Data, Annotation Schemes and Coding Tools for Natural Interactivity

Laila Dybkjær and Niels Ole Bernsen

Natural Interactive Systems Laboratory
Science Park 10, 5230 Odense M, Denmark
Email: {laila, nob}@nis.sdu.dk
Tel: +45 65 50 35 53, Fax: +45 63 15 72 24

ABSTRACT

This paper briefly presents results of three surveys of natural interactivity and multimodal resources carried out by a Working Group in the ISLE project on International Standards for Language Engineering. Information has been collected on a large number of corpora, coding schemes and coding tools world-wide.

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. The idea of multimodality is to improve human-system interaction in various ways by using novel combinations of (unimodal) input/output modalities [1]. Natural interactivity is by nature (mostly) multimodal.

Across the world, researchers and companies are beginning to focus on investigating and exploiting the potential of natural interactive and multimodal systems. An important foundation for work on such systems is resources, i.e. data (corpora), corpus annotation schemes and annotation tools. Given the fundamental role of speech in natural interactive communication, the speech processing community is set to have a key role in the development of natural interactive and multimodal systems. It may be expected, therefore, that our community has a strong interest in up-to-date information about existing natural interactivity resources.

This paper presents substantial and, to our knowledge, unprecedented groundwork on resources carried out in the European Natural Interaction and Multimodality (NIMM) Working Group of the EU-HLT/US-NSF project International Standards for Language Engineering (ISLE). The NIMM Working Group 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 resource descriptions which are systematic, follow standard formats, and are sufficient for providing interested parties in research and industry with the information they need to decide if a particular resource matches their interests. Each resource comes with contact information on its creator(s). The three surveys are available in html and pdf format at the ISLE NIMM website isle.nis.sdu.dk.

The report on NIMM data resources [6] reviews a total of 64 resources world-wide, 36 of which are facial resources and 28 are gesture resources. Several corpora 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 [5] reviews 7 descriptions of coding schemes for facial expression and speech, and 11 descriptions of annotation schemes for gesture and speech.

The survey of NIMM corpus coding tools [2] describes 12 annotation tools and tool projects most of which support speech annotation combined with gesture annotation, facial expression annotation, or both.

In the following, we briefly present the resources identified in the surveys (Sections 2-4).

2. DATA RESOURCES

The approach adopted for producing the NIMM data resources survey [6] was to (i) first identify a common set of criteria for selecting the data resources to be described and decide upon issues of quality of content as well as of presentation; then (ii) establish a common template for describing each data resource; (iii) identify relevant data resources world-wide based on the web, literature, contacts among researchers in the field, etc.; and, finally (iv), interact with the data resource creators to the extent possible in order to gather information on their resources and ask them to verify the data resource descriptions produced. Figure 1 lists the reviewed data resources. The overall division is into data resources which have their main focus on facial expression, possibly including speech and other modalities, and data resources which have their main focus on gesture, possibly combined with speech and other modalities.

Dynamic Facial Data Resources with Audio

1. Advanced Multimedia Processing Lab
2. ATR Database for bimodal speech recognition
3. The BT DAVID Database
4. Data resources from the SmartKom project
5. FaceWorks

6. M2VTS Multimodal Face Database
7. M2VTS Extended Multimodal Face Database – (XM2VTSDB)
8. Multi-talker database
9. NITE Floorplan Corpus (Natural Interactivity Tools Engineering)
10. Scan MMC (Score Analysed MultiModal Communication)
11. VIDAS (VIDeo ASsisted with audio coding and representation)
12. /'VCV/ database

Dynamic Facial Data Resources without Audio

1. LIMSI Gaze Corpus (CAPRE)

Static Facial Data Resources

1. 3D_RMA: 3D database
2. AR Face Database
3. AT&T Laboratories Database of Faces
4. CMU Pose, Illumination, and Expression (PIE) database
5. Cohn-Kanade AU-Coded Facial Expression Database
6. FERET Database Demo
7. Psychological Image Collection at Stirling (PICS)
8. TULIPS 1.0
9. UMIST Face Database
10. University of Oulu Physics-Based Face Database
11. VASC – CMU Face Detection Databases
12. Visible Human Project
13. Yale Face Database
14. Yale Face Database B

