/SC4 has defined a number of sub-areas for its work: (1) descriptors and mechanisms for language resources; (2) structural content of language; (3) SC semantic content of multimodal data; (4) discourse-level content; (5) multilingual text representation; (6) lexical databases,

and (7) workflow of language resource management. ISO/TC37/SC4 recognises the fact that various initiatives have worked on the topics mentioned for years already, producing many valuable contributions. Therefore, ISO/TC37/SC4 aims to summarise the different approaches in order to propose unified standards that are based on year-long experience and that have a potential to be used broadly by the large language resources community, in particular, of course, industry.

Two examples may help clarify the aims of ISO/TC37/SC4 in more detail. In the area of descriptive metadata for use in resource discovery, various suggestions have been made during the past two years. DublinCore (DC) (http://dublincore.org/) has suggested to use a descriptor set of 15 (by intention) vaguely specified elements to describe almost any web-localisable resources. Of course, any such set will be missing many elements that are relevant to language resources specialists. The OLAC initiative (http://www.language-archives.org/) started from the DC set and suggested several refinements to meet the most urgent needs, such as the inclusion of an element *language,* which is the language of the resource. By contrast, the IMDI descriptor set developed within the ISLE NIMM framework (http://www.mpi.nl/ISLE/) started from an overview of metadata and header information suggestions and came out with a more elaborated and structured set of descriptors to meet the needs of professionals. In addition to these, a number of other initiatives have produced proposals in highly overlapping application areas. MPEG-7 is an example. The MPEG-7 standard, motivated mainly by the film and media industry, includes descriptive metadata for resource discovery. It is obvious that the standard proposals just mentioned address the resource discovery metadata issue in different ways and that they all leave gaps, such as the definition of a schema for open vocabularies. TC37/SC4 wants to gather the experience made with those proposals and define a more general metadata framework which can serve as an umbrella.

Another example of the activities of TC37/SC4 is the definition of a proposed standard for annotation structures for multimedia/multimodal resources. Also in this area, excellent contributions have been made already, such as the ATLAS/MAIA (see Section 3.7), GATE (http://gate.ac.uk/), MATE (http://mate.nis.sdu.dk), and EUDICO (http://www.mpi.nl/world/tg/lapp/eudico/eudico.html) annotation formats. These different proposals each have their strengths and weaknesses, and also in this area a number of essential aspects still remain to be dealt with in a satisfactory manner. ISO/TC37/SC4 aims to gather the accumulated information and experience in order to work out a more complete and stable solution.

As mentioned, TC37/SC4 began its activities rather recently. A first meeting about "A Linguistic Annotation Framework" under the ISO umbrella was held in November 2002. A requirements specification document for metadata will be produced in early 2003. For all its activities a large number of specialists are already participating to guarantee a broad coverage and success.

3.4.2 Important websites and other information sources

The ISO TC37/SC4 website: www.tc37sc4.org.

3.4.3 Area covered by ISO TC37/SC4

ISO/TC37/SC4 will work out a number of coding schemes at different levels. Details are still missing at this point. It will deal with aspects of structural encoding, such as which kind of XML schema can be recommended to be general enough to include all structural phenomena that can occur in multimodal annotations associated with multimedia records, but also with aspects of linguistic encoding, i.e. what tag labels can be recommended for encoding various linguistic aspects such as morphology and what type of encoding should be used for identifying languages. TC37/SC4 will take into account standards defined in other ISO bodies and build upon the valuable work carried out within various projects.

3.4.4 Standardisation

Due to its broad coverage in terms of international experts from the various initiatives that are already active in ISO TC37/SC4, it is expected that the ISO standards to be developed will cover many relevant aspects and find a broad usage.

3.4.5 Evaluation of ISO/TC 37/SC4

ISO/TC 37/SC4 is still to produce results, so evaluation is not possible at this stage.

3.4.6 Tools support

It is expected that the emerging ISO standards will be supported by a large variety of tools when they will be released. The reason is that the specialists involved will modify their existing high-quality tools stepwise dependent on the progress of the ISO discussions. So, ISO TC37/SC4 will not develop tools itself, however, it expects that tool builders will adapt their tools to the forthcoming standards.

3.5 MPEG-4 SNHC - Moving Pictures Expert Group, Synthetic/Natural Hybrid Coding

3.5.1 Brief description

MPEG-4 is an object-based multimedia compression standard which allows to encode different audio-visual objects (AVOs). The MPEG-4 SNHC group has proposed an architecture for the efficient representation and coding of synthetically and naturally generated audio-visual information.

3.5.2 Important web-sites and other information sources

The Moving Pictures Expert Group's Homepage: http://mpeg.telecomitalialab.com/

A description of MPEG-4: http://ligwww.epfl.ch/mpeg4

A description of parameter based facial animation:
http://www.research.att.com/~osterman/AnimatedHead/index.html

A description of MPEG-4 compliant facial animation and Hybrid Video Coding:
http://www-dsp.com.dist.unige.it/~pok/RESEARCH/index.htm

Tutorial issue on the MPEG-4 standard: Elsevier (http://leonardo.telecomitalialab.com/icjfiles/mpeg-4_si/)

P. Doenges ,F. Lavagetto, J. Ostermann, I.S. Pandzic and E. Petajan: MPEG-4: Audio/Video and Synthetic Graphics/Audio for Mixed Media. Image Communications Journal, vol. 5(4), May 1997.

