Data Resources and Annotation Schemes for Natural Interactivity: Purposes and Needs

Laila Dybkjær and Niels Ole Bernsen

Natural Interactive Systems Laboratory
University of Southern Denmark
Science Park 10, 5230 Odense M, Denmark
{laila, nob}@nis.sdu.dk

Abstract

This paper reports on work carried out in the ISLE project on natural interactivity and multimodal resources. Information has been collected on a large number of corpora, coding schemes and coding tools world-wide. The paper focuses on corpora and coding schemes and the purposes for which they were developed or which they could serve.

1. Introduction

The long-term vision of natural interactivity envisions that humans communicate, or exchange information, with machines (or systems) in the same ways in which humans communicate with one another, using thoroughly co-ordinated speech, gesture, gaze, facial expression, head movement, bodily posture, and object manipulation [Bernsen 2001]. The idea of multimodality is to improve human-system interaction in various ways by using novel combinations of (unimodal) input/output modalities [Bernsen 2002]. Natural interactivity is by nature (mostly) multimodal. Across the world, researchers and companies are beginning to tap the potential of natural interactive and multimodal systems. This emerging community needs information about what is already there, how they might access it, what they might use it for, etc., in order that fewer people try to re-invent the wheel than would otherwise risk being the case. In many ways, we are only at the start of what could be a revolution in human-system interaction. It will be some time before a new community of researchers and developers, coming from what is currently an archipelago of widely dispersed areas and specialties, has consolidated in this most exciting field of exploration.

This paper provides an overview of selected aspects of the information on data resources (corpora) and annotation schemes that was collected in the European Natural Interactivity and Multimodality (NIMM) Working Group of the joint EU-HLT/US-NSF project International Standards for Language Engineering (ISLE).

ISLE is the successor of EAGLES (European Advisory Group for Language Engineering Standards) I and II and includes three working groups on lexicons, machine translation evaluation, and NIMM, respectively. The NIMM Working Group (isle.nis.sdu.dk) began its work in early 2000 and has now completed three comprehensive surveys. The surveys address NIMM data, annotation schemes, and annotation tools, respectively. Focus has been on producing descriptions which are systematically organised, follow standard formats, have been verified by the resource creators themselves, and provide interested parties in research and industry with the information they need to decide if a particular resource matches their interests. Each resource (data, coding scheme or tool) comes with contact information on its creator(s) and on how to get access to it. To our knowledge, the surveys significantly contribute to our common knowledge of the state of the art in data, coding schemes, and tools for natural interactivity and multimodal interaction. It appears that no other published work has produced comparatively large collections of information on NIMM resources.

The survey of NIMM data resources [Knudsen et al. 2002a] includes a total of 64 resources world-wide, 36 of which are facial resources and 28 are gesture resources. Several data resources combine speech with facial expression and/or gesture. The report also includes a survey of market and user needs produced by ELRA (the European Language Resources Agency) and 28 filled questionnaires collected at the Dagstuhl workshop on Coordination and Fusion in Multimodal Interaction held in late 2001.

The survey of NIMM corpus annotation schemes [Knudsen et al. 2002b] includes 7 descriptions of annotation schemes for facial expression and speech, and 14 descriptions of annotation schemes for gesture and speech. In addition, the survey draws some conclusions on current coding best practices based on the collected material.

The survey of NIMM corpus coding tools [Dybkjær et al. 2001a] describes 12 annotation tools and ongoing tool development projects, most of which support speech annotation combined with gesture annotation, facial expression annotation, or both. Conclusions on requirements to be met by a general-purpose NIMM annotation tool are made and further refined in [Dybkjær et al. 2001b].

Based on the above ISLE NIMM reports, in particular [Knudsen et al. 2002a and 2002b], this paper reviews the purposes for which the surveyed data resources and coding schemes have been used or are intended to be used, and discusses annotation best practices.

2. Purposes of data resources

This section provides an overview of the purposes for which, according to their creators, the data resources collected in ISLE NIMM have been applied or are intended to be applied (Section 2.1). A summary is then presented of selected results from a market study performed by ELRA and included in [Knudsen et al. 2002a] (Section 2.2).