Lesser Known/Used Facial Data Resources

1. 3D Surface Imaging in Medical Applications
2. ATR Database for Talking Face
3. Audio-Visual Speech Processing Project
4. Facial Feature Recognition using Neural Networks
5. Image Database of Facial Actions and Expressions
6. JAFFE Facial Expression Image Database
7. Multi-modal dialogue corpus
8. Photobook
9. Video Rewrite

Gesture Data Resources

1. ATR Multimodal human-human interaction database
2. CHCC OGI Multimodal Real Estate Map
3. GRC Multimodal Dialogue during Work Meeting
4. LIMSI Multimodal Dialogues between Car Driver and Copilot Corpus
5. LIMSI Pointing Gesture Corpus (PoG)
6. McGill University, School of Communication Sciences & Disorders, Corpus of gesture production during stuttered speech
7. MPI Experiments with Partial and Complete Callosotomy Patients Corpus
8. MPI Historical Description of Local Environment Corpus
9. MPI Living Space Description Corpus
10. MPI Locally-situated Narratives Corpus
11. MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 1
12. MPI Narrative Elicited by an Animated Cartoon "Canary Row" Corpus 2
13. MPI Narrative Elicited by an Animated Cartoon "Maus" and "Canary Row" Corpus
14. MPI Natural Conversation Corpus
15. MPI Naturalistic Route Description Corpus 1
16. MPI Naturalistic Route Description Corpus 2
17. MPI Traditional Mythical Stories Corpus
18. MPI Traditional Mythical Stories with Sand Drawings Corpus
19. National Autonomous University of Mexico, DIME multimodal corpus
20. National Center for Sign Language and Gesture Resources
21. RWC Multimodal database of gestures and speech
22. University of Chicago Origami Multimodal corpus
23. VISLab Cross-Modal Analysis of Signal and Sense Data and Computational Resources for Gesture, Speech and Gaze Research

Lesser Known/Used Gesture Data Resources

1. ATR sign language gesture corpora
2. IRISA Georal Multimodal Corpus
3. LORIA Multimodal Dialogues Corpus
4. University of California Video Series on Nonverbal Communication
5. University of Venice Multimodal Transcription of a Television Advertisement

Figure 1. Data resources surveyed.

2.1. Conclusions on data resources

The reviewed data resources reflect a multitude of coding needs and purposes, including: 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 characters, including, e.g., talking heads; automatic person identification; training of speech, gesture and emotion recognisers; multimodal and natural interactive systems specification and development.

Significantly, across all the collected data resources, re-use is a rare phenomenon. If a data 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. However, the lack of re-use may also to some extent be due to the fact that existing resources may be difficult to locate. On the other hand, it should be mentioned that vendors of data resources exist (e.g. ELRA and LDC). See [3] for a more detailed description of the intended and actual use of the surveyed data resources.

3. ANNOTATION SCHEMES

The approach adopted for producing the ISLE NIMM annotation schemes survey [5] was basically the same as the one described for data resources in section 2. Thus, the four steps of (i) identifying selection criteria and deciding on issues concerning quality of content and of presentation, (ii) establishing a common template for describing each coding scheme, (iii) identifying relevant coding schemes, and (iv), interacting with the coding scheme creators, were also followed in the coding schemes description process. Figure 2 lists the reviewed coding schemes. The overall division is into facial coding schemes and gesture coding schemes. It should be noted that four of the entries in Figure 2 (no. 2 under lesser know facial coding schemes, no. 11 under gesture schemes, and no's 2 and 3 under lesser known gesture schemes) are not coding

schemes proper but rather more general descriptions of coding schemes for particular modalities.