J. Ostermann: Animation of synthetic faces in MPEG-4. Computer Animation'98, Philadelphia, USA, 49-51, June 1998.

E. Petajan: Facial Animation Coding, Unofficial Derivative of MPEG-4. Work-in-Progress, Human Animation Working Group, VRML Consortium, 1997.

Igor S. Pandzic and Robert Forchheimer (Eds.): MPEG-4 Facial Animation - The standard, implementations and applications, John Wiley & Sons, 2002.

3.5.3 Area covered by MPEG-4

MPEG-4 provides a coding scheme for media objects that are organised hierarchically. At the leaves are primitive media objects of the forms:

- Still image

- Video object
- Audio object

Moreover MPEG-4 defines coded representations of:

- Text and graphics
- Virtual agents (talking head or animated body)
- Synthetic sound

A coded object is defined independently of its surroundings or background (e.g. the video of a talking person without the background).

MPEG-4 enables the composition of media objects to create a scene (see Figure 3.5.1) by providing a standardised way to represent a scene and manipulate it by:

- Placing media objects anywhere in a given coordinate system;
- Applying any transformation to change the geometric or acoustic properties of objects;
- Defining compound media objects;
- Applying streamed data to media objects (e.g. animation parameters driving a synthetic face);
- Changing, interactively, the user's viewing and listening points anywhere in the scene.

In particular, MPEG-4 defines a coding scheme for driving the animation of synthetic facial and body models by defining several sets of parameters:

- Facial Animation Parameters (FAPs): describe the facial movements at a low level (e.g. move horizontally the outer right lip corner) and at a high level (visemes and expressions of emotions).
- Facial Definition Parameters (FDPs): define the structure of the face. These parameters may be used to modify the geometry of the facial model or to encode the necessary information to transmit a new model.
- Body Animation Parameters (BAPs): define the joints values of the body. The body joints may follow the H-Anim specification.

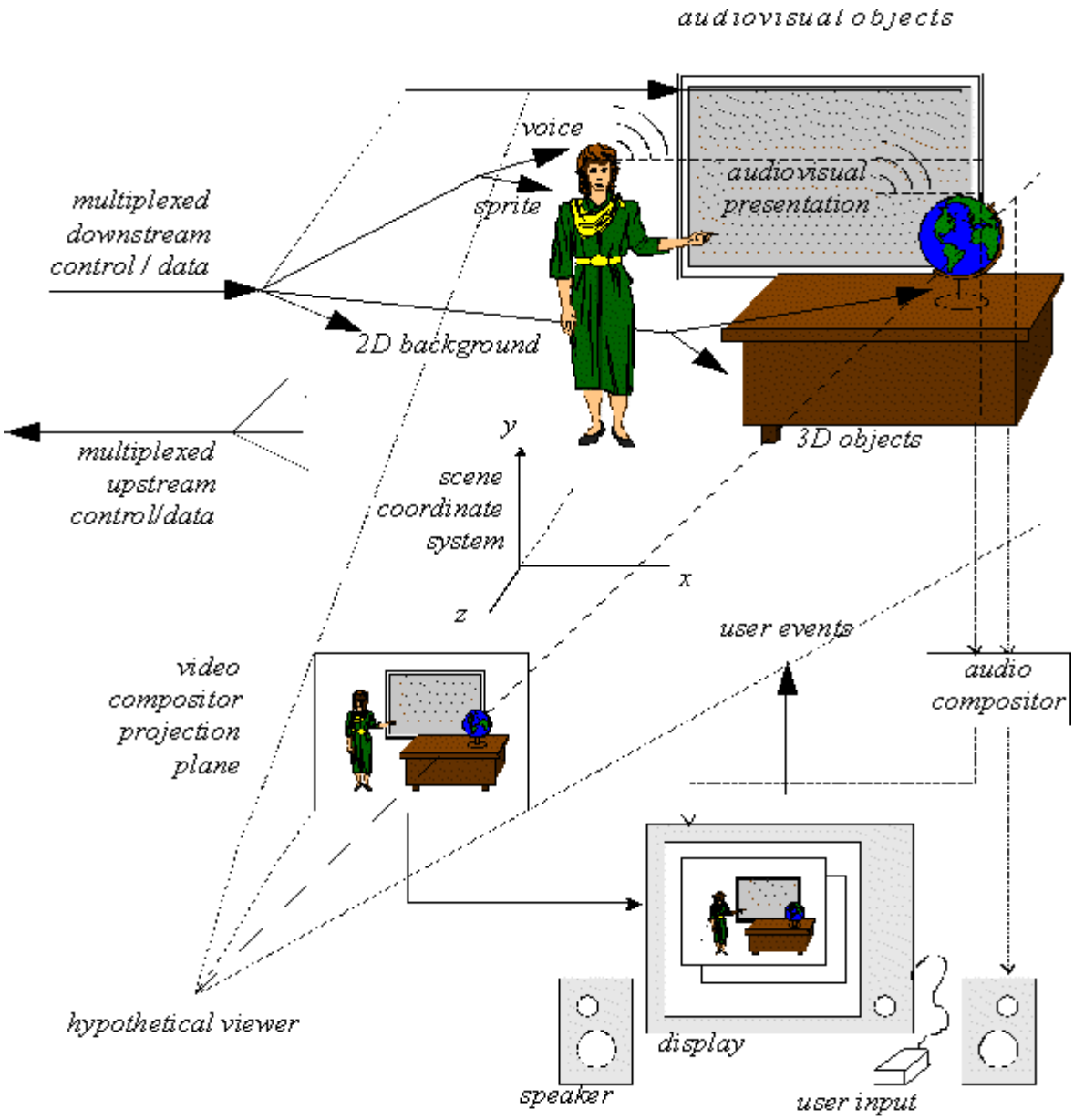

Figure 3.5.1. MPEG-4 composition of media objects.

