

We thank The MITRE Corporation for their  
administrative support  
in the organization of the workshop

The MITRE Corporation  
202 Burlington Road  
Bedford, MA 01730

MITRE

[www.mitre.org](http://www.mitre.org)

# Multimodal Resources and Multimodal Systems Evaluation

## Workshop Program

Saturday, June 1, 2002

Palacio de Congreso de Canarias

8:00 a.m. Welcome  
Mark Maybury (*MITRE, USA*) and Jean-Claude Martin (*LIMSI-CNRS, France*)

### Resources and Annotation: Multimodal

8:30 a.m. Data Resources and Annotation Schemes for Natural Interactivity  
*Laila Dybkjær and Niels Ole Bernsen*  
*University of Southern Denmark, Denmark*

8:50 a.m. Metadata Set and Tools for Multimedia/Multimodal Language Resources  
*P. Wittenburg, D. Broeder, Freddy Offenga, and Don Willems,*  
*Max Planck Institute for Psycholinguistics, The Netherlands*

### Resources and Annotation: Gesture and Speech

9:10 a.m. FORM: A Kinematic Annotation Scheme and Tool for Gesture Annotation  
*Craig Martell, Chris Osborn, Jesse Friedman, and Paul Howard,*  
*University of Pennsylvania, USA*

9:30 a.m. Cross-Linguistic Studies of Multimodal Communication  
*P. Wittenburg, S. Kita, and H. Brugman,*  
*Max Planck Institute for Psycholinguistics, The Netherlands*

### Resources and Annotation: Facial Expressions, Speech, Integration

9:50 a.m. Development of User-State Conventions for Multimodal Corpus in SmartKom  
*Silke Steininger, Susen Rabold, Olga Dioubina, and Florian Schiel,*  
*Ludwig-Maximilians University, Munich, Germany*

10:10 a.m. Integration of multi-modal data and annotations into a simple extendable form: the extension of the BAS Partitur Format  
*Florian Schiel, Silke Steininger, Nicole Beringer,*  
*Ulrich Tuerk, and Susen Rabold,*  
*University of Munich, Germany*

10:40 a.m. Multimodal Resources Group Discussion  
*All*

**11:00 – 11:20 a.m. Morning Break**

## **Annotation Tools**

- 11:20 a.m.     Multimodal Corpus Authoring System  
*Anthony Baldry, Univ. of Pavia, Italy, and Christopher Taylor, Univ. of Trieste, Italy*
- 11:40 a.m.     The Observer Video-Pro: Professional System for Collection,  
Analysis and Presentation of Observational Data  
*Niels Cadée, Erik Meyer, Hans Theuws, and Lucas Noldus,  
Noldus Information Technology, The Netherlands*
- 11:20 a.m.     Data Resources and Annotation Schemes for Natural Interactivity  
*Laila Dybkjær and Niels Ole Bernsen  
University of Southern Denmark, Denmark*
- 11:40 a.m.     Metadata Set and Tools for Multimedia/Multimodal Language Resources  
*P. Wittenburg, D. Broeder, Freddy Offenga, and Don Willems,  
Max Planck Institute for Psycholinguistics, The Netherlands*

**13:00 p.m.     Lunch**

## **Multimodal Fusion**

- 14:30 p.m.     Prosody based co-analysis of Deictic Gestures and Speech  
in Weather Narration Broadcast  
*Kettebekov Sanshzar, Yeasin Mohammed, Krahnstoever Nils, SharmaRajeev,  
Dept. of CS and Engineering, Pennsylvania State University, USA*
- 14:50 p.m.     A Generic Formal Description Technique for Fusion Mechanisms of  
Multimodal Interactive Systems  
*Philippe Palanque and Amélie Schyn, LIIHS – IRIT, Université Toulouse, France*

## **Research Infrastructure**

- 15:10 p.m.     Eye Bed  
*Ted Selker, Winslow Burlison, Jessica Scott, and Mike Li,  
MIT Media Lab, Cambridge, USA*
- 15:40 pm.     MUMIN: A Nordic Network for MultiModal Interfaces  
*Patrizia Paggio, Center for Sprogteknologi, Copenhagen,  
Kristiina Jokinen, University of Art and Design, Helsinki, and  
Arne Jönsson, University of Linköping*

## **System Evaluation**

- 16:00 pm     PROMISE - A Procedure for Multimodal Interactive System Evaluation

*Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel, Uli Türk,  
University of Munich, Germany*

16:30 - 17:00 p.m. **Afternoon Break**

17:00 p.m. Final Group Discussion  
*All*

18:00 p.m. Close



## Table of Contents

	<b><u>Page</u></b>
Preface .....	xiii
Author Index.....	xix
 <b><i>Resources and Annotation</i></b>	
Data Resources and Annotation Schemes for Natural Interactivity: Purposes and Needs	
<i>Laila Dybkjær and Niels Ole Bernsen</i> .....	1
Metadata Set and Tools for Multimedia/Multimodal Language Resources	
<i>P. Wittenburg, D. Broeder, F. Offenga and D. Willems</i> .....	9
The FORM Gesture Annotation System	
<i>Craig Martell, Chris Osborn, Jesse Friedman and Paul Howard</i> .....	15
Sample Annotated Video Using Anvil and FORM	
<i>Craig Martell</i> .....	23
Cross-Linguistic Studies of Multimodal Communication	
<i>P. Wittenburg, S. Kita, and H. Brugman</i> .....	25
Development of the User–State Conventions for the Multimodal Corpus in SmartKom	
<i>Silke Steininger, Susen Rabold, Olga Dioubina, and Florian Schiel</i> .....	33
Integration of Multi-Modal Data and Annotations into a Simple Extendable Form: the Extension of the BAS Partitur Format	
<i>Florian Schile, Silke Steininger, Nicole Beringer, Ulrich Tuerk, and     Susen Rabold</i> .....	39
 <b><i>Annotation Tools</i></b>	
Multimodal Corpus Authoring System: multimodal corpora, subtitling and phasal analysis	
<i>Anthony Baldry and Chris Taylor</i> .....	45
The Observer <sup>®</sup> Video-Pro: a Versatile Tool for the Collection and Analysis of Multimodal Behavioral Data	
<i>Niels Cadée, Erik Meyer, Hans Theuws and Lucas Noldus</i> .....	53

## Table of Contents (Concluded)

### ***Multimodal Fusion***

Prosody Based Co-analysis of Deictic Gestures and Speech in Weather Narration Broadcast <i>Sanshzar Kettebekov, Mohammed Yeasin, Nils Krahnstoever, and Rajeev Sharma.....</i>	57
A Generic Formal Description Technique for Fusion Mechanisms of Multimodal Interactive Systems <i>Philippe Palanque and Amélie Schyn .....</i>	63

### ***Gaze Interaction***

Eye-Bed <i>Ted Selker, Winslow Burlison, Jessica Scott, and Mike Li .....</i>	71
--	----

### ***Multimodal System Evaluation***

PROMISE - A Procedure for Multimodal Interactive System Evaluation <i>Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel and Uli Türk.....</i>	77
--	----

### ***Research Infrastructure***

MUMIN A Nordic Network for MultiModal INterfaces <i>Patrizia Paggio, Kristiina Jokinen, and Arne Jönsson.....</i>	81
--	----

# Preface

## **Motivation**

Individual organizations and countries have been investing in the creation of resources and methods for the evaluation of resources, technologies, products and applications. This is evident in the US DARPA HLT programme, the EU HLT programme under FP5-IST, the German MTI Program, the Francophone AUF programme and others. The European 6th Framework program (FP6<sup>1</sup>), planned for a start in 2003, includes multilingual and multisensorial communication as major R&D issues. Substantial mutual benefits can be expected from addressing these issues through international cooperation. Nowhere is this more important than in the relatively new areas of multimedia (i.e., text, audio, video), multimodal (visual, auditory, tactile), and multicodal (language, graphics, gesture) communication.

Multimodal resources are concerned with the capture and annotation of multiple modalities such as speech, hand gesture, gaze, facial expression, body posture, graphics, etc. Until recently, only a handful of researchers have been engaged in the development of multimodal resources and their application in systems. Even so, most have focused on a limited set of modalities, custom annotation schemes, within a particular application domain and within a particular discipline. Until now, the collection and annotation of multimodal corpora has been made on an individual basis; individual researchers and teams typically develop custom coding schemes and tools within narrow task domains. As a result, there is a distinct lack of shared knowledge and understanding in terms of how to compare various coding schemes and tools. This makes it difficult to bootstrap off of the results and experiences of others. Given that the annotation of corpora (particularly multimodal corpora) is very costly, we anticipate a growing need for the development of tools and methodologies that enable the collaborative building and sharing of multimodal resources.

## **Increased International Attention**

Recently, several projects, initiatives and organisations have addressed multimodal resources with a federative approach:

- At LREC2000, a workshop addressed the issue of multimodal corpora, focusing on meta-descriptions and large corpora  
<http://www.mpi.nl/world/ISLE/events/LREC%202000/LREC2000.htm>
- NIMM is a working group on Natural Interaction and Multimodality under the IST-ISLE project (<http://isle.nis.sdu.dk/>). Since 2001, NIMM has been engaged with conducting a survey of multimodal resources, coding schemes and annotation tools. Currently, more than 60 corpora are described in the survey. The ISLE project is developed both in Europe and in the USA (<http://www ldc.upenn.edu/sb/isle.html>).
- In November 2001, ELRA (European Language Resources Association) conducted a survey of multimodal corpora including marketing aspects (<http://www.icp.inpg.fr/ELRA/>).
- In November 2001, a Working Group at the Dagstuhl Seminar on Multimodal Fusion and Coordination received 28 completed questionnaires from participating researchers; 21 announced their intention to collect and annotate multimodal corpora in the future.  
([http://www.dfki.de/~wahlster/Dagstuhl\\_Multi\\_Modality/](http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality/))
- Several recent surveys have focused specifically on multimodal annotation coding schemes and tools (COCOSDA, LDC, MITRE).

---

<sup>1</sup> <http://www.cordis.lu/rtd2002/fp-debate/fp.htm>

Other recent initiatives in the United States include:

- NIST Automatic Meeting Transcription Project ([http://www.nist.gov/speech/test\\_beds/mr\\_proj](http://www.nist.gov/speech/test_beds/mr_proj)): "The National Institute of Standards and Technology (NIST) held an all-day workshop entitled "Automatic Meeting Transcription Data Collection and Annotation" on 2 November 2001. "The workshop addressed issues in data collection and annotation approaches, data sharing, common annotation standards and tools, and distribution of corpora. ... To collect data representative of what might be expected in a functional meeting room of the future, [NIST has] created a media- and sensor-enriched conference room containing a variety of cameras and microphones."
- ATLAS (<http://www.nist.gov/speech/atlas>): Also at NIST, "ATLAS (Architecture and Tools for Linguistic Analysis Systems) is a recent initiative involving NIST, LDC and MITRE. ATLAS addresses an array of applications needs spanning corpus construction, evaluation infrastructure, and multimodal visualisation."
- TALKBANK (<http://www.talkbank.org>): TALKBANK is funded by the National Science Foundation (NSF). Its goal "is to foster fundamental research in the study of human and animal communication. TalkBank will provide standards and tools for creating, searching, and publishing primary materials via networked computers." One of the six sub-groups is concerned with communication by gesture and sign.

## **Objective**

The primary purpose of this one day workshop (feeding into a subsequent half day Multimodal Roadmap workshop) is to report and discuss multimodal resources, annotation standards, tools and methods, and evaluation metrics/methods, as well as strategize jointly about the way forward. The workshop consists of short presentations and facilitated sessions with the intent of jointly identifying grand challenge problems, a shared understanding of and plan for multimedia resources and applications, and identification of methods for facilitating the creation of multimedia resources.

## **Scope**

The workshop focuses on multimodal resources, annotation and evaluation. Workshop participants were encouraged to annotate multimodal corpora samples using their own coding scheme or tool and report results at the workshop. Topics in the call for papers, listed in its entirety at <http://www.lrec-conf.org/lrec2002/lrec/wksh/Multimodality.html>, included but were not limited to:

- Guidelines, standards, specifications, models and best practices for multimedia and multimodal LR
- Methods, tools, and procedures for the acquisition, creation, management, access, distribution, and use of multimedia and multimodal LR
- Methods for the extraction and acquisition of knowledge (e.g. lexical information, modality modelling) from multimedia and multimodal LR
- Integration of multiple modalities in LR (speech, vision, language)
- Ontological aspects of the creation and use of multimodal LR
- Machine learning for and from multimedia (i.e., text, audio, video), multimodal (visual, auditory, tactile), and multicodal (language, graphics, gesture) communication
- Exploitation of multimodal LR in different types of applications (information extraction, information retrieval, meeting transcription, multisensorial interfaces, translation, summarisation, www services, etc.)
- Multimodal information presentation
- Multimedia and multimodal metadata descriptions of LR
- Applications enabled by multimedia and multimodal LR
- Benchmarking of systems and products; use of multimodal corpora for the evaluation of real systems
- Processing and evaluation of mixed spoken, typed, and cursive (e.g., pen) language processing
- Evaluation of multimodal document retrieval systems (including detection, indexing, filtering, alerting, question answering, etc.)
- Automated multimodal fusion and/or multimodal generation (e.g., coordinated speech, gaze, gesture, facial expressions)

Table 1 below lists the papers included in the workshop, the primary task focus of the article, the kinds of modalities focused on, and the multimodal research issues addressed in the papers.

**TABLE 1. Overview of Contributions**

<b>Focus</b>	<b>Contribution</b>	<b>Author(s)</b>	<b>Modality</b>	<b>Research Issues</b>
<i>Resources and Annotation</i>	Data Resources and Annotation Schemes for Natural Interactivity: Purposes and Needs	Laila Dybkjær and Niels Ole Bernsen	multimodal	Natural interactivity, data resources, coding schemes, coding purposes, coding needs
<i>Resources and Annotation</i>	Metadata Set and Tools for Multimedia/Multimodal Language Resources	P. Wittenburg, D. Broeder, Freddy Offenga, Don Willems	multimodal	Metadata
<i>Resources and Annotation</i>	FORM: A Kinematic Annotation Scheme and Tool for Gesture Annotation	Craig Martell, Chris Osborn, Jesse Friedman	gesture and speech	Gesture, Gesture Annotation, Multimodal Annotation, Annotation Tools, Annotation Graph Formalism
<i>Resources and Annotation</i>	Multimodal Annotation Sample	Craig Martell	gesture	Gesture annotation
<i>Resources and Annotation</i>	Cross-Linguistic Studies of Multimodal Communication	P. Wittenburg, S. Kita, H. Brugman	Gesture and speech	Cross-linguistic studies of multimodal communication
<i>Resources and Annotation</i>	Development of the User–State Conventions for the Multimodal Corpus in SmartKom	Silke Steininger, Susen Rabold, Olga Dioubina, Florian Schiel	multimodal (facial expressions and speech prosody)	multi–modal, annotation, user–states, human–machine interaction, coding conventions
<i>Resources and Annotation</i>	Integration of multi-modal data and annotations into a simple extendable form: the extension of the BAS Partitur Format	Florian Schiel, Silke Steininger, Nicole Beringer, Ulrich Tuerk, Susen Rabold	multimodal	integration, multimodal, annotation Quick Time, BAS Partitur Format
<i>Annotation Tools</i>	Multimodal Corpus Authoring System	Anthony Baldry, Christopher Taylor	multimodal	Multimodality, cocordancing, text, resources, translation
<i>Annotation Tools</i>	The Observer Video-Pro: Professional system for collection, analysis and presentation of observational data	Niels Cadée	multimodal	methods, tools, and procedures for the acquisition, creation, management, access, distribution, and the use of multimedia and multimodal language resources

**TABLE 1. Overview of Contributions (Continued)**

<b>Focus</b>	<b>Contribution</b>	<b>Author(s)</b>	<b>Modality</b>	<b>Research Issues</b>
<i>Multimodal fusion</i>	Prosody based co-analysis of Deictic Gestures and Speech in Weather Narration Broadcast	Kettebekov Sanshzar, Yeasin Mohammed, Krahnstoever Nils, SharmaRajeev	speech and gesture	Multimodal, gesture, prosody, modality integration, speech gesture co-occurrence
<i>Multimodal fusion</i>	A Generic Formal Description Technique for Fusion Mechanisms of Multimodal Interactive Systems	Philippe Palanque, Amélie Schyn	multimodal	Formal description techniques, multimodal systems engineering, fusion mechanisms.
<i>Gaze interaction</i>	A Test-Bed for Intelligent Eye Research	Ted Selker	gaze	Gaze interaction system
<i>Multimodal System Evaluation</i>	PROMISE - A Procedure for Multimodal Interactive System Evaluation	Nicole Beringer, Ute Kartal, Katerina Louka, Florian Schiel*, Uli Türk	multimodal	Multimodality, SmartKom, dialogue system evaluation, evaluation framework
<i>Research Infrastructure</i>	MUMIN: A Nordic Network for MUltiModal INterfaces	Patrizia Paggio, Kristiina Jokinen, Arne Jönsson	multimodal	Multimodal integration, cognitive and usability studies, multimodal dialogue, multimodal research and resources in the Nordic Countries

Any international workshop demands the selfless contributions of many individuals. We first thank the authors and participants for their important contributions. We next thank the Organizing Committee for their time and effort in providing detailed and high quality reviews and counsel. And we thank Paula MacDonald at MITRE for her excellent administrative workshop support.

Mark Maybury and Jean-Claude Martin  
Workshop Co-chairs

## Workshop Organizers

*Mark Maybury (Co-chair)*  
The MITRE Corporation  
Bedford, MA USA  
[maybury@mitre.org](mailto:maybury@mitre.org)

*Jean-Claude Martin (Co-chair)*  
LIMSI-CNRS, LINC-University Paris 8  
Orsay, France  
[martin@limsi.fr](mailto:martin@limsi.fr)

## Workshop Program Committee

*Niels Ole Bernsen*  
NISLab  
University of Southern Denmark  
Odense, Denmark  
[nob@nis.sdu.dk](mailto:nob@nis.sdu.dk)

*Dybkjaer Laila*  
NISLab  
University of Southern Denmark  
Odense, Denmark  
[laila@nis.sdu.dk](mailto:laila@nis.sdu.dk)

*Harry Bunt*  
Tilburg University  
[Harry.Bunt@kub.nl](mailto:Harry.Bunt@kub.nl)

*Catherine Pelachaud*  
University of Rome "La Sapienza"  
Italy  
[cath@dis.uniroma1.it](mailto:cath@dis.uniroma1.it)

*Lisa Harper*  
The MITRE Corporation  
USA  
[lisah@mitre.org](mailto:lisah@mitre.org)

*Oliviero Stock*  
IRST  
[stock@irst.itc.it](mailto:stock@irst.itc.it)

*Michael Kipp*  
DFKI  
Germany  
[kipp@dfki.de](mailto:kipp@dfki.de)

*Wolfgang Wahlster*  
DFKI  
Germany  
[wahlster@dfki.uni-sb.de](mailto:wahlster@dfki.uni-sb.de)

*Steven Krauwer*  
ELSNET  
[steven.krauwer@elsnet.org](mailto:steven.krauwer@elsnet.org)

*Antonio Zampolli*  
Consiglio Nazionale delle Ricerche  
[pisa@ilc.pi.cnr.it](mailto:pisa@ilc.pi.cnr.it)



## Author's Index

	<u>Page</u>
Baldry, Anthony.....	45
Beringer, Nicole.....	39,77
Bernsen, Niels Ole .....	1
Broeder, D. ....	9
Brugman, H. ....	25
Burleson, Winslow.....	71
Cadée, Niels .....	53
Dioubina, Olga.....	33
Dybkjær Laila .....	1
Friedman, Jesse.....	15
Howard, Paul.....	15
Jokinen, Kristiina .....	81
Jönsson, Arne.....	81
Kartal, Ute.....	77
Kita, S. ....	25
Kettebekov, Sanshzar.....	57
Krahnstoever, Nils .....	57
Li, Mike.....	71
Louka, Katrina .....	77
Martell, Craig .....	15,23
Meyer, Erik .....	53
Noldus, Lucas .....	53
Offenga, D. F .....	9
Osborn, Chris.....	15
Paggio, Patrizia.....	81
Palanque, Philippe.....	63
Rabold, Susen .....	33,39
Selker, Ted.....	71
Schiel, Florian.....	33,39,77
Schyn, Amélie.....	63
Scott, Jessica .....	71
Sharma, Rajeev .....	57
Steininger, Silke.....	33,39
Taylor, Christopher.....	45
Theuws, Hans .....	53
Türk, Ulrich.....	39,77
Willems, D. ....	9
Wittenburg, P. ....	9,25
Yeasin, Mohammed .....	57



# 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 coordinated 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
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, midsagittal 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 Listerri, 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., Listerri, 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](http://isle.nis.sdu.dk)

# Metadata Set and Tools for Multimedia/Multimodal Language Resources

P. Wittenburg, D. Broeder, F. Offenga, D. Willems

Max-Planck-Institute for Psycholinguistics  
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands  
peter.wittenburg@mpi.nl

## Abstract

Within the ISLE Project about International Standards for Language Engineering the IMDI Metadata Initiative developed a complete environment for creating, maintaining and using metadata descriptions for multimedia/multimodal language resources. This environment includes a proposal for a suitable metadata set, tools to create, browse and search in IMDI metadata domains and suggestions about how to organize centers acting as metadata repositories. By using the IMDI approach a formulation in RDF is intended which enable the IMDI set to be integrated in Semantic Web activities.

## 1. Introduction

In 1999 the Max-Planck Institute for Psycholinguistics started using metadata to organise its multi-media corpora [1]. This project was called “Browsable Corpus” (BC) because it not only used metadata for resources in order to make them locatable by automatic procedure, but it also used metadata for creating a hierarchical structure that can be browsed for the purpose of corpus exploitation. This was achieved by recursively structuring corpora in ever-smaller sub-corpora structures with each one described by its own metadata description pointing to the metadata descriptions of its sub-corpora. Creating browsable structures this way which creates space to integrate many other types of information such as project notes, also formed a basis for efficient corpus management.

The basic concepts of BC were used as one of the inputs to the ISLE Metadata Initiative (IMDI) [2] founded in early 2000. IMDI aims to reach consensus within a representative part of the linguistic community on a standard for metadata descriptions for multimedia/multimodal language resources. The IMDI metadata set is currently being applied within projects such as DOBES [3], the CGN corpus [4] and, of course the MPI’s own corpora. Its relevance was checked for several other multimedia corpora such as the SmartKom [5] corpus. A preliminary showcase combined corpus data from 6 European institutions into one browsable and searchable domain.

## 2. Using Metadata Descriptions

A key issue in the IMDI approach is that a metadata set should be used for corpus discovery and corpus management as well as corpus exploitation. This implies that the metadata set should be able to describe the resources in sufficient detail to allow the resolution of relevant queries for the domain. It also implies that linked networks of metadata descriptions should be available, generated either automatically or manually and that it should be possible to include human readable texts or files with the metadata descriptions that can assist the user when browsing through a corpus. Corpora organized in this way can be easily integrated into bigger domains and they are an extremely useful facility for corpus managers to group all relevant information and knowledge together to facilitate corpus management. In this domain of linked

metadata descriptions the user would be able to browse and search and as a result find a single resource or a sub-corpus to work on. Consequently the user is likely to want to start a suitable tool for analysis, i.e. the metadata must contain information which indicates which operations can be executed on the resources found. Within IMDI it was anticipated that each user has his own view on corpora, therefore it was concluded that the IMDI environment should provide users the possibility of creating their own hierarchies so that several views can co-exist in parallel.

Of course, metadata will always exist as a source of information distributed via Internet, therefore all resources including the metadata descriptions themselves have to be specified as URLs. In this way metadata descriptions and connected resources can be accessed on the Internet by using standard HTTP. This simplifies the connection of different corpus domains to one super-domain. To support global searches via, for example, Dublin Core [6] based service providers, the IMDI domain is available for metadata harvesting in compliance with the Open Archives Initiative protocol [7].

Although the concept of metadata descriptions is still fairly new, the community is becoming aware that metadata descriptions will facilitate re-usage of valuable resources. Currently, most of the many resources are hidden in the storage containers of the various institutions and companies. Only few of them are visible via web-sites each having its own style of description. Since metadata are available to everyone, a domain of unified descriptions form an ideal way of informing others about available data even if the resources themselves are not directly accessible.

## 3. IMDI Metadata Set

IMDI’s guiding principles when defining a metadata set have been that the best way to describe linguistic resources is to be able to describe the events and/or performances that are involved in their creation and usage by the community. The descriptions need to contain as much detail as necessary for a user who needs to easily discover resources, quickly check their usefulness and immediately exploit them. This bottom up approach can be compared with the approach in the media and film community which defined the MPEG7 standard [8]. It can and will lead to a more extensive and structured set than, for instance, the Dublin Core set. In taking such an

approach, the metadata set found can be seen as a first step towards a more complex domain ontology.

Some argue that it is necessary to have a low-overhead metadata set, since users may not want to spend too much time in providing all the information defined by the proposed IMDI element set. For IMDI the solution is that efficient tools are provided and that almost all fields are optional. So the overhead argument in case of more elaborate metadata sets does not hold, if elements are optional as in the IMDI case. Flexibility of the set of elements was one of the recurrent requirements, since we deal with a large number of different projects all recording multimedia material. In IMDI, flexibility was introduced by allowing user definable keyword/value pairs at several levels in the metadata structure.

The IMDI set for sessions<sup>1</sup> contains the necessary elements to describe the project a resource belongs to, the responsible scientists who created it, date and location of the recording, its content, its media files and annotations and if available the its derivative source. In the following a list of all elements is given. It is not the purpose of this paper to explain in detail what all the elements represent. For this we refer to the IMDI web-site: <http://www.mpi.nl/ISLE>. An attribute specifies whether the element is just a string, constrained (c), associated with a closed vocabulary (ccv) as in the case of “Continents” or with an open vocabulary (ov) which is open for extensions, or refers to a sub-block of information (sub).

### Session

Name	str
Title	str
Date	c
Location	
Continent	ccv
Country	ccv
Region +	str
Address	str
<u>Description</u> <sup>2</sup>	sub
<u>Keys</u> <sup>3</sup>	sub
Project	
Name	str
Title	str
ID	str
<u>Contact</u>	sub
<u>Description</u> +	sub

### Collector

<sup>1</sup> Sessions are the leaves in a corpus tree and cover units of linguistic analysis or performance including their media and annotation files. The IMDI initiative has defined a few other very similar metadata sets for corpus nodes, published corpora and lexica. They are not discussed in this paper.

<sup>2</sup> Descriptions are a field which the annotator can use to enter prose text intended for quick inspection by the user.

<sup>3</sup> Keys are those fields which guarantee flexibility. Each project or even user can define extensions in form of key-value pairs.

Name	str
<u>Contact</u>	sub
<u>Description</u> +	sub
Content	
CommunicationContext	
Interactivity	ccv
PlanningType	ccv
Involvement	ccv
Genre	
Interactional	ovl
Discursive	ovl
Performance	ovl
Task	ovl
Modalities	ovl
Languages	
<u>Description</u>	sub
<u>Language</u> +	sub
<u>Description</u> +	sub
<u>Keys</u>	sub
Participants	
<u>Description</u> +	sub
Participant+	
Type	ov
Name+	str
FullName	str
Code	str
Role	ov
<u>Language</u> +	sub
EthnicGroup	str
Age	c
Sex	ccv
Education	str
Anonymous	ccv
<u>Description</u> +	sub
<u>Keys</u>	sub
Resources	
MediaFile+	
ResourceLink	c
Size	c
Type	ccv
Format	ov
Quality	c
RecordingCondition	str
Position	c
<u>Access</u>	sub
<u>Description</u>	sub
AnnotationUnit+	
ResourceLink	c
MediaID	c
Annotator	str
Date	c
Type	ov
Format	ov
ContentEncoding	str
CharacterEncoding	str
<u>Access</u>	sub
<u>Language</u>	sub
Anonymous	ccv
<u>Description</u>	sub
Source+	
ID	str

Format	ov
Quality	ccv
Position	c
<u>Access</u>	sub
<u>Description</u>	sub

## References

It is important to mention here how multimedia and multimodality can be described in IMDI. The IMDI set allows the user to describe the *Content* of a session which refers to a unit of analysis in the corpus. Each session is associated with the media and annotation resources belonging together. The IMDI set has elements to describe the *Communication Context*, the *Genre*, the *Task*, the *Modalities*, the *Languages involved*, and to add other useful project specific elements.

In most instances the associated vocabularies clarify what the definition of the element is although IMDI has already provided careful definitions. The element *Task* stands for typical experimental tasks occurring in language engineering and field-linguistics such as *info-kiosk situation*, *route description*, *wizard-of-oz experiment*, *frog-story*. The element *Modalities* has, of course, a vocabulary which includes, amongst others, *speech*, *gesture*, *sign*, *facial expression*.

As can be seen, the IMDI set has elements not only to describe content, but also to describe the *Media Files* (type of data, format of file, quality of material, conditions of recording, etc), the available *Annotations* (type of annotation, format of file, etc), and the *Original Media* (cassette, MD, etc) if available. To give the user immediate feedback on accessibility, IMDI contains elements to describe the access rights and whom to contact to obtain the resources.

As already indicated, Controlled Vocabularies (CVs) associated with elements are an important component of the IMDI metadata set and its tools, since they will guarantee that elements are used coherently by researchers and that search operations will provide the correct resources.

To achieve interoperability with Dublin Core (a more general set of 15 partially vaguely defined elements used to describe web resources used by the general public) a mapping document was created. Based on DC, another set (OLAC [9]) was created to achieve interoperability in the language resource domain. IMDI repositories will be open to OAI [7] type of metadata harvesting to implement the interoperability with DC and OLAC.

The IMDI set is defined in all respects through an XML Schema which is available at the IMDI web-site. All tools generate and operate on these XML files.

## 4. IMDI Tools

The tools that support the IMDI metadata set and infrastructure are:

- ? The IMDI BCEditor that is used to create IMDI metadata descriptions.
- ? The IMDI BCBrowser. A viewer for the IMDI metadata descriptions that allows navigating the universe of connected IMDI metadata descriptions.
- ? The IMDI Search tool that allows the user to specify a query for specific resources in the IMDI universe.
- ? A number of scripts allowing to work efficiently

All tools were programmed in Java and Perl for platform independence and are downloadable from the web-site: <http://www.mpi.nl/tools>.