2.1. Data resources

Many of the 64 reviewed NIMM data resources were found via the web. Others were found through proceedings of specialised conferences and workshops [Knudsen et al. 2002a]. When a resource can be downloaded from the web, this is indicated in the report. For each data resource, contact information is provided so that the resource creators can be contacted and asked how to obtain the resource if it is not directly accessible.

The collected data resources reflect a multitude of needs and purposes, including the following (in random order):

- automatic analysis and recognition of facial expressions, including lip movements;
- audio-visual speech recognition;
- study of emotions, communicative facial expressions, phonetics, multimodal behaviour, etc.;
- creation of synthetic graphical interface characters, including, e.g., talking heads;
- automatic person identification;
- training of speech, gesture and emotion recognisers;
- multimodal system specification and development.

In many cases, the people working with the data, in particular those working with static image analysis, have created their own resource databases. Algorithms for image analysis are sometimes dependent on lighting conditions, picture size, subjects' face orientations, etc. Thus, computer vision research groups may have had to create their own image databases with good reason. Image analysis using computer vision techniques remains a difficult task, and this may be the reason why we have primarily found static image resources produced by workers in this field.

In other areas, (dynamic) video recordings - mostly including audio - are needed. For example, studies of lip movements during speech, co-articulation, audio-visual speech recognition, temporal correlations between speech and gesture, and relationships among gesture, facial expression, and speech, all require video recordings with audio.

Across the collected data resources, re-use is a rare phenomenon. If a resource has been created for a specific application purpose, it has usually been tailored to satisfy the particular needs of its creators, highlighting, e.g., particular kinds of interaction or the use of particular modality combinations. Figure 1 provides an overview of the data resources reviewed, including the purpose(s) for which they were created or have been used.

Modalities	**Name of data resource**	**Purpose(s)**
Dynamic face	LIMSI Gaze Corpus (CAPRE)	Track face, nose and eyes.
Dynamic face, audio	Advanced Multimedia Processing Lab	Lip reading, speech-reading techniques for higher speech recognition accuracy.
	ATR Database for bimodal speech recognition	Research, speech recognition and speech-to-lip generation (animated agents, talking face), observations on the differences in lighting conditions, size of lips, and inclination of a face.
	The BT DAVID Database	Research on audio-visual technologies in speech or person recognition, synthesis, and communication of audio-visual signals.
	Data resources from the SmartKom project	Collect data for the training of speech, gesture and emotion recognisers, to develop dialogue and context models and to investigate how users interact with a machine that has far greater communication skills than at present.
	FaceWorks	Enable multimedia developers to create digital personalities.
	M2VTS Multimodal Face Database	User authentication, lip tracking, face recognition, extend the scope of application of network-based services by adding novel and intelligent functionalities enabled by automatic verification systems combining multimodal strategies (secured access based on speech, image and other information).
	M2VTS Extended Multimodal Face Database – (XM2VTSDB)	Lip tracking, eye coordinate determination, face and speech authentication. Large multi-modal database, which will enable the research community to test their multi-modal face verification algorithms on a high-quality large dataset.
	Multi-talker database	Quantitatively characterize optical speech signals, examine how optical phonetic characteristics relate to acoustic and physiological speech production characteristics, study what affects the intelligibility of optical speech signals, and apply the knowledge obtained to optical speech synthesis and automatic speech recognition.