3.5.4 Standardisation

The goal of MPEG-4 is to provide standards for AV formats and players to satisfy the needs of end-users, authors and service providers. MPEG-4 defines "media objects" as units of visual or acoustic data. The data may be synthetic data (i.e. computer generated images or sounds) or recorded images or sounds from real data. MPEG-4 provides means to compose these objects. It also offers synchronisation and multiplexer schemes to ensure the transmission over networks of compound media objects while maintaining their relationship. Moreover, MPEG-4 defines means of interaction with the audio-visual scene generated at the receiver's end.

It is difficult to state how widely used the MPEG-4 standard is. Many academic groups have developed tools following the MPEG-4 standard. MPEG-4 has been deployed in several industries at a large scale. For instance, Microsoft's Windows Media software includes an MPEG-4 encoder and decoder; Real Networks and Apple's Quicktime version 6 supports MPEG-4; the Internet Streaming Media Alliance (ISMA), founded by Apple, Cisco, IBM, Kasenna, Philips and Sun, has specified a fully end-to-end interoperable system for internet streaming, including MPEG-4 Visual and MPEG-4 Audio profiles.

3.5.5 Evaluation of MPEG-4

Several evaluation studies have been done for specific tasks, such as the coding of facial expressions of emotion, the calibration scheme for facial models, etc.

3.5.6 Tools support

MPEG-4 provides an "object profile" for describing the syntax and the coding/decoding tools for a given "media object". Profiles exist for various types of media content (audio, visual and graphics) and for scene descriptions. The "Face and Body Animation" tools allows one to transmit parameters that can define, calibrate or animate models. The models themselves are not standardised in MPEG-4. Only the parameters and their transmission are part of the standard.

The MPEG-4 tools include [ISO/IEC JTC1/SC29/WG11 - N4668, March 2002]:

- Definition and coding of face and body animation parameters (model independent):
    - Feature point positions and orientations to animate the face and body definition meshes.
    - Visemes, or visual lip configurations, corresponding to speech phonemes.
- Definition and coding of face and body definition parameters (for model calibration):
    - 3-D feature point positions.
    - 3-D head calibration meshes for animation.
    - Personal characteristics.
    - Facial texture coding.

MPEG-4 defines three "object profiles" for facial animation:

- Simple Facial Animation Object Profile: given a proprietary facial model, this tool decodes the FAP stream to drive the animation.
- Calibration Facial Animation Object Profile: on top of the previous profile, the decoder must use a sub-set of the transmitted FDP (called "features points") to calibrate the proprietary facial model, i.e. to adapt the model geometrically.
- Predictable Facial Animation Object Profile: to fully predict the model and animation from the bit stream.

3.6 MPEG-7 - Moving Pictures Expert Group, Multimedia Content Description Interface

3.6.1 Brief description

MPEG-7 is a recently developed standard from the Moving Pictures Experts Group, a large group of specialists representing the requirements originating from the film industry. While the MPEG-1 and MPEG-2 standards are about the decoding of streams that contain audio and video information, and MPEG-4 is about the decoding of different types of media objects, such as video streams or animations, MPEG-7 (Multimedia Content Description Interface) is about the annotation of media streams. The application area is motivated primarily by the MPEG-4 scenario that is meant to allow the user to select and combine various streams and media objects on the screen according to the user's own wishes. Only annotations at various levels will allow the user to retrieve, select, and filter those objects that are of interest at a certain moment in time. The MPEG-7 standard is based on the experience with SMPTE (http://www.smpte.org/) which is a currently used standard in the film industry.

The major items in the MPEG-7 standard are the Description Definition Language (DDL), the Descriptors (D) and the Description Schemas (DS). The DDL is used to define the descriptors and schemas to be used within the MPEG-7 domain. It is based on XML Schema, but adds a few elements such as array, matrix and time primitive data types, which are especially useful for dealing with media streams. Descriptors specify the data categories used in the MPEG-7 domain. They can refer to low-level features, such as the definition of the colour index of a video frame, but can also refer to high-level features, such as the author of an object. Description Schemas relate such descriptors in a structured way and constitute the framework for describing the content of media objects. Currently, more than 100 DSs have already been developed within MPEG-7 to define various types of objects and aspects of objects. There are schemas for describing the content of objects, typical management events such as creation dates, navigation and access information, and usage information. Recently, a first suggestion was made for an MPEG-7-compliant simple description schema for linguistic annotations.

Description Schemas can be part of a hierarchical system of DSs, since a media document may consist of several segments, each of which being described with its own characteristics, but where of course the whole object can be described as well. In summary, a DS defines the structural and semantic relations between its components. The components can be either Descriptors that refer to primitive data categories, such as specific video features, or Description Schemas that describe segments.

The MPEG-7 descriptors are mainly designed to describe low-level audio and visual (AV) features. The description of such features is expected to be done automatically. On the other hand, MPEG-7 DSs are primarily designed to describe high-level AV features, as well as meta-information regarding production, use, etc. of the AV. Moreover, the DSs provide information on the relationship among the different descriptors. The details of the MPEG-7 standard are described at the following website: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm.