Facial Coding Schemes

1. The Alphabet of eyes: formational parameters of gaze
2. Facial Action Coding System – FACS
3. The Maximally Discriminative Facial Movement Coding System (MAX)
4. MPEG-4 SNHC (Moving Pictures Expert Group, Synthetic/Natural Hybrid Coding)
5. ToonFace

Lesser Known/Used Facial Coding Schemes

1. BABYFACS – Facial Action Coding System for Baby Faces
2. General description of coding schemes for hand annotation of mouth and lip movements and speech

Gesture Coding Schemes

1. DIME: National Autonomous University of Mexico, Multimodal extension of DAMSL
2. HamNoSys - Hamburg Notation System for Sign Languages
3. HIAT -- Halbinterpretative Arbeitstranskriptionen
4. LIMSI Coding Scheme for Multimodal Dialogues between Car Driver and Copilot
5. MPI GesturePhone
6. MPI Movement Phase Coding Scheme
7. MPML - A Multimodal Presentation Markup Language with Character Agent Control Functions
8. SmartKom Coding scheme
9. SWML (SignWriting Markup Language)
10. TUSNELDA Corpus Annotation standard
11. General description of coding schemes for prosody, gestures and speech

Lesser Known/Used Gesture Coding Schemes

1. LIMSI TYCOON scheme for analysing cooperation between modalities
2. W3C Working Draft on Multimodal Requirements for Voice Markup Languages
3. The New England Regional Leadership Non-Verbal Coding scheme

Figure 2. Annotation schemes surveyed.

3.1. Conclusions on coding schemes

Nearly all the reviewed coding schemes are aimed at markup of video, sometimes including audio. A couple of schemes are used for static image markup. See [3] for a more detailed description of the intended and actual use of the surveyed coding schemes.

Based on the collected material, it may be concluded that there is still a long way to go before we will be able to code, on a scientifically sound basis, natural interactive communication and multimodal information exchange in all their forms, at any relevant level of analytic detail, and in all their cross-level and cross-modality forms. This observation is already true for the coding of spoken dialogue at several important levels of analysis, such as dialogue acts or co-reference, as shown in [4]. When we move beyond spoken dialogue annotation to considering facial coding, we do find a couple of general and substantially evaluated coding schemes for different aspects of the facial expression of information (eyes, facial muscles), i.e. MPEG-4 and FACS, cf. Figure 2. It seems clear, however, that we still need a number of higher-level facial 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, moreover, the state of the art is even further from the ideal described above. General coding schemes which go beyond the classification of gesture into few broad categories, and as opposed to coding schemes designed for the study of particular kinds of task-dependent gesture, are hard to find at all, the only exception being in the specialised field of sign languages. Also, the state of evaluation of particular gesture coding schemes is generally poor. Finally, when it comes to the most complex, and perhaps ultimately the most significant, of all areas of natural interactive behaviour annotation, i.e. that of cross-level and cross-modality coding, no coding scheme of a general-purpose nature would seem to exist at all. Even special-purpose coding schemes are hard to come by as yet in this area.

4. ANNOTATION TOOLS

A number of tools in support of natural interactivity and multimodal data annotation, i.e. tools which support annotation of spoken dialogue, facial expression, gesture, bodily posture, or cross-modality issues, were reviewed in the ISLE NIMM coding tools survey [2]. For this survey, and in view of the expected scarcity of NIMM coding tools world-wide, no particular selection criteria were set up except that it should be possible to somehow get access to the tools reviewed. With this exception, the same approach was taken as for the descriptions of NIMM data resources and coding schemes, i.e. (i) a common template was established for describing each coding tool, (ii) relevant coding tools were identified, and (iii) coding tool creators were contacted. Figure 3 lists the reviewed NIMM coding tools and tool projects. A couple of tools had not yet been implemented at the time of the review. Two tools are commercial (The Observer and SyncWriter). The rest are research tools (or projects). MATE is a limiting case in another sense, because the MATE Workbench only supports spoken dialogue and text annotation. The tool is included because of its advanced properties for multi-level and cross-level annotation, which may show the way towards building a general-purpose natural interactivity coding tool. The CLSU Toolkit coding tool is for output generation. Finally, so far, at least, the SmartKom project is a user rather than a provider of NIMM coding tools.