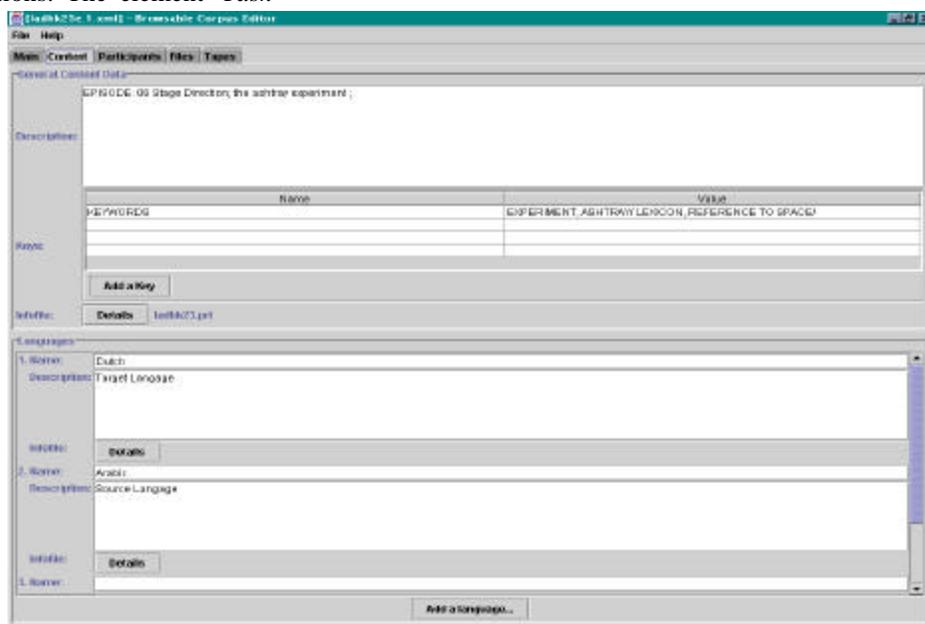


Figure 1 shows a screenshot from the IMDI Editor

The editor presents all the IMDI metadata elements in a structured GUI to the user. It supports the use of Controlled Vocabularies and user definable keyword/value pairs that the IMDI set allows for user or project specific extensions. Also it enforces constraints on the values for some metadata elements where applicable and practical. To aid working efficiency the editor allows the re-usage of a number of element blocks which will recur in many metadata descriptions such as biographical data of the informants and collectors. The editor is programmed to synchronize with repositories providing controlled vocabularies on user command if the computer the editor is running on is connected to the web. This mechanism ensures that the user can download and use the most recent definitions, e.g. of the names of countries. Internationally agreed notation conventions allow differences between different vocabularies. For example, the ISO language lists contain only a few hundred language names and the Ethnologue list [10] contains more than 4000 names. In fact users can add their own

lists but searching would become a problem if there is no mapping definition.

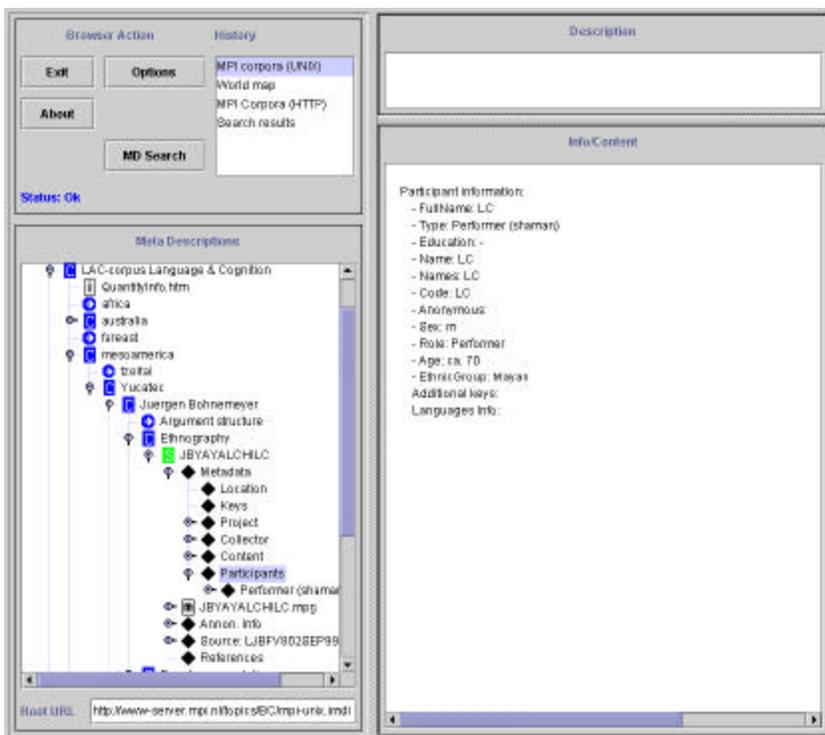


Figure 2 shows a screenshot of the IMDI browser.

The IMDI BCBrowser is the central tool for exploiting the IMDI metadata infrastructure. It allows navigation in the domain of linked IMDI metadata descriptions by clicking on corpus links. The browser keeps track of its position in the browsable corpus structure and displays the metadata and human readable descriptions associated with the sub-corpus in focus. It allows the user to set bookmarks so that easy navigation is facilitated.

The browser is also capable of displaying HTML formatted or PDF files that are often provided as extra documentation for corpora. It is possible to link in such HTML pages or PDF files in the corpus tree. From the HTML pages there may be links back to metadata descriptions making it possible to mix classical HTML browsing with browsing the IMDI corpus universe.

An interesting application of this is a world map that was created as a portal of the MPI corpora. This world map is viewable as an HTML file but has, at the appropriate places, links to metadata descriptions for corpora that correspond to those locations. We are presently engaged in trying to incorporate a professional geographic information system since the HTML world map is not completely satisfactory. The worldmap is just one other alternative view on a corpus since it is organized according to geographical principles.

One of the very important functions of the browser is that it offers the user a set of appropriate tools for further analysing resources once they have been located and it allows for operation in a distributed scenario where all resources are indicated by URLs. Each user or group of users can create a configuration file containing information on how to immediately start a tool and pass over the necessary parameters to start the tool with the discovered resource(s). The browser offers a selection from which the user can choose.

The search tool is the most recent IMDI development. It allows the user to specify a query for sessions whose metadata complies with the specified constraints. The UI offers the user an easy way to specify a query compliant with the IMDI element set, the elements value constraints and CVs used.

Results are presented in the form of URLs for the session metadata description files that comply with the query. The user may make these sessions visible in the IMDI-BCBrowser for further inspection or a special corpus label can be created containing all these

sessions that can be saved for future reference and processing. The search tool can, of course, be started from the IMDI-BCBrowser. The search tool has to be extended to support the distributed architecture underlying the IMDI concept and it has to be checked as to how it can support harvesting of other metadata repositories by, for instance, using the OAI protocol. Currently, two teams are working on an improved search tool working in fully distributed scenarios.

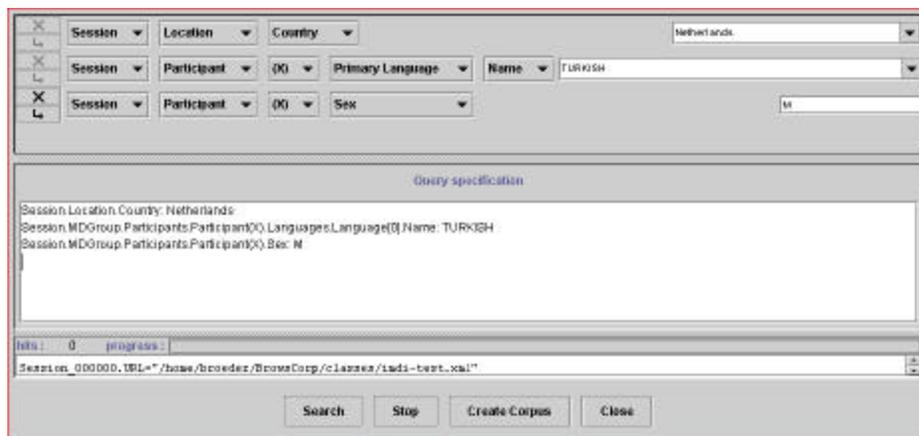


Figure 3 shows a screenshot from the search component.

The IMDI team also created a number of scripts which allow users to efficiently work with IMDI type of metadata descriptions. One such tool is provided to add or change element values in a whole range of MD descriptions by one command. Another allows the user to create metadata descriptions from spreadsheet documents, although this has proved problematic. Spreadsheet entries are not guided by constraints or controlled vocabularies

therefore conformity has to be checked very carefully. There are a few other minor scripts which will hopefully become obsolete when the editor or browser have been extended.

## 5. IMDI Corpora

At present we have available as IMDI tagged corpora:

- ? the MPI corpora of the “Acquisition” and “Language and Cognition” group which contains more than 2 TB of media data and more than 7000 multimedia sessions;
- ? a large second learner language acquisition corpus also containing audio recordings;
- ? the data of the DOBES project about endangered languages where also audio and video recordings form the basis;
- ? the data of the CGN (Spoken Dutch Corpus) project.

Furthermore we have been experimenting with converting parts of existing corpora to see if the IMDI set is applicable. These tests range from the well-known “Childes” corpora [11] to language engineering corpora as “TIMIT” [12] and “SmartKom”. An interesting project was also the construction of a distributed corpus with examples of (parts of) corpora of six different European institutes. This was demonstrated as a first distributed IMDI scenario during the official opening ceremony of the “European Year of the Language” in Lund in 2001.

## 6. Future Developments

As a preliminary solution and part of the IMDI showcase, the MPI serves as a focal point maintaining the IMDI web portal as a starting point for the IMDI universe and maintaining the IMDI metadata Schema and CV definitions. However, the MPI does not have ambitions to perform this task in the long run. Such hosting activities are better performed by organisations such as BAS [13], ELRA and LDC. The maintenance of the IMDI set and the related tools by the MPI has been secured for many years by using them in different long-term projects. Besides these organisational problems, there is also a need for further tool development, such as a tool offering users a graphical interface for creating alternative “personal” corpus trees. Maintenance tools are required that allow users to copy parts of corpus trees to other portable media such as CDROM and DVD. In this way they can work under field conditions or make personal archive copies.

A major revision of the IMDI metadata set is expected to occur in 2002, therefore comments on how to improve it are welcome. According to the most recent discussions, it can be concluded that the MD set in general is very mature and stable with the exception of a very few elements such as “Anonymous”. But the elements and vocabularies which were defined to describe the content of the

resources have to be modified after a year of experience. Here, the elements define the dimensions of descriptions and the vocabularies the values along these dimensions. Although the current definitions are based on linguistic experience, it is obvious that not all contents can be described equally well with them.

Currently, the IMDI definitions are specified with the help of an XML Schema, i.e. the relations between concepts are implicitly defined in the structured IMDI set. To open up the way to the Semantic Web these implicit relations will be explicitly defined with the help of RDF [14]. All RDF Schemas will be put into open RDF repositories so that they can be re-used. It has to be checked whether it will be possible to make use of already existing descriptions within the IMDI set.

## 7. References

- [1] Broeder, D.G., Brugman, H., Russel, A., and Wittenburg, P., (2000), A Browsable Corpus: accessing linguistic resources the easy way. In *Proceedings LREC 2000 Workshop*, Athens.
- [2] ISLE/IMDI: <http://www.mpi.nl/ISLE> & [http://www.mpi.nl/world/ISLE/documents/papers/white\\_paper\\_11.pdf](http://www.mpi.nl/world/ISLE/documents/papers/white_paper_11.pdf) & [http://www.mpi.nl/ISLE/documents/draft/ISLE MetaData 2.5.pdf](http://www.mpi.nl/ISLE/documents/draft/ISLE_MetaData_2.5.pdf)
- [3] DOBES: <http://www.mpi.nl/DOBES>
- [4] CGN: <http://www.now.nl/gw/introductie>
- [5] SmartKom: <http://smartkom.dfki.de>
- [6] DC: <http://www.dublincore.org/>
- [7] OAI: <http://www.openarchives.org/>
- [8] MPEG7: <http://mpeg.telecomitalia.com/standards/mpeg-7/mpeg-7.htm>
- [9] OLAC: <http://www.language-archives.org/OLAC/>
- [10] Ethnologue Language List: <http://www.ethnologue.com>
- [11] ChilDes: <http://childes.psy.cmu.edu>
- [12] TIMIT: <http://www ldc.upenn.edu/Catalog/LD93S1.html>
- [13] BAS: <http://www.phonetik.uni-muenchen.de/Bas/BasHomeen.html>
- [14] RDF: <http://www.w3.org/RDF> & <http://www.w3.org/sw>

# The FORM Gesture Annotation System

Craig Martell<sup>\*†</sup>, Chris Osborn<sup>\*†</sup>, Jesse Friedman<sup>\*</sup>, Paul Howard<sup>\*</sup>

<sup>\*</sup>Linguistic Data Consortium  
University of Pennsylvania  
3615 Market Street, Suite 200  
Philadelphia, PA 19104  
{cmartell, cosborn, jessef2, pch}@unagi.cis.upenn.edu

<sup>†</sup>Department of Computer and Information Sciences  
University of Pennsylvania  
200 S. 33rd St  
Philadelphia, PA 19104

## Abstract

The Friedman Osborn Martell (FORM) system for annotating gestures has been created for the purpose of producing a corpus of speech and its corresponding gestures. The corpus is open source and will be available to all researchers who wish to use it in their work. FORM attempts to capture the kinematics of gesture using *quasi*-geometric descriptions of the locations/shapes and movements of the arms and hands. Currently, we have a pilot corpus of 22 minutes of gesture-annotated video of Brian MacWhinney teaching a Research Methods course at Carnegie Mellon University. There are plans to extend the corpus to include not only speech transcriptions and syntactic information, but also body-movement and intonational information as well. We are currently gathering other data of various types and in various settings to add to the corpus. All of these data will be published under the TalkBank project (<http://www.talkbank.org>).

## 1. Introduction

In “An Agenda for Gesture Studies” (Kendon, 1996), Adam Kendon outlines a long-term research agenda for a better understanding of gesture and its relationship to the communicative process. A major aspect of that agenda is the development of what Kendon calls the “Kinetics of Gesture”:

Such a programme of work could be linked to, and would contribute importantly, to research on what might be called the ‘kinetics’ of gesture (in parallel to ‘phonetics’). We really have little explicit knowledge about how gestures are organized as physical actions. . . . An important part of the ‘kinetics’ research should include a study of just how gesture phrases are organized in relation to speech phrases.

The FORM project began, in large part, as a response to this challenge<sup>1</sup>. FORM is an annotation scheme designed both to describe the kinematic information in a gesture, as well as to be extensible in order to add speech and other conversational information.

Our plan, then, is to build an extensible corpus of annotated videos in order to allow for general research on the relationship among the many different aspects of conversational interaction. Additionally, further tools and algorithms to add these annotations and evaluate inter-annotator

agreement will be developed. The end result of this work will be a corpus of annotated conversational interaction, which can be:

- extended to include new types of information concerning the same conversations; as new tag-sets and coding schemes are developed—discourse-structure or facial-expression, for example—new annotations could easily be added;
- used to test scientific hypotheses concerning the relationship of the paralinguistic aspects of communication to speech and to meaning;
- used to develop statistical algorithms to automatically analyze and generate these paralinguistic aspects of communication (e.g., for Human-Computer Interface research).

## 2. FORM

### 2.1. The Annotation Scheme

FORM is designed as a series of tracks representing different aspects of the gestural space. Generally, each independently moved part of the body has two tracks, one track for Location/Shape/Orientation, and one for Movement. When a part of the body is held without movement, a Location object describes its position and spans the amount of time the position is held. When a part of the body is in motion, Location objects with no time period are placed at the beginning and end of the movement to show where the gesture began and ended. Location objects spanning no period of time are also used to indicate the Location information at critical points in certain complex gestures. See Figure 1 for a snapshot of FORM implemented using the Anvil tool (Kipp, 2001).

---

<sup>1</sup>The authors wish to sincerely thank Adam Kendon for his input on the FORM project. He has provided not only suggestions as to the direction of the project, but also his unpublished work on a kinematically-based gesture annotation scheme was the FORM project’s starting point (Kendon, 2000).

An object in a movement track spans the time period in which the body part in question is in motion. It is often the case that one part of the body will remain static while others move. For example, a single hand shape may be held throughout a gesture in which the upper arm moves. FORM's multi-track system allows such disparate parts of single gestures to be recorded separately and efficiently and to be viewed easily once recorded. Once all tracks are filled with the appropriate information, it is easy to see the structure of a gesture broken down into its anatomical components.

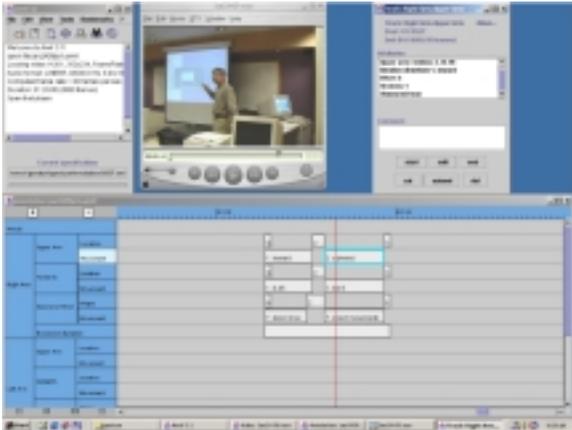


Figure 1: FORM annotation of Jan24-09.mov, using Anvil as the annotation tool

At the highest level of FORM are groups. Groups can contain subgroups. Within each group or subgroup are tracks. Each track contains a list of attributes concerning a particular part of the arm or body. At the lowest level (under each attribute), all possible values are listed. The structure, then, is as follows:

**Group**

*Subgroup*

*Track*

ATTRIBUTE

Value

The following descriptions will follow this structure. The groups described are Right/Left Arm, Gesture Obscured, Excursion Duration, and Two-Handed Gesture. Not described are Head and Torso Movement/Location. These will be implemented in a later version of FORM.

**Right/Left Arm**

*Upper Arm* (from the shoulder to the elbow).

*Location*

UPPER ARM LIFT (from side of the body)

- no lift
- 0-45
- approx. 45
- 45-90
- approx. 90

- 90-135
- approx. 135
- 135-180
- approx. 180

**RELATIVE ELBOW POSITION:** The upper arm lift attribute defines a circle on which the elbow can lie. The relative elbow position attribute indicates where on that circle the elbow lies. Combined, these two attributes provide full information about the location of the elbow and reveal total location information (in relation to the shoulder) of the upper arm.

- extremely inward
- inward
- front
- front-outward
- outward (in frontal plane)
- behind
- far behind

The next three attributes individually indicate the direction in which the biceps muscle is pointed in one spatial dimension. Taken together, these three attributes reveal the orientation of the upper arm.

**BICEPS: INWARD/OUTWARD**

- none
- inward
- outward

**BICEPS: UPWARD/DOWNWARD**

- none
- upward
- downward

**BICEPS: FORWARD/BACKWARD**

- none
- forward
- backward

**OBSCURED:** This is a binary attribute which allows the annotator to indicate if the attributes and values chosen were "guesses" necessitated by visual occlusion. This attribute is present in each of FORM's tracks.

*Movement*

The next three attributes individually indicate the direction of elbow movement in one spatial direction. When diagonal movement occurs, a non-*none* (i.e. not *none*) value for more than one of the attributes is chosen. Each attribute has combination values so repeated or back-and-forth motions can be annotated as such.

**LINEAR MOVEMENT (HORIZONTAL PLANE):** Indicates the direction(s) of inward or outward elbow movement.

- none
- inward
- outward
- inward-outward
- outward-inward

LINEAR MOVEMENT (MEDIAN PLANE): Indicates the direction(s) of upward or downward elbow movement.

none  
up  
down  
up-down  
down-up

LINEAR MOVEMENT (FRONTAL PLANE): Indicates the direction(s) of elbow movement towards or away from the body.

none  
towards  
away  
towards-away  
away-towards

UPPER ARM ROTATION: The degree of change of bicep direction. Ranges are exclusive. Direction of change is not included, as it can be inferred from the information in the Location track.

0-45  
approx. 45  
45-90  
approx. 90  
90-135  
approx. 135  
135-180  
approx. 180  
greater than 180

ARC-LIKE MOVEMENT: This boolean attribute indicates whether or not the elbow movement was arc-like. When checked, Location objects will co-occur to note the location of the elbow at the beginning, apex, and end of the movement.

CIRCULAR MOVEMENT: A non-none value indicates that elbow movement is circular in shape and notes the plane in which the movement is performed as well as its direction (clockwise or counter-clockwise). As was the case for arc-like movements, the Location track will be simultaneously utilized, in this case noting the location of the elbow at the start and halfway mark of the circle. This convention allows the size of the circle to be inferred.

parallel to horizontal plane (c=clockwise)  
parallel to horizontal plane (cc=counter-clockwise)  
parallel to median plane (c)  
parallel to median plane (cc)  
parallel to frontal plane (c)  
parallel to frontal plane (cc)

EFFORT: Indicates the effort of the movement on a 1 to 5 scale.

1  
2  
3  
4

5

STROKES: Indicates the number of strokes of a movement.

1 ... 20  
More than 20  
Indeterminate

OBSCURED

**Forearm:** the part of the arm extending from the from elbow to wrist)

*Location*

ELBOW FLEXION: The angle made by the bend in the elbow.

0-45  
approx. 45  
45-90  
approx. 90  
90-135  
approx. 135  
135-180  
straight

FOREARM ORIENTATION: Describes the orientation of the forearm if the upper arm were to be by the side and the elbow flexed at 90 degrees.

supine  
supine/neutral  
neutral  
neutral/prone  
prone  
prone/inverse  
inverse

OBSCURED

*Movement*

ELBOW FLEXION CHANGE: The amount of change in elbow flexion measured in degrees. Direction of flexion change is not indicated, as it can be inferred from information in the Location track.

0-45  
approx. 45  
45-90  
approx. 90  
90-135  
approx. 135  
135-180  
approx. 180

FOREARM ROTATION: Direction of change of forearm orientation. Amount of change is not indicated, as it can be inferred from information in the Location track.

none  
inward  
outward  
inward-outward  
outward-inward

EFFORT

STROKES

OBSURED

**Hand and Wrist**

*Shape:* Information about the static shape of the hand and orientation of the wrist.

The next two attributes give values to describe the shape of the hand. The values are represented in a catalog of hand-shapes (Figure 2), which is organized as a two-dimensional matrix. This method is employed because the complexity of the hand would make purely physicalistic descriptions too unwieldy.

**HAND-SHAPE GROUP:** Indicates the group (organized by number of extended fingers with 0 representing fist and 6 referring to miscellaneous shapes) in the hand shape catalog.

- 0
- 1
- 2
- 3
- 4
- 5
- 6

**HAND-SHAPE LETTER:** Indicates the appropriate hand-shape within the selected group.

- A
- B
- C
- D
- E
- F
- G
- H
- I
- J
- K
- L
- M

**TENSION:** Describes the amount of tension apparent in the performer's hand. An average amount of tension corresponds to the "slightly tense" variable.

- relaxed
- slightly tense
- very tense

**WRIST BEND: UP AND DOWN:** How far the wrist is bent towards the upper side or under side of the forearm.

- up
- up-neutral
- neutral
- down-neutral
- down



Figure 2: Catalog of Handshapes. Based on the HamNoSys catalog (<http://www.sign-lang.uni-hamburg.de/Projects/HamNoSys.html>)

**WRIST BEND: SIDE TO SIDE:** How the wrist is bent towards the thumb or little finger.

- towards thumb
- neutral
- towards little finger
- extremely towards little finger

**PART OF BODY TOUCHED:**

- none
- top of head
- eye (same)
- eye (opposite)
- ear (same)
- ear (opposite)
- temple (same)
- temple (opposite)
- nose
- cheek (same)
- cheek (opposite)
- chin
- neck (same side)
- neck (center)
- neck (opposite side)
- chest
- groin

OBSURED

*Movement*

**HAND MOVEMENT:** Describes type of hand movement (if any). The A joint refers to the knuckle furthest from the fingertip and the B joint refers to the first joint above the A joint. Information about the C joint (the joint closest to the fingertip) is not recorded because C joint movement is usually dependent upon movement of the B joint. The numbering scheme of the first three variables is explained in the Finger Coordination attribute.

none  
1) A joint movement  
2) B joint movement  
3) A and B joint movement  
wrist circular  
thumb rubbing index finger  
thumb rubbing multiple fingers  
direct movement between two shapes

**WRIST UP-DOWN MOVEMENT:** Describes the up-down movement (to the underside or upper side of the arm) of the wrist.

up  
down  
up-down  
down-up

#### WRIST SIDE-TO-SIDE MOVEMENT

towards little finger  
towards thumb  
towards little finger-towards thumb  
towards thumb-towards little finger

**FINGER COORDINATION:** Describes the motion of the fingers in relationship to each other. A non-none value is only applicable if one of the choices labeled 1, 2, or 3 was selected from the Hand movement attribute.

parallel movement without thumb  
random movement, without thumb  
parallel movement, with thumb  
random movement, with thumb  
movement in sequence

EFFORT

STROKES

OBSCURED

**Excursion Duration:** Marks the length of the excursion of the arm from a resting position to another resting position. Since there is ambiguity about what constitutes a single gesture, this convention for grouping was adopted.

**Gesture Obscured:** Similar to above except this attribute refers to the entire gesture duration, rather than just one track.

### Two-handed Gestures

#### RIGHT-HAND CONTACT

none  
thumb  
index finger  
middle finger

ring finger  
little finger  
palm  
back of hand  
more than one digit  
holding

**LEFT-HAND CONTACT:** The list of values is identical to that of the Right-hand Contact attribute.

The following seven attributes are all boolean-valued.

MOVING IN PARALLEL

MOVING APART

MOVING TOWARDS

MOVING AROUND ONE ANOTHER

MOVING IN ALTERNATION

CROSSED

OBSCURED

## 2.2. Ambiguities/Imprecisions in FORM

There are two known ambiguities/imprecisions in the current version of the FORM system.

The first concerns the *Upper Arm:Location* attributes that specify biceps direction. While anatomically it seems more accurate to describe the upper arm rotation by degrees of rotation rather than by using the direction of the biceps in free space, a problem arises when defining the neutral position of the arm rotation. For example, we could define *normal* as the position when the arm is held at the side with palm facing towards the body and the elbow flexed to make a 90-degree angle with the upper arm. If one then lifts the upper arm to the side so it is at 90 degrees with the body and still in the frontal plane, the upper arm has not rotated at all. Let's call this position 1. If, however, one returns to the starting position, raises the upper arm forward so it is at 90 degrees with the body but parallel to the median plane, and then moves the upper arm 90 degrees to the side so that it is in the frontal plane again, it can be seen that this position is also reached without rotating the upper arm. Let's call this position 2. It is clear that position 1 is not the same as position 2, but both were reached by keeping the upper-arm in the *normal* position.

To solve this issue we could define a normal that is rotated 45 degrees when the Upper arm lift is at what we've deemed "approx. 90" and Relative elbow position is "front-outward." This convention, however, is hard to conceptualize by annotators and thus we opted to use the direction of biceps in free space since it is more intuitive. The downside to this approach is that it allows for a large range of positions for each combination of values. Many positions could be called "forward-inward-upward," for example.

The second area of concern is in the *Upper arm:Movement* track. This track describes the movement of the upper arm independent of the forearm, elbow flexion, and hand-and-wrist. This movement can be described either as a combination of linear movement in different planes or as arc-like movement (using Location points to denote points along the curve). Since the upper arm is only

able to move on a partial sphere with the shoulder as the center, it does not make sense anatomically to describe its movement as linear. However, since most movements are small enough not to appear as distinct arcs, linear values sufficiently approximate the movement.

### 3. The Current FORM Corpus<sup>2</sup>

#### 3.1. Pilot Corpus

We currently have a pilot corpus of about 22 minutes of Brian MacWhinney teaching a Research Methods course at Carnegie Mellon University. These data were chosen since they were freely available via the the TalkBank project (<http://www.talkbank.org>). They have been very useful for the pilot phase of the project as people often gesture in a clear and exaggerated fashion while teaching. See (Martell, 2002) for a further description of the data format (Annotation Graphs (Bird and Liberman, 1999)), as well as examples of the video and of tool currently being used (Anvil (Kipp, 2001)).

#### 3.2. Annotation Complexity

An experienced annotator can create approximately 3 seconds of annotation per hour. He/she can annotate at most for 6 hours per day, generating 18 seconds/day. Accordingly, it will take an experienced annotator 5 work days to annotate a 90-second video of conversational interaction.

Generating only 90 seconds of annotation per work week makes such an annotation project seem a daunting task. However, the amount of information contained in conversational gesturing is substantial—on the order of 3500 distinct ATTRIBUTE:Value pairs per minute. This underscores the potential value of such a corpus, viz. there is seemingly much more information in 90 seconds of communicative interaction than we are currently capturing by only transcribing speech.

#### 3.3. Preliminary Inter-Annotator Agreement Results

Preliminary results from FORM show that with sufficient training, agreement among the annotators can be very high. Table 2 shows preliminary interannotator agreement results from a FORM pilot study.<sup>3</sup> The results are for two trained annotators for approximately 1.5 minutes of Jan24-09.mov, the video from Figure 1. For this clip, the two annotators agreed that there were at least these 4 gesture excursions. One annotator found 2 additional excursions. Precision refers to the decimal precision of the time stamps given for the beginning and end of gestural components. The *SAME* value means that all time-stamps were given the same value. This was done in order to judge agreement with having to judge the exact beginning and end of an excursion factored out. *Exact* vs. *No-Value* percentage refers to whether both the attributes and values matched exactly or whether just the attributes matched exactly. This distinction is included because a gesture excursion is defined as

<sup>2</sup>Most of this section is taken from (Martell, 2002)

<sup>3</sup>Essentially, all the arcs for each annotator are thrown into a bag. Then all the bags are combined and the intersection is extracted. This intersection constitutes the overlap in annotation, i.e., where the annotators agreed. The percentage of the intersection to the whole is then calculated to get the scores presented.

all movement between two rest positions of the arms and hands. For an excursion, the annotators have to judge both which parts of the arms and hands are salient to the movement (e.g., upper-arm lift and rotation, as well as forearm change in orientation and hand/wrist position) as well as what values to assign (e.g., the upper-arm lifted 15-degrees and rotated 45-degrees). So, the *No-Value%* column captures the degree to which the annotators agree just on the structure of the movement, while *Exact%* measures agreement on both structure and values.

The degree to which inter-annotator agreement varies among these gestures might suggest difficulty in reaching consensus. However, the results on *intra*-annotator agreement studies demonstrate that a single annotator shows similar variance when doing the same video-clip at different times. Table 3 gives the intra-annotator results for one annotator annotating the first 2 gesture excursions of Jan24-09.mov.

Gesture Excursion	Precision	Exact%	No-Value%
1	2	3.41	4.35
	1	10.07	12.8
	0	29.44	41.38
	SAME	56.92	86.15
2	2	37.5	52.5
	1	60	77.5
	0	75.56	94.81
	SAME	73.24	95.77
3	2	0	0
	1	19.25	27.81
	0	62.5	86.11
	SAME	67.61	95.77
4	2	10.2	12.06
	1	25.68	31.72
	0	57.77	77.67
	SAME	68.29	95.12

Table 1: Inter-Annotator Agreement on Jan24-09.mov

Gesture Excursion	Precision	Exact%	No-Value%
1	0	5.98	7.56
	1	20.52	25.21
	0	58.03	74.64
	SAME	85.52	96.55
2	2	0	0
	1	25.81	28.39
	0	89.06	95.31
	SAME	90.91	93.94

Table 2: Intra-Annotator Agreement on Jan24-09.mov

For both sets of data, the pattern is the same:

- the less precise the time-stamps, the better the results;
- *No-Value%* is significantly higher than *Exact%*.