Due to its open definition, MPEG-7 can be extended by Description Schemas for sub-disciplines annotating special phenomena that go along with media streams. Therefore, it is difficult to describe the status of MPEG-7 as a whole. The basic principles and the Description Definition Language (DDL) have been accepted as a standard within ISO. It is not at all clear yet whether MPEG-7 will be accepted by the community. One of the ideas to promote MPEG-7 was the development of so-called smart recording devices, such as digital cameras that immediately allow the creation of relevant elements of the media object description. But the economical realisation of such concepts remains to be demonstrated. For complex linguistic annotations of multimedia/multimodal resources, such as those discussed in the linguistic and language engineering community, there is no acceptable Description Schema available. This would have to be designed by experts.

3.6.2 Important web-sites and other information sources

The major web-site for MPEG-7 is http://ipsi.fhg.de/delite/Projects/MPEG7/

The major web-site for the DDL is http://archive.dstc.edu.au/mpeg7-ddl/

A good overview about MPEG-7 is available at http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm

3.6.3 Area covered by MPEG-7

The MPEG-7 standard encodes the meaning of data. The encoded data is passed on to a device, or a computer, that needs to decode and interpret the incoming information streams. An example is image understanding where recognised objects are associated with physical measures and time information. A possible use could be to ensure that some criteria are met by the visual data in view of triggering a suitably programmed PVR (Personal Video Recorder) or an alarm if a particular visual event happens.

Other examples of use are information retrieval (efficient search for information from multimedia data), media conversion (speech-to-text, picture-to-speech, speech-to-picture, etc.).

MPEG-7 can be used for any application using multimedia data, ranging from broadcast media selection to digital libraries (video archive, multimedia dictionaries), education (personalised multimedia courses), and E-commerce (online search, use of catalogues).

3.6.4 Standardisation

MPEG-7 still has to demonstrate its usefulness and appropriateness in practical applications.

3.6.5 Evaluation of MPEG-7

Some research labs were involved in designing and testing MPEG-7, but only the MPEG-4 scenario made up by individual media objects that can be flexibly combined will require an annotation such as proposed by MPEG-7.

3.6.6 Tools support

The institutions that were heavily involved in the definition of MPEG-7 have developed specialised tools. The above-mentioned websites provide tools details.

MPEG-7 Description tools allows to create descriptions, i.e., a set of instantiated Description Schemes and their corresponding Descriptors at the user's will, of content that may include (from website: http://mpeg.telecomitalialab.com/standards/mpeg-7/mpeg-7.htm):

- Information describing the creation and production processes of the content.
- Information related to the use of the content.
- Information of the storage features of the content.
- Structural information on spatial, temporal or spatio-temporal components of the content.
- Information about low-level features in the content.
- Conceptual information of the reality captured by the content.
- Information about how to browse the content in an efficient way.
- Information about collections of objects.
- Information about user interaction with the content.

3.7 ATLAS - Architecture and Tools for Linguistic Analysis Systems

3.7.1 Brief description

The Architecture and Tools for Linguistic Analysis Systems initiative (ATLAS) started in 1999 and was originally based on LDC (Linguistic Data Consortium – http://www.ldc.upenn.edu/) research on Annotation Graphs. Annotation Graphs provide a data model for working with linear signals, such as text and audio, indexed by intervals. Annotation Graphs are a limited sub-set of the more generic ATLAS data model.

ATLAS Level 0, also known as Annotation Graphs, provides a data model, interchange format, and application programming interfaces for working with linear signals indexed by intervals. ATLAS Level 1 is a generalised model, suitable for annotating signals of essentially arbitrary dimensionality with annotations having essentially arbitrary structure. An early application of ATLAS Level 1 is the OCR (optical character reading) annotation, where textual images are indexed using bounding boxes.

ATLAS has four main components:

- a data model,

- an Application Programming Interface (API),
- the ATLAS Interchange Format (AIF), and
- the Meta-Annotation Infrastructure for ATLAS (MAIA).

The ATLAS data model provides the abstractions upon which the rest of the framework is built. These abstractions can be implemented using any full-featured programming languages. NIST (the US Institute for Standards in Technology) has created a Java instantiation of the data model. This implementation provides a set of objects that can be used to quickly develop linguistic applications. These objects each publish operations via which their data can be manipulated and behaviour-controlled. The ensemble of these operations defines the ATLAS API. ATLAS annotations can be serialised to XML using AIF to facilitate their exchange and reuse. Finally, MAIA was added to permit constraining of ATLAS' generic constructs for specific needs.

The philosophy behind ATLAS is not to make any assumptions on annotation schemes or signals that people would use. Focus is on providing a generic and expressive data model based on a simple annotation ontology. This data model provides basic constructs, or building blocks, that can be combined to create more complex constructs. Users can create new constructs that can still be understood by the framework via MAIA's type definition language. The creation of new annotation schemes in ATLAS is a matter of understanding the data model and then creating a MAIA definition describing how constructs of the data model are combined to create scheme-specific constructs. By leveraging the expressiveness and genericity of the ATLAS framework, developers can create generic linguistic applications and tools. MAIA allows users to customise these generic applications to their specific needs by writing an XML type definition describing the entities of their corpus. AIF is intended to be a flexible and extensible file format that will facilitate exchange and reuse of annotation data. AIF is a representation of the ATLAS data model, which employs a general notion of annotation.