	VIDAS (VIDeo ASsisted with audio coding and representation)	Devise suitable methodologies and algorithms for time-correlated representation, coding and manipulation of digital A/V bit streams.
	/'VCV/ database	Study lip shape characterisation during speech.
	ATR Database for Talking Face	Research.
	Audio-Visual Speech Processing Project	Research.
	Video Rewrite	Facial animation system to automate all the labelling and assembly tasks required to resynchronise existing footage to a new soundtrack.
Dynamic face, audio, gesture	NITE Floorplan Corpus (Natural Interactivity Tools Engineering)	Test resource for cross level, cross modality analysis of natural interactive communication.
	Scan MMC (Score Analysed MultiModal Communication)	Research on facial expression and gesture.
	Multi-modal dialogue corpus	Research on multi-modal dialogue.
Static face	3D_RMA: 3D database	Validation of facial 3D face acquisition by structured light, recognition experiments by 3D comparison.
	AR Face Database	Create a better resource for face recognition and expression recognition.
	AT&T Laboratories Database of Faces	Face recognition research.
	CMU Pose, Illumination, and Expression (PIE) database	Collect material for the design and evaluation of face recognition algorithms (facial expression detection, temporal issues of facial expressions and other kinds of analysis of facial expressions).
	Cohn-Kanade AU-Coded Facial Expression Database	Develop and test algorithms for facial expression analysis.
	FERET Database Demo	Face recognition.
	Psychological Image Collection at Stirling (PICS)	Psychological research (visual perception, memory and processing).
	TULIPS 1.0	Test lip-tracking algorithms.
	UMIST Face Database	Examine pose-varying face recognition.
	University of Oulu Physics-Based Face Database	Face recognition under varying illuminant spectral power distribution.
	VASC – CMU Face Detection Databases	Train and test face detection algorithms.
	Visible Human Project	Studies of anatomy, creation of synthetic models and test image segmentation algorithms.
	Yale Face Database	Research on face recognition.
	Yale Face Database B	Face recognition under various poses and illumination.
	3D Surface Imaging in Medical Applications	Medical applications.
	Facial Feature Recognition using Neural Networks	Face recognition.
	Image Database of Facial Actions and Expressions	Train neural networks to classify facial behaviours based on FACS.
	JAFFE Facial Expression Image Database	Research on facial expression.
	Photobook	Tool for performing queries on image databases based on image content.
Gesture	MPI Experiments with Partial and Complete Callosotomy Patients Corpus	Research on split-brain patients.
	National Center for Sign Language and Gesture Resources	Support research on sign language.
	ATR sign language gesture corpora	Creation of an inventory of the most important words of Japanese sign language as a basis for the development and evaluation of gesture recognition systems.
Gesture, audio	ATR Multimodal human-human interaction database	Provide a source for analysing the relation between speech and gesture.

	CHCC OGI Multimodal Real Estate Map	Compare the linguistic differences and relative ease of processing multimodal input compared with unimodal input.
	GRC Multimodal Dialogue during Work Meeting	Study the patterns of multimodal communication during a work session about collaborative conception.
	LIMSI Pointing Gesture Corpus (PoG)	Basis for specification of a recognition system
	McGill University, School of Communication Sciences & Disorders, Corpus of gesture production during stuttered speech	Study relations between gesture and stuttered speech.
	MPI Historical Description of Local Environment Corpus	Research.
	MPI Living Space Description Corpus	Research.
	MPI Locally-situated Narratives Corpus	Research.
	MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 1	Research.
	MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 2	Research.
	MPI Narrative Elicited by an Animated Cartoon "Maus" and "Canary Row" Corpus	Research.
	MPI Natural Conversation Corpus	Research.
	MPI Naturalistic Route Description Corpus 1	Research.
	MPI Naturalistic Route Description Corpus 2	Research.
	MPI Traditional Mythical Stories Corpus	Research.
	MPI Traditional Mythical Stories with Sand Drawings Corpus	Research.
	National Autonomous University of Mexico, DIME multimodal corpus	Build and test an interactive multimodal Spanish spoken - graphics system to assist human users in a geometric design task (kitchen design).
	RWC Multimodal database of gestures and speech	Build a speech and video database that can be shared among different research groups pursuing similar work that will promote research and development of multimodal interactive systems integrating speech and video data.
	University of Chicago Origami Multimodal corpus	Study origami, study learner gestures (with and without speech, collaborative gestures), learner gestures in relation to instructor gestures.
	IRISA Georal Multimodal Corpus	Study how people use speech and gestures on a tactile screen to interact with a graphical tourist map.
	LORIA Multimodal Dialogues Corpus	Research.
Gesture, gaze, audio	VISLab Cross-Modal Analysis of Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research	Understanding relationships between speech and gesture.
	LIMSI Multimodal Dialogues between Car Driver and Copilot Corpus	Study of multimodal communication between a driver and a co-pilot in different settings.
	University of Venice Multimodal Transcription of a Television Advertisement	Understanding the properties and functions of dynamic genres, including verbal and written discourse, gesture, gaze, colour, voice quality.
Gesture, face, audio	University of California Video Series on Nonverbal Communication	Research on non-verbal communication, including facial expressions, tones of voice, gestures, eye contact, spatial arrangements, patterns of touch, expressive movement, cultural differences, and other "nonverbal" acts.