It is also important to note that Gesture Excursion 1 is far more complex than Gesture Excursion 2. And, in both sim-

ple and complex gestures, inter-annotator agreement is approaching intra-annotator agreement. Notice, also, that for Excursion 2, inner-annotator agreement is actually better than intra-annotator agreement for the first two rows. This is a result of the difficulty for even the same person over time to precisely pin down the beginning and end of a gesture excursion. Although the preliminary results are very encouraging, all of the above suggests that further research concerning training and how to judge similarity of gestures is necessary. Visual information may need very different similarity criteria.

#### 4. Future Directions and Open Questions

As mentioned in the introduction, the goal of the FORM project is to create a corpus to be used for both scientific and technological research concerning gesture and its relationship to the rest of the communicative process. However, in order to build a corpus suitable for these goals, a number of issues have to be addressed.

**Augmentation of the FORM corpus** with other aspects of communication is necessary. Over the next 3 years, as we continue to build the FORM corpus, we will also be augmenting it with:

- Speech transcriptions and syntactic information;
- Body movement, in the form of head and torso information; and
- Intonation and pitch contour information.

Much research is needed to discover the best annotation schemes for each of these aspects, as well as to discover which algorithms are best for uncovering the correlations among them.

**Better methods of annotation** need to be developed. Although we believe it will prove necessary to continue to annotate at a fine-grained level of detail, it is currently too expensive to make using FORM practical. We intend to use the hand-annotated corpus, as it grows, to explore automatic or semi-automatic methods of annotation.

**Visualization and Animation Tools** which will “play back” an annotation are needed to allow an annotator to better judge the correctness of his/her annotation. Additionally, these visualization tools may be able to help ease the annotation process. If we are able to develop a close enough mapping between the animated character and the annotation scheme, we may be able to use a movable animated character as a means to input the data. Research is needed to see if this will indeed speed up the process.

**New Metrics for Inner-Annotator Agreement** need to be explored. As mentioned in Section 3.3, above, our current numbers are based on the bag-of-arcs technique. However, as the scores there indicate, often annotators agree to a large degree on structure, but differ only on exact beginning or ending timestamp, or on the value of an attribute. Unfortunately, small differences in timestamp and value are judged incorrect to the same degree as large differences. Visual feedback, as just described, will allow us to discover whether small differences in coding actually have little difference visually. If this proves to be the case, then we will

need to experiment with more geometrically-based measures of similarity, e.g., distance in n-dimensional space.

**Statistical Experiments using FORM** are already underway. If FORM is to be successful, it must be shown that our fine-grained analysis sufficiently captures the phenomena in question. To do this we are conducting two sets of experiments.

- We have annotated some of the corpus with Preparation-Stroke-Retraction information. Using standard training-set/test-set methods, we are building Preparation-Stroke-Retraction recognizer system. If the results of these experiments are sufficiently high, we will have demonstrated that FORM captures at least as much information as a more coarse-grained annotation scheme.
- However, only showing that FORM is as good as coarse-grained annotation scheme is not sufficient justification for using FORM. Accordingly, we are also working on a Statistical Gesture Generation System (SGGS). Given some input set of sentences, the SGGS, if successful, will be able to output those sentences augmented with a FORM description of valid accompanying gestures. This, then, could be used with the above described animation tool to automatically generate animated gesture excursions.

#### 5. Conclusion

The FORM project has developed a geometrically-based gesture annotation scheme and a 22-minute pilot corpus of gesture-annotated video. Over the next few years, the corpus will be augmented and new tools and algorithms will be developed. The envisioned goal of the project is a large-scale corpus of multi-modal annotations suitable for both scientific and technological research concerning the relationships among different aspects of communicative interaction.

#### 6. References

- Steven Bird and Mark Liberman. 1999. A formal framework for linguistic annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Sciences, University of Pennsylvania, Philadelphia, Pennsylvania. <http://citeseer.nj.nec.com/article/bird99formal.html>.
- Adam Kendon. 1996. An agenda for gesture studies. *Semiotic Review of Books*, 7(3):8–12.
- Adam Kendon. 2000. Suggestions for a descriptive notation for manual gestures. Unpublished.
- Michael Kipp. 2001. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of Eurospeech 2001*, pages 1367–1370.
- Craig Martell. 2002. Form: An extensible, kinematically-based gesture annotation scheme. In *Proceeding of the International Conference on Language Resources and Evaluation*. European Language Resources Association. <http://www ldc.upenn.edu/Projects/FORM>.

## Sample Annotated Video Using Anvil and FORM

Craig Martell  
Linguistic Data Consortium  
University of Pennsylvania  
3615 Market Street Suite 200  
Philadelphia, PA 19104-2608, USA  
cmartell@unagi.cis.upenn.edu  
<http://www.cis.upenn.edu/~cmartell>

### Video Sample:

<http://www ldc.upenn.edu/annotation/gesture/Jan24-05.mov>

### Coding Scheme:

This is a video of Brian MacWhinney lecturing, and we coded his gestures using an annotation scheme FORM. This was described in a separate submission for the workshop -- or you can see the same description at

<http://www ldc.upenn.edu/Projects/FORM>

FORM is a kinematic annotation scheme which describes gestures by their physical movements. The idea is to later add speech, and other paralinguistic information, to the data set to better understand the relationship of gesture to speech.

### Annotation File:

<http://www ldc.upenn.edu/annotation/gesture/Jan24-05pch.anvil>

This is an XML file for use with Anvil, described below

### Annotation Tool:

We currently use Michael Kipp's tool Anvil (<http://www.dfki.de/~kipp/anvil/>). To use this with FORM, you will need the specification file, which contains the entire coding scheme. This can be found at:

<http://www ldc.upenn.edu/annotation/gesture/gestureAnnotation0807.xml>

# Cross-Linguistic Studies of Multimodal Communication

P. Wittenburg, S. Kita, H. Brugman

Max-Planck-Institute for Psycholinguistics  
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands  
peter.wittenburg@mpi.nl

## Abstract

Gestures are culture specific forms of arm movements which are used in communication to transfer information to the listener, to guide the planning of the speech production process and to disambiguate the incoming speech. To understand the underlying mechanisms gestures have to be analyzed in cross-linguistic processes. Large projects are necessary covering speakers from various cultural background and many recordings. Such projects can only be successfully carried out, when suitable gesture encoding schemes, generic annotation schemes, powerful tools supporting the schemes and efficient methods for easy resource discovery and management are available. At the Max-Planck-Institute all aspects were tackled.

## 1. Introduction

The MPI for Psycholinguistics has a long history of research on the synchronization between different modalities in human communication. In the 1980s eyetracking signals and signals about pointing gestures produced important information about the mental processes responsible for speech production [1, 2]. Such signals were typically recorded in relation to spoken utterances. The equipment used was designed to make automatic fine grained temporal analysis possible. For gesture registration IR-light based methods were used. More recently, ultrasonic equipment was used for this purpose identifying the location of maximally 8 sources. This tradition is still continued in the baby labs where eye tracking is recorded to study, for example, the focus of childrens' attention during linguistic tasks. In recent years brain imaging methods (EEG, MEG, PET, MRI) have often been added to get online information about brain activities during speech production and perception task.

In the last few years, research using multimodality shifted towards observational methods in communicative situations of various sorts. Child-caretaker interaction is studied with the help of extensive video recordings to better understand how childrens' language learning is influenced by input and environmental factors. The use of various types of gestures (pointing, iconic and emblematic) is studied in different situations. The following studies should be mentioned in particular: (1) ethnography of pointing gestures; (2) gestural facilitation of speaking or understanding; (3) gestural expression of motion events; (4) speech dysfluencies and gestures; (5) influence of gestures on recipients' gaze movement; (6) hemispheric specialization of types of gestures [3, 4, 5, 6, 7, 8]. In addition, studies about sign language and their comparison to gestural patterns were carried out. The goal of these recordings is fundamental research about the relation between language and thought and the role of gesture in human communication. Since gestures are very much dependent on language and culture, most of the recordings are cross-linguistic, i.e. various countries and cultures are included.

Nowadays the study of multimodal communication based on video recordings is much easier. Information

technology allows science to work with digitized video greatly facilitating the analysis work. For the last two years, all recordings at the MPI have been digitized, yielding an online multimedia corpus consisting of more than 7000 sessions (units of linguistic analysis). Gesture studies form a substantial part of these recordings. Powerful corpus management with the help of metadata descriptions and multimodal annotation tools were developed at the institute to enable the type of research explained. Annotations are stored in well-documented formats well adapted to capturing the complexity of the annotation which are typical of multimodal studies.

## 2. Multimodality Research

Multi-modal records allow us not only to approach old research problems in new ways, but also open up entirely new avenues of research. An old issue, for example, is just how 'modular' language processing is, that is to what extent non-linguistic processes can intervene in the course of linguistic processing. This can be studied by looking at the interaction between two entirely different behaviour streams, gesture and speech. A large multi-media corpus of natural dialogue shows, for example, that when speakers self-edit speech, gesture inhibition actually occurs earlier, suggesting interaction between the speech and gesture execution systems. Similarly, in the comprehension process it can be shown that gesture content is incorporated into the immediate 'message'. Eye-tracking shows that speakers can manipulate the likelihood of this by looking at their own gestures, which are then more often fixated by listeners. More fundamentally, we can look at the role of the two cerebral hemispheres in the production of the two behaviour streams, speech and gesture. Careful studies of the gestures of split-brain patients show that gesture production is largely driven from the right hemisphere, while language of course is normally processed in the left.

In addition to contributing to such long-standing theoretical issues, annotated multimedia records also make possible entirely new lines of research. For example, we have been interested in whether the semantic character of a specific language leads to a special construal of a scene to be described. The study of gesture during online production shows that the way a language 'packages' information has a demonstrable effect on the depiction of

a scene in gestures. Turkish for example packages movement with direction in a single clause but puts manner of motion into a separate adverbial clause ('The ball descended, rolling') – while English allows manner and direction to occur in the same simple clause ('The ball rolled down'). Turkish speakers tend to produce separate gestures for direction and manner, while English speakers tend to fuse them. In a similar way, we have been able to study spatial thinking as it occurs in non-spatial domains, by examining the gestures of speakers talking about e.g. kinship relations.

Sign languages are another domain which has been opened up by multi-media technology. Sign languages are fully-expressive languages which utilize not only the hands, but also the face, gaze and even body-posture to construct complex utterances with phonology, morphology, syntax and 'prosody'. These different 'articulators' express different distinctions in overlapping time windows, where the offset can indicate e.g. the scope of a question. Even the simplest description of a signed utterance therefore requires a multi-tiered annotation of a video-record, and the development of such annotation tools make possible systematic databases for sign language research for the first time. Fascinating questions can now be pursued about effects of modality on language – for example does the spatial nature of the visual-gestural channel have profound effects on the nature of sign languages, and give sign languages an underlying commonality? Most deaf signers are exposed to the gestural systems of the surrounding spoken language, and we can also ask to what extent these gestural systems are recruited into the sign language. Preliminary results from the study of a sign language in the process of standardization (Nicaraguan sign language) suggests that there is such an interaction.

These examples should serve to indicate just what a revolution in our understanding of language and its relation to other aspects of cognition is being made possible by the new technologies. There are also fundamental advantages to archiving multi-media records for all branches of the language sciences. For example, studies of the acquisition of language are hugely enriched by having available the very scene available to the infant language user – we now know for example that unexpressed arguments (e.g. subjects and objects) in Inuit care-takers' speech are often recoverable by the child just because they are most likely in the child's field of view at the moment of utterance. Similarly, records of dying or endangered languages are greatly enhanced by having visual information correlate with the language use. In all these cases, richly annotated multi-media records make possible the extraction of systematic information about the correlation of linguistic and non-linguistic events.

### 3. Gesture Encoding Schemes

#### General

This variety of studies all based on observational methods (i.e. audio and video, sometimes also gaze) required many different gesture encoding schemes on the different

linguistic levels, efficient procedures and powerful tools. Since our researchers are involved in international projects broad agreements on the methods for encoding multimodal behavior are very important. Yet for international standards it seems to be too early, the discipline is too young, although it would facilitate integrating and comparing the data of all the scholarly work.

Most of the studies require careful encoding of the articulator movements<sup>1</sup> and their global timing pattern. Naturally, we are faced with similar problems to those for identifying the articulator movements in the case of speech production. The articulator movements form a continuum, are overlapping and have tolerances dependent on the situation. Therefore, it is not only difficult to make proper time segmentation, but also to classify them.

For gestures which are movements of the arms and its parts accompanying verbal communication acts, it is sufficient to annotate their type and meaning in addition to the articulators. The type of a gesture is a taxonomic classification of its principle purpose and role in communication. It is widely accepted to separate between pointing, iconic and emblematic gestures. Pointing gestures refer to a spatial point or a movement. They appear either as isolated gestures where the meaning is obvious to the listener or mostly in overlap with verbal utterances where the gestures are much more simple to generate and interpret than verbal descriptions. Their meaning is easy to describe by the object they refer to and their intrinsic purpose. Also iconic gestures appear spontaneously as co-speech activities while emblematic gestures stand alone. Iconic gestures have a culturally bound meaning since they are widely accepted within an area.

Gestures often correlate with emotional state, are used to facilitate the planning of speech production and to facilitate speech perception due to their disambiguation capability. Emotional state can be described, although there are no clear conventions yet.

#### Articulators in Gestures

The basis of all scientific work when studying gestures is an encoding scheme for the articulator movements. It was soon perceived that an exhaustive gesture encoding including all relevant characteristics would be ideal but impossible (except for small segments). On the other hand the recordings were perceived as so valuable that re-usage for various research questions was anticipated. To cope with this contradiction it was realised that only an iterative encoding approach would suffice where the needs of primary research projects do not hinder the addition of gesture encodings dedicated to completely different research interests. To support research, the underlying

---

<sup>1</sup> For gestures we have as articulators the arms and its parts up to the fingers. Characteristic movements of the head and the eyes in communicative situations are not treated as part of the gesture although they have similar purposes.

scheme should be exhaustive to define a grid allowing easy computational comparison. Therefore, for a number of recordings focused on in the Institute's gesture project, a thorough study was carried out to attain a general gesture encoding scheme that would allow comparative analysis to be made easily.

Based on Kendon's work a more accurate scheme was developed by v. Gijn, vd Hulst and Kita [9] to separate various phases in a gesture. A *MovementUnit* therefore can exist of several *MovementPhrases*. Basically, each of these can be seen as a sequence of a *Preparation* phase, an *ExpressivePhase* and a *Retraction* phase. An *ExpressivePhase* which covers the meaningful nucleus of a gesture is either an *IndependentHold* or a sequence of a *DependentHold*, a *Stroke*, and another *DependentHold*.

*MovementUnit* = *MovementPhrase*\*  
*MovementPhrase* = (*Preparation*) => *ExpressivePhase* => (*Retraction*)  
*ExpressivePhase* = *IndependentHold*  
*ExpressivePhase* = (*DependentHold*) => *Stroke* => (*DependentHold*)  
*Preparation* = (*LiberatingMovement*) => *LocationPreparation* >>  
*HandInternalPreparation*  
*Retraction* (if subsequent movement) = *PartialRetraction*  
= consists of, \* one or several, => discrete transition, () optional,  
>> normally blended out, occasionally discrete transition

The authors developed a set of descriptive criteria to identify the phases and their usefulness was shown in several studies which were successfully annotated by student assistants.

v. Gijn, vd Hulst and Kita also developed an encoding scheme to describe mainly the articulator movements in the *ExpressivePhase* [10]. It is this phase where annotators are confronted with all the about 60 degrees of freedom and where not only the location and shape has to be described but also for example changes in motion and direction. The following aspects are described: *PathMovementShape* (*straight, circle, round, iconic, 7-form, ?-form, x-form, +form, z-form*), *PathMovement Direction* (*[up/down], [front/back], [ipsilateral/contralateral]*), *HandOrientationChange* (*[supination/pronation], rotation, [flexion/extension], nodding, [ulnar flexion/radial flexion], lateral flexion*), *HandShape Change* (*[opening/closing], [abduction/adduction], [hinging/dehinging], [clawing/declawing], wiggling, opening wave, closing wave, rubbing, cutting*), *HandOrientation* (*[up/down], [front/back], [ipsilateral/contralateral]*), and *HandShape*. For the latter basically the HamNoSys scheme was re-used.

To support the various gesture related research activities simple encoding schemes are most often derived from this exhaustive scheme. The reference back to the unified exhaustive scheme together with the online availability of the annotated multimedia document allows easy re-usage and an enhancement of the annotations. This can either be corrections of the existing or the addition of new tiers.

When encoding gestures it is of great importance to understand the exact time relationships with the verbal utterances. This is not part of the gesture annotation scheme, but the annotation structure scheme has to provide adequate mechanisms.

#### 4. Annotation Structures

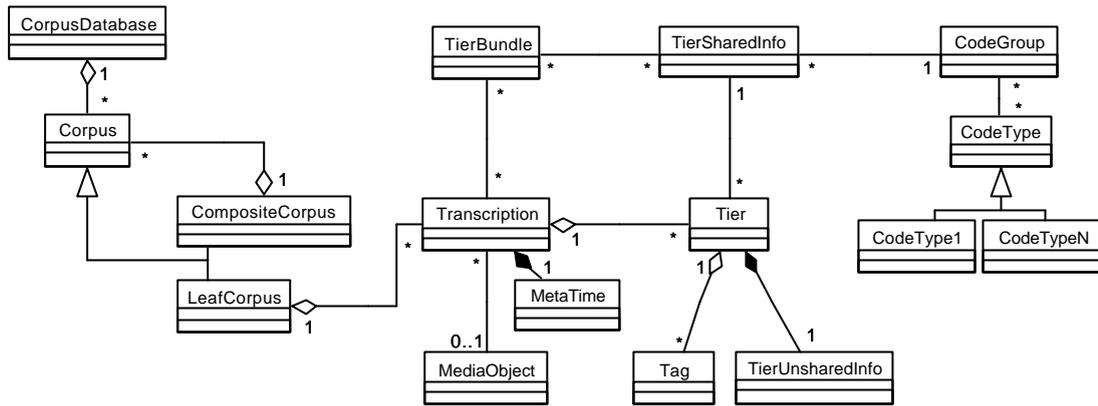
While the encoding scheme describes how to encode the linguistic phenomena (a close handshake in gestures is encoded as "close"), the annotation structure scheme describes the expressive power in structural respects. It has to provide mechanisms for all possible structural phenomena. From our long experience with gesture and sign language studies we know that the annotations can become very complex. There are projects which try to solve this complexity by merging the annotations associated with different linguistic levels into one tier. This method, which is known especially from traditional annotation schemes such as CHAT [11], is also used in new projects. The resulting annotation includes many relations implicitly, i.e. it is the tool which has to include all the knowledge. At the MPI this method was not seen as useful for the future. Different linguistic levels should be separated and all relations such as interruptions, parallelism, semantic correlation should be made explicit.

This is the only way to easily modify the coding later.

In many cases different linguistic interpretations of a gesture are possible. The annotation scheme has to take this into account. Essentially, we follow the indicated way: add another tier which can be used by a new annotator. If only adaptations of the existing annotations are intended, a copy action may be useful for bootstrapping the tier.

The structural phenomena which can occur in annotations are described in detail in [12]. We can summarize the main points:

- The number of tiers can become comparatively high and cannot be seen in advance. It will increase due to various annotators and due to new research goals which require additional information.
- There are all kinds of temporal relations between gesture components and especially between annotations associated with different streams like gestures, speech, facial expression, gaze and others. The complexity makes it necessary to link annotations to periods of time and not to encode overlap and other phenomena in the annotations as older formats require.
- In some occasions spatial relations have to be encoded. They can be encoded as other annotations, i.e. individual or group of coordinate pairs can be linked to time periods.
- In many types of annotations hierarchical relationships have to be included to express linguistic phenomena. These can be token or type oriented. Type specific dependencies are defined at the level of



tier type definitions. Token specific dependencies occur randomly and are defined per linguistic unit.

- Cross-references are very relevant in many cases of linguistic annotation. They describe certain relations which the user wants to draw between two different linguistic units which can be on the same tier or on a completely different one. Comments on some annotation can be interpreted as such cross-reference.

## 5. Abstract Corpus Model

To design the Abstract Corpus Model informal use-case driven method was chosen. In addition a number of existing and well-known annotation formats were analyzed and discussions with linguists about their requirements were carried out. The resulting model defined in UML is more of an operational model than a mere data model.

ACM is realized in first instance as a set of abstract classes that implement common behavior. These abstract classes each have concrete subclasses, one for each of the annotation file formats that ACM currently supports (CHAT, Shoebox [13], relational database [14], Tipster [15], several varieties of XML).

The method calls from ACM's interfaces can be used by a range of annotation related tools. The interfaces are uniform to the tools although the actual objects that implement those interfaces may be instantiated from differently formatted files or even from a relational database. For example, the tools are not aware whether they work on a CHAT file or on a set of database records. Most ACM objects are implemented as remote objects using Java's RMI facilities (Remote Method Invocation). This means that these objects can exist on a central annotation server while the annotation related tools that use their services run on local clients on the network. Method calls to a set of remote interfaces, with arguments and return value, offer a natural way to organize protocols for an annotation server. This type of support for remote objects is efficient since only data that is asked for is sent over the network, i.e. a tier name instead of a complete tier or annotation document. It also forms the basis for a collaborative annotation environment since remote objects can be simultaneously accessed by multiple users. For a class diagram of the first generation of the ACM see figure 1.

It is not the intention of this paper to discuss the part of the class diagram depicted in figure 1 in detail. For this we refer to [16]. But an example can demonstrate how to read it. In this version of ACM, *Tags* have begin and end times that can be specified or unspecified. To make this possible the order of all unaligned *Tags* (i.e. tags which have no specified time marks yet) in a *Transcription* has to be stored explicitly. The object responsible for this is called *MetaTime* and is associated with *Transcription*.

### ACM Revision

Recently, the ACM was revised considerably to include new features. Merging the more elaborated BC (BrowsableCorpus) [17] and EUDICO models of corpora required the introduction of a Session class in ACM. The direct association between Transcriptions and MediaObjects is now administered by a Session object. The composite Corpus structure in ACM is maintained, but as an alternative to BC Corpus hierarchies. There was also a need to introduce Metadata, MetadataContainer and LanguageResource interfaces into ACM as a way to merge in behavior that is needed for BC.

In the first version of ACM, new objects were usually instantiated by their direct ancestors in the corpus tree e.g. Transcription objects were instantiated from LeafCorpus objects. The exact type of the LeafCorpus determined the exact type of the Transcription to be instantiated. In the case of instantiation of a Transcription from a browser over generic corpus trees (like the BC browser) we needed another way to specify the exact type of the Transcription object, and a separate mechanism for creation of this object has to be available.

We were also confronted with a number of related cases where the issue of specifying type and location, and subsequent instantiation of the proper object played a role. For example, in the case of the Spoken Dutch Corpus, currently all digital audio data is delivered on a number of CDROMs. Pointing at and accessing this data, including prompting for the proper CDROM, can be solved by a similar mechanism. For the same corpus, a variation of stand-off annotation is used for annotation documents, where separate annotation tiers are kept in separate XML files in separate directories. Instantiation of an annotation document requires pointing at and combining of these separate files.

To solve this range of problems a design was finished that makes use of the standard mechanisms that Java offers to deal with URLs. Based on a generalization of URL syntax and content type the required access mechanisms (like login prompt, prompt for media carrier) are triggered automatically and the proper type of object is instantiated. In case of ordinary URLs and content types everything automatically falls back on Java's built-in URL handling.

As said, new projects required more complex relations between annotations than the ACM could deal with in its original form. For example, for the Spoken Dutch Corpus both utterances and individual words can be (but don't have to be) time aligned, and each word can have a number of associated codes on different tiers. The Spoken Dutch Corpus also required support for syntactic trees.

For the DoBeS project a wide range of legacy material has to be incorporated in the archive and the EUDICO based archive software has to be able to cope with that. Much of this data is Shoebox or Shoebox-style MS Word data. Therefore interlinear glossing formats have to be supported at the level of ACM. Within the DoBeS community, the maximal format requirements are well described by Lieb and Drude in their Advanced Glossing paper [18].

To support all of these structures two basic types of Annotations were added: `AlignableAnnotations` and `ReferenceAnnotations`. While `AlignableAnnotations` has the necessary characteristics to link annotations to time periods, `ReferenceAnnotations` provide the necessary mechanisms to draw relations between annotations independent of their tier.

In almost every annotation system or format the concept of a tier exists as a kind of natural extension of the concept of a database field applied to time-based data. It is an old idea to "put different things in different places". A tier is the place to put similar things. A tier is a group of annotations that all describe the same type of phenomenon, that all share the same metadata attribute values and that are all subject to the same constraints on annotation structures, on annotation content and on time alignment characteristics.

Metadata attributes for example can be a participant, coder, coding quality, or reference to a parent tier. Constraints on annotation structures can be aspects such as that annotations on the tier refer to exactly one associated parent annotation on a parent tier ( $I - n$ ) or that Annotations on the tier must be ordered in time.

Also annotation content can be constrained by for example a specific closed vocabulary and by a range of possible characters such as Unicode IPA. Constraints on time alignment can also be of various sort such as: Annotations on this tier may not overlap in time.

Explicitly including these types of constraints in the ACM makes tool support for a wide range of use cases and for user interface optimizations possible. For example, known begin or end times of annotations can be reused for new annotations or as constraints on the time segment of other annotations. Text entry boxes can be set up automatically

with the proper input method for IPA, annotation values can be specified using popup menus.

Tier metadata, with attribute values specified or not specified, combined with the tier constraints could be reused as a template for the creation and configuration of new tiers, either in the same document or in another. One step further, a set of tier templates could be part of a document template, making it possible to reuse complete configurations of tiers for other documents.

## 6. Interchange Format

A direct consequence of the ACM is the definition of a suitable and powerful enough annotation interchange format. It is seen as a framework allowing to make ACM content persistent. Here our intentions are fairly comparable with what is currently worked out especially at NIST - called the ATLAS Interchange Format (AIF) [19]. Since AIF could not yet handle all necessary requirements (AIF did not yet support a tier concept) a EUDICO Interchange Format was defined (EAF, see Appendix). However, we would like to join the AIF train to achieve a high degree of interoperability world-wide. Its main structural components are: (1) Time slot values referring to as many as needed concrete time values; (2) information about the tier types and (3) as many `AlignableAnnotations` or `ReferenceAnnotations` as necessary. While the first refer to time slots, the latter refers to annotation IDs.

## 7. Tools

To provide researchers with an efficient annotation and analysis environment, the Institute began early on to setup digitization lines and to build true multimedia tools. The first was the MAC-based MediaTagger annotation tool [20] built in 1994. Consequently, the Institute decided to fully rely on all-digital techniques, i.e. all video and audio signals were digitized. For video it was decided to rely on MPEG1 (after an initial phase of using MJPEG and CINEPAK). Due to its limited resolution, for example, to identify facial expressions in field recordings, it was now decided to change to MPEG2 as a basis for the multimedia archive which has a factor of about 3 more data and bandwidth.

The development of the Java-based EUDICO Tool Set for annotating and exploiting multimedia signals was begun in 1998 and has now reached a flexibility and functionality which makes it one of the most advanced tools for multimodal work. Its nucleus is based on ACM, i.e. it has a comprehensive internal representational power. It has a flexible and easy-to-use annotation and time linking component which allows the user to define his tier setup, which can work with audio and/or video signals in the same way and which makes it possible to do the annotation in various writing systems. It has input methods, for example, for IPA, Chinese, Cyrillic, Hebrew and Arabic. Annotations can either be linked to moments in time in the media stream or to other annotations. It is possible to include hierarchical annotations which is necessary, for example, for an interlinearized representation of morphology.

The EUDICO tool set also provides various views on the multimedia data which can be sound, video, or annotation tracks or other types of signals such as eye tracking tracks. There are a number of stereotypic views on the annotations scientists prefer, therefore EUDICO supports different views and more views can be added according to individual scientists' needs. An important feature is that researchers can easily select and arrange the data tracks they want to see. All viewers in EUDICO are synchronized, i.e. whenever the cursor in a viewer is set to a certain time or segment, all other viewers will move to that instance. The tool set also has a flexible search interface which allows the user to define patterns and associate them with annotation tiers (including all supported input methods) making it possible to enter complex patterns covering several tiers and distances between the patterns. The EUDICO tool set can work in a fully distributed environment where annotation and media tracks are at different locations and support media streaming of fragments. An XML-based generic interchange format was defined (EUDICO Annotation Format), but other formats such as rDBMS, CHAT and Shoebox are also supported.

understand the basic mechanisms of the speech production and comprehension processes. Further the usage of gestures in various cultures could help clarifying the relationship between language and thought. Gestures are very much dependent on the culture and the languages spoken in these cultures.

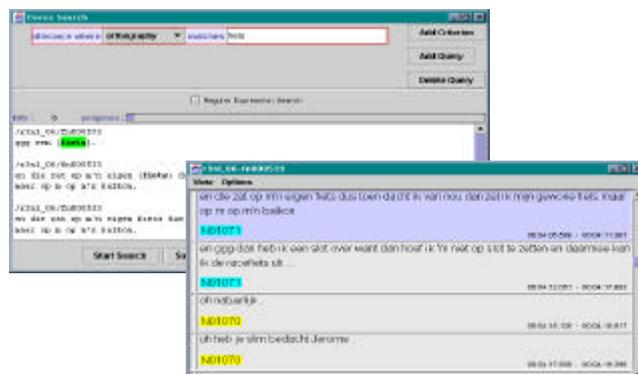


Figure 9 gives an impression of the search feature. It basically allows the user to define search patterns, associate them with tiers and logically combine these patterns to a complete query where also distances can be specified. The result is a list of hits which can be clicked to directly yield the corresponding fragment.

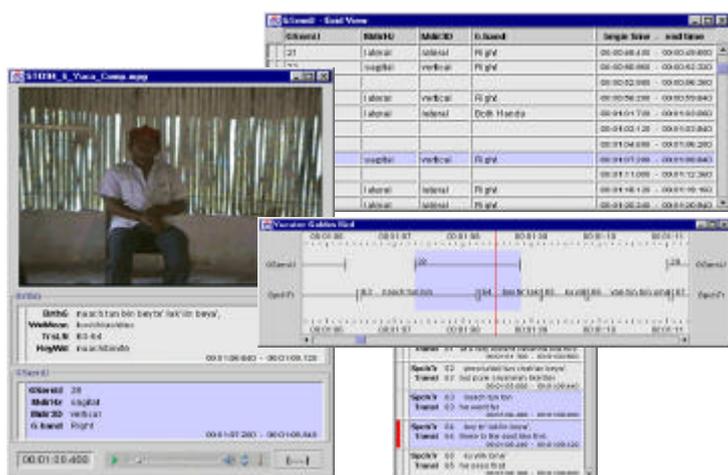


Figure 8 shows the visualization power of EUDICO. Dependent on the project different stereotypic visualizations of the material can be selected. The type of output, the tiers and the order of tiers can be selected by the user. The range of viewers covers dynamic subtitles, a time line view and text viewers with compressed texts.

Tier types can be defined including controlled vocabularies and constraints. Pixel management is very important when dealing with complex tier structures. The user can define the tiers he wants to see and specify the order of presentation. Currently, MPEG1 streaming is supported. MPEG2 is also supported, however downsizing of the video widget is absolutely necessary in order to see the annotations as well.