3.7.2 Important websites and other information sources

The ATLAS website: http://www.nist.gov/speech/atlas/index.html

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C. and Liberman, M.: ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation. Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000), Athens, 1699-1706.

Publications on ATLAS: http://www.ldc.upenn.edu/sb/home/publications.html#lrec00-atlas

3.7.3 Area covered by ATLAS

ATLAS defines a meta-annotation scheme allowing users to represent their annotation schemes using ATLAS abstractions and leverage existing ATLAS tools for free.

The ATLAS framework aims at facilitating the development of linguistic applications. The primary goal of ATLAS is to provide an abstraction over the diversity of linguistic annotations. The abstraction, which expands on Bird and Liberman's Annotation Graphs, is able to represent complex annotations on signals of arbitrary dimensionality, including text, audio, images, and video.

3.7.4 Standardisation

ATLAS is not standard. However ATLAS addresses the important problem of how to create tools that are usable across coding schemes and annotation formats.

3.7.5 Evaluation of ATLAS

Developers expect that the advantages of ATLAS will be that it is generic and not tied to a particular annotation scheme and/or signal modality. It is extensible and flexible, so that people can create new annotation schemes and new classes of signals and have them handled by ATLAS-compliant

applications without problems. Furthermore, it will provide the basic services needed to create full-fledged linguistic applications. However, evaluation results are still to be awaited.

3.7.6 Tools support

The ATLAS framework allows for the development of applications and tools which work on ATLAS data. The free availability of ATLAS coupled with the defined interchange format position it as a candidate for tool developers by supporting easy creation of reusable tools and components.

3.8 NITE - Natural Interactivity Tools Engineering

3.8.1 Brief description

NITE is a European HLT (Human Language Technologies) project which started in 2001 and ends by mid-2003. The primary objective of NITE is to build an integrated best practice workbench, or toolset, for multi-level, cross-level, and cross-modality annotation, retrieval, and exploitation of multi-party natural interactive human-human and human-machine dialogue data. The NITE software will implement and support use of the NITE markup framework [Dybkjær et al. 2002]. The NITE markup framework builds on, and extends, the MATE markup framework to accommodate the needs of current and emerging coding modules for natural interactivity coding.

The notion of a coding module was introduced in the MATE markup framework for the coding of spoken dialogue corpora [Dybkjær et al. 1998] as an elaboration of the notion of a coding scheme. A coding module includes or describes everything that is needed in order to perform a certain kind of markup of a particular natural interactivity corpus. It describes itself and prescribes what constitutes a coding. The core of a coding module is the set of phenomena in natural interactive communication which we are interested in marking up when using this particular coding module, including the corresponding tag set. However, a tag set is not very useful without accompanying information about origin, purpose, scope, semantics, intended use, best coding practice, etc. This additional information serves to inform a user on the tag set and its use. From a usability point of view, this additional information is just as important as the tag set itself. Without the information, the tag set is unlikely to be useful to anyone except perhaps its creator at creation time.

A coding module includes two types of information. One is mainly intended for the user, the other for the tool system. The information intended for the user may be viewed as the tag set concepts together with their meta-data information, see also Chapter 4. The information meant for the system is the tag set as represented by the underlying data structure.

3.8.2 Important websites and other information sources

The NITE website: http://nite.nis.sdu.dk/

Dybkjær, L., Bernsen, N. O., Carletta, J., Evert, S., Kolodnytsky, M. and O'Donnell, T.: The NITE Markup Framework. NITE Report D2.2, 2002.

Dybkjær, L., Bernsen, N.O., Dybkjær, H., McKelvie, D. and Mengel, A.: The MATE Markup Framework. MATE Report D1.2, November 1998.

3.8.3 Area covered by NITE

The NITE software aims to enable users to enter coding modules and use them for annotating any kind of natural interactive communication, as exchanged in whatever modalities and however complex, including cross-modality interrelationships between multiple classes of phenomena. The NITE software will include functionality for inspection and analysis of codings. Focus is on handling audio and video data.

3.8.4 Standardisation

The recent NITE proposal for a markup framework which includes a proposal for standardisation of coding scheme documentation in terms of a coding module, is not a standard. What makes the NITE markup framework a standard candidate is the fact that it embodies current ideas of the necessity of providing information about a coding scheme sufficient for its easy retrieval on the internet as well as for supporting decisions on coding scheme reuse. Thus, NITE coding modules may have the potential for satisfying the need for uniform and extensive documentation of coding schemes.

3.8.5 Evaluation of NITE

There is not yet any evaluation of the NITE proposal for standardisation of coding scheme documentation.

3.8.6 Tools support

Three tools development strands are being pursued in NITE. One strand, the NITE Workbench for Windows (NWB), will be open source software, is based on a Windows platform, and aims at users who want an easy-to-use interface that requires no programming skills. The second strand, the NITE XML Toolkit (NXT), will be open source software as well, is cross-platform and Java-based, and builds on MATE (mate.nis.sdu.dk) and ANVIL (www.dfki.de/~kipp/anvil). NXT focuses on users who are able and willing to do some programming to use the tool. The third strand will enable future versions of the commercial Noldus Observer software (www.noldus.com) to support some amount of annotation of natural interactive communication. At least NWB and Noldus Observer will support users in entering and using coding modules.