Figure 1. The reviewed data resources and their purposes.

2.2. Market study

A market study on data resources and user needs was performed by ELRA. A questionnaire was sent to more than 150 people, including ELRA members and people from both industry and academia. 25 responses were received. Among others, the questionnaire included questions on (1) the types of data resources needed, used by, or offered by, respondents, (2) the kinds of task for which data resources are well suited, and (3) the areas in which data resources are being used.

2.2.1. Types of data resources needed or offered

The NIMM data resources in which the respondents seem most interested include audio, video and image resources. Audio is most popular (mentioned by 84% of the respondents) followed by video (mentioned by 52%) and image (mentioned by 28%). If a data resource has also been annotated, this is considered an advantage since value has been added. In many cases, the users of data resources produce the resources they need themselves. Sometimes these resources are also offered to other users.

Authentication: Speech verification (8), Face verification (6), User authentication (5). Other: finger print and signature, biometric authentication (speech, signature).

Recognition: Speech recognition (14), Face recognition (7), Person recognition (3), Expression recognition (3). Other: mimic, music and other sounds, gesture recognition, gestures on a touchscreen.

Analysis: Speech/lips correlation (7), Body movements tracking (lips, hands, head, arms, legs, etc.) (6). Other: co-operation between gesture and speech; acoustics, video, 3D optical, midsagital magnetometry; written language analysis.

Synthesis: Multimedia development (6), Talking heads (5), Humanoid agents (5), Avatars (2). Other: text generation.

Control: Voice control (7), Speech-assisted video (1).

Other: Information retrieval (14), Other: multimodal command languages (speech + gesture), research into cross-modality issues, multimodal dialogue (speech + gesture), linguistic research, information extraction, text summarisation.

Figure 2. Resource application list from the ELRA report in [Knudsen et al. 2002a, chapter 8]. Numbers in parentheses indicate how many respondents gave a particular answer.

2.2.2. What can data resources be used for

The questionnaire mentioned six general task categories for which data resources may be used. For each category, a number of more specific possibilities were listed. Respondents were supposed to indicate the kinds of applications they were interested in. Responses are shown in Figure 2. The primary applications of data resources are information retrieval and speech recognition, each of which were mentioned by 14 respondents. Then follows speech verification mentioned by 8, and face recognition, speech/lips correlation, and voice control, each mentioned by 7 respondents.

2.2.3. Application areas

To get an idea of the overall application or market areas for data resources, the questionnaire listed five possibilities (including "other") among which respondents were asked to choose the ones they found appropriate to their work. The area mentioned most frequently was research (21). Then follows information systems development (e.g. banking, tourism, telecommunication) (14), web applications development (10), education/training (9), and edutainment (6). Other areas proposed include security, control of consumer devices, and media archiving for content providers.

3. Purposes of annotation schemes

This section provides an overview of the purposes for which the reviewed coding schemes [Knudsen et al. 2002b] have been created or used (Section 3.1). Then follows a brief description of practices and best practices as these emerged from the collected material (Section 3.2).

3.1. Annotation schemes

There probably exists a wealth of NIMM annotation schemes most of which are tailored to a particular purpose and used solely by their creators or at the creators' site. Such coding schemes tend not to be very well described. They also tend to be hard to find. The reviewed material includes such coding schemes many of which were created by ISLE participants or people known to ISLE participants, this being the main reason why we were aware of them. Other coding schemes included are fairly general ones, in frequent use, or even considered standards in their field, cf. Section 3.2.

Nearly all the reviewed coding schemes are aimed at markup of video, possibly including audio. A couple of schemes can be used for static image markup.

The collected material comprises schemes for markup of a single modality as well as schemes for markup of modality combinations. Figure 3 provides an overview of the majority of the schemes reviewed, including the annotation purpose for which they were created. The coding scheme descriptions which have not been included below are of a more general nature and do not concern any particular coding scheme and its purpose(s).

3.2. Practices and best practices

In most cases, a coding scheme has been created because a person or site had a particular need, e.g. related to systems development.

In the area of facial expression, MPEG-4 is considered a standard and is being widely used. FACS is also used by many people but is not really well suited for markup of lip movements. ToonFace is good for 2D caricature but not for real (or life-like) facial expression. Other reviewed facial expression schemes seem to have been used by a single person or by a few people only.