Further details about the EUDICO Tool Set can be seen on the web-page [21].

## 8. Conclusions

At the MPI for Psycholinguistic the study of gestures has a long tradition. Gesture recordings are used to better

To support this research a large cross-linguistic gesture corpus had to be built including annotations of the speech acts and the gestures. Currently, large international projects have been setup to further investigate the scientific questions raised in this paper.

Such research was only possible by a consequent digitization policy of the institute, by building efficient multimodal annotation and exploitation tools and by powerful mechanisms which help the user to manage large corpora. With the EUDICO and Browsible Corpus technology which was extended within the ISLE project the researchers can rely on tools which will be supported for many years. Since the file formats of both technologies is XML based it can be expected that they will be widely used.

## 9. References

- [1] W.J.M. Levelt (1980). Online processing constraints on the properties of signed and spoken language. In Biological Constraints on linguistic form. U. Bellugi, M. Studdert-Kennedy (eds.). Vgl. Chemie, Weinheim.
- [2] G. Richardson (1984). Word recognition under spatial transformation in retarded and normal readers. Journal of Experimental Child Psychology 38, 220-240.
- [3] S. Kita, J. Essegbey (to appear). Pointing left in Ghana: How a taboo on the use of the left hand influences gestural practice. Gesture.
- [4] S. Kita (1998). Expressing a turn at an invisible location in route direction. In Ernest Hess-Lüttich, J.E. Müller & A. vanZoest (eds.), Signs & SPace. 159-172. Tübingen: Narr.
- [5] A. Özyürek, S. Kita (1999). Expressing manner and path in English and Turkish: Differences in speech, gestures, and conceptualization. In M. Hahn and C. Stones

(eds.), Proceedings of the 21 st Annual Meeting of the Cognitive Science Society. 507-512. Amsterdam.

[6] M. Gullberg, K. Holmqvist (2001). Eye tracking and the perception of gestures in face-to-face interaction vs. on screen. In C. Cave, I. Guaitella, S. Santi (Eds.), *Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication* (pp. 381-384). Paris: L'Harmattan.

[7] H. Lausberg, S. Kita (2001). Hemispheric specialization in spontaneous gesticulation investigated in split-brain patients. In C. Cave, I. Guaitella, S. Santi (Eds.), *Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication* (pp. 431-434). Paris: L'Harmattan.

[8] M. Seyfeddinipur, S. Kita (2001). Gesture and dysfluency in speech. In C. Cave, I. Guaitella, S. Santi (Eds.), *Oralite et gesturalite: Interactions et comportemetns multimodaux dans la communication* (pp. 266-270). Paris: L'Harmattan.

[9] S. Kita, I. v. Gijn, H. vd. Hulst (1998). Movement Phases in Signs and Co-speech Gestures, and their Transcription by Human Coders. In I. Wachsmuth and Martin Frühlich (eds.), *Gesture and Sign Language in Human-Computer Interaction*, Vol. 1371: 23-35. Proceedings of the International Gesture Workshop Bielefeld, Lecture Notes in Artificial Intelligence. Berlin: Springer Verlag.

[10] S. Kita, I. v. Gijn, H. vd. Hulst (2000). *Gesture Encoding*. MPI Internal Report.

[11] B. MacWhinney (1999). *The CHILDES Project: tools for analyzing Talk*. Second ed. Hillsdale, NJ: Lawrence Erlbaum.

[12] S. Levinson, S. Kita, P. Wittenburg, H. Brugman (2002). Multimodal Annotations in Gesture and Sign Language Studies. In *Proceedings of the LREC 2002 Conference*, Las Palmas.

[13] [www.sil.org/computing/catalog/shoebox.html](http://www.sil.org/computing/catalog/shoebox.html)

[14] [www.mpi.nl/world/tg/CAVA/CAVA.html](http://www.mpi.nl/world/tg/CAVA/CAVA.html)

[15] [www.cs.nyu.edu/cs/faculty/grishman/tipster.html](http://www.cs.nyu.edu/cs/faculty/grishman/tipster.html)

[16] H. Brugman, P. Wittenburg (2001). The application of annotation models for the construction of databases and tools. In *Proceedings of the Workshop on Linguistic Databases*. Philadelphia.

[17] [www.mpi.nl/ISLE](http://www.mpi.nl/ISLE)

[18] H. Lieb, S. Drude (2000). *Advanced Glossing: A language documentation format*. Unpublished working paper.

[29] [www.nist.gov/speech/atlas](http://www.nist.gov/speech/atlas)

[20] [www.mpi.nl/world/tg/lapp/mt/mt.html](http://www.mpi.nl/world/tg/lapp/mt/mt.html)

[21] [www.mpi.nl/world/tg/lapp/eudico/eudico.html](http://www.mpi.nl/world/tg/lapp/eudico/eudico.html)  
[www.mpi.nl/tools](http://www.mpi.nl/tools)

## 10. Appendix

This appendix contains the DTD for the EUDICO Annotation Format (EAF).

<!-- edited with XML Spy v4.1 U (<http://www.xmlspy.com>) by Hennie Brugman (Technical Group) -->

<!--

Eudico Annotation Format DTD  
version 0.1  
July 5, 2001

```
-->
<!ELEMENT ANNOTATION_DOCUMENT (HEADER,
    TIME_ORDER, TIER*, LINGUISTIC_TYPE*, LOCALE*)>
<!ATTLIST ANNOTATION_DOCUMENT
    DATE CDATA #REQUIRED
    AUTHOR CDATA #REQUIRED
    VERSION CDATA #REQUIRED
    FORMAT CDATA #FIXED "1.0"
>
<!ELEMENT HEADER EMPTY>
<!ATTLIST HEADER
    MEDIA_FILE CDATA #REQUIRED
    TIME_UNITS (NTSC-frames | PAL-frames | milliseconds)
    "milliseconds"
>
<!ELEMENT TIME_ORDER (TIME_SLOT*)>
<!ELEMENT TIME_SLOT EMPTY>
<!ATTLIST TIME_SLOT
    TIME_SLOT_ID ID #REQUIRED
    TIME_VALUE CDATA #IMPLIED
>
<!ELEMENT TIER (ANNOTATION*)>
<!ATTLIST TIER
    TIER_ID ID #REQUIRED
    PARTICIPANT CDATA #IMPLIED
    LINGUISTIC_TYPE_REF IDREF #REQUIRED
    DEFAULT_LOCALE IDREF #IMPLIED
    PARENT_REF IDREF #IMPLIED
>
<!ELEMENT ANNOTATION (ALIGNABLE_ANNOTATION |
    REF_ANNOTATION)>
<!ELEMENT ALIGNABLE_ANNOTATION
    (ANNOTATION_VALUE)>
<!ATTLIST ALIGNABLE_ANNOTATION
    ANNOTATION_ID ID #REQUIRED
    TIME_SLOT_REF1 IDREF #REQUIRED
    TIME_SLOT_REF2 IDREF #REQUIRED
>
<!ELEMENT REF_ANNOTATION (ANNOTATION_VALUE)>
<!ATTLIST REF_ANNOTATION
    ANNOTATION_ID ID #REQUIRED
    ANNOTATION_REF IDREF #REQUIRED
    PREVIOUS_ANNOTATION IDREF #IMPLIED
>
<!ELEMENT ANNOTATION_VALUE (#PCDATA)>
<!ELEMENT LINGUISTIC_TYPE EMPTY>
<!ATTLIST LINGUISTIC_TYPE
    LINGUISTIC_TYPE_ID ID #REQUIRED
>
<!ELEMENT LOCALE EMPTY>
<!ATTLIST LOCALE
    LANGUAGE_CODE ID #REQUIRED
    COUNTRY_CODE CDATA #IMPLIED
    VARIANT CDATA #IMPLIED
>
```

# Development of the User–State Conventions for the Multimodal Corpus in SmartKom

Silke Steininger<sup>†</sup>, Susen Rabold<sup>†</sup>, Olga Dioubina<sup>†</sup>, Florian Schiel<sup>\*</sup>

<sup>†</sup>Institute of Phonetics and Speech Communication

<sup>\*</sup>Bavarian Archive for Speech Signals (BAS)

Ludwig–Maximilians–University, Schellingstr.3, 80799 Munich, Germany  
{kstein, rabold, olga, schiel}@phonetik.uni-muenchen.de

## Abstract

This contribution deals with the problem of finding procedures for the labeling of a multimodal data corpus that is created within the SmartKom project. The goal of the SmartKom project is the development of an intelligent computer–user interface that allows almost natural communication with an adaptive and self–explanatory machine. The system does not only accept input in form of natural speech but also in form of gestures. Additionally the facial expression and prosody of speech is analyzed.

To train recognizers and to explore how users interact with the system, data is collected in so–called Wizard–of–Oz experiments. Speech is transliterated and gestures as well as user–states are labeled. In this contribution we will describe the development process of the User–State Labeling Conventions as an example for our strategy of functional labeling.

Key–words: multi–modal, annotation, user–states, human–machine interaction, coding conventions.

## 1. Introduction

The goal of the SmartKom project is the development of a multimodal dialogue system that allows the user to interact almost naturally with the computer. Among other things the emotions of the user are taken into account by the system. Since not much is known about the role emotions play in a human–machine dialogue, data is collected in Wizard–of–Oz experiments. The analysis of the interaction of the users with the simulated system can reveal which emotions occur in such a situation, in which way the emotions are expressed and in what connection. For such an analysis the data has to be labeled<sup>1</sup>.

This contribution deals with the problem of how to define a labeling procedure for emotions, respectively. user–states<sup>2</sup>. We will first describe shortly how the data was collected that was used for the development of the labeling procedure. Then we describe the requirements the procedure had to meet. After that we give an overview over the steps of the development process of the procedure and some open questions.

## 2. Collection Of Multimodal Data

The data collection is done with the Wizard–of–Oz technique: The subjects think that they interact with an existing system but in reality the system is simulated by two humans from another room.

In each Wizard–of–Oz session spontaneous speech, facial expression and gestures of the subjects are recorded with different microphones, two digital cameras

(face and sideview hip to head) and an infrared sensitive camera (from a gesture recognizer: SIVIT/Siemens) which captures the hand gestures (2–dimensional) in the plane of the graphical output. Additionally, the output to the display is logged into a slow frame video stream. Each subject is recorded in two sessions of about 4.5 minutes length each. For more information on technical details of the data collection see Türk (2001).

## 3. Developing the Labeling Procedures – Starting Point

### 3.1 Goals

The labeling of user–states in SmartKom serves two main functions:

1. The training of recognizers.

2. The gathering of information how users interact with a multimodal dialogue system and which user–states occur during such an interaction.

These two goals had to be satisfied with the labeling procedures we had to define. For practical and theoretical reasons we decided against a specific system like the "Facial Action Coding System" of Ekman (1978) where the precise morphological shape of facial expressions is coded, but used a simplified, practice–oriented system. The user–states are defined with regard to the subjective impression that a human communication partner would have, if he would be in place of the SmartKom system. This is a functional definition: Not the user–state per se is coded, but the impression the communicated emotion or state generates.

In Steininger, Lindemann & Paetzold (2002a) we already discussed this approach with regard to gestures<sup>3</sup>. The next paragraphs explain our approach relating to user–states.

<sup>1</sup> The development and structure of the gesture labeling is described in detail in Steininger, Lindemann & Paetzold (2002a). The transliteration conventions can be found in Oppermann et al. (2000). The special problem of combining the information of the different labeling steps and the transliteration is discussed in Schiel et al. (2002) at this workshop.

<sup>2</sup> The name "emotion labeling" was changed in "user–state labeling" because the targeted episodes in the data comprise not only emotional, but also cognitive states.

<sup>3</sup> Our gesture coding system also defines hand gestures functionally (not morphologically). A labeled unit is coded with regard to the intention of the user, i.e. with regard to his (assumed) discrete goal.

### 3.2 Practical Requirements

To satisfy the two goals of the labeling process mentioned above the following requirements had to be met. They apply to transliteration, gesture and user-state labeling.

1. The labels should refer to the functional level<sup>4</sup>, not the morphological level. For theoretical reasons we want to use a functional coding system (see below). However, the decision is also made for practical reasons since the structural coding of e.g. facial expressions is exceedingly time consuming.

2. The labels should be selective. Functional codes (as indirect measurements) are not as exact as direct methods, therefore exceptional care has to be taken to find labels that are well-defined, easy to observe and unproblematic to discriminate by means of objective (communicable) criteria. This is even more true for user-states than for gestures because communicable criteria for the discrimination of functional user-state categories are hard to find.

3. The coding system should be fast and easy to use.

4. The resulting label file should facilitate automatic processing (a consistent file structure, consistent coding, non-ambiguous symbols, ASCII, parsability) and preferably should be easy to read.<sup>5</sup>

5. The main categories and most of the modifiers should be realized as codes and not as annotations, in order to heighten consistency. Annotations (free comments and descriptions that don't follow a strict rule) are more flexible, but codes (predefined labels from a fixed set) increase the conformity between labelers.

### 4. Definition of the User-State Coding System

The questions that have to be solved to detect user-states automatically are: Which features of the face and of the voice contribute to an emotional impression – and in which degree does each feature contribute to the impression? Which of these features can be detected automatically?

If we already knew the answers it would make sense to define coding conventions that mark these features in the data. But since we are far from answering these questions conclusively we decided to use another strategy: The labelers mark beginning and end of a user-state sequence and sort it into one of several subjective categories.

A human in a conversation with another human is able to judge which emotion or user-state his or her communication partner shows. Therefore he or she should be able to discriminate relevant user-states in a video. Of course the labeler does not know which emotion is truly present in his communication partner/a human in a video and he or she will make mistakes. But he or she should be good enough to use his emotion-

<sup>4</sup> "Functional code" or "functional unit" is sometimes defined differently by different authors. We use the term in accordance with Faßnacht (1979) for a unit that is defined with regard to its effect or its context.

<sup>5</sup> Many of the practical criteria were adopted from the transliteration conventions for speech in SmartKom, see Oppermann et al. (2000).

detection capability to keep the conversation smooth. This goal is the same for the system – it should be able to detect which user-state is present in its communication partner to keep the conversation smooth.

This consideration we used for the definition of the user-state coding system.

### 4.1 First Step: Pretest – Labeling with some defined subjective categories

First we decided to look for several categories that were deemed interesting for user-state recognition: "anger/irritation", "boredom/lack of interest", "joy/gratification (being successful)", "surprise/amazement", "neutral/anything else". A few sessions were labeled with these categories. Beginning and end were defined by an observable change in the emotional state of the user. It was marked if the user-state seemed "weak" or "strong".

In the first step each session was labeled by at least two different labelers. After the labeling the categories were discussed. "Boredom/lack of interest" was excluded because it could not be distinguished from "neutral". "Neutral" and "anything else" were separated into two different categories because many sequences were found where the users definitely did not show a neutral expression but no meaningful label could be given. Two new categories were included to describe user-states that occurred quite often in the data and are important in the context of human-computer interaction: helplessness and pondering/reflecting.

The label "anything else" comprises three cases:

1. Grimaces with no emotional content, for example playing with the tongue in the cheek, twitching muscles etc. (about 65%).

2. Emotional sequences that have no label in our system, for example disgust (about 5%).

3. States that seem to have an emotional or cognitive meaning, but cannot be decided upon by the labelers (about 30%).

The three cases were put together into one category because they all comprise sequences that are not suited as training material.

Cases like number 2 (disgust etc.) are very uncommon in our context and because of this an extra category was not deemed worthwhile. Cases like number 1 (grimaces for physiological reasons) sometimes look very similar to user-states, but have a different meaning – therefore they have to be distinguished from neutral.

Cases like number 3 would be interesting to analyze further because they comprise complex or difficult to understand user-states. They are sorted into the "anything else" category simply for practical reasons: The other labels should be selective, therefore any label that cannot be categorized for certain has to be sorted into "anything else".

### 4.2 Second Step: Holistic labeling with the conventions

In a second step the sessions were labeled with the following fixed set of categories:

- joy/gratification (being successful)
- anger/irritation
- helplessness

- pondering/reflecting
- surprise
- neutral
- unidentifiable episodes

Consistency was achieved by two correction steps. Final correction was done by the same corrector for every session. Difficult episodes were discussed.



Figure 1: Example of the front view that is used for the holistic and the facial expression labeling. The picture was taken from an episode that was labeled as "anger/irritation" in the holistic labeling step.

### 4.3 Third step: Finding features

The categories are assigned according to the subjective impression of the labelers. Nevertheless the goal is to find detectable features. Additionally the categories have to be describable with observable criteria – otherwise no one else apart from the labelers will be able to understand the content of the labels.

Therefore, for each category some characteristic features were listed. A feature was included in the list if it occurred regularly or if it seemed very distinctive of a category for some subjects.

This step of the development process is still in progress. At the moment the features are simply an aid for labeling. However, the feature list could be studied with objective methods to judge which features are good candidates to be "indicators" for a category.

### 4.4 Fourth Step: Overcoming some limitations

With the holistic labeling system we were relatively sure to catch all relevant user–state episodes and to sort them into selective categories. However, a serious problem had to be solved: For the recognition of facial expressions the coding system was not well suited. Because of the holistic approach the labels included not only information from the facial expression, but also from the voice and from the context. This is a problem because a facial expression recognizer derives information only from the facial expressions and a prosody recognizer derives information only from the voice.

First, we tried to solve the problem with a special marker of the source for a category: voice or face. But it turned out that it was very difficult to make the judgment with regard to the source. Additionally, only very few episodes with the source "voice" could be found.

We abandoned the source marker and included two different labeling steps: Labeling of the facial expression without audio and prosodic labeling.

For the facial expression labeling a different labeler–group watched the videos without audio. The labelers started with a pre–segmented file (from the holistic labeling) to avoid missing subtle episodes that are hard to perceive without audio and context information. This pre–segmentation was derived from the holistic labeling – the names of the categories (apart from "neutral") were deleted, the borders were retained.

Since it seemed to be difficult to use the functional approach with regard to the voice, we adopted a formal coding system that was used in Verbmobil (Fischer, 1999) and changed it to suit our needs in SmartKom.

For the prosodic labeling the transliteration files are filtered: Only the orthographic transcript remains so that the transliteration labels don't divert the prosodic labelers. For the labeling prosodic features like pauses, irregular length of syllables and other prosodic features which could reveal the emotional state of the particular user are marked. There are nine categories for the prosodic labeling:

1. Pauses between phrases
2. Pauses between words
3. Pauses between syllables
4. Irregular length of syllables
5. Emphasized words
6. Strongly emphasized words
7. Clearly articulated words
8. Hyperarticulated words
9. Words overlapped by laughing

The labels were chosen according to the requirements for the User–State recognition group in SmartKom and are thought to represent prosodic features that are indicative of emotional speech. Hyperarticulated words for example, can be indicative of anger. However, it is still not known very well which prosodic features occur during which emotional states. Nevertheless, by the comparison between the holistic labeling and the prosodic labeling it should be possible to detect relevant user–states in speech. For more information on the usage of prosodic features as indicators of emotional speech please refer to Batliner et al. (2000).

For a detailed description of the labels and concrete examples for the labeling procedure please refer to our paper at the main conference (Steininger, Schiel & Glesner, 2002b).

### 4.5 Open Questions

We have to state clearly that the user–state labeling procedure is work in progress. The description of the categories, along with some formal criteria to help differentiate categories that can be mixed easily is not complete. After it's completion, the intercoder agreement has to be measured. At the moment, we can only use the extent of corrections that are done in each correction step

as a rough indicator how reliable the labeling procedure probably is:

**Holistic labeling:** About 20% of all labels are changed with regard to content. About 10% of the segment borders are changed. This is the case for correction step 1 as well as 2.

**Facial Expression labeling:** Only one correction step exists. Segments borders have to be corrected almost never. Changes of labels with regard to content occur in about 20% of the cases.

**Prosodic labeling:** Only one correction step exists. Changes of labels with regard to content occur in about 20% of the cases. Changes of time markers occur in about 50% of the cases.

One other problem that remains are mixed emotions. Since there is no category for mixed emotions, all such cases have to be sorted into "anything else". However, the problem is not as big as it seems: Since we use categories that are defined mainly by subjective impression not mainly by formal criteria, it is rare that a labeler has the impression of a mixed emotion<sup>6</sup>. As already mentioned, the labeler take the viewpoint of a communication partner and try to discern which state his opponent is in. On this level, there almost always is an integrated impression of only one emotion at a time. Many emotional states are mixed of course if one analyses them closely. With a formal system like FACS (Ekman, 1978), mixed emotions correspond to mixed expressions: The face may show anger (for example with a frown) and surprise (for example with an open mouth). In a functional system like ours the viewpoint is taken that it is not known if a frown always means anger and an open mouth always means surprise. If the frown and the open mouth leave the observer (labeler) with the impression of reflecting then this label is given. That is to say that a mixed state on the formal level can lead to a new (holistic) impression on the functional level. Actually this is quite often the case. In most instances there is a clear message for a communication partner. We label only this "clear message", not the subtle undercurrents.

Of course the overall impression can also be of a mixed state. In this case the label "anything else" is given since only very few mixed states were found. Since for the voice a formal system is used and in one labeling step the facial expression is judged without the audio information mixed states for speech and facial expression can occur. In some cases they will be real mixed states but in some cases they will occur because of labeling mistakes.

In our view, formal and functional systems can complement each other, but cannot replace each other because they refer to different levels.

A third important open question is the "anything else" category. For practical reasons some of the most interesting cases "disappear" into this category, namely the episodes that cannot be categorized neatly. Of course it would be of great interest to analyze these difficult episodes further. How could this be done? It is no option

---

<sup>6</sup> With the exception of "sarcasm": Cases where the user is smiling and laughing, but it can be suspected that he is also scornful are labeled as "joy/gratification". Sarcasm is hard to detect reliably, therefore we decided against a special label.

to ask the subjects what they felt in the case of an unidentifiable user-state, because with the functional approach the emotions are labeled that are transmitted to a communication partner. Introspective evaluation of the emotion by the user will give a different picture because of effects of social conventions (among other things). To include recordings of other modalities could be helpful: Hesitant movements for example could give hints about the user-state "helplessness". However, we decided against using additional visual context information because we wanted to focus the labelers on the face and on changes in the voice accepting that some episodes remain unidentifiable. Adding such information later can change the impression (which is highly context dependent), therefore the whole labeling process has to be done again. An interesting option would be to have the unidentifiable episodes judged by a group of naive, untrained labelers (without giving them predefined categories). In this way it could be analyzed if the unidentifiable episodes are episodes that are difficult to understand by a communication partner or if at least some of them form a user state not yet identified as important.

## 5. Conclusion

With the example of the user-state labeling we show a way to handle the problem of finding a labeling system that is consistent, fast and catches the most important episodes in a human-machine dialogue. Since as yet there is not known enough about good indicators for user-state recognition we decided against a formal/morphological system. Instead we define the labels after practical experience with the data, in this way circumventing the danger of missing important aspects by making assumptions about indicators for automatic detection that cannot be justified very well yet.

Additionally, by combining holistic labeling, labeling of the facial expression and a formal system for the speech we can make up for the disadvantages a purely holistic, functional coding system would have. Through comparing the different label files it is possible to analyze and process the data from many different points of view, looking at the whole or at parts at will.

It is also possible to combine the user-state labels with the gesture labels or the speech transliterations. It could be interesting to analyze which kinds of gestures occur during which kinds of user-states. During helplessness there should be less interactional gestures and more searching gestures, for example. The comparison between the gesture labels and the transliterations is especially interesting with regard to reference words that are possibly uttered. A combination of all three modalities could be useful to analyze the question if there are more hesitations and aborts in the speech and gestures during angry and/or helpless episodes.

With the traditional way of annotating input modalities separately such comparisons are not possible. The labeling of data of multimodal systems allows new ways of studying human-machine interaction. However, this will be successful only if the coding conventions allow the combination of the labeling of the different modalities with ease.

## 6. References

- Batliner, A., Fischer, K., Huber, R., Spilker, J., & Nöth, E., 2000. Desperately Seeking Emotions Or: Actors, Wizards, and Human Beings. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.): *Proc. of the ISCA Workshop on Speech And Emotion*. Belfast: Textflow.
- Ekman, P., & Friesen, W. V., Facial Action Coding System (FACS), 1978. *A technique for the measurement of facial action*. Palo Alto, Ca.: Consulting Psychologists Press.
- Faßnacht, G., 1979. *Systematische Verhaltensbeobachtung*. München: Reinhardt.
- Fischer, K., 1999. Annotating Emotional Language Data. *VerbMobil Report 236*.
- Oppermann, D., Burger, S., Rabold, S., & Beringer, N., 2000. Transliteration spontanprachlicher Daten-Lexikon der Transliterationskonventionen-SmartKom. *SmartKom Technisches Dokument Nr. 2*.
- Schiel, F., Steininger, S., Beringer, N., Türk, U., & Rabold, S., 2002. Integration of multi-modal data and annotations into a simple extendable form: the extension of the BAS Partitur Format. To appear in the *Proc. of the 3<sup>rd</sup> Int. conf. on Language Resources and Evaluation, Workshop On Multimodal Resources And Multimodal Systems Evaluation*, Las Palmas, Spain.
- Steininger, S., Lindemann, B., and Paetzold, T., 2002a. Labeling of Gestures in SmartKom – The Coding System. To appear in *Proc. of the Gesture Workshop*, London: Springer.
- Steininger, S., Schiel, F., & Glesner, A., 2002b. User-State Labeling Procedures For The Multimodal Data Collection Of SmartKom. To appear in the *Proc. of the 3<sup>rd</sup> Int. conf. on Language Resources and Evaluation*, Las Palmas, Spain.
- Türk, U., 2001. The technical processing in the SmartKom data collection: A case study. *Proc. of Eurospeech*, Scandinavia, p. 1541–1544.

# Integration of Multi-modal Data and Annotations into a Simple Extendable Form: the Extension of the BAS Partitur Format

Florian Schiel\*, Silke Steininger†, Nicole Beringer†,  
Ulrich Türk†, Susen Rabold†

\*Bavarian Archive for Speech Signals (BAS)

†Institut für Phonetik und Sprachliche Kommunikation

University of Munich, Schellingstr. 3, 80799 München, Germany  
{schiel,kstein,beringer,tuerk,rabold}@phonetik.uni-muenchen.de

## Abstract

Multi-modal resources typically consist of very different data in terms of content and format. This paper discusses a practical solution for the integration of different physical signals as well as associated symbolic data into a common framework. There are ongoing efforts like for instance the ISLE project to develop guidelines and best-of-practice for the standardized representation of such data collections. Since these efforts have not yet converged into a widely accepted concept, we suggest as a starting point to use two different already existing frameworks that can be easily combined for this purpose: The QuickTime format for the handling of synchronized multi-modal signals and the (extended) BAS Partitur Format for the handling of all symbolic data. We can show that with this simple approach it is already possible to integrate the rather complex data streams of the SmartKom Corpus into an easy-to-use format that will be distributed via the Bavarian Archive for Speech Signals (BAS) starting in July 2002.

## 1. Introduction

The last years have seen quite a number of projects starting to work on the processing / recognition / output of multi-modal data in man-machine-interaction systems. However, a quick survey in the Web sites of LDC<sup>1</sup>, ELDA<sup>2</sup>, CSLU<sup>3</sup> as well as in general search engines shows that such data are not widely available to the scientific community outside of dedicated project groups<sup>4</sup>. On the other hand projects like ISLE<sup>5</sup> started with the aim to extend the EAGLES initiative with guidelines and standards for multi-modal data, but has not produced any recommendations yet. Although standards and role models do not exist, in most scientific projects people had to get started collecting data for their special needs, in most cases gathering material for training and evaluation of multi-modal input devices. Almost like twenty years ago when the creation of language resources started to get going the concerned scientists nowadays collect and annotate data to their needs and with the tools and standards available.

So did we when we started to collect data for the German SmartKom project<sup>6</sup> beginning of 2000. Unfortunately, this MO will very likely aggravate the future use of these corpora, which is a shame considering the very high efforts (and costs) that are invested into these resources.

Meanwhile the SmartKom group at BAS has collected a vast amount of multi-modal data (about 1500 GByte) and has solved most of the technical problems that come with such a task. As reported elsewhere (Tuerk, 2001)

the SmartKom data collection consists of 9 different audio channels, two high resolution video streams, one infrared video stream (black and white) and a screen capture (very low frame rate), a HID input and a pen input. Within the last year we were faced with the problem to integrate all these different modalities (signals) together with the various annotation of data streams into a common framework that may be used for the final distribution of the corpus (starting in July 2002 with the first release of SK Public). The two main problems here are that on the one hand different modalities are recorded by different non-synchronized capture devices, on the other hand annotations to different modalities are produced with the use of different – sometimes even self-written – software tools. All this results in a huge variety of resolutions, time bases, file formats that will hinder the easy usage of the corpus by others.

## 2. Practical solution

In this contribution we would like to give a proposal (to be precise: two independent proposals) how to handle these problems with existing frameworks. We do not claim that our proposal will be the ultimate and best solution. However, it could act as an intermediate step that allows the immediate work with multi-modal data and might make the conversion of multi-modal resources into a future standard (whatever it might be) less painful.

Let us first list a few basic requirements denoting the intended characteristics of the framework for multi-modal resources (FMMR). Our intended FMMR

- should be extensible and flexible.

In almost all cases a fixed format for data resources is bad news for the scientist or developer, because he then uses a lot of unnecessary time to solve data format problems. Although this has been true for mono-modal resources as well, the problem multiplies when

<sup>1</sup><http://www ldc.upenn.edu/>

<sup>2</sup><http://www.elda.fr/>

<sup>3</sup><http://cslu.cse.ogi.edu/corpora/>

<sup>4</sup>The only exception being the M2VTS biometrical corpus available at ELDA

<sup>5</sup><http://isle.nis.sdu.dk/>

<sup>6</sup><http://smartkom.dfki.de/>

it comes to multi-modal data. Therefore the framework should not be a fixed definition for different kinds of modalities and how to treat them but rather an extensible framework that can be easily adapted to upcoming needs.

- should be easy to process.

The reason for this key point is obvious. The conclusion is that we may use a well developed format for which tools are available (for instance XML) or that we use such a simple format that it may be processed with standard tools on the operation system level.

- should not integrate signals and annotations in one file format.

According to our experience in many cases users of a data resources do not need to access all signals or all annotations at the same time. To simplify handling and distribution we therefore strongly recommend that signal and annotation data are separated in storage but linked together via the time base (like it was done in the SAM and BAS Partitur File (BPF) standards).

With these basic requirements in mind our proposed method can be summarized as follows:

1. To integrate the raw data we use QuickTime (QT)<sup>7</sup> for all data that are measured signals or events.
2. To integrate annotations we use BPF or a similar flexible framework (e.g. annotation graphs (Bird, 2001)).
3. We link both representations through the physical time base only.
4. We use what ever necessary relational/hierarchical linking only between the annotation layers.

Note that although we use the BPF in the following examples, this is exchangeable to any other equally qualified format. The point we want to stress here is not the format but that the symbolic (annotation) data should be kept separate from the signals, but be grouped into a single framework for easier analysis.