4. ISLE recommendations

The previous chapters provide an overview of the state-of-the-art in the field of NIMM annotation schemes. Chapter 3 shows that (de facto) standards exist mainly for speech and text, especially in the area of transcription, and for media production related issues (MPEG-4 and MPEG-7). For other NIMM sub-areas no real standards seem yet to exist. The existing standards have been brought forward by projects or international groups of people with a shared interest in an area and sufficient need and momentum to get the consensus-building process started. Most existing standards are accompanied by supporting software, which makes them even more attractive to use since their use is facilitated by the software.

The present chapter presents and discusses the ISLE NIMM Working Group's recommendations for guidelines for the development of NIMM annotation schemes.

In Chapter 1 we mentioned that standardisation of coding schemes could concern

- how to create NIMM coding schemes;
- how to document NIMM coding schemes;
- how to represent NIMM coding schemes and annotations in a computer readable format;
- how to locate and select an appropriate existing coding scheme;
- how to adapt an appropriate existing coding scheme.

The present chapter also discusses coding scheme evaluation which is an important parameter both for the location and selection and for the creation process.

4.1 Coding scheme creation

A coding scheme is designed to enable corpus tagging of instances of a particular class of phenomena (or set of types of tokens) expressed in one or several modalities. Coding scheme creation involves, at least, conceptual/theoretical work, tag set creation, and coding scheme testing and evaluation. Coding scheme creation often serves a particular initial purpose but this does not exclude, of course, that once created, the coding scheme could benefit many other coders and many different coding purposes.

The following rules of thumb address conceptual/theoretical work and tag set creation. Testing and evaluation is discussed in Section 4.4.

The coding scheme creator should at least consider points such as the following:

- what is/are the coding purpose(s), what will the annotations be used for, etc.;
- which modality/modalities should be marked up;
- which phenomena are of interest;
- is the identified class of phenomena sufficient for the purpose(s) for which it is intended;
- is the class of phenomena kept as general as allowed by the coding purpose(s);
- often but not always, the class of phenomena to be coded is based on a theory which claims closure for the class, such as, for instance, that the class of phenomena includes all possible, different human facial expressions. This theory needs testing and validation;
- sometimes the coding scheme is merely intended to capture a subset of some larger class of phenomena for some purpose, such as when speech transcribers often only use a subset of a larger set of transcription tags. In such cases, there should be clear rules for how to add new phenomena to the coding scheme, should that be needed later, so that these will be consistent with the already existing ones;

- each phenomenon must be clearly described (assigned a clear semantics) so that both the coding scheme creator and others are always able to decide, given a certain token in a corpus, whether or not that token is an instance of that phenomenon. This point is crucial to inter-coder agreement on how to apply the coding scheme to a given corpus, cf. Section 4.4. Semantic weaknesses in the coding scheme translate into reduced intercoder agreement, reduced consistency of codings, and quickly into a coding scheme which is too unreliable for practical use;
- each phenomenon must be assigned a syntactic tag whose presence in the corpus, or whose reference to a particular token in the corpus, indicates the presence of the phenomenon;
- the tag set representing the relevant class of phenomena should preferably be defined using some kind of standard format for coding tool use, such as XML. The tag set to be interpreted by machine does not have to have the same format as the tag set used by the human coder, one-to-one correspondence is sufficient (se also Section 4.3);
- the tag set should be extensible following well-defined rules.

The guidelines above are closely connected with coding scheme documentation and coding scheme formats, as discussed in the following Sections 4.2 and 4.3.

4.2 Coding scheme documentation

Experience shows that many coding schemes are poorly documented, which makes their retrieval and reuse very difficult. There is not yet any standards as regards which kind of documentation to include with a coding scheme. However, the MATE and NITE projects (Section 3.8) have proposed the concept of a coding module which extends the notion of a coding scheme with documentation that should be sufficient for colleagues to understand and use the coding scheme. At the same time, this documentation is structured in such a way that it would be easy to search through if available on the web. The contents of a coding module is listed below:

- name of coding module
  (E.g. my_gestures.)
- author(s) of coding module
  (E.g. Tom Jones.)
- version
  (E.g. v1.2.)
- notes
  (References to literature, validation information, comments, etc.)
- purpose of the coding module
  (Description of the purpose for which the coding module was first created.)
- coding level(s) covered by the coding module
  (E.g. dialogue acts, hand gesture, nose wrinkles, …)
- description of data source type(s) required for use of the coding module
  (Description of what is required in order for the coding scheme to be used. For instance an orthographic transcription may be a pre-condition for applying a particular coding scheme.)
- explanation of references to other coding modules
  (If the coding module assumes that there are references to other levels of markup then these references should be explained.)
- coding procedure
  (Description of how the coding module should be applied to a corpus in order to produce a

reliable new coding. The coding procedure is important to ensure the reliability of the coding and thus to its quality. The coding procedure should include, cf. [Dybkjær et al. 1998]:

- Description of the coders: their number, roles and required training.
- The steps to be followed in the coding.
- Intermediate results, such as temporary coding files.
- Quality measures (the non-satisfaction of which may require re-coding).

- coding example showing the coding scheme markup in use
  (This could be a snippet from an actually annotated file or it could be a constructed example. The purpose is to give users of the coding module an idea of what the markup looks like when applied.)
- clear description of each phenomenon, example(s) of each phenomenon
  (The descriptions provided here are essential to communicating the semantics of the concepts of the coding scheme. It should be explained as clearly as possible how each concept-tag pair should be applied during markup. Any uncertainty left by the descriptions and examples provided will translate into unreliable coding, inter-coder disagreement, etc.)
- a markup declaration, possibly hierarchically ordered, of the tags for the (individually named) phenomena which can be marked up using the coding module
  (The tag set declaration can be presented in several different ways, e.g. as a DTD, cf. Section 4.3.)