In the area of gesture, the picture seems considerably more varied than for facial expression. Where facial expression is often the sole point of focus, gesture often seems to be studied along with other modalities. Only when it comes to the highly specialised area of sign

languages, the schemes we looked at focused solely on gesture. Many other gesture schemes were created to study gesture in combination with one or several other modalities with the purpose of supporting the development of a multimodal system. There are no real standards for gesture markup. HamNoSys seems to be the most frequently used among the schemes we looked at as regards gesture annotation-only. For gesture in combination with other modalities there are many schemes – mostly used by few people - but no standardisation.

The picture, provided by the survey, of a proliferation of home-grown coding schemes is supported by the 28 questionnaires in [Knudsen et al. 2002a], asking people at a multimodal interaction workshop, e.g., which coding scheme(s) they had used or planned to use for data markup. Some people did not answer the question or had not made a decision yet as to which coding scheme to use. However, in no less than 15 cases the answer indicated that a custom-made scheme would be, or was being, used. Only a few respondents also mentioned more frequently used annotation schemes, such as TEI, BAS, or HamNoSys.

Intended for markup of	**Name of coding scheme**	**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 image.
	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 image.
	MPEG-4	Define a set of parameters to define and control facial models.
	ToonFace	Code facial expression with limited detail. Developed for easy creation of 2D synthetic interface agents.
Gesture	HamNoSys	Designed as a transcription scheme for (different) sign languages.
	SWML (SignWriting Markup Language)	Code utterances in sign languages written in the SignWriting System.
	MPI GesturePhone	Transcribe 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) 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. gestural) 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 Copilot	Annotation of a resource which contains multimodal dialogues between drivers and copilots 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 web site visitor through a web site.
Speech, gesture, facial expression	SmartKom Coding scheme	Provide information about the intentional information contained in a gesture.

Figure 3. Reviewed coding schemes and their purposes.

4. Conclusion

Even if we have reviewed a large number of data resources and coding schemes, there probably exist many other NIMM corpora and coding schemes which we did not manage to identify. Many resources are not publicly accessible and their creators do not want to share them with others. Thus, they can be very hard to find. But also, our primary focus has been on resources which are accessible to people other than their creators. We believe that the collected information and resulting reports, although probably far from being exhaustive, reflect quite well the state-of-the-art in the NIMM resources area.

If this is indeed the case, some conclusions are: to a large extent, people still create their own single-purpose data resources and coding schemes without any strong guidance by best practice and standards, and hence without any strong purpose of sharing their resources with others. However, vendors of data resources exist, such as ELRA and LDC, and standards will emerge eventually and become applied. The standardisation process seems to be further advanced for facial expression than for gesture, and for gesture combined with other modalities there is still a long way to go.

In the ISLE project we do not have the resources required for regularly extending the information collected with new data, coding schemes or coding tools. Therefore, a web-based facility will be set up which will enable any interested colleague to upload information about a NIMM resource which has not been included already. We hope that our colleagues in the emerging NIMM community will use the facility to help each other by sharing their information with others and contribute to maintaining an up-to-date and valuable pool of NIMM resource information.

5. Acknowledgements

We gratefully acknowledge the support of the ISLE project by the European Commission's Human Language Technologies (HLT) Programme. We would also like to thank all ISLE NIMM participants for their report contributions which have made the present presentation possible. In particular, we have in this paper drawn on information provided by ELRA, Catherine Pelachaud, Isabella Poggi and Jean-Claude Martin.

6. References

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 (to appear).

Bernsen, N. O.: Natural human-human-system interaction. In Earnshaw, Rae, Guedj, Richard, van Dam, Andries, and Vince, John (Eds.): Frontiers of Human-Centred Computing, On-Line Communities and Virtual Environments. Berlin: Springer Verlag 2001, Chapter 24, 347-363.

Dybkjær, L., Berman, S., Bernsen, N. O., Carletta, J., Heid, U. and Llisterri, J.: Requirements Specification for a Tool in Support of Annotation of Natural Interaction and Multimodal Data. ISLE Deliverable D11.2, 2001b.

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, 2001a.

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, 2002b.

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, 2002a.

The reports referenced above are available at the website for the European ISLE NIMM Working Group at isle.nis.sdu.dk