We will discuss the pro and cons of our approach in the following section using the SmartKom corpus as an example.

### 3. Example SmartKom

To demonstrate that our proposal does actually work we show as an example the integration of a complex data collection in the SmartKom project where a wide range of signals and annotations are currently used.

#### 3.1. Integration of signals in QT

Let us first look at the integration of signals into a QT frame. QT allows the integration of several kinds of media into a single multi-media file. Theoretically every signal format that describes physical measurements (signals or events) may be incorporated, if you provide the necessary interface to QT. Fortunately, interfaces for most of the

common file formats do already exist. Therefore, it is possible to integrate for instance video, audio, images, vector graphic and even text into a QT frame without the need to transform the single modalities from their original format; since they remain in their original files, it is also possible to access to the data via other tools than the QT player, if necessary. The only problem is the synchronisation of different time bases, e.g. the synchronisation of a video stream with 25 frames per sec on one computer with an audiostream captured at 48 kHz on another system. We have not found yet an elegant solution to synchronize automatically. At the moment we use a technique quite similar as in movie productions: we synchronize manually with regard to a significant acoustical and visual event at the beginning of each recording. Even more difficult is the synchronization of 2D spatial data with the video signals. In the SmartKom corpus the output of the gesture analyzer consists of a stream of coordinates in the working area indicating pointing gestures of the user. We solved this problem by converting the two-dimensional data into so-called sprites – that are little bit maps that move in the visual plane – and then overlap both pictures to synchronize the infrared picture of the hand with the sprite. Please refer to (Tuerk, 2001) for a detailed discussion of the synchronization problem.

In SmartKom a typical session file contains the following tracks:

- video of the face, frontal, DV format.
- video of upper body, from left, DV format.
- video of infrared camera directed on display to capture hand gestures, from top, DV format.
- audio in 10 channels (microphone array (4), directed mic, headset (2), background noise (2), system output) captured by a 10-channel audio card with 48 kHz
- graphical system output captured by a screen capture application at 4fps, AVI format.
- combined video frame with face, upper body, system output and infrared, AVI format.
- coordinate logfiles: output of either the gesture recognition system (finger tip) or the output of the graphic tableau (pen tip)

For performance reasons all streams are captured on different computers. Coordinate logfiles are transformed into a sprite track to make coordinates visible in the video signals. Then all raw signals are synchronised and integrated into a QT frame.

#### 3.2. Pros and Cons of QT

As mentioned above QT is an open format that serves some of our intended purposes: it is quite easy to use, it is extensible to new, yet unknown formats, and data are accessible via the QT standard library. The synchronization is still a problem but solvable. The alternative would be a fully synchronized capturing hardware, but that was far out of our budget range. The original formats of the data are still accessible on the distribution media which makes the

<sup>7</sup><http://developer.apple.com/techpubs/quicktime/quicktime.html>

access easy for people that do not want to use QT. Furthermore, parts of the synchronized stream may be used across different data collections.

When the SmartKom project started we also discussed other possible formats than QT. The Java Media Framework (JMF) was already out at that time and would have had the advantage to run completely in JAVA. However, this also caused a very low performance compared to QT which is coded in C++ (encapsulated in a JAVA class library). Also, we could not get necessary drivers in JMF for our intended platforms, for instance no recording drivers for Mac and no DV codec.

The other alternative would have been the Microsoft Media Format (MMF, nowadays mostly replaced by AVI). MMF was only available for MS platforms and – being a mere format definition and no consistent system like JMF or QT – was not flexible enough for our needs.

One major drawback of QT is the still missing QT library and QT player for Linux OS (we managed to get a QT player running in a Win emulation environment, but the performance is very bad). We hope that with the further spreading of QT this will be solved in the near future.

Depending on how many video streams are integrated into the QT frame it is sometimes necessary to spread the frame over more than one DVD-5 which makes working with the data difficult. Also the time deviation between the time bases of the capturing devices is getting significant in longer recording sessions. We avoid this by restricting the length of one recording session to 300 sec.

Figure 1 shows four data streams of a SmartKom recording within a single flattened video frame. In the upper left quadrant the video signal of the face camera is shown; in the upper right quadrant the video signal of the body from the left; in the lower left quadrant the displayed output of the system, in the lower right quadrant the output of the system and as an overlay the video signal of the infrared camera that captures the user's gestures. The shown frame is actually from a video stream that was calculated from the original QT frame; the QT Player Pro is principally capable to show many video streams simultaneously, however the performance on a standard Intel platform is still unsatisfying.

### 3.3. Integration of Annotations into BPF

During the last 5 years we have shown that the BAS Partitur Format (BPF) developed at the Bavarian Archive for Speech Signals in 1995 is very successful to integrate so called 'symbolic information' (that is in most cases some kind of annotation) of speech recordings into a simple text based format (see for instance (Schiel et al., 1998)). A BPF is a simple text file very similar to the first SAM label file standard, but has no fixed format concerning the syntax and semantics of the contained tier information blocks. Therefore it is quite easy to extend the format to new needs as long as the meta structure is followed to. Based on the UNIX filter concepts it is possible to add new tier information blocks to a BPF without the need to re-write existing application software (as long as this software does not need to access to the new tier information, of course). A simple chaining mechanism within the different tiers al-

lows the integration of annotations without any direct link to the physical time base; by following the chaining to such a tier all remaining tiers are automatically projected to their right position within the signal.

Let us have a closer look at the structure of the BPF<sup>8</sup>: A BPF file is a simple ASCII file in which each line has a three character key followed by a colon at the beginning that defines the syntax and semantic of this particular line. A BPF consists of a mandatory header structure (compatible to SAM) that must contain a minimum of descriptors, for instance:

```
LHD: Partitur 1.2.11
REP: Muenchen
SNB: 2
SAM: 16000
SBF: 01
SSB: 16
NCH: 1
SPN: ABZ
LBD:
```

Most important entry in this context is 'SAM' which denotes the sampling frequency for all time references in the following annotation tiers.

After this header block an arbitrary number of tier blocks may follow marked by their respective line key. Registered BPF tiers together with their syntax and semantics can be found on the BAS Web pages. For instance the tier block

```
ORT: 0 all
ORT: 1 right
ORT: 2 Mister
ORT: 3 Durante
ORT: 4 <uh>
```

transcribes the pure lexical words of a short utterance. The numbers in the second column are 'links' between different tiers. In principle there may any sort of links units defined (for instance chunks, words, syllables, events etc.). At the moment the BPF standard uses only one type of link that is the word unit counted from the beginning of the recording. Therefore BPF tiers come in only 5 basic types:

1. Events attached to a word, a group of words or the time slot between two words.
2. Events that denote a segment of time without a relation to the word structure.
3. Events that denote a singular time point without a relation to the word structure.
4. Events that denote a segment of time associated with a word, a group of words or the time slot between two words.
5. Events that denote a singular time point associated with a word, a group of words or the time slot between two words.

<sup>8</sup><http://www.bas.uni-muenchen.de/Bas/BasFormatseng.html>

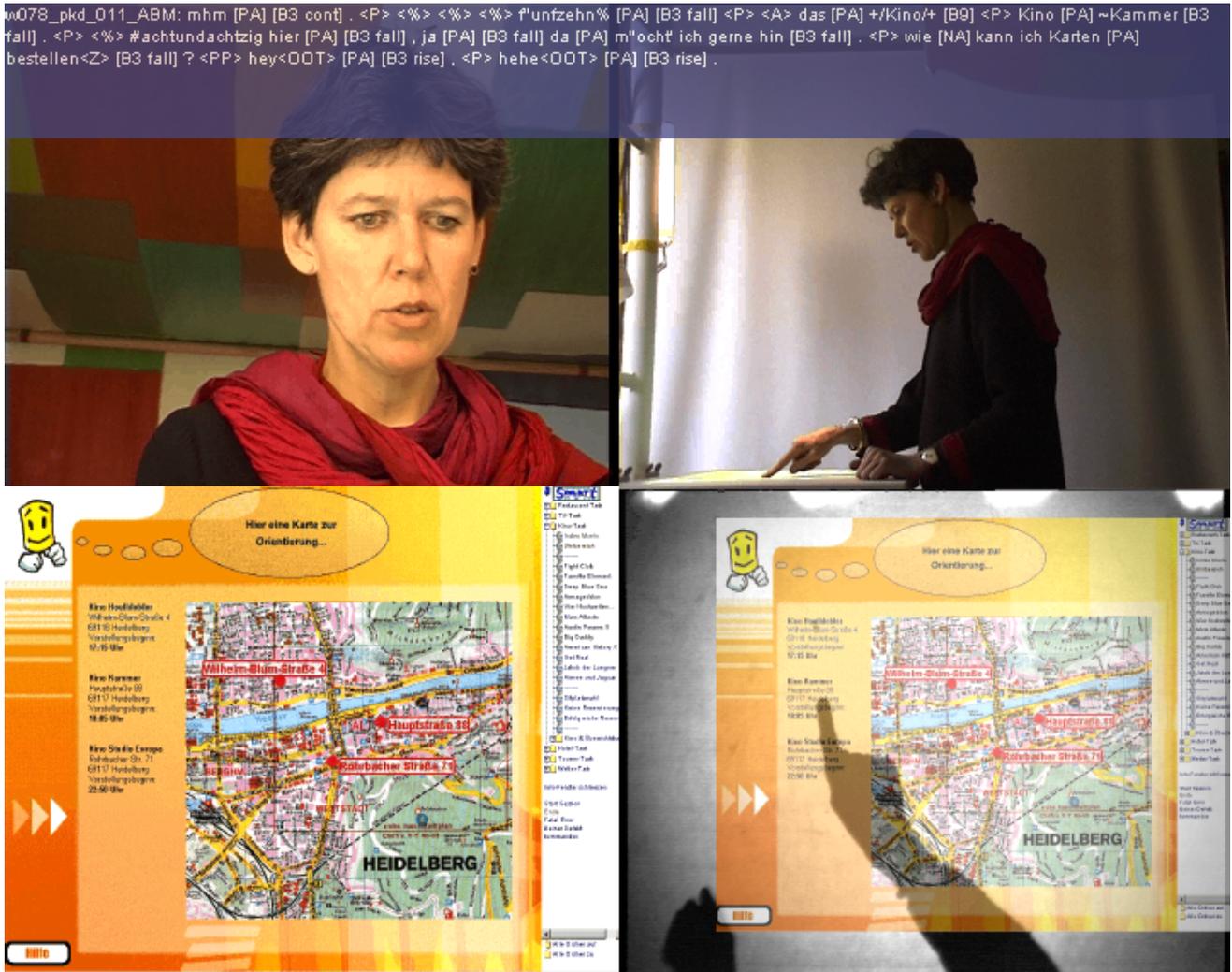


Figure 1: Four synchronized video streams extracted from a SmartKom QT file (see text)

The tier blocks have no preference in order<sup>9</sup> nor hierarchical structure. It is therefore quite easy to cut and paste BPF tiers with standard UNIX tools.

We have shown that the BPF is capable to integrate a variety of symbolic information that was produced within the German Verbmobil project corpus. These data range from simple word alignment over complex syntactic-prosodic tagging up to syntax tree structures. A total of 21 different tiers to the speech signal were used in the Verbmobil corpus (Weilhammer et al., 2002).

Encouraged by this success we started to think about the possibility of integrating symbolic information of multi-modal data as well. Surprisingly enough we managed without changing the meta structure of BPF to integrate the following tier information into an BPF (in brackets the corresponding BPF tier keys):

- SmartKom Transliteration of audio channels (TRS,SUP,NOI,ORT,KAN)
- Turnsegmentation (TRN)

<sup>9</sup>not even within one tier, although the readability is better if the entries follow the time flow

- Segmentation and labeling of gestures in the 2D plane (GES)
- Segmentation and labeling of user state (facial and speech) (USH)
- Segmentation and labeling of user state from facial expression only (USM)
- Segmentation and labeling of complex prosodic features to recognize 'emotions' (USP)

Please note that the above annotations are produced with a variety of different software tools (eg. USS, CLAN, Interact). Simple Perl scripts are used to transform the label and segmentation information into the BPF tier information block and add them by concatenation to the existing BPF.

The following example shows an extract from a SmartKom BPF. For better readability the file is abbreviated to the first 12 words of the dialogue and the header block is omitted.

```

TRS: 0      <"ah> [NA] [B2]
TRS: 1      hallo [PA] [B3 fall] . <A> <P>
TRS: 2      kennst [NA]
TRS: 3      du

```

```

TRS: 4 den [B2]
TRS: 5 Wetterbericht [PA]
TRS: 6 f"ur
TRS: 7 heute
TRS: 8 abend [B3 fall] ? <P>
TRS: 9 <:<#> na:> [NA] [B2] ,
TRS: 10 vergi"s [PA]
TRS: 11 es [B3 fall] . <#>
...
SUP: 42,43 w104_mt_SMA.par @1m"ochtest @1du
SUP: 55 w104_mt_SMA.par Pl"atze . <P>2@>
SUP: 56 w104_mt_SMA.par <:<#> hier3@:>
SUP: 61 w104_mt_SMA.par bitte . <P>4@>
ORT: 0 <"ah>
ORT: 1 hallo
ORT: 2 kennst
ORT: 3 du
ORT: 4 den
ORT: 5 Wetterbericht
ORT: 6 f"ur
ORT: 7 heute
ORT: 8 abend
ORT: 9 na
ORT: 10 vergi"s
ORT: 11 es
...
KAN: 0 QE:
KAN: 1 hal'o:
KAN: 2 k'Enst
KAN: 3 d'u:+
KAN: 4 d'e:n+
KAN: 5 v'Et6#b@r"ICT
KAN: 6 f'y:6+
KAN: 7 h'OYt@
KAN: 8 Q'a:b@nt
KAN: 9 n'a+
KAN: 10 f6g'Is
KAN: 11 Q'Es+
...
TRN: 66560 197888 0,1,2,3,4,5,6,7,8,9,10,11 002
TRN: 377984 43776 12,13,14,15 004
...
NOI: 1;2 <A>
NOI: 9 <#>
NOI: 11;12 <#>
...
USH: 0 244480 Neutral
USH: 244480 519040 "Uberlegen/Nachdenken
USH: 517760 25600 Hand im Gesicht
...
USM: 0 515840 Neutral
USM: 515840 216960 "Uberlegen/Nachdenken
USM: 517760 25600 Hand im Gesicht
...
USP: 1364144 3936 27 CLEAR_ART
USP: 1377776 3536 30 CLEAR_ART
USP: 3437728 5856 63 EMPHASIS
USP: 3983392 14992 73 PAUSE_SYLL
...
GES: 265600 32000 U-Geste U - "uberleg - \
p re Stift nicht erkennbar 640
GES: 376320 30080 I-Geste I - tipp + \
re Stift nicht erkennbar
GES: 515200 29440 R-Geste R - emot - \
re Hand 393600 8320 "Uberlegung/Nachdenken
...

```

In this example the following tier blocks are contained (see references for details about labeling systems and conventions):

- TRS : SmartKom transliteration (Oppermann et al., 2000)
- SUP : Labeling of cross talk between user and system
- ORT : Lexical entity
- KAN : Citation form in SAM-PA
- TRN : Turn segmentation
- NOI : Noise labeling

- USH : User state labeling using video and audio (Steininger et al., 2002b)
- USM : User state labeling using video only (Steininger et al., 2002a)
- USP : Prosodic labeling of features for user state detection
- GES : Labeling of 2D gestures (Steininger et al., 2001)

### 3.4. Pros and Cons of BPF

BPFs of Smartkom are fully compatible to BPFs of mono-modal resources. For instance we can easily train a speech recognizer with the data of Smartkom as well as the data of Verbmobil together, since the BPFs tier information blocks for this purpose are identical.

Since the BPF is an open format it is very simple to extend it, for instance by a new tier that contains the time synchronized coordinates of the finger tip delivered by an early stage of the gesture recognizer.

As defined in the BPF format the link to the actual physical signals is solely achieved by reference to the physical time base. It is clear that by doing this the format of the individual signals is arbitrary. It may be the QT format that we use; it may be another format or it may be even just an extraction of a certain modality, as long as the time synchrony is maintained.

Software tools that read only a specific tier information do not need to be adapted when the BPF is extended to a new tier (except of course that the tool needs to process the new tier blocks).

Since the BPF is a simple ASCII file it is usable across platforms.

The BPF does not allow free hierarchical structuring as for instance in the EMU system.

There is no provision in BPF to use UNICODE for special languages or for IPA.

There is no general purpose viewer available for BPF. Up to now we use Praat<sup>10</sup> or SFS<sup>11</sup> to view traditional mono-modal BPFs resources. For the SmartKom corpus we use the QT library that allows to blend in time-aligned text labels as can be seen in figure 1.

There is no dedicated databank system for the BPF. Although we have developed a PROLOG based databank system for the Web that allows simple and complex queries, this is not a general purpose tool. However, it is quite easy to import BPF files into any data bank system.

Last but not least: BPF is not XML. We have started to use parsers that convert BPF tiers into XML. However, it turns out that BPF is easier to read by humans than the XML version.

## 4. Conclusion

Our approach to use two existing data frameworks, QuickTime (QT) and BAS Partitur Format (BPF) for multi-modal data collections was borne out of the need to get started without having any role models and/or applicable

<sup>10</sup><http://www.praat.org/>

<sup>11</sup><http://www.phon.ucl.ac.uk/resource/sfs/>

standards. We recognize that our current mode of operation is a compromise with some drawbacks. On the other hand it is quite surprising that the integration of multi-modal signal data together with their annotations went rather smoothly. We hope that our experiences will help other researchers that face similar logistic problems as well as researchers that are in the process of defining best-of-practice procedures in the field of multi-modal speech resources. The SmartKom corpus will be made accessible for the public beginning July 2002. Following our policies with monomodal speech resources we will provide a free access to the symbolic data of the corpus via simple FTP download from the BAS server<sup>12</sup>. To obtain the QT files on DVD-5 media please contact [bas@bas.uni-muenchen.de](mailto:bas@bas.uni-muenchen.de) or consult the general BAS Web documentation<sup>13</sup>.

## 5. References

- St. Bird. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33(1,2):23–60.
- D. Oppermann, S. Burger, S. Rabold, and N. Beringer. 2000. Transliteration spontanprachlicher Daten - Lexikon der Transliterationskonventionen. TechDok 02-V4, The SmartKom Project.
- F. Schiel, S. Burger, A. Geumann, and K. Weilhammer. 1998. The Partitur Format at BAS. *Proc. of the 1st Int. Conf. on Language Resources and Evaluation, Granada, Spain*, pages 1295–1301.
- S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in SmartKom - The coding system. *Springer "Gesture Workshop 2001", London*, page to appear.
- S. Steininger, S. Rabold, O. Dioubina, and F. Schiel. 2002a. Development of the user-state conventions for the multimodal corpus in SmartKom. *LREC Workshop on "Multimodal Resources", Las Palmas, Spain*, page to appear.
- S. Steininger, F. Schiel, and A. Glesner. 2002b. Labeling procedures for the multimodal data collection of SmartKom. *Proceedings of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Spain*, page to appear.
- U. Tuerk. 2001. The technical processing in the SmartKom data collection: A case study. *Proceedings of EU-ROSPEECH Scandinavia*, pages 1541–1544.
- K. Weilhammer, F. Schiel, and U. Reichel. 2002. Multi-tier annotations in the Verbmobil corpus. *Proc. of the 3rd Int. Conf. on Language Resources and Evaluation, Las Palmas, Spain*, page to appear.

---

<sup>12</sup><ftp://ftp.bas.uni-muenchen.de/pub/BAS>

<sup>13</sup><http://www.bas.uni-muenchen.de/Bas>

# Multimodal Corpus Authoring System: multimodal corpora, subtitling and phasal analysis

Anthony Baldry\* and Chris Taylor†

\*Dip. LLSM, University of Pavia,  
Strada Nuova 106/c I-27100 Pavia  
baldry@gemini.unipv.it

†SSLMIT, University of Trieste,  
Via F. Filzi 14 - 34132 Trieste  
taylor@sslmit.univ.trieste.it

## Abstract

Designed by Baldry and Thibault and constructed by Beltrami and Caglio, MCA is a multimodal concordancer that identifies recurrent patterns in films. As an authoring tool, it enables researchers, however imperfectly, to view short pieces of film and simultaneously to write multimodal descriptions of them. Using MCA's editing tool, researchers can segment film into functional units and, while viewing these units, type out detailed annotations relating both to the semiotic resources they deploy and the functions they perform within the film. The incorporated relational database allows researchers to search the corpora thus created and identify patterns in them, all of which leads to a further round of hypothesis formulation, segmentation, description and comparison of results. As exemplified by the work carried out by Baldry and Thibault on a corpus of TV car ads, MCA was initially conceived as part of research into the applicability of the systemic-functional approach to multimodal description (Halliday, Kress and van Leeuwen) and in particular Gregory's concept of phase and transition. MCA has since been experimented in various projects within *LINGUATEL* ([claweb.cla.unipd.it/Linguatel/Pavia/MCA.htm](http://claweb.cla.unipd.it/Linguatel/Pavia/MCA.htm)). As the article explains, one such project, namely corpus-driven screen translation, has led the MCA interface to be partly redesigned.

## 1. Introduction

If you are planning your summer holidays abroad this year, you may well decide to learn the local language, or at least just enough to understand what people around you are saying. If you have no time for evening classes, then you will want to do this in your own home. You could, of course, watch a DVD film in the chosen language, switching on the subtitling in that language so as to be able to identify at least some of the words being spoken. But you will soon realise that a DVD presenting your favourite film will not normally allow you to select *all the cases* of a specific activity or the ways in which that activity 'translates' into grammatical structures, whether those of your own language or the foreign language you have chosen to study. Nor will it allow you to check how a particular word or combination of words is typically used in the film. Thus, while DVD may be a great advance over the VHS cassette, when it comes to language learning it has so far provided only limited forms of access to film and video texts. Without the possibility of confirming your intuitions about the way your chosen foreign language works in relation to your own language, you will soon give up. The end of your language learning plans!

Now consider instead watching, and listening to, film texts in a foreign language using an Internet-based multimodal concordancer that carries out targeted searches in film corpora. By definition, such a tool allows you to carry out multiple 'incursions' into film texts, some of which are likely to correspond to your preferred associative patterns and learning strategies. You might, for example, want to take a strictly grammatical approach, searching, for example, for all the cases in a corpus of utterances that correspond to English "you can" carried out in MCA by a query of the type: *AD contains you can*.

Presented, as in any concordance, as a series of rows, a major difference between multimodal concordancing and the linguistic variety lies in the fact that by selecting the player symbol on the left-hand side of each row, you can see and hear *exactly* that part of the film which contains all (and only) the foreign language expressions that correspond to the concordance query (in this case Italian equivalents for "you can"). Moreover, the returned search also transcribes the words used in the foreign language thus assisting word recognition – so important in the initial stages of language learning. The search also indicates the number the text has in the corpus, so that a further query will allow you to listen to your chosen expression in the context of the entire text. Such a query will be of the type: *AD contains n* (where *n* is the number indexing the specific text).

However, your language learning strategies might be such that you tend to shy away from an overtly grammatical approach. You may well prefer listening to a more extensive piece of text using dual-language subtitling. Given that a multimodal concordancer is likely to be based on a *relational* database (in the case of MCA, *Microsoft Sequel Server*), more complex searches can be made – *de facto* a combination of several searches. A set of dual-language subtitles will returned by a query of the type (see Fig. 1): *AD contains text + English subtitle contains + Italian subtitle contains*. Pursuing this approach, you might well decide to select specific grammatical patterns that illustrate and compare, for example, the way questions are formed in the languages under consideration: queries would be of the type: *AD contains text + Italian subtitle contains questions + English subtitle contains questions*. You could, of course, mix the

|                  |                       |     |      |
|------------------|-----------------------|-----|------|
| ▶    1           |                       | 1   | 3.3  |
| English Subtitle | Honda has created     |     |      |
| Italian Subtitle | Honda ha creato       |     |      |
| ▶    2           |                       | 3.3 | 4    |
| English Subtitle | a car so advanced     |     |      |
| Italian Subtitle | un'auto così avanzata |     |      |
| ▶    3           |                       | 4   | 4.6  |
| English Subtitle | that it does without  |     |      |
| Italian Subtitle | che fa men            |     |      |
| ▶    4           |                       | 4.6 | 5.1  |
| English Subtitle | fuel                  |     |      |
| Italian Subtitle | del carburante        |     |      |
| ▶    9           |                       | 9   | 10.4 |
| English Subtitle | today it costs        |     |      |
| Italian Subtitle | oggi costa            |     |      |



Fig. 1 Dual-language subtitling generated by a relational database

two requirements: *AD contains text + Italian subtitle contains + English contains questions*. And you could also decide that instead of car adverts you want to listen to (and watch) something else – hence a query of the type *TV News contains...* etc. Flexible as this may seem, as your search skills increase you may well want to integrate the use of subtitling, translation and grammatical patterns with other strategies such as those that explore specific human activities or textual properties. Here a slightly different type of query can be applied. A search of the type: *AD contains text + SLOGAN contains YES* (or + *SONG contains YES* or + *HIDING contains YES*) will respectively find all the cases in the corpus exemplifying written and/or spoken slogans, songs and examples of hiding. And as your comprehension of the target language increases, you might also want to go beyond this, associating a linguistic approach to language learning with explorations of meanings made in other semiotic modalities – for example, hand and body movements that couple with language to make multimodal meanings in a way that even the casual observer will recognize as fundamental to a film's overall meaning: e.g. touching something or somebody, possible with a search of the type: *AD contains text + touches contains: YES*.

This brief illustration exemplifies how a multimodal concordancer can be used to achieve specific applicative functions (such as language learning) within a multimodal approach to text analysis. Indeed, at the time of writing MCA is still a prototype that is constantly being redesigned – for example, to make it suitable for the learning of minority European languages using the principle of query-generated screen overlays (subtitling, captioning and other more visually-oriented overlays). Other applications include the use of a multimodal concordancer within University courses to help students to understand the multimodal organization of texts, including, as Taylor explains in the following section, efforts undertaken by the Trieste *LINGUATEL* research group to guide students in their learning about screen translation (for a bibliography see Gottlieb).

Indeed, the next section provides a summary of the thinking that led to the original development of MCA and its constant redefinition, a matter discussed in more detail by Baldry in the subsequent section, which also describes MCA's technical specification in relation to research into texts as consisting of phases and transitions between phases, an approach ultimately concerned with defining the typical characteristics of specific multimodal genres.

## 2. MCA in Trieste

The University of Trieste *LINGUATEL* research unit, as part of a wider national research initiative in Italy sponsored by the Ministry, specialises in multimodal text analysis and the devising of strategies for the translation and subtitling of video text. An example of the work carried out by the unit provides the opportunity to describe how MCA can work in practice. Many types of dynamic text have so far been analysed by the Trieste group (feature films, TV soap operas, cartoons, advertisements, documentaries, news broadcasts, etc.) in particular by using the device of the multimodal transcription, originally devised by Thibault and Baldry (Baldry, 2000, Kress and van Leeuwen, 1996). The multimodal transcription technique consists of breaking a film down into single frames of, say, one second duration and minutely analysing their component parts (visual image, kinesic action, soundtrack, dialogue, etc.) thereby providing an approach that really gets to grips with the *multimodal* side of screen translation (see Fig. 2). It provides an ideal tool for analysing the multimodal text in its entirety and drawing the relevant conclusions in terms of how meaning can be successfully conveyed by the various semiotic modalities in operation, and thus how dispensable or indispensable the verbal element is in different sets of circumstances. From this premise it is possible to make informed choices regarding the translation strategies to adopt in subtitling a film.

One type of video text subjected to this multimodal approach was the television comedy series 'Blackadder'. Humour is notoriously difficult to translate, especially apparently British humour, as it involves a large number of interweaving factors: word play, register shifts, timing, characterisation – and creating the humorous effect through subtitles is doubly difficult. The episode examined here, from the Elizabethan era series, features a highly implausible plot involving Lord Blackadder and his scatter-brained assistants, the dreadful Baldrick and Lord Percy, who have inadvertently executed the wrong man, while temporarily in charge of the royal prison. The wife of the unfortunate victim, Lady Farrow, insists on seeing her husband, who she believes is still awaiting trial in the prison. Blackadder's scheme to extricate himself from this situation is to impersonate Lord Farrow at the meeting with his wife by wearing a bag over his head. Lord Percy has the job of explaining this to the unsuspecting lady.

|   |   |  |   |   |
|---|---|--|---|---|
| 1 |  | <p>Shot 1<br/> CP: stationary/ HP: frontal/ VP: median/ D: MLS; VC: interior of the jail; Percy; Blackadder; Baldrick; Mr. Ploppy/ VS: the bag, exactly in the middle of the scene/ CO: artificial set; VF: distance: median; orientation: Blackadder's and Baldrick's gaze towards Percy<br/> <b>Kinesic Action:</b> Blackadder orders Percy out by shouting to him/ Tempo: M</p> | <p>{RG} [ ] Blackadder: (**)<i>Go on, (NA)go on!</i><br/> Pause/ Volume: f/ Tempo: F</p>  | <p>Sbrigati! Sbrigati!</p>  |
| 5 |  | <p>↓<br/> VF: orientation: Percy with closed eyes, avoiding Lady Farrow's gaze; Lady Farrow staring at him<br/> <b>Kinesic Action:</b> Percy turns his head and closes the door, keeping his left hand on the handle/ Tempo: M</p>   | <p>{RG} [ ] Lord Percy: <i>Em (#) (*sorry about the delay (NA)madam!</i> Pause/ Volume: n/ Tempo: M<br/> {RG} [ ] <i>eh (#) as you know (#) you're about to meet your (NA)husband, whom you'll recognise on account of the fact that (#) he has got a (*bag over his head!</i> Pause/ Volume: n/ Tempo: M</p> | <p>Ehm, scusate il ritardo. Tra qualche istante potrete vedere vostro marito. Lo riconoscerete dal sacco in testa</p> |

**Fig. 2: An example of a multimodal transcription**

For reasons of space only two of the rows in the transcription of the subphase (the first and last 1-second frames) have been included. In the first row, Blackadder can be seen among his henchmen preparing to put the bag on his head. Lord Percy is nervously getting ready to meet Lady Farrow who is waiting outside (Row 5). Blending an interpersonal interpretation onto this ideational description, the viewer sees the participants from the same conspiratorial level. Indeed, the viewer has a better perspective than any of the characters in that he/she has an unhindered view of all of them as they are arrayed on the screen. Blackadder is recognised as the boss – he has central position and the others occupy the margin, to use Kress and van Leeuwen's terminology (1996: 206). From a textual point of view, the scene is the thematic element for the whole phase (Gregory, *in press*) covering the fateful meeting, and marks a cohesive element with the third sub-phase when Percy re-enters the room. The set is an obvious mock-up of a prison, the costumes instantly recognisable as Elizabethan period, from Percy's fancy ruff to the rags that Baldrick wears as a member of the lowest social order, the colours and lack of colour playing an important role. This already prepares the audience for the incongruous actions that are to follow, which are at the heart of all humour. The audience subconsciously knows that humour is based on this premise and generally makes every effort to make sense of the text somehow, however bizarre it may be. They are helped by their intertextual knowledge of similar texts they have previously been exposed to. Even a patchy and scholastic knowledge of Elizabethan England prepares the viewer for the setting, knowledge of past BBC comedies, the style of Rowan Atkinson, and indeed past Blackadder series, prepare him or her for the kind of parody that will take place. The foreign audience, however, may need a little more priming, especially pre-planning in the form of prior publicity, articles in other media, etc., but the basic mechanisms come into play just the same. Otherwise, how could we account for the massive popularity of Brazilian 'telenovelas' on Russian television?