ISLE recommends that coding scheme documentation follows the guidelines proposed by NITE as listed above.

4.3 Coding scheme representation

This section addresses which formats to use for the representation of coding schemes. We need to distinguish between computer-readable formats and human-readable formats.

As for computer-readable formats, there is a strong trend today towards using XML. Coding scheme definitions are most often provided via an XML DTD (Document Type Definition) or via XML Schemas. We recommend to follow this de facto standard since XML, DTDs and Schemas are machine readable, extensible, and widespread. Also for annotated data, XML is widely used. This means that using XML for this purpose as well will facilitate the exchange of annotated corpora. For more information on XML, see, e.g., http://www.w3.org/XML/

Whereas XML DTDs and Schemas are excellent for computers, they are not so easy to read and write for humans. If tool support is available when one makes a markup declaration, it may be possible to use a format which is more friendly and easy for humans to use without special programming skills. Behind the user interface, the tool may then, e.g., convert the markup declaration into an XML DTD. Seen from the user's side, however, the markup declaration may just be in terms of, e.g., well-defined form-filling. The special XML tags are then added behind the scene by the tool.

We recommend and support the development of tools which facilitate the indication of markup declarations and support the use of an underlying standard representation format.

4.4 Coding scheme evaluation

Coding scheme evaluation follows coding scheme creation and documentation. The purpose of evaluation is to test the quality of the coding scheme and of the results produced by using the coding scheme as intended. Precise and informative evaluation results provide very useful information to those looking for an existing coding scheme to use, cf. Section 4.5.

The coding scheme should be applied according to the prescriptions in the coding procedure, cf. Section 4.2. This means, e.g., that the annotators must have the background and expertise recommended and that the number of annotators prescribed must be used to ensure the quality of the coding.

The ease-of-use and reliability of the coding scheme may be measured by:

- asking coders their opinion (interview, questionnaire);
- checking if different coders use tags consistently;
- measuring the time taken to code;
- measuring the quality of the annotations, cf. below.

The ease-of-use of coding tools may also be evaluated by asking coders their opinion and by measuring the time it takes them to code. Measuring quality of codings is also relevant for tools evaluation if markup is done semi-automatically or automatically.

Coding scheme quality is a research area of its own. A coding scheme may be evaluated by:

- comparing different corpus samples coded by means of the scheme to assess *coverage;*
- comparing the results produced by different coders to assess *intercoder reliability;*
- comparing the results produced by the same coder on the same corpus sample at different times, for instance with a one-week delay, to assess *consistency*.

Coding scheme quality may be evaluated

- qualitatively through discussion of the choices made by coders when they differ;
- quantitatively through scoring measures.

A very frequently used method to compare the results produced by different coders (intercoder agreement) is called *kappa*:

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

P(A) is the proportion of times that the coders agree and *P(E)* is the proportion of times that they are expected to agree by chance. The problem with this method is that there is no sound interpretation of which kappa values are good enough. Moreover, kappa presupposes independent events which is far from always the case in NIMM contexts, see also [Dybkjær and Dybkjær 2002].

Two other measures, precision and recall, may be used if there is an 'authoritative source' to which the codings may be compared. *Precision* expresses the proportion of the occurrences found that have been correctly coded:

$$\text{precision} = \frac{\text{found} - \text{incorrect}}{\text{found}}$$

Found represents everything that was marked by the coder, and *incorrect* represents the incorrect markups made by the coder, as determined by the authority. *Recall* expresses the proportion of occurrences that have been found:

$$\text{recall} = \frac{\text{all} - \text{missing}}{\text{all}}$$

All represents all occurrences present in the corpus, as determined by the authority, and *missing* represents those occurrences that were not identified by the coder [Bernsen et al. 1998].

We recommend that

- any evaluation made of a coding scheme is referenced from the documentation of the coding scheme so that it is easy to find;

- evaluation methods used and evaluation process are clearly described;
- evaluation results are clearly documented.

4.5 Coding scheme selection and adaptation

We have now discussed recommendations related to the creation, documentation, and evaluation of coding schemes. However, it is of course much easier if there is already a well-documented and – evaluated coding scheme available somewhere which fits the needs one may have. And it is even better is this coding scheme comes with tools support.

No matter if one is going to create a coding scheme or select an already existing scheme, one should consider the issues listed in Section 4.1. Moreover, one should know who will be doing the coding, i.e. which level of expertise is available for this task.

When this is done, we recommend to start looking for a coding scheme which satisfies the identified constraints before a possible decision is made to create one's own coding scheme. Locating existing coding schemes is not necessarily easy to do for the moment since there are many sources which one may consult, including, e.g., ISLE NIMM Report D9.1 [Knudsen et al. 2002a], proceedings of conferences such as LREC (Language Resources and Evaluation Conference), the ELRA/ELDA website (http://www.elda.fr/), and free-style web search.

The checking of which coding schemes exist could be greatly facilitated if coding schemes are:

- well-documented, following the recommendations in Section 4.2;
- available on the web in the form of collections maintained at a small number of sites.

Documentation following the recommendations above would also greatly facilitate comparison of different coding schemes.