1. ***Anvil:*** Annotation of Video and Language Data, is a Java-based tool for annotating digital video files. See www.dfki.de/~kipp/anvil
2. ***ATLAS:*** Architecture and Tools for Linguistic Analysis Systems. No tool was available at the time of the review. See www.itl.nist.gov/iaui/894.01/atlas
3. ***CLAN:*** Computerized Language Analysis, is a program designed specifically for analysing data transcribed in the format of the Child Language Data Exchange System (CHILDES). Transcriptions can be linked to audio or video files. See childes.psy.cmu.edu
4. ***CSLU Toolkit:*** Center for Spoken Language Understanding Toolkit, is a suite of tools including the Rapid Application Developer, BaldiSync (for facial animation), SpeechView, OGIsable (an annotation tool),

speech recognition tools, and a programming environment (CSLUsh). OGIsable is the only annotation tool included. It allows the user to attach properties to a text before it is spoken (via the Festival TTS engine), e.g. to synthesise facial expression synchronised with speech output. See cslu.cse.ogi.edu/toolkit/

5. ***MATE:*** Multilevel Annotation Tools Engineering. The MATE Workbench is a Java-based tool in support of multi-level annotation of spoken dialogue corpora and information extraction from annotated corpora. See mate.nis.sdu.dk
6. ***MPI tools:*** CAVA and EUDICO/Computer Assisted Video Analysis and European Distributed Corpora. Both tools support annotation of audio-visual files and information extraction. See www.mpi.nl/world/tg/CAVA/CAVA.html, www.mpi.nl/world/tg/lapp/eudico/eudico.html
7. ***MultiTool*** was developed in a project on a Platform for Multimodal Spoken Language Corpora. It is a Java-based tool in support of the creation and use of multimodal spoken language corpora (audio and video). See www.ling.gu.se/multitool
8. ***The Observer*** is a commercial system for the collection, analysis, presentation and management of video data. It can be used to record activities, postures, movements, positions, facial expressions, social interactions or any other aspect of human or animal behaviour as time series of tagged data. See www.noldus.com/products/index.html?observer/
9. ***Signstream*** was developed as part of the American Sign Language Linguistic Research Project. It is a database tool for analysis of linguistic data captured on video. See web.bu.edu/asllrp/SignStream
10. ***SmartKom*** is a large-scale project which aims to merge the advantages of spoken dialogue-based communication with the advantages of a mixture of graphical user interfaces and gesture and mimetic interaction. SmartKom uses tools developed elsewhere: in Verbmobil for audio annotation, Anvil for mimics and gesture coding. See www.smartkom.org
11. ***SyncWriter*** is a commercial tool for transcription and annotation of synchronous "events" such as speech and video data. See www.sign-lang.uni-hamburg.de/software/software.html
12. ***TalkBank*** is a project which aims to provide standards and tools for creating, searching, and publishing primary materials via networked computers. No tool had been developed in the project at the time of the review. However, further development of Transcriber (for orthographic transcription) and CLAN (see above) was ongoing. See www.talkbank.org

Figure 3. Annotation tools and projects surveyed.

4.1. Conclusions on coding tools

Figure 3 confirms our initial expectations as to the present scarcity of coding tools for natural and multimodal interactive behaviour. Current needs for more general-purpose NIMM annotation tools may be viewed as being reflected in the nature of the reviewed tools many of which are intended to be somewhat general-purpose rather than supporting one particular project's needs. When one inspects the properties of the tools in more detail [2], it becomes quite clear that it is far from easy to build adequate and robust, general-purpose NIMM coding tools. Moreover, tools originating from research projects are usually research demonstrators with what that entails in terms of fragile and buggy software. Today's users are aware that the tools which are currently available are far from optimal. It is up to the research and development community to try to meet their needs.

5. REFERENCES

[1] 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).

[2] 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, 2001.

[3] Dybkjær, L. and Bernsen, N. O.: Data Resources and Annotation Schemes for Natural Interactivity: Purposes and Needs. Proceedings of the LREC'2002 Workshop on Multimodal Resources and Multimodal Systems Evaluation, 2002.

[4] 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 Deliverable D1.1, 1998. MATE reports are available at mate.nis.sdu.dk

[5] 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.

[6] 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 ISLE NIMM surveys presented in this paper are available in html and pdf format at isle.nis.sdu.dk under reports. The surveys include links to the reviewed data resources, annotation schemes and coding tools.

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 European ISLE NIMM participants for their contributions to the surveys described in this paper: CNRS-LIMSI (Paris, France), UROME (Rome, Italy), MPI (Nijmegen, the Netherlands), DFKI (Saarbrücken, Germany), IMS (Stuttgart, Germany), ELRA (Paris, France), ILC (Pisa, Italy), HCRC (Edinburgh, UK), and DFE (Barcelona, Spain).