To turn now to the question of the translation, the only thing that is said in this brief scene is Blackadder's impatient injunction to Percy – **Go on! Go on!** – and would thus not seem to tax the powers of the translator unduly. But conflicting pressures come to bear. In the interests of condensation, the obvious first step would be to remove the repetition, but this would overlook the importance of interpersonal elements; here the repeated order is designed to express Blackadder's contempt for Percy and intense irritation at Percy's constant incompetence. He almost snarls the words. So do we keep the repetition? At this point, the question of the audience arises. A minimum knowledge of the source language would equip any viewer with the necessary resources to interpret the text. And it is true that even those with no knowledge of the source language would still catch the aggressive intonation and the head movements expressing the feelings of the speaker. However, repetition of a word or short expression puts less pressure on the receptive capacities of a viewer than new material, and repeating the order would probably be the best option. This to-ing and fro-ing between competing solutions reflects the thought processes of the translator as various options are considered, a process well illustrated by Krings' 'thinking aloud protocols' (Krings, 1987). But the problem still remains of what actual words to use. A literal translation into Italian would provide something like – **Avanti! Avanti!** – but if the interpersonal elements are to be integrated, namely the contempt and the irritation, then a version incorporating a fairly colloquial verb plus the second person singular intimate pronoun (expressing the superior to inferior relationship), might be preferred: **Sbrigati! Sbrigati!**

The time taken to discuss this first minimum utterance is an indication of how much thought is required to translate a film for subtitles, but also shows how the multimodal transcription enables the translator to focus his efforts. Proceeding in this vein, the analyst/translator/adaptor/subtitler (who may or may not be the same person) gets a very clear picture of how meaning is being

expressed and therefore to what extent s/he can intervene on the purely verbal element.

Analysing and subtitling large numbers of different kinds of texts, some of which purely for research purposes, others for student thesis preparation, others for language teaching modules, others still for genuine practical use, puts a large burden on computer storage facilities and databank management. Access to MCA allows the researcher/student to plug into a large selection of on-line filmed material which can be experimented with, in Trieste, without downloading material unless and until necessary. Secondly, in a reciprocal light, Trieste users can add to the stock of material on the MCA corpus which then becomes available to other members of the research team, the research community at large, and other selected participants.

This kind of symbiosis is already a reality within the *LINGUATEL* structure. In this way research into limitless genres and subgenres of video text can continue apace and at the same time feed back into the system material already analysed (even tagged) which can be used for other purposes. It is, of course, hoped to extend this service to all interested parties. The potentialities of the system have already far exceeded original expectations and are destined to produce ever more interesting avenues of use.

### 3. MCA in Pavia

#### 3.1 Why we decided to build MCA

The multimodal transcription illustrated above and originally developed by Baldry in relation to the comparison of scenes from different medical texts (Baldry, 2000) and by Thibault in terms of a complete system for multimodal annotation of a bank advert (Thibault, 2000) has many limitations. Essentially a multimodal transcription is a static representation of something that is quintessentially dynamic, providing an *in vitro* frame-by-frame analysis of the component parts of a film. This is fine as far as it goes. But if we want to understand how films make meaning we need instead to develop instruments that examine texts in terms of an *in vivo* analysis treating them as if they were living objects which, as they unfold in time, present constantly changing patterns of semiotic selections. Dynamic texts need to be seen for what they are: a constant weaving and foregrounding of different constellations and integrations of meaning-making resources such as space, gesture, language, ambient sounds, music and gaze.

As Thibault (2000:320-321) points out, multimodal text analysis does not accept either in theory or in practice the notion that the meaning of the text can be divided into a number of separate semiotic 'channels' or 'codes'. The meaning of the text is the composite product/process of the ways in which different resources are co-deployed. A text can be segmented into a series of phases and transitions between phases. This will tell us how the selections of resources from different semiotic systems achieve a consistency of co-patterning. Phases, according to Thibault, are the enactment of the locally foregrounded

selections of options which realise the meaning which is specific to a given phase of the text. Phases and subphases refer to salient local moments in the global development of the text as it unfolds in time. A given phase will be marked by a high level of metafunctional consistency or homogeneity among the selections from the various semiotic systems that comprise that particular phase in the text. Thibault also observes that the points of transition between phases have their own special features that play an important role in the ways in which observers or viewers recognise the shift from one phase to the next. Generally speaking, transition points are perceptually more salient in relation to the phases themselves. Thus, Thibault concludes, viewers of texts have no difficulty in perceiving particular textual phases thanks to their ability to recognise the transition points or boundaries between phases.

Of course, the multimodal transcription can be a useful starting point for an understanding of the ways in which resources such as gaze, gesture and language combine in typical phasal patterns. In the early stages of this work Baldry and Thibault developed a dynamic version of the static multimodal transcription, a forerunner of MCA, which allowed the user to generate the individual rows of a transcription through a query mechanism, and which facilitated understanding of how visual objects and their movements could be analysed in terms of Halliday's metafunctions (Halliday, 1994:38-144).

In an extension to their original conception Baldry and Thibault also devised a form of multimodal transcription that incorporated a multimodal tagging system based on Halliday's description of transitivity (Halliday, 1994: 101-144) but which also included the gestural semiotic (Baldry and Thibault: 2001:94-98). This kind of work can be particularly useful in understanding how (for example in a lecture) a speaker will typically use combinations of language, voice prosodics and gesture to express the point of view and/or circumstances of another person. All this helps us to understand how gesture and language combine to instantiate projection (Baldry and Thibault, 2000:96-98). This approach fits in with the notion of a specialised corpus highlighting specific kinds of textual phenomena such as projection, visual collocation, visual metaphor and so on within in multimodal approach to textuality.

But if we are to pursue our understanding of the codeployment of semiotic resources any further we need to understand how dynamic texts typically unfold in time and to ensure that this unfolding in time can be captured by *in vivo* rather than by *in vitro* multimodal analysis. In order to be able to identify typical patterns, the research process requires us to build corpora that can be analysed in terms of various textual phenomena, including in particular a study of the typical phasal organisation of a specific genre.

These then are the premises that led Baldry and Thibault to construct a corpus TV of car ads (currently 100) that a multimodal concordancer could help analyse. While sketching out some of the very preliminary results of this analysis (the corpus is still being constructed), we may describe the current organisation of MCA (the Beta test version released on 24.03.2002).

### 3.2 How it works

MCA is an XML-based multimodal concordancer whose user interface presents a series of rectangular green buttons (Fig. 3) which make up the *Projects Menu* and which represent the complete set of multimodal texts in the database.

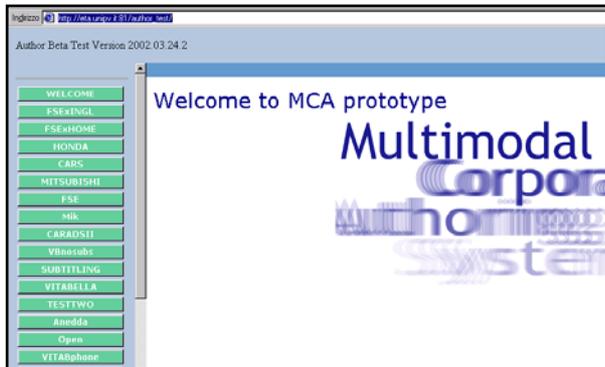


Fig. 3 Home page

When one is selected a specific project will open containing a video text. Using the scroll bar to browse through the *Projects Menu*, two light-blue buttons appear at the bottom of the list which respectively allow the user to create a *new* project in MCA or define *new* query-and-analysis parameters for use within the existing projects. Clearly, the system can be interpreted in one of two ways. While, on the one hand, it may be seen as an archive that can be consulted permitting detailed study of texts, on the other hand, MCA also represents a new area of distance work, given that users can enrich the relational database (which is the heart of MCA) with *their own* projects that are created and modified on-line and which the Server will immediately update and make available in Internet. When a particular MCA project is opened, a web page is loaded with a series of six light-blue buttons (grouped together on the left-hand side of the page) designed to ‘handle’ multimodal texts in ways described below.



Fig. 4: Analysis Inquiry and the Query Page

A user merely interested in viewing a film (such as the language learner posited in the *Introduction*) need only click on the last button, *Analysis Inquiry*, which will open up the *Query Page* and which when searched, using the *Parameters tool*, will return a film. In the example shown in Fig. 4, the query takes the form: *AD contains text 7*.

This will automatically be opened and shown in *Windows Media Player* in the upper right-hand side of the web page. Clicking on the two leftmost *Media Player* buttons (Fig. 4), will, of course, stop and start the film. When we observe the open document, however, we quickly realise that construing MCA merely as a system for the reproduction of film texts would be reductive.

MCA in fact merges a relational database with streaming video technology, that allows specific sequences in a much longer film to be identified (and viewed) and to associate a description to each sequence. The basic idea is that the user can, in this way, consult a series of film sequences which share common characteristics. For example, a scholar concerned with an analysis of soap operas might want to find all the cases in which there is a dialogue between three, as opposed to two, speakers. Although in its current stage of development MCA is not able to show more than one film sequence at a time, the user can, as Fig. 5 indicates, nevertheless identify, with complete accuracy and certainty, all the cases sharing a particular feature, in the case in point all the Audi ads in the corpus.

| Sequence          | Start                          | End  |
|-------------------|--------------------------------|------|
| Parameter         | Brief Description              |      |
| Comments          |                                |      |
| Text 1: Audi (a)  | 1                              | 34   |
| AD                | Text 1: Audi A6; Hong Kong     |      |
| AD                | Ad set in HK at the time of th |      |
| Text 33: Audi (b) | 1119                           | 1148 |
| AD                | Text 33: Audi A6; Guggenheim   |      |
| Text 40: Audi (c) | 1390                           | 1405 |
| AD                | Text 40: Audi A6               |      |
| Text 46: Audi (d) | 1551                           | 1582 |
| AD                | Text 46: Audi (d) Symphony     |      |
| Text 47: Audi (e) | 1586                           | 1621 |
| AD                | Text 47: Audi (e) The Fan      |      |
| Text 48: Audi (f) | 1621                           | 1653 |
| AD                | Text 48: Audi (f) Moving Man   |      |

Fig. 5 Finding a subcorpus within the corpus

In this way, for example, the various sequences can be compared in such a way as to understand how body movements and gestures accompany these linguistic acts in characteristic patterns. To give one example (not shown here), we may care to analyse the way hand-and-arm movements are (as is always the case with gesture) crucially co-deployed with space in the construction of meaning. In a car ad for the New Mini, zombies appear from under the earth and prepare to lay their hands on a couple quietly kissing in a remote spot in their new car. The ad is constructed around the notion of space below and above ground (the zombies pop up from the world below) and inside and outside the car. A series of hand-arm movements are correlated to these notions creating the meaning, recurrent in contemporary European car ads, that the car is a place of safety. Indeed, the transition points in the advert coincide with the camera selecting parts of the car which divide the space in the Mini from the outside world (thereby reinforcing the message of safety and protection): the doors centrally locking from

inside, the front and rear windscreens. They also focus on a cohesive chain of hand movements: zombies outstretched hands (ready for attack), couple's hands raised upwards in fear and defence; driver's hands yawning; one zombie's hand replacing the windscreen wipers and patting the windscreen in a gesture of reassurance and leave-taking. All this is created in tandem with language which takes the form of an off-screen narrator (a voiceover) commenting on the zombies failed attack on the car.

In order for the user to be able to carry out such analyses and to make comparisons with other texts, preliminary operations (segmentation, indexing and tagging) need to be carried out. In particular, the user-author must:

- define a new project associating a film to MCA's descriptive tools (*Project Definition*)
- select the parameters used to tag and describe individual sequences (*Parameters Selection*)
- break the film up virtually into various sequences (*Video Indexing*)
- describe the characteristics of the individual sequences (*Sequence Analysis*) with a view to obtaining finely detailed information when queries are made.

When this has been done the final tool – *Analysis Inquiry* – can be used to produce the results shown in the various figures.

The user who wishes to create a new project can do so by clicking on the *New Project* button but can also modify an existing one (having selected it from MCA's home page). Creation and modification require completion (or redefinition) of the *Project Definition* menu. It is in this phase that the researcher associates a film (previously converted to the \*.wmv format) with the project. Input and output film are the same in MCA. In fact the film remains in its original form. All the work of segmentation, description tagging and retrieval is carried out within the relational database and associated tools.

When this first phase has been completed the parameters relevant to the research project need to be selected through the *Parameters Selection* menu. In the case that appropriate parameters are not available they can be added via a page accessible through the *Parameters Definition* menu. The list of parameters selected can be seen in *Sequence Analysis* page, through which a detailed description of the video text can be made. But before tagging and describing data, the film needs to be split up *virtually* into sequences. This operation is carried out in *Video Indexing*. The research and development cycle is completed with the use of *Analysis Inquiry*, from whose menu various queries and comparisons can be made.

#### 4. Discussion

Although the system is relatively simple to use, nevertheless like any other software program, the MCA system is the result of specific design work which allows a limited degree of flexibility, on the one hand, but, on the other, allows the user to carry out investigations at a speed which would be hard to achieve by other means or which

would be so demanding as not to be worth the candle. Thus while the system requires the use of Microsoft Internet Explorer browser and preferably release 5.5 or higher and a suitably updated release of Windows Media Player, on the other hand, the user is spared the constant need to wind backwards and forwards as is the case with video-cassettes or the wasteful dead-times associated with transferring, reproducing and downloading film.

Moreover, the user/author is able not only to play movie samples but also to add further annotations, providing he/she is authorised to do so (a system of authorisations and passwords is currently being added). Thus two or more researchers working in different locations can work on the description and/or tagging of the same corpus.

In one year's use many initial difficulties have been overcome, the system functioning reliably and responding to requirements for which it was not originally designed

Originally, conceived as an instrument to support research it has proved to be a useful means for teaching and on-line thesis preparation, the user base now including the following categories:

- the researcher who wishes to carry out his/her own work using MCA
- the teacher who needs to hold a language lesson supported by multimedia files
- the student who wishes to follow a self-access language learning course from his/her own home
- the thesis student who must carry out multimodal descriptions of texts
- the user who wishes to show the results of his/her research inserting them into a database.

For each of these user categories, specific needs have emerged which have required further work on the system so as to update it periodically to meet new demands. Feedback from users, who have required help in overcoming problems relating to minimum system requirements, has enabled us to perfect the film-coding technique, in such a way as to optimise the connection via modem and via LAN, avoiding, for example, lack of synchronisation between audio and image in the streaming video. A particular note needs to be made relating to the creation of new materials with subtitles, given the difficulties that users have encountered. For this reason the suggestion has been made that it would be appropriate to introduce a "text box" below the *Media Player* area in future MCA prototypes in which to introduce the subtitle corresponding to the sequence shown, which would be memorised in the database by means of a procedure that would extend the virtual approach adopted in MCA (i.e. subtitles would appear to be printed on the film, whereas in fact they are generated separately from the film).

Like any prototype MCA needs to be improved on and a fully-fledged second prototype is under production which, in addition to what has been outlined above, will introduce account-based security and privacy features respecting different user needs and user typologies more fully and introducing appropriate customisations. On the basis of the experience so far acquired, which has indicated a wider user base than at first expected, we can assume that other user categories, including institutional

users such as Language Centres and Libraries, will make use of MCA for the conversion/distribution of analogical or paper-based data, which may lead to the identification of new criteria for use of MCA. To date, most users have been closely associated only with Italian Universities, and mostly with the University of Pavia. Nevertheless, it has been exciting to follow the progress of graduating students who have used the system as an integral part of their graduation theses. We can therefore expect a growth in the number of graduating and postgraduate students who will use this system and are thus actively seeking inter-University ties and inter-University development projects that will help stimulate this goal.

## 5. Conclusion

Born in the text linguistics sector, MCA is an instrument for analysing dynamic multimodal texts, i.e. film and video texts which, as they unfold in time, display different and constantly varying constellations of sound, image, gesture, text and language (Baldry, 2000, Thibault 2000). Much of this work has already been reported elsewhere but this paper has described a new version of MCA as well as some of the results of one year's use of the tool. The growth in MCA's user base is evidence, apart from the growing interest in the description of multimodal texts, of the desire to learn about the potential and characteristics of this instrument, (including, of course, the need to understand how it works). Designed initially as a support for researchers dealing with the multimodal text analyses of texts, and specifically to provide them with the possibility of examining and comparing multiple contexts and texts in real time, it has proved a useful self-access distance language-learning and text analysis tool, since it provides students with the possibility of listening to, and watching, film clips, that are played and stopped at will. But the system has not yet benefited from critical comparison, one reason why we have decided to present it in various congresses. MCA has been built in virtual isolation vis-à-vis other systems and, in this respect, needs to grow considerably.

## 6. References

- Baldry, A. P. (2000) 'Introduction', *Multimodality and multimediality in the distance learning age*, ed. A. P. Baldry, Campobasso: Palladino Editore, pp. 11-39.
- Baldry, A.P. and Thibault, P.J. (2001) 'Towards multimodal corpora' in *Corpora in the description and teaching of English*, G. Aston and L. Burnard, eds. Bologna: CLUEB, pp. 87-102
- Gottlieb, H (1997) *Subtitles, Translation and Idioms*, Copenhagen: Department of English, University of Copenhagen
- Gregory, M (*in press*) 'Phasal analysis within communication linguistics: two contrastive discourses'. Relations and functions in language and discourse. P. Fries, M. Cummings, D. Lockwood and W. Sprueill, eds New York & London: Continuum
- Halliday, M.A.K. (1994[1985]) *Introduction to Functional Grammar*. Second Edition. London and Melbourne: Arnold.
- Kress, G. and van Leeuwen, Th. (1996) *Reading Images. The Grammar of Visual Design*. London and New York: Routledge.
- Krings H.P., 1987. The Use of Introspective Data in Translation, in Faerch & Kasper 1987: 159-176.
- Taylor, C. (2000) 'The subtitling of film; reaching another community' in *Discourse and community; doing functional linguistics*, E. Ventola, ed. (Tübingen: Gunter Narr Verlag, , pp. 309-327.
- Taylor, C. and Baldry, A. P. (2001) 'Computer assisted text analysis and translation: a functional approach in the analysis and translation of advertising texts' in *Exploring translation and multilingual text production: beyond content*, ed. E. Steiner and C. Yallop, (Berlin: Mouton de Gruyter, pp. 277-305).
- Thibault, P.J. (2000) The multimodal transcription of a television advertisement: theory and practice. In *Multimodality and multimediality in the distance learning age*, ed. A. P. Baldry (pp. 311-38). Campobasso: Palladino Editore.

# The Observer<sup>®</sup> Video-Pro: a Versatile Tool for the Collection and Analysis of Multimodal Behavioral Data

Niels Cadée, Erik Meyer, Hans Theuws & Lucas Noldus

Noldus Information Technology bv  
P.O. Box 268, 6700 AG Wageningen, The Netherlands  
n.cadee@noldus.nl, <http://www.noldus.com>

## Abstract

The Observer is a professional tool for the collection, management, analysis and presentation of observational data. The user can record activities, postures, movements, positions, social interactions or any other aspect of behavior. The Observer can be used either for live scoring, or for scoring from analog or digital video material. With The Observer's generic configuration utility, detailed coding schemes can be designed for observing hand gestures, body postures, and facial expression. In the current version of the software, speech transcription is supported through free-format comments with time stamps. Improvements in the area of speech annotation are currently under way in the framework of the EU-funded NITE project. This includes more advanced coding schemes for speech annotation, as well as interfacing with other linguistic data collection and analysis tools via XML.

## 1. Introduction

The Observer (Noldus et al., 2000) has originally been developed as a tool to support observational studies in ethology. However, over the years it has become clear that the generic nature of the data collection and analysis functions of The Observer make it suitable for almost any observational study. The Observer is currently in use at thousands of universities, research institutes and industrial laboratories worldwide. Applications are found in a wide range of disciplines, including psychology, psychiatry, human factors and ergonomics, usability testing, industrial engineering, labor and time studies, sports research, consumer behavior and market research. Recently, we have noticed an increasing interest from researchers in the area of multimodality and speech annotation. We are in the process of extending our software to cater for the specific demands of researchers in this field, through our active collaboration in the EU-funded NITE project (Natural Interactivity Tools Engineering, <http://nite.nis.sdu.dk/>). Some research groups are already using The Observer to study, for example, turn taking in dialogues between two persons, verbal and non-verbal communication between mothers and toddlers, and for making a dictionary of everyday gestures.

## 2. Program features

The Observer can be used either for live scoring (Basic version), or scoring from analog or digital video material (Video-Pro version). The Observer can control a video recorder from the pc, and by use of a video overlay board, display the video image of a tape on the computer screen, within the program (Fig. 1). Digital video files can be played directly within The Observer. There is a direct coupling between time code in the data files with scored annotations and the video material. This allows for accurate scoring, even when playing the video at slow or fast speeds. Advanced search functions allow the user to find particular events or time stamps on the videotape or media file. A search for events is always based on elements from the coding scheme. A special highlights function allows the user to select specific episodes based on the scored events, and make analog or digital video clips for presentation purposes.

## 3. Program demonstration

During the workshop, The Observer 4.0, the latest release, will be demonstrated. Compared to previous editions, version 4 features improved usability, especially for design of the coding scheme. Data selection has been completely redesigned, and allows for the most complex filtering of annotation results. For example, one can define time intervals of variable length based on actual scored events, to answer questions like 'How often did Peter grin between the time when John entered the room, and the time when John left the room again?' Finally, The Observer 4.0 has an intuitive new layout that shows projects and their content in a tree view (Fig. 2).

## 4. Designing coding schemes

With The Observer's generic configuration utility, detailed coding schemes can be designed for observing hand gestures, body postures, and facial expression. Gaze can be scored manually, or recorded with additional eye-tracking equipment. The coding scheme is based on behavioral classes, which each contain a set of mutually exclusive behaviors. In a simple example, you can have one behavioral class with hand gestures, and another class with types of speech. Aggressive and normal speech could be two mutually exclusive behaviors in the speech class, while pointing and waving could both be in the hand gesture class. A pointing gesture in the hand gesture class can be scored at the same time as aggressive speech in the speech class, but it could also be scored during normal speech. The user can further detail the coding scheme by attaching one or two modifiers to the behaviors. These can indicate for example the intensity of a behavior, or the person or object the behavior is aimed at. For example, for a pointing gesture, you can also score the object that the person is pointing at.

## 5. Speech annotation

Speech and other audio signals can also be annotated with The Observer. The current form of speech transcription is as free-format comments with time stamps. We are working on improvement in the speech area through participation in the NITE project. We are currently looking into options for XML export, and for allowing more structure for speech annotation. The

Observer and prototypes of other software developed in this project will be shown in a demonstration session during the LREC conference.

## 6. Data analysis

The Observer has extensive features for data analysis. The user can filter the data with the data selection function, and for example select variable time intervals based on scored events. Data can be visually examined in tables and plots of events against time (Fig. 2). A range of elementary statistics can be calculated. Reliability analysis is another important feature, where users can

check the consistency of several different people's annotations of the same video material. With lag sequential analysis, temporal relations and patterns can be discerned.

## 7. References

Noldus, L.P.J.J., Trienes, R.J.H., Hendriksen, A.H.M., Jansen H., & Jansen, R.G. (2000). The Observer Video-Pro: new software for the collection, management, and presentation of time-structured data from videotapes and digital media files. *Behavior Research Methods, Instruments & Computers* 32, 197-206.



Figure 1: The Observation Module of The Observer. This example shows a project on children's playing behavior. The user can customize the size and position of the windows, and select which ones to display. In this case, the screen shows the codes for annotation, the event log with the data file with times and scored events, the video image, video controls, and timers.

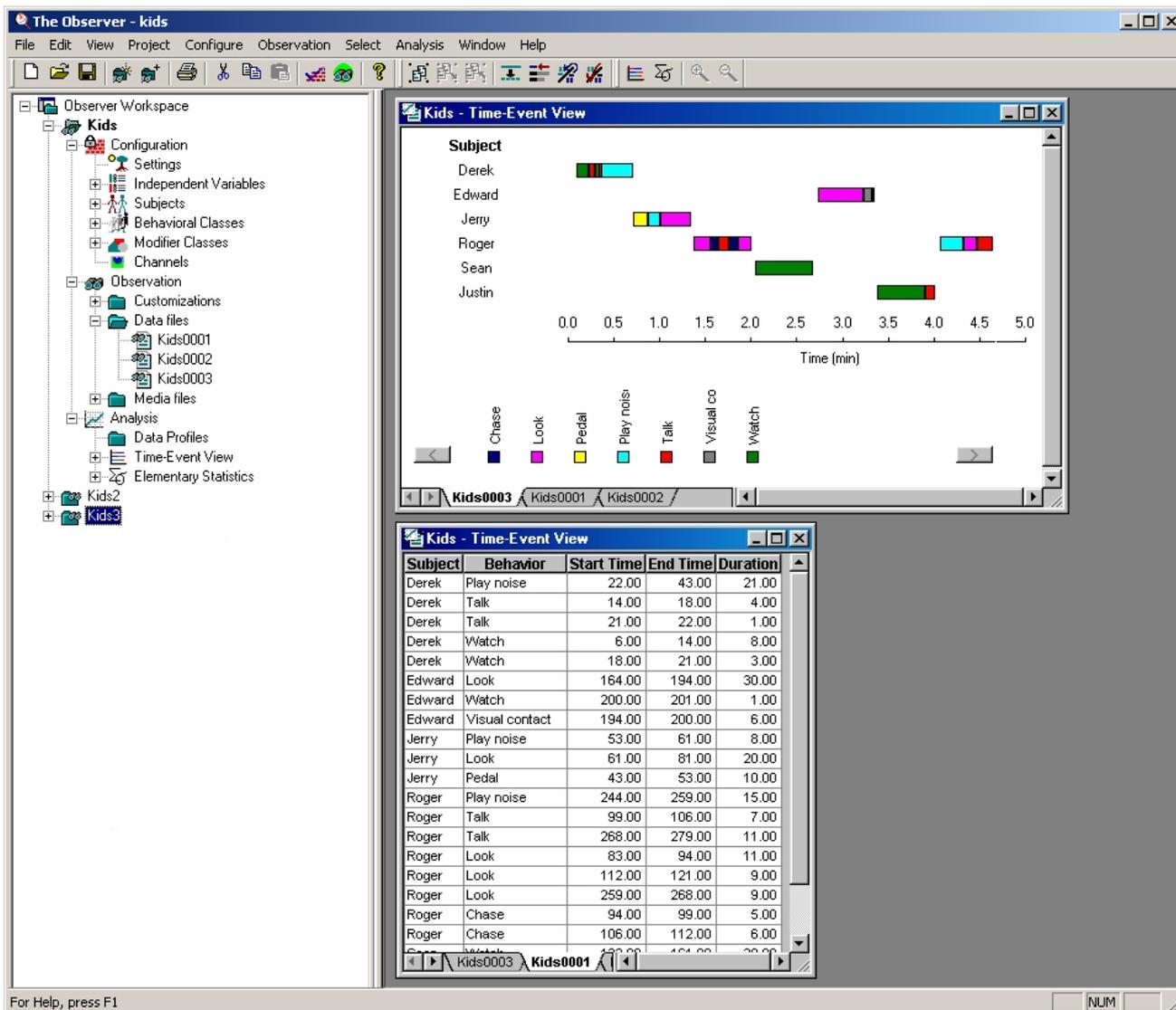


Figure 2: The Main Module of The Observer. In the Explorer view on the left side, a workspace with three projects is shown. One project (on children's playing behavior) is expanded to show its contents. The Configuration contains all settings of the coding scheme. On the right side of the screen, two types of analysis results are shown: a time-event plot and a time-event table.

# Prosody Based Co-analysis of Deictic Gestures and Speech in Weather Narration Broadcast

Sanshzar Kettebekov, Mohammed Yeasin, Nils Krahnstoever, Rajeev Sharma

Department of Computer Science and Engineering  
Pennsylvania State University  
220 Pond Laboratory  
University Park, PA 16802, USA  
[kettebek; yeasin; krahnsto; rsharma]@cse.psu.edu

## Abstract

Although speech and gesture recognition has been studied extensively all the successful attempts of combining them in the unified framework were semantically motivated, e.g., keyword co-occurrence. Such formulations inherited the complexity of natural language processing. This paper presents a statistical approach that uses physiological phenomenon of gesture and speech production process for improving accuracy of automatic segmentation of continuous deictic gestures. The prosodic features from the speech signal were co-analyzed with the visual signal to create a statistical model of co-occurrence with particular kinematical phases of gestures. Results indicated that the above co-analysis improves continuous gesture recognition. The efficacy of the proposed approach was demonstrated on a large database collected from the weather channel broadcast. This formulation opens new avenues for bottom-up frameworks of multimodal integration.

## 1. Introduction

In combination, gesture and speech constitute the most important modalities in human-to-human communication. People use large variety of gestures either to convey what cannot always be expressed using speech only or to add expressiveness to the communication. Motivated by this, there has been a considerable interest in incorporating both gestures and speech as the means for Human-Computer Interaction (HCI).

To date, speech and gesture recognition have been studied extensively but most of the attempts at combining them in an interface were in the form of a predefined signs and controlled syntax such as “*put <point> that <point> there*”, e.g., (Bolt, 1980). Part of the reason for the slow progress in multimodal HCI is the lack of available sensing technology that would allow non-invasive acquisition of natural behavior. However, the availability of abundant processing power has contributed to making computer vision based continuous gesture recognition in real time to allow the inclusion of natural gesticulation in a multimodal interface (Kettebekov and Sharma, 2001, Pavlovic et al., 1997, Sharma et al., 2000).

State of the art in *continuous gesture recognition* is far from meeting the requirements of a multimodal HCI due to poor recognition rates. Co-analysis of visual gesture and speech signals provide an attractive prospect of improving continuous gesture recognition. However, lack of fundamental understanding of speech/gesture production mechanism restricted implementation of the multimodal integration at the semantic level, e.g. (Kettebekov and Sharma, 2001, Oviatt, 1996, Sharma et al., 2000). Previously, we showed somewhat significant improvement in co-verbal gesture recognition when those were co-analyzed with keywords (Sharma et al., 2000). However, the implications of using a top-down approach has augmented challenges with those of natural language and gesture interpretation and made automatic processing challenging.

The goal of the present work is to investigate co-occurrence of speech and gesture as applied to continuous gesture recognition from a bottom-up perspective. Instead of keywords, we employ a set of prosodic features from speech that correlate with deictic gestures. We address the general problem in multimodal HCI research, e.g., availability of valid data, by using narration sequences from the weather channel TV broadcast. The paper is organized as follows. First, a brief overview of the types of gestures that occur in the analysis domain is presented. The synchronization hierarchy of gestures and speech is also reviewed. In section 3 we discuss a computational framework for continuous gesture acquisition using a segmental approach. Section 4 presents a statistical method for correlating visual and speech signals. There, acoustically prominent segments are detected and aligned with segmented gesture phases. Finally, results are discussed within the framework for continuous gesture recognition.