As regards NIMM coding scheme availability, we have collected information about 21 coding schemes in ISLE Report D9.1. This documentation is available at the ISLE NIMM website at isle.nis.sdu.dk. At this same website, a form is available for acquiring information about coding schemes which are not yet present at the site. The slots in the form correspond to those used in describing the 21 coding schemes which are presented at the website already. We would like to take this opportunity to encourage readers of this report to fill in the form if they are willing to make their coding scheme known and available to colleagues.

If one or several coding schemes are found which could be candidates for selection, we recommend to consider at least the following criteria before selection and to weight the criteria according to their importance in the specific case.

- coding scheme documentation;
- coding scheme evaluation;
- coding scheme extensibility, if applicable (cf. Section 4.1);
- coding scheme adaptability.

By extensibility we understand that new tags and their conceptual descriptions can easily be added. Extensibility becomes easier if the coding scheme includes a description of how this should be done. Adaptation of a coding scheme may include coding scheme extension but may also include other forms of changes to the original scheme, such as partial replacement of the tag set, a different coding procedure, or other/more coding files referenced. Whether adaptation - which is typically a larger operation than extending a scheme - is the right choice, depends at least on:

- how many changes are needed to make the coding scheme fit one's purpose;
- how easy will it be to make the adaptation;
- what will be gained from making the adaptation compared to creating a new coding scheme.

Ease of adaptation depends on the coding scheme itself and the available documentation.

The gain from making adaptation may range from not having to create an entirely new coding scheme and not having to do all the documentation of a coding scheme from scratch, to getting access to tools support which may greatly facilitate the annotation and analysis process. If the gain is small, it may, in fact, pay off to create a new coding scheme instead, one which fits one's purposes one hundred percent. Available tools support, on the other hand, means a great advantage and may make adaptation the optimal choice.

Acknowledgements

We gratefully acknowledge the support of the ISLE project by the European Commission's HLT Programme. We would also like to thank Laurent Romary for his contribution to the description of the ISO TC37/SC4 initiative.

5. References

This reference list includes literature referenced in Chapters 1, 2, and 4. For literature referenced in Chapter 3 see this chapter.

AAMAS Workshop on "Embodied conversational agents - let's specify and evaluate them!" Marriot, A., Pelachaud, C., Rist, T., Ruttkay, S., and Vilhjalmsson, H. (Eds.). http://www.vhml.org/workshops/AAMAS/papers.html. In conjunction with The First International Joint Conference on Autonomous Agents and Multi-Agent Systems, Bologna, Italy, 16 July, 2002.

Bernsen, N. O.: Multimodality in language and speech systems - from theory to design support tool. In Granström, B. (Ed.): *Multimodality in Language and Speech Systems*. Dordrecht: Kluwer Academic Publishers 2002.

Bernsen, N.O., Dybkjær, H. and Dybkjær, L.: Designing Interactive Speech Systems. From First Ideas to User Testing. Springer Verlag 1998.

Dybkjær, H. and Dybkjær, L.: Measuring Transaction Success in Spoken Dialogue Information Systems. Proceedings of Nordtalk Symposium on Relations between Utterances, Copenhagen, December 2002.

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 Report D11.1, January 2001.

Dybkjær, L., Bernsen, N.O., Dybkjær, H., McKelvie, D. and Mengel, A.: The MATE Markup Framework. MATE Report D1.2, November 1998.

HF2002 Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges. Melbourne, Australia, http://www.vhml.org/workshops/HF2002/papers.shtml, 29th November, 2002.

Klein, M., Bernsen, N.O., Davies, S., Dybkjær, L., Garrido, J., Kasch, H., Mengel, A., Pirrelli, V., Poesio, M., Quazza, S. and Soria, S.: Supported Coding Schemes. MATE Report 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 Report 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 Report D8.1, 2002b.

Martell, C., Osborn, C., Friedman, J. and Howard. P.: FORM: A Kinematic Annotation Scheme and Tool for Gesture Annotation. Proceedings of the Workshop on "Multimodal Resources and Multimodal Systems Evaluation". During the Third International Conference on Language Resources and Evaluation (LREC'2002), Las Palmas, Canary Islands, Spain, 2002.

McNeill, D.: Hand and mind - what gestures reveal about thoughts. University of Chicago Press, 1992.

Pirker, H. and Krenn, B.: Report D9c of the NECA project: Report on the assessment of existing markup languages for avatars, multimedia and multimodal systems on the WWW. OFAI.http://www.ai.univie.ac.at/NECA/publications/publication_docs/d9c.pdf, 2002.

Piwek, P., Krenn, B., Schröder, M., Grice, M., Baumann, S. and Pirker, H.: RRL: A Rich Representation Language for the Description of Agent Behaviour in NECA. Workshop on Embodied conversational agents - let's specify and evaluate them!. In conjunction with The First International Joint Conference on Autonomous Agents and Multi-Agent Systems, Bologna, Italy, 2002.

PRICAI International Workshop on Lifelike Animated Agents Tools, Affective Functions, and Applications. In conjunction with Seventh Pacific Rim International Conference on Artificial

Intelligence. Tokyo, Japan. http://www.miv.t.u-tokyo.ac.jp/~helmut/pricai02-agents-ws.html, August 19, 2002.

VHML Working Draft v0.3, October 21st 2001, http://www.vhml.org/downloads/VHML/vhml.pdf, 2001.

W3C 2002. Multimodal Interaction Activity: http://www.w3.org/2002/mmi/