## 2. Co-verbal Gesticulation for HCI

McNeill (1992) distinguishes four major types of gestures by their relationship to the speech. *Deictic* gestures are used to direct a listener’s attention to a physical reference in course of a conversation. These gestures, mostly limited to the pointing, were found to be co-verbal, cf. (McNeill, 1992). From our previous studies, in the computerized map domain (*iMAP*, see Figure 1) (Kettebekov and Sharma, 2000), over 93% of deictic gestures were observed to co-occur with spoken nouns, pronouns, and spatial adverbials.

*Iconic* and *metaphoric* gestures are associated with abstract ideas, mostly peculiar to subjective notions of an individual. *Beats* serve as gestural marks of speech pace. In the weather channel broadcast the last three categories roughly constitute 20% of all the gestures exhibited by the narrators. We limit our current study to the *deictic* gestures for a couple of reasons. First, they are more suitable for manipulation of a large display, which becomes more common for HCI applications. Second, this

type of gestures exhibits relatively close coupling with speech.

## 2.1. Gesture and Speech Production

The issue of how gestures and speech relate in time is critical for understanding the system that includes gesture and speech as part of a multimodal expression. McNeill (1992) distinguishes three levels of speech and gesture synchronization: semantic, phonological, and pragmatic. The pragmatic level synchrony is common for metaphoric and iconic gestures and therefore is beyond the scope of the present work.

Semantic synchrony rule states that speech and gestures cover the same idea unit supplying complementary information when they occur synchronously. The current state of HCI research provides partial evidence to this proposition. Previous co-occurrence analysis of weather narration (Sharma et al., 2000) revealed that approximately 85% of the time when any meaningful gestures are made, it is accompanied by a spoken keyword mostly temporally aligned during and after the gesture. Similar findings were shown in the pen-voice studies (Oviatt et al., 1997). The implication of the semantic level synchronization rule was successfully applied at the keyword level co-occurrence in the previous weather narration study (Sharma et al., 2000).

At the phonological level, Kendon (1990) found that different levels of movement hierarchy are functionally distinct in that they synchronize with different levels of prosodic structuring of the discourse in speech. For example, the peaking effort in a gesture was found to precede or end at the phonological peak syllable (Kendon, 1980). These findings imply a necessity for viewing a continuous hand movement as a sequence of kinematically different segments of gestures. This approach is reflected in the next section. Issue of using the phonological peak syllables is associated with the complexity of the nature of the tonal correlates, e.g., pitch of the voice. Pitch accent, which can be specified as low or high, is thought to reflect a phonological structure in addition to the tonal discourse, cf. (Beckman et al., 1992). We address this issue by proposing a set of correlate point features in the pitch contour that can be associated with the points on the velocity and acceleration contours of the moving hand (section 4).

## 3. Gesture Acquisition

Building human computer interfaces that can use gestures involves challenges that range from low-level signal processing to high-level interpretation. A wide variety of methods had been introduced to create gesture driven interfaces. With the advances in technology there has been a growing interest in using vision-based methods (Pavlovic et al., 1997). The advantage of these is in their non-invasive nature. The idea of a natural interface comes from striving to make HCI as close as communicating in ways we are accustomed to. Vision-based implementation therefore can be very useful for a natural interface.

One could expect that the meaning encoded in multimodal communication is somehow distributed across speech and gesture modalities. A number of recent implementations used predefined gesture syntax, e.g., (Oviatt, 1996). A user is confined to the predefined gestures for spatial browsing and information querying.

As a result, a rigid syntax is artificially imposed. Therefore the intent of making interaction natural is defeated. However, with imprecise recognition of non-predefined gestures, it may be harder to argue for replacing more precise HCI devices, e.g., electronic pen with fixed predefined functions.

The key problem in building such interface, e.g., using statistical techniques, is the lack of existing *natural* multimodal data. Studies from human-to-human communication do not automatically transfer over to HCI due to artificially imposed paradigms. This controversy leads to a "chicken-and-egg" problem.

While the use of the weather narration domain as a bootstrapping analysis offers virtually unlimited bimodal data it can be assumed as a reasonable simplification of an HCI domain. In the series of the previous studies we employed the weather narration broadcast analysis (Sharma et al., 2000) to bootstrap *iMAP* framework (Figure 1) (Kettebekov and Sharma, 2001). It showed that the gesticulative acts used in both domain have similar kinematical structure as well as gesture and keyword co-occurrence patterns. However, the key aspect for choosing the weather domain for the current study is in a possibility of applying simple processing techniques for extraction of prosodic information from uninterrupted narration.



Figure 1. *iMAP* testbed in the context of a computerized map. The cursor is shown within the circle.

Over 60 minutes of the selected weather narration data was used in the analysis. The video sequences contained uninterrupted monologue of 1-2 minutes in length. The subject pool was presented by 5 men and 3 women.

### 3.1. Kinematics of Continuous Gestures

A continuous hand gesture consists of a series of qualitatively different kinematical phases such as movement to a position, hold, and transitional movement. We adopt Kendon's framework (Kendon, 1990) by organizing these into a hierarchical structure. He proposed a notion of gestural unit (*phrase*) that starts at the moment when a limb is lifted away from the body and ends when the limb moves back to the resting position. The *stroke* is distinguished by a peaking effort and it is thought to constitute the meaning of a gesture (Kendon, 1990). After extensive analysis of gestures in weather narration and *iMAP* (Kettebekov and Sharma, 2001, Sharma et al., 2000) we consider following strokes: *contour*, *point*, and *circle*.

Kita (1997) suggested that a *post-stroke hold* was a way to temporally extend a single movement stroke so that the *stroke* and *post-stroke hold* together will

synchronize with the co-expressive portion of the speech. It is thought that a *pre-stroke hold* is a period in which gesture waits for speech to establish cohesion so that the stroke co-occurs with the co-expressive portion of the speech. Therefore, in addition to our previous definitions we also include *hold* as a functional primitive.

### 3.2. Continuous Gesture Segmentation

Sixty minutes of weather domain gesture data for training and testing was collected from broadcast video using a semi-automatic gesture analysis tool (GAT) (see Figure 2). The tool provides a convenient user interface for rapid and consistent collection of positional data and a easily configurable set of pattern classification tools. GAT is integrated with PRAAT software for phonetics research (Boersma and Weenink, 2002) for speech processing and visualization.



Figure 2. Gesture analysis tool (GAT) interface

The task of positional data ground truthing involves initialization the head and hand tracking algorithms (described in 2.3.1) at the beginning of each video sequence and in the events of self-occlusions of the hands.

#### 3.2.1. Motion Tracking

The algorithm for visual tracking of the head and hands is based on motion and skin-color cues that are fused in a probabilistic framework. For each frame and each tracked body part, a number of candidate body part locations are generated within a window defined by the location of the body part in the previous frame and the current estimate of the predicted motion. The true trajectories of the body parts are defined as the most probable paths through time connecting candidate body part locations. The Viterbi algorithm is used to efficiently determine this path over time. This approach effectively models the hand and head regions as skin-colored moving blobs (Figure 3).

#### 3.2.2. Kinematical Analysis

To model the gestures, both spatial and temporal characteristics of the hand gestures (phonemes) were considered. The time series patterns of gesture phases can be viewed as a combination of ballistic and guided motion of the hand reflected on the skewedness of the velocity profile. In the current study, a gesture phoneme is defined as a stochastic process of 2D positional and time differential parameters of the hand and head over a suitably defined time interval.

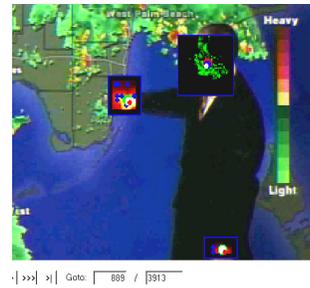


Figure 3. Semi-automatic ground truthing process employing a tracking algorithm;

A Hidden Markov Model (HMM) framework was employed for continuous gesture recognition, as described in (Sharma et al., 2000). The total of 446 phoneme examples extracted from the segmented training video footage were used for HMM training. The results of the continuous gesture recognition showed that only 74.2 % of 1876 were classified correctly. Further analysis indicated that phoneme pairs of preparation-pointing and contour-retraction constitute most of the substitution errors. This type of error, which can be attributed to the similarity of the velocity profiles, was accounted for the total of 33% of all the errors. The deletion<sup>1</sup> errors were mostly due a relatively small displacement of the hand during a pointing gesture. Those constituted approximately 58% of all the errors.

Although purpose of this work was not to introduce a robust algorithm with a high recognition rate there is an inherent limitation with the current acquisition method. I.e., 2D projected motion data can potentially introduce spurious variabilities that can have a detrimental effect on the recognition rate. The gesture model is based on the observed end-effector motion of the hands and the motion of the head projected into the camera plane and is only and indirect measurement of the true body

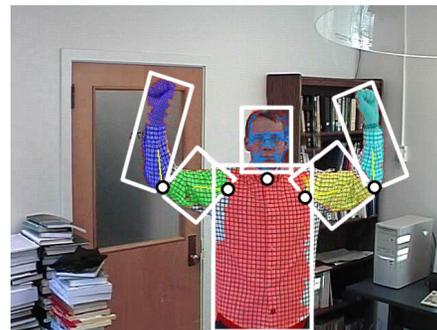


Figure 4. Model based tracking for future extraction of direct kinematical gesture parameters.

<sup>1</sup> Deletion type of errors occur when a gesture phoneme is recognized as part of another adjacent gesture.

kinematics. This observation model can hence introduce distortions and additional spurious variabilities that complicate the differentiation between gestures. Current work in progress, cf. (Krahnstoeber et al., 2002), has the goal of visually extracting the true 3D kinematical parameters such as body pose and angles of the shoulder and arm joints (see Figure 4).

#### 4. Prosody Based Co-analysis

Both psycholinguistic, e.g., (McNeill, 1992), and HCI, e.g., *iMAP* (Kettebekov and Sharma, 2000), studies suggest that deictic gestures do not exhibit one-to-one mapping of form to meaning. Previously, we showed that the semantic categories of strokes (derived through the set of keywords), not the gesture phonemes, correlate with the temporal alignment of keywords, cf. (Kettebekov and Sharma, 2000). This work distinguishes two types of gestures: referring to a static point on the map and to a moving object (i.e., moving precipitation front). Due to the homogeneity of the context and trained narrators in the weather domain we can statistically assume (mismatch <2%) that pointing gesture is the most likely to refer to the static and contour stroke to the moving objects. Therefore, for simplicity we will use *contour* and *point* definitions.

The purpose of the current analysis is to establish a framework by identifying correlate features in visual and acoustic signals. First we will separate acoustically prominent segments. A segment is defined as a voiced interval on the pitch contour that phonologically can vary from a single phone/foot<sup>2</sup> to intonational phrase units, see (Beckman, 1996) for details. Then we will analyze alignment of the prominent segment with the gesture phonemes. This framework was implemented in GAT.

##### 4.1. Detecting Prosodically Prominent Segments

Pitch accent association in English underlines the discourse-related notion of focus of information. Fundamental frequency ( $F_0$ ) is the correlate of pitch defined as the time between two successive glottis closures (Hess, 1983). We employed PRAAT software to extract  $F_0$  contour, as described in (Boersma, 1993).

Prominent segments were defined as segments which were relatively accentuated (or perceived as such) from the rest of the monologue. We considered combination of the pitch accent and the pause before each voiced segment to detect abnormalities in spoken discourse. Maximum and minimum of  $F_0$  contour represent features for high pitch and low pitch accents. Maximum gradient of the pitch slope was also considered. A statistical model of prosodic discourse for each narration sequence was created (Figure 5), see (Kettebekov et al., 2002) for details.

To find an appropriate level of threshold to detect prominent segments we employed a bootstrapping technique involving a perceptual study. A control sample set for every narrator was labeled by 3 naïve coders for auditory prominence. The coders had access only to the wave form of speech signal. The task was to identify at least one acoustically prominent sound within the window of 3 seconds. The moving window approach was considered to account for abnormally elongated pauses in

the spoken discourse. Allowing 2% of misses, the threshold was experimentally set for each narrator (Figure 5). If a segment appeared to pass the threshold value it was considered for co-occurrence analysis with the associated gesture.

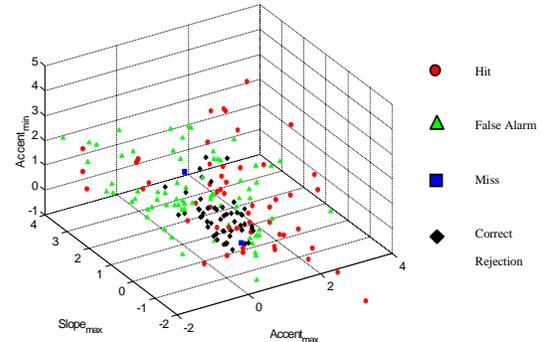


Figure 5. A sample distribution of auditory prominence for a female narrator with the decision boundary from the perceptual study.

##### 4.2. Co-occurrence Models

A statistical model of the temporal alignment of active hand velocity and a set of features of the prominent pitch segments was created for every gesture phoneme class (Figure 6). The features on the pitch profile included max, min, beginning, and max of derivative of  $F_0$ , see (Kettebekov et al., 2002) for details. Present formulation

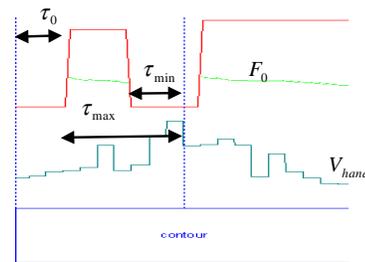


Figure 6. A set of features used for co-occurrence modeling of the hand velocity ( $V_{hand}$ ) and a pitch ( $F_0$ ) segment. Red contour represents prominence level of corresponding segments

accounts for the two levels of possible prosodic co-occurrence: discourse and phonological. The onset between a gesture and the beginning of a prominent segment is to model discourse cohesion (pauses). The onset of the peaks in the  $F_0$  and peaks in the velocity profile of the hand addresses phonological level synchronization. All of 446 phonemes that have been used for training gesture phonemes were utilized for training of the co-occurrence models. Analysis of the resulted models indicated that there was no significant difference between *retraction* and *preparation* phases. Peaks of *contour* strokes tend closely to coincide with the peaks of the pitch

<sup>2</sup> Foot is a phonological unit that has a "heavy" syllable followed by a "light" syllable(s).

segments. *Pointing* appeared to be quite silent, however, most of the segments were aligned with the beginning of the *post-stroke hold* interval.

Figure 7 summarizes findings of the co-analysis framework. At the first level we separate co-verbally meaningful gestures (*strokes*) from *auxiliary* phonemes that included *preparation* and *retraction* phases. Also, we exclude strokes that are re-articulate previous gestures such as a stroke can be followed by the identical stroke where the second movement does not have associated speech segment. At the second level co-verbal strokes can be further classified according to their deixis, cf. (Kettebekov and Sharma, 2001). As it was noted before, in the context of the weather narration we can statistically consider those to be represented by *point* and *contour* phonemes without further definitions. *Preparation* and *retraction* phases were eventually collapsed into the same category and were not differentiated.

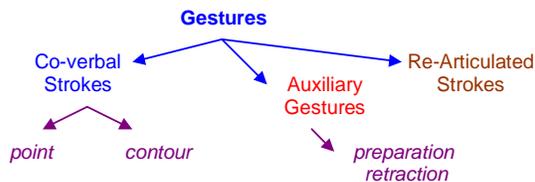


Figure 7. Prosodic co-analysis framework

The co-analysis models for co-verbal strokes were merged with the beginning of the post stroke-*hold* phases for classification purposes. Such redefinition of the co-verbal strokes for the purpose of co-analysis was motivated by the results associated with the *pointing* strokes and it was included into the computational framework.

#### 4.3. Continuous Gesture Recognition with Co-occurrence Models

We employed Bayesian formulation to fuse the gesture framework and the co-occurrence models at the decision level, see (Kettebekov et al., 2002). The resulted segmentation showed significant improvement in the overall performance with the correct recognition of 81.8% (versus 72.4%). Subsequently, there was a significant reduction of deletion (8.6% versus 16.1%) and substitution errors (5.8% versus 9.2%). The deletion type of errors were minimized due to the inclusion of small point gestures, which are quite salient when correlated with prominent acoustic features. Figure 8 shows example of elimination of a deletion error after applying co-analysis. White trace on the figure illustrates visually negligible hand movement trajectory. Improvement of substitution errors can be attributed to the differentiation between the auxiliary gesture phases and the strokes in the co-occurrence analysis.

#### 5. Conclusions

We presented an alternative approach for combining gesture and speech signals from the bottom-up perspective. Unlike commonly controlled gesture

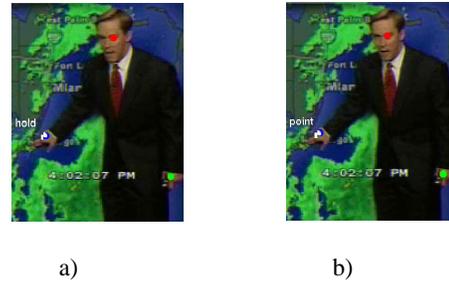


Figure 8. Example of deletion error using: a) visual-only signal resulted in *hold* gesture; b) with co-occurrence model *point* was recognized as a part of preceding *hold* (case a.);

recognition domains, we address this problem in the weather broadcast domain, which can be characterized by relatively unrestricted narration. Such formulation is more favorable for automated recognition of continuous deictic gestures than the semantic based (keyword co-occurrence). The current results demonstrate the concept of improving recognition of co-verbal gestures when combined with the prosodic features in speech. This is a first attempt which requires further improvement. The issues of portability to an HCI setting, e.g., *iMAP* framework, are currently under investigation.

Applicability of the current formulation for the other types of gestures is probably possible if the segmental approach is considered for the gesture acquisition. In a domain with more spontaneous behavior, e.g., in a dialogue (e.g., *iMAP*) (versus monologue as presented in the present work) the methodology of prosodically prominent feature extraction is more complex. It would require acquisition of an improved kinematical model (see section 3.2.2.) that considers additional visual cues such as turn of head (direction of the gaze), and etc.

#### 6. Acknowledgements

The financial support of this work in part by the National Science Foundation CAREER Grant IIS-97-33644 and NSF IIS-0081935 is gratefully acknowledged. We thank Ryan Poore for his help with the data processing and implementation.

#### 7. References

- Beckman, M. E., Dejong, K., Jun, S. A., and Lee, S. H. 1992. The Interaction of Coarticulation and Prosody in Sound Change [JAN-JUN]. *Language and Speech* 35:45-58.
- Beckman, M. E. 1996. The parsing of prosody [FEB-APR]. *Language and Cognitive Processes* 11:17-67.
- Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Paper presented at *Institute of Phonetic Sciences of the University of Amsterdam*.
- Boersma, P., and Weenink, D. 2002. PRAAT. Amsterdam, NL: Institute of Phonetic Sciences. University of Amsterdam, NL.
- Bolt, R.A. 1980. Put-that-there: Voice and gesture at the graphic interface. In *SIGGRAPH-Computer Graphics*.

- Hess, W. 1983. Pitch Determination of Speech Signals. In *Springer Series of Information Sciences*. Berlin: Springer-Verlag.
- Kendon, A. 1980. Gesticulation and speech: Two aspects of the process of the utterance. In *The relation between verbal and non-verbal communication*, ed. M.R. Key, 207-227. Hague: Mouton.
- Kendon, A. 1990. *Conducting Interaction*: Cambridge: Cambridge University Press.
- Kettebekov, S., and Sharma, R. 2000. Understanding gestures in multimodal human computer interaction. *International Journal on Artificial Intelligence Tools* 9:205-224.
- Kettebekov, S., and Sharma, R. 2001. Toward Natural Gesture/Speech Control of a Large Display. In *Engineering for Human Computer Interaction*, eds. M.R. Little and L. Nigay, 133-146. Berlin Heidelberg New York: Springer Verlag.
- Kettebekov, S., Yeasin, M., and Sharma, R. 2002. Prosody based co-analysis for continuous recognition of co-verbal gestures, submitted to ICME'02.
- Kita, S., Gijn, I.V., and Hulst, H.V. 1997. Movement phases in signs and co-speech gestures, and their transcription by human coders. Paper presented at *Intl. Gesture Workshop*.
- Krahnstoever, N., Yeasin, M., and Sharma, R. 2002. Automatic Acquisition and Initialization of Articulated Models. Paper presented at *To appear in Machine Vision and Applications*.
- McNeill, D. 1992. *Hand and Mind*: The University of Chicago Press, Chicago IL.
- Oviatt, S. 1996. Multimodal interfaces for dynamic interactive maps. Paper presented at *Conference on Human Factors in Computing Systems (CHI'96)*.
- Oviatt, S., Angeli, A. De, and Kuhn, K. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. Paper presented at *Conference on Human Factors in Computing Systems (CHI'97)*.
- Pavlovic, V. I., Sharma, R., and Huang, T. S. 1997. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans on Pattern Analysis and Machine Intelligence* 19:677-695.
- Sharma, R., Cai, J., Chakravarthy, S., Poddar, I., and Sethi, Y. 2000. Exploiting Speech/Gesture Co-occurrence for Improving Continuous Gesture Recognition in Weather Narration. Paper presented at *International Conference on Face and Gesture Recognition (FG'2000)*, Grenoble, France.

# A Generic Formal Description Technique for Fusion Mechanisms of Multimodal Interactive Systems

Philippe Palanque, Amélie Schyn

LIHS - IRIT, Université Toulouse III, 118, route de Narbonne, 31062 Toulouse Cedex 4  
Tel : +33 (0)561 556 359 - Fax : +33 561 556 258  
palanque@irit.fr, schyn@irit.fr  
<http://lihs.univ-tlse1.fr/palanque>, <http://lihs.univ-tlse1.fr/schyn>

## Abstract

Representing the behaviour of multimodal interactive systems in a complete, concise and non-ambiguous way is still a challenge for formal description techniques. Indeed, multimodal interactive systems embed specific constraints that are either cumbersome or impossible to capture with classical formal description techniques. This is due to both the multiple facets of a multimodal system (in terms of supported modes) and the strong temporal constraints usually encountered in this kind of systems. This position paper presents a formal description technique dedicated to the engineering of interactive multimodal systems. The formal description technique is then used for the modelling and analysis of two fusion mechanisms. Lastly, benefits and limitations of the approach are discussed.

## 1. Introduction

Despite some efforts for providing toolkits for the construction of multimodal interactive systems (Bederson et al. 2000, Chatty 94), the actual engineering of multimodal interactive systems remains a cumbersome task usually carried out in a rather crafty process. Indeed, while the design (Coutaz & Nigay 1993, Nigay & Vernier 2000) and the evaluation (Coutaz et al. 1996) of multimodal interactive systems have been thoroughly studied, the process of going from a given design to an actual functional system has been the focus of very little research work.

An important aspect of this development process is the reuse of work done from a previous design to another application. Some work on toolkits (Bederson & al. 2000) and architectures (Nigay & Coutaz 95) address this problem at a very low level of abstraction thus making the solution bounded either to modalities or to development platforms.

We believe that the use of an adequate formal description technique can provide support for a more systematic development of multimodal interactive systems. Indeed, formal description techniques allows for describing a system in a complete and non-ambiguous way thus allowing for an easier understanding of problems between the various persons participating in the development process. Besides, formal description techniques allow designers to reason about the models by using analysis techniques. Classical results can be the detection of deadlock or presence or absence of terminating state. A set of properties for multimodal systems have been identified (Martin 1999, Coutaz et al. 1995) but their verification over an existing multimodal system is usually impossible to achieve. For instance it is

impossible to guarantee that two modalities are redundant whatever state the system is in.

The paper is structured as follows. Next section is dedicated to related work dealing with specification of multimodal interactive systems. Section 3 is dedicated to the informal presentation of the ICO (Interactive Cooperative Objects) formalism. Section 4 presents extensions to ICO called the MICO formalism (Multimodal Interactive Cooperative Objects) which is dedicated to the formal description of multimodal interactive systems. This formalism is applied to two fusion mechanisms. The first one integrates voice and gesture while the second one features two handed interaction. Last section (section 5) presents the advantages and limitations of the approach as well as future and ongoing work.

## 2. Related work

Work in the field of multimodal can be sorted in four main categories. Of course the aim of this categorization is not to be exhaustive but to propose an organisation of previous work in this field.

- Understanding multimodal systems
  - ✍ (Coutaz & Nigay 1993) typology of multimodal systems, refined in (Bellik 1995)
  - ✍ (Coutaz et al. 1995), (Martin 1999) presenting properties of multimodal systems
- Software construction of multimodal systems
  - ✍ (Chatty 94, Bederson et al. 2000) propose toolkits for the construction of multimodal systems
  - ✍ (Nigay & Coutaz 95) proposes a generic software architecture for multimodal systems
- Analysis and use of novel modalities

- ✍ (Bolt 1980) presents the first use of voice and gesture as combined modalities.
- ✍ (Buxton & Myers, 1986) introduces two handed interaction.
- ✍ (Bolt & Herranz 92) introduces the use of the two handed interaction in virtual reality applications.
- ✍ (Vo & Wood 1996) presents Jeanie, a multimodal application, to test the use of eyes tracking and lips movements' recognition.
- Multimodal systems description.
  - ✍ (Nigay & Coutaz 95) presents a Multimodal Air Traffic Information System (MATIS) using both voice and direct manipulation interaction.
  - ✍ (Cohen et al. 1997) presents QuickSet a cooperative user interface to military systems using both voice and gesture.
  - ✍ (Bier et al. 1993) presents a drawing systems featuring two handed interaction through a trackball and a mouse.

While formal description techniques have been defined and used for interactive systems since the early work from Parnas (Parnas 69), their extension and use for multimodal systems is still relatively rare. We can quote for instance work from (Duke & Harrison 1997) or (MacColl & Carrington 98) where they present how software engineering techniques such as Z and CSP can be used for the modelling of MATIS the multimodal air traffic information system developed by Nigay (Nigay & Coutaz 1995).

We believe that multimodal interactive systems feature intrinsic characteristics that make formal description techniques used in software engineering not directly suitable for multimodal systems. First, multimodal interactive systems are, by definition, interactive and thus behave in an event-driven way, usually hard to capture and represent in state based descriptions such as Z. Second, the temporal constraints are at the core of these systems which are more often than not real time and highly concurrent. Indeed, users' actions may occur simultaneously on several input devices and the fusion mechanism must process those input in real-time. Formal description techniques with an interleaving semantics (such as CSP, CCS or LOTOS) are not capable of representing such truly concurrent behaviours. Lastly, the use of temporal windows in fusion mechanisms requires, from a formal description technique, the possibility to represent time in a quantitative way by expressing for instance that an event must be received within 100 milliseconds.

Petri nets is one of the few formal description techniques that allows for representing the behaviour of such systems. Indeed, they feature true-concurrency semantics, they are able to deal both with events and states and they provide several ways to represent quantitative time (Bastide & Palanque 1994).

For space reasons we do not present in detail the notation here but next section shows how these characteristics are used while modelling two fusion mechanisms.

### 3. Informal Description of ICOs

The Interactive Cooperative Objects (ICOs) formalism is a formal description technique dedicated to the specification of interactive systems (Bastide et al. 1998). It uses concepts borrowed from the object-oriented approach (dynamic instantiation, classification, encapsulation, inheritance, client/server relationship) to describe the structural or static aspects of systems, and uses high-level Petri nets (Genrich 1991) to describe their dynamic or behavioural aspects.

ICOs are dedicated to the modelling and the implementation of event-driven interfaces, using several communicating objects to model the system, where both behaviour of objects and communication protocol between objects are described by Petri nets. The formalism made up with both the description technique for the communicating objects and the communication protocol is called the Cooperative Objects formalism (CO and its extension to CORBA COCE (Bastide et al. 2000)).

In the ICO formalism, an object is an entity featuring four components: a cooperative object with user services, a presentation part, and two functions (the activation function and the rendering function) that make the link between the cooperative object and the presentation part.

**Cooperative Object (CO):** a cooperative object models the behaviour of an ICO. It states how the object reacts to external stimuli according to its inner state. This behaviour, called the Object Control Structure (ObCS) is described by means of high-level Petri net. A CO offers two kinds of services to its environment. The first one, described with CORBA-IDL (OMG 1998), concerns the services (in the programming language terminology) offered to other objects in the environment. The second one, called user services, provides a description of the elementary actions offered to a user, but for which availability depends on the internal state of the cooperative object (this state is represented by the distribution and the value of the tokens (called marking) in the places of the ObCS).

**Presentation part:** the Presentation of an object states its external appearance. This Presentation is a structured set of widgets organized in a set of windows. Each widget may be a way to interact with the interactive system (user  $\rightarrow$  system interaction) and/or a way to display information from this interactive system (system  $\rightarrow$  user interaction).

**Activation function:** the user  $\rightarrow$  system interaction (inputs) only takes place through widgets. Each user action on a widget may trigger one of the ICO's user services. The relation between user services and widgets is fully stated by

the activation function that associates to each couple (widget, user action) the user service to be triggered.

**Rendering function:** the system  $\rightarrow$  user interaction (outputs) aims at presenting to the user the state changes that occurs in the system. The rendering function maintains the consistency between the internal state of the system and its external appearance by reflecting system states changes.

ICO are used to provide a formal description of the dynamic behaviour of an interactive application. An ICO specification fully describes the potential interactions that users may have with the application. The specification encompasses both the "input" aspects of the interaction (i.e. how user actions impact on the inner state of the application, and which actions are enabled at any given time) and its "output" aspects (i.e. when and how the application displays information relevant to the user).

An ICO specification is fully executable, which gives the possibility to prototype and test an application before it is fully implemented (Navarre et al. 2000). The specification can also be validated using analysis and proof tools developed within the Petri nets community and extended in order to take into account the specificities of the Petri net dialect used in the ICO formal description technique.

#### 4. Fusion Mechanisms Modelling

This section presents how MICO formalism can be used for the modelling of two fusion mechanisms. As explained in previous section this formalism is able to capture all the elements that are embedded in fusion mechanisms.

##### 4.1. Voice and Gesture Interaction

Our first example is Bolt's system (Bolt 80). Bolt was the first to have the idea to use voice and

gesture recognition synergistically for multimodal input. This idea had been implemented in a drawing application, in which user can specify a command orally and give its arguments with either a precise oral description or with a deictic word (this, here, there, ...) and a designation gesture

In this system, five different commands are allowed: create, name, delete, make and move. Each command features a given number of arguments. As long as the command is incomplete, the system waits for the missing argument(s). When a deictic is uttered, user's gesture is taken in account.

As an input for the modelling of this system, we have taken the informal description that can be found in Bolt's papers. Of course, as this application has been presented in natural language and implemented, but not formally described, it is difficult to perfectly understand the functioning of the integration between deictic and gesture. We have supposed that the analysing of a deictic word is at the origin of the triggering of the gesture recognition. Similarly, fusion criteria between command and its potential arguments are not detailed. In the model, this has been represented by the use of typing constraints. Figure 2 and Figure 2 present the formal description of the system according to the assumption presented above. This model describes in a non-ambiguous way the behaviour of the fusion mechanism. In the model rectangles (called transitions) represent actions the system can perform while ellipses (called places) represent state variable of the system. Places can hold tokens and the distribution of tokens in the places represent the current state of the system. The Petri model used in the MICO formalism is called a high-level Petri net model as token can hold values.

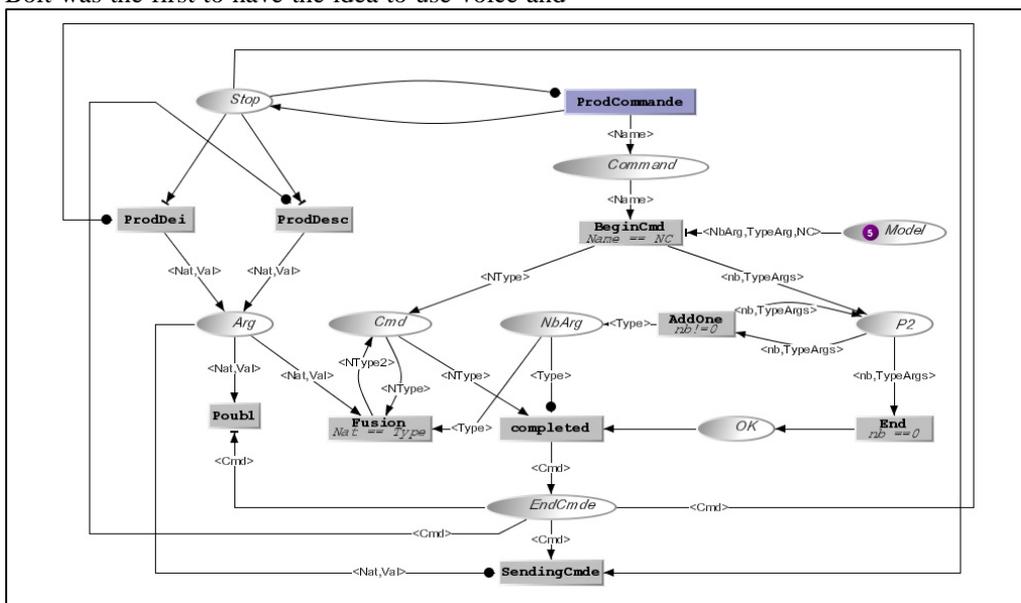


Figure 1. A formal description of the fusion mechanism in Bolt's system (behavioural part)

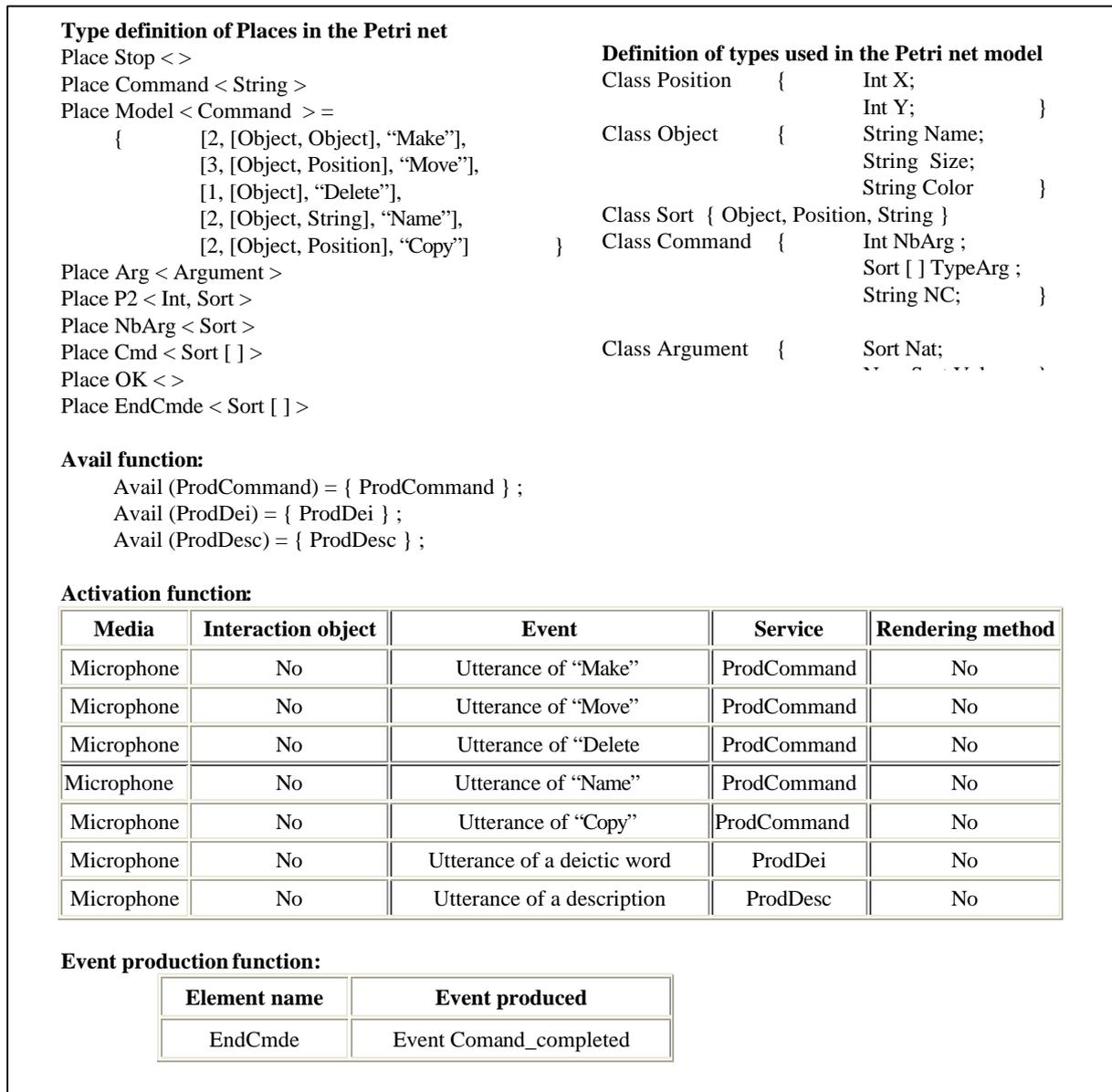


Figure 2. A formal description of the fusion mechanism in Bolt's system(interaction part)

The initial state of the system presented in Figure 2 is thus the presence of 5 tokens in place *Model* and no token in the other places of the net. The values of these tokens are the description of the commands the system can interpret i.e. their number of arguments, the type of these arguments (object, position or name) and their name (create, name, delete, make or move). Places and transitions are related by arcs. A transition can be fired (i.e. an action performed) if and only if each input place of the transition holds at least one token. When a transition is fired, the tokens are removed from the input places and one token is deposited in each output place. The model in Figure 2 features two specific kinds of arcs. Tests arcs model the fact that a token is tested i.e. it is not removed or changed by the firing of the transition but its existence is necessary for the actual firing of the transition. Such an arc is

represented between transition *ProdDesc* and place *Stop* meaning that in order for the system to process a description (transition *ProdDesc*) the system must be in the *Stop* state i.e. place *Stop* holds at least one token. Inhibitor arcs model the zero test in a Petri net. For instance the arc between place *Stop* and transition *ProdCommand* is an inhibitor arc (the end of the arc is a black dot) meaning that this transition can only be fired if there is no token in place *Stop*. Relationship between transitions and events is done by means of dedicated transitions called synchronized transitions. A synchronized transition can only be fired if it is fireable (according to the current marking of the net) and the associated event is triggered (for instance after a corresponding user action on a dedicated input device). In Figure 2, "ProdCommand", "ProdDei" and "ProdDesc" are synchronized with user events (utterance of a

command, deictic word or univocal description). As voice modality is dominant in this system, gestures are taken in account only if a deictic word is uttered. So there is no gesture event. Formal analysis of the Petri net of Figure 2 guarantees that whatever state the system is in, there is always at least one transition in the model that is fireable which means that the model is live. For space reasons we don't explain in details other properties that can be proven on the model and how the formal analysis is performed.

#### 4.2. Two handed interaction

The models in Figure 4 and Figure 3 describe the interaction level events policy for two handed interaction. There is no assumption about the type of the devices except that they are graphical and produce the same set of low level events. It tells

when and how those events are produced according to the user's actions on the devices. In this Petri net, the policy works like a transducer: each time a physical event is accepted, the Petri net fires a transition and creates higher level events.

For example, in the policy represented in Figure 4, the physical mouse-move (m) is transformed into a higher level mousemove (M), e.g. the transition between places *One\_Click* and *Idle* reacts to the event m by generating another event M plus an event click (C). However; all physical (low level) events are not immediately translated into interaction level events. For example, each event d (down) received while the system is in the initial state, is consumed without any production.

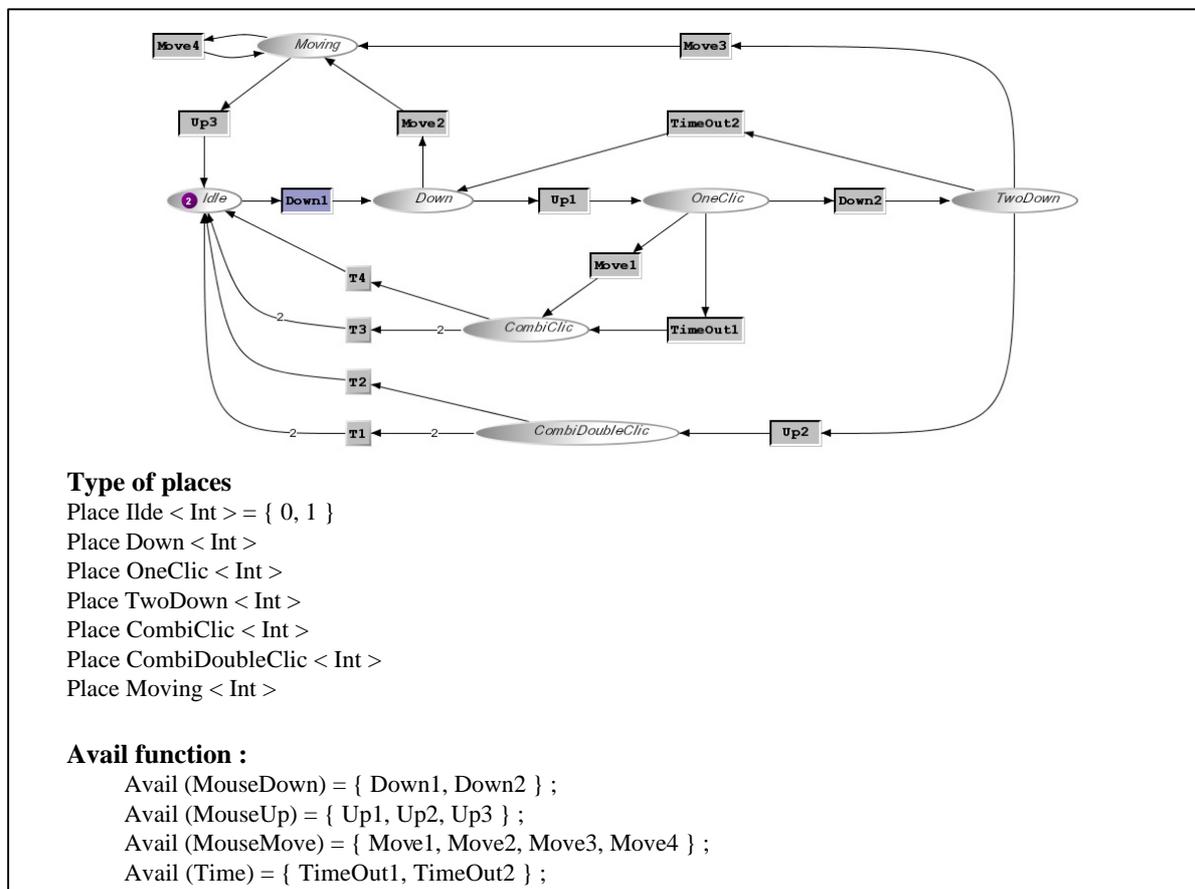


Figure 3. Formal model of a fusion mechanism for two handed interaction (behavioural part)

| <b>Activation function:</b> |                           |              |                |                         |
|-----------------------------|---------------------------|--------------|----------------|-------------------------|
| <b>Media</b>                | <b>Interaction object</b> | <b>Event</b> | <b>Service</b> | <b>Rendering method</b> |
| Mouse button                | All                       | Down         | MouseDown      | Internal                |
| Mouse button                | All                       | Up           | MouseUp        | Internal                |
| Mouse                       | All                       | Move         | MouseMove      | Move_Mouse              |
|                             |                           | TimeOut      | Time           | Internal                |

| <b>Event production function:</b> |                       |
|-----------------------------------|-----------------------|
| <b>Element name</b>               | <b>Event produced</b> |
| T1                                | Event CDC             |
| T3                                | Event CC              |
| Up2                               | Event DC              |
| Move1                             | Events C and M        |
| TimeOut1                          | Event C               |
| TimeOut2                          | Event C               |
| Move2                             | Event B               |
| Move4                             | Event D               |
| Up3                               | Event E               |
| Move3                             | Events C and B        |

Figure 4. Formal model of a fusion mechanism for two handed interaction(interaction part)

We have already presented this example in (Accot et al. 1996) while presenting how transducer can be modelled using Petri nets. The model in Figure 4 presents a way to integrate information provided by two different mice. The system can react to the following set of events produced through users' actions on the physical devices: Button Down (d), Button Up (u), Mouse Move (m) and Time Out (t). If a precise sequence of event is performed on the mice, multimodal events are produced by the model. Such events are: "CombiDoubleClick" and "CombiClick" corresponding to the arrival, in the model, of tokens in place CombiClick and CombiDoubleClick.

## 5. Conclusion and future work

The MICO formalism is an extension of ICO formalism that is formalism dedicated to the design specification, verification and prototyping of interactive systems. Its formal underpinnings make it especially suitable for safety critical interactive systems. ICO formalism has been applied to various kinds of systems including business, Air Traffic Management and command and control applications. The continuously increasing complexity of the information manipulated by such systems calls for new interaction techniques increasing the bandwidth between the system and the user. Multimodal interaction techniques are considered as a promising way for tackling this problem. However, the lack of engineering techniques and processes for such systems makes

them hard to design and to build and thus jeopardises their actual exploitation in the area of safety critical application.

This position paper has presented a formal description technique that can be used for the modelling and the analysis of multimodal interactive systems. This work is part of a new project on the evaluation and use of multimodal interaction techniques in the field of command and control real time systems.

## 6. References

- (Accot et al. 1996) Accot, J. Chatty, S. and Palanque, P. (1996). A Formal Description of Low Level Interaction and its Application to Multimodal Interactive Systems. In *3rd EUROGRAPHICS workshop on "design, specification and verification of Interactive systems"* (pp. 92-104). Springer Verlag,.
- (Bastide & Palanque 1994) Bastide, Rémi and Palanque, Philippe. Petri Net based design of user-driven interfaces using the Interactive Cooperative Objects formalism. in: Paternò, Fabio (Ed.). *Interactive systems: design, specification, and verification, DSV-IS'94*. Springer-Verlag; 1994; pp. 383-400.
- (Bastide et al. 1998) Rémi Bastide, Philippe Palanque, Duc-Hoa Le, Jaime Muñoz. Integrating Rendering Specifications into a Formalism for the Design of Interactive Systems. 5th Eurographics workshop on "design, specification and verification of Interactive

- systems", DSV-IS'98, U.K., 3-5 June 1998, Springer Verlag.
- (Bastide et al. 2000) Bastide, Rémi; Sy, Ousmane; Palanque, Philippe, and Navarre, David Formal specification of CORBA services: experience and lessons learned. ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA'2000); Minneapolis, USA. ACM Press; 2000: 105-117. ACM SIGPLAN Notices. v. 35 (10).
- (Bederson et al. 2000) Bederson, B. B. Meyer, J. and Good L. (2000). Jazz An Extensible Zoomable User Interface Graphics Toolkit in Java. In *UIST2000, ACM Symposium on User Interface Software and Technology, CHI Letters* (2(2), p. 171-180).
- (Bellik 1995) Bellik, Y. (1995). Interfaces Multimodales: Concepts, Modèles et Architectures. Thèse de Doctorat en Informatique. Université de Paris-XI.
- (Bier et al. 1993) Bier, E. Stone, M. Pier, K. Buxton, W. and DeRose, T. (1993). Toolglass and Magic Lenses: the See-Through Interface. In *Proceedings of ACM SIGGRAPH93* (pp. 73-80). Anaheim, ACM Press.
- (Bolt & Herranz 1992) Bolt, R and Herranz, E. (1992). Two-Handed Gesture in Multi-Modal Natural Dialog. In *Proceedings of the fifth annual ACM symposium on User interface software and technology* (p 7-14). Monterey, California. ACM Press.
- (Bolt 1980) Bolt, R. (1980). Put That There: Voice and Gesture at the Graphics Interface. In *SIGGRAPH'80* (Vol 14, p262-270).
- (Buxton & Myers 1986) Buxton, W. and Myers, B. (1986). A Study in Two-Handed Input. In *Proceeding of the ACM CHI* (p 321-326) Addison-Wesley.
- (Chatty 1994) Chatty, S. (1994). Extending a Graphical Toolkit for Two-Handed Interaction. In *Proceedings of the ACM symposium on User Interface Software and Technology*, (p195-204) Marina del Rey, California. ACM Press.
- (Cohen et al. 1997) Cohen, P. Johnston, M. McGee, D. Oviatt, S. Pittman, J. Smith, I. Chen, L. and Clow, J. (1997). QuickSet : Multimodal Interaction for Distributed Applications. In *Proceedings of the fifth ACM international conference on Multimedia* (p 31-40) Seattle, Washington. ACM Press.
- (Coutaz & Nigay 1993) Coutaz, J. and Nigay L. (1993). A Design Space for Multimodal Systems Concurrent Processing and Data Fusion. In *Human Factors in Computing Systems, INTERCHI'93 Conference proceedings* (p 172-178) Amsterdam, The Netherlands.
- (Coutaz et al. 1995) Coutaz, J. Nigay, L. Salber, D. Blandford, A. May, J. and Young, R. (1995). Four Easy Pieces for Assessing the Usability of Multimodal in Interaction the CARE Properties. In *Human Computer Interaction, Interact' 95* (p115-120). Lillehammer, Norway.
- (Coutaz et al. 1996) Coutaz, J. Salber, D. Carraux, E. and Portolan, N. (1996). Neimo, a Multiworkstation Usability Lab for Observing and Analysing Multimodal Interaction. In *Human Factors In Computing Systems CHI'96 Conference Companion* (p 402-403) Vancouver, British Columbia, Canada. ACM Press.
- (Duke & Harrison 1997) Duke, D. and Harrison, M. D. (1997). Mapping User Requirements to Implementations. In *Software Engineering Journal* (Vol 10(1), p 54-75).
- (Genrich 1991) Genrich, H.J. *Predicate/Transition Nets*, in K. Jensen and G. Rozenberg (Eds.), High-Level Petri Nets: Theory and Application. Springer Verlag, Berlin, pp. 3-43.
- (MacColl & Carrington 1998) MacColl and Carrington, D. (1998). Testing MATIS: a Case Study On Specification-Based Testing of Interactive Systems. In *FAHCI98* (p57-69). ISBN 0-86339-7948.
- (Martin 1999) Martin, J.C. (1999). TYCOON six Primitive Types of Cooperation for Observing, Evaluating and Specifying Cooperations. In *Working notes of the AAAI Fall 1999 Symposium on Psychological Models of Communication in Collaborative Systems*. Sea Crest Conference Center on Cape Cod, North Falmouth, Massachusetts. <http://www.limsi.fr/Individu/martin/aaai99/html/martin-final-v7.html>
- (Nigay & Coutaz 1995) Nigay, L. and Coutaz, J. (1995). A Generic Platform for Addressing the Multimodal Challenge. In *Conference proceedings on Human factors in computing systems* (p 98-105). Denver, Colorado.
- (Nigay & Vernier 2000) Nigay, L. and Vernier, F. A Framework for the Combination and Characterization of Output Modalities. In *Proceeding of the DSV-IS '2000 Conference* (p 35-50). P. Palanque, F. Paterno (Eds) Springer.
- (OMG 1998) OMG. The Common Object Request Broker: Architecture and Specification. CORBA IIOP 2.2 /98-02-01, Framingham, MA (1998).
- (Parnas 69) Parnas, D. L. (1969). On the Use of Transition Diagram in the Design of a User Interface for Interactive Computer System. In *Proceedings of the 24th ACM Conference*, (p. 379-385).
- (Vo & Wood 1996) Vo, M.T. and Wood, C. (1996). Building an Application Framework for Speech and Pen Input Integration in Multimodal Learning Interface. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) IEEE*, (Vol 6, p 3545-3548)

# Eye-Bed

Ted Selker, Winslow Burleson, Jessica Scott, Mike Li

MIT Media Lab  
20 Ames St. Cambridge, MA 02139  
{selker, win}@media.mit.edu; {jbscott, mli}@mit.edu

## Abstract

The Eye-bed prototype introduces new ergonomic language scenarios. This paper focuses on developing a demonstration eye gesture language for intelligent user interface and computer control. Controls are integrated with a user model containing a history of previous interactions with the interface. Context recognition enables the Eye-Bed environment to continually adapt to suit the user's inferred and demonstrated preferences. Staring, gazing, glancing, blinking, etc... are part of person-to-person visual communication. By creating a computer interface language composed of exaggerated eye gestures, we create a multi-modal interface that is easy to learn, use, and remember.

## 1. Introduction

Computer human interfaces have long been applied to everyday situations. These interfaces are often trapped in a user-directed model, relying on the user to know and use a language to directly specify what she wants from the computer. More recently computers are finding their way into everyday things. These days our appliances seem to need to have their computers booted before they work. Cars, phones, record players even house locks come to a grinding halt when their computers don't work. We have been programming our thermostats, watches, and videotape recorders for so long that it seems reasonable to spend hours learning to use an MP3 music player that fits in your hand. They are about to get simpler.

At the very least hospital beds have some control for comfort, height, angle, and temperature. The bed area has another control to call an assistant. Usually each bed has a control for a television. How should we position ourselves optimally to watch a movie from bed? Automation of communication and media in bed could be very useful. The Media-Bed scenario adds new rich command and control integration opportunities to the computer human interface (Shelley, R., 2001).

If one integrated environmental controls, educational materials, and entertainment media into a bed interface how would a user communicate with it? Imagine imagery projected on the ceiling over a bed. (Figure 2.) Consider functions presented spatially as an integrated ecological user interface. This would require a person to select things on the spatial interface. Spatial selection has many dexterity problems. Historically the control theory issues and other obstacles of the components of Fitts and Steering laws (Hinckley, K. et al, 1994) have paled in comparison to the difficulty of learning command languages. We consider new languages that are trivial to learn and require low cognitive overhead to use. Through mimicry and extension of the social communication people employ nonverbally, we explore the realm of reduced consciousness communication.

Graphical interfaces are wonderful in that they allow a user to recognise something they might not have been able to remember otherwise, like a place in a file hierarchy. If they want an item they simply point at it to select it. Still, graphical interfaces have long been

cumbersome and frustrating. People control 3D interfaces with analogue devices that change the rate and angle of motion as though the ultimate way to interact with a computer would be some sort of hovering gravity-independent helicopter. It is hard to learn to fly a helicopter. Many novice users of 3D interfaces have the constant feeling of listing dangerously as they walk into walls and can't stop the scenes from rotating.



Figure 1: Multimedia Bed with Ceiling Projection

A provocative area of user input design has been eye control. It has seemed like one of the ultimate interface

approaches since the 1960s when the first rudimentary mechanical Perkinji trackers were demonstrated. Indeed psychologists and marketing people have used eye position to understand people's interests. (Yarbus, L. 1967). Unfortunately eye tracking utility has been stymied: the head moves; the eyes don't want to look at one thing; the tracking devices work in lab settings better than in an office; etc.... For the past 40 years, people have been improving eye-tracking technology using Electro-oculogram eye trackers, contact lenses, tracking infrared cameras and dual cameras to get to lightweight camera based systems. Dual camera systems like the Autostereoscopic user-computer interface (Pastoor, S., Skerjanc, R., 1997) and multiple multiplexed structured light source camera systems like the Blue Eyes™ system at IBM have become excellent tools (Flickner, 2001). Unfortunately, these "better eye tracking systems" have made it even more obvious that the eye is not simply looking for interesting things that it wants to effect.

The eye is not a cursor control device. The eye notices movement in the periphery and has to attend to it vigilantly searching for danger. The eye is a guard dog; it has a job to do. Using eyes and trackers to move a cursor precisely is like using a security officer in a bank to show people the bathroom. The officer could do it but not as well as a concierge; at the same time the officer would risk being remiss in the primary security duties.

## 2. Eye motion as language

Large areas of the brain are devoted to interpreting visual input and controlling the eye (Carpenter, M., 1976). The sensitivity of the eye itself makes it a strange choice for a pointing device. The eye, after all, seeks to understand anything in its view. The area centralis is some 3 degrees wide; anything in this visual area is well known to the mind. One of the most difficult issues with eye-tracking scenarios is that the eye-tracking computer demands "eye contact". This is the very thing people are most used to devoting to scanning for safety, acknowledging other people, and to expressing their feelings non-verbally (Clark, H., Wilkes-Gibbs, D., 1986).

Interest Tracker (Maglio et al, 2000) lets people use generalized directional gaze to select information content by demonstrated interest, much as a person does when meeting a new acquaintance. This stands in contrast to the standard eye-tracking interface in which a user is asked to stare at a specific thing until it is selected: the physical difficulty of doing so; the social inappropriateness; and the uncomfortable feeling of the interface is significant. In contrast if the user is asked to look at the general area of an item to be selected these interface obstacles are diminished.

More recent work, demonstrated "Magic Pointing" (Shumin, Z. 1999), an approach that uses eye gaze to make a non-linear jump or "warp" a cursor to where the eye is looking on a screen. Subsequent GUI control is done through the standard cursor control device. It is quite easy to use eye tracking to identify areas of interest. Of value to interface design is the fact that the eye is a course output device and a fine input device. The most important notion however is that Interest Tracker and Magic Pointing take advantage of the fact that the eye

wants to look in the area of interest. The syntax that the action of looking at a work area changes the spatial position of the cursor is a powerful one. Using a dwell time of just 0.3 seconds was more than adequate to allow a user to distinguish things they wanted to select. It was also found to be much faster than a mouse can select the same area (Maglio et al., 2000) Interest Tracker, introduced above, is a system that shows another simple and productive use of gaze interpretation. It augments a person's natural gaze at an area of interest with additional information or content of a similar nature.

Invision (Li, M.; Selker, T., 2001), takes this one step further, based on evidence that shows that the paths that people's eyes follow demonstrate what they are thinking (Yarbus, 1967). When people rapidly transit from one place to another they are more likely to be making a selection of a familiar item. When people's eyes move slowly around the field-of-view they are taking in information, and making decisions, but not selections.

The pattern based Invision interface made two contributions to eye tracking. It demonstrated that eye tracking accuracy could in many cases be improved by interpreting eye movement as the endpoints of the trajectory (i.e. knowing where the eye had moved from and too helps to understand user intent and focus). In the second, and more interesting case, the relationships between objects that the eye gazes at and the order that they are gazed at become the language that drives the computer. The system showed a set of objects representing the various sponsors of the Media Lab. As a person traversed them with their eyes, it used the path to notice their interests. As a person's eye went back and forth, between two things, the objects they were looking at moved closer to one another. In this way, as a user shows interest in a group of items the interface literally brings these items together. This has been explored as an interface for a kitchen as well (opening the refrigerator, oven, cabinets and dishwasher). These pieces of research all focus our attention on the information that comes out of an eye.

### 2.1.1. Gaze vs. Stare Detection:

The Eye-aRe Project took this further. Eye-aRe is a simple system that consists of a glasses mounted infrared LED and photodiode that detect reflected infrared light from the eye's cornea and sclera. (Selker et al., 2001) A small PIC can detect when a person is staring and when their eyes stay relatively fixed. It is not hard to separate simple eye gaze intent. This approach can separate out intended versus unintended selection events. Even without a camera, Eye-aRe has successfully been used to send business card information when a user stares at (or is engaged in conversation with) another person, to bring up information about a display when a person looks at it, and detect closed and opened eyes and individual blink signatures.

If the actions used to interact with a computer mimic the normal use of human eye gesture language, this synergy could assist user's learning and memory. Can such an eye gesture based language be the basis of an ecological interface? Can such a natural control language

be integrated without being difficult to learn or generating confusion? Can reasoning, learning, and representation of intelligence be employed to give users more control?

Complex social dynamics are traceable to eye motion (Clark, 1986). These can be used to enhance human computer communication. Eye motion demonstrates a social gesture language. These are significantly easier to record than eye position. With this thesis we will describe the ways that eye gestures and task modeling have been experimented with in the Eye-Bed to reduce reliance on direct manipulation in the interface.



Figure 2: Ecological display projected on Ceiling

### 3. Media-Bed & Eye-Bed

The bed is a place where the average person will spend approximately one third of their life. Once made of plant fiber and then synthetic materials, we have now made the bed digital. The Media-Bed and Eye-Bed (Figure 1.) are a response to the challenges of integrating environmental, educational, and entertainment controls in a universal interface. The Media-Bed and Eye-Bed could simplify the controls of a hospital bed while adding new features that integrate these domains [good morning america].

The Media-Bed and Eye-Bed are part of a growing body of language based interface development. (Selker, T.; Burleson, W., 2000) The thesis is that replacing explicit spatial selection with a language-based interaction may provide interfaces that are easy to learn, use, and remember. One novel control approach in this direction has been the use of eye tracking. The social language of the eye (i.e. “wink, wink.... Know what I mean” as said again and again in ... Monty Python’s Holy Grail) can be used as a natural easily understood language. In the Bed projects we overlay and map expected characterized ocular responses such as stare, gaze, wink, etc... with a language to communicate interface intentions between the user and the computer.

The Media-Bed and Eye-Bed are a computer systems that recognizes and remembers what a person is doing in bed to provide useful information and environmental modifications. They “listens” to many information channels to enhance the semantics of a language. The Eye-Bed extends language recognition of the Media-Bed

to include eye-tracking semantics: blinking, winking, staring, and gazing. Both create a user model which includes time stamps, interface states, knowledge of the position and sound of the user, in addition to the traditional direct user input channel.

The Media-Bed and Eye-Bed are a place for us to experiment with new scenarios for using a computer in our live. They are also a place to experiment with new multi-modal input devices. For example, eye tracking in a bed has advantages. The person’s head is supported and can be stabilized. This naturally reduces the difficulty of finding and tracking the eye position. The bed consists of an integrated multimedia personal computer and video projector. It runs a Macromedia Director movie projected onto the ceiling above a standard bed. This projection creates a virtual world that provides the user with a space for interaction and reactive input.

#### 3.1. Prototype Scenario

A person is lying in bed. Many simple activities can be computer-facilitated making lying in bed more pleasurable and productive. A scene appears, projected on the ceiling above the user’s head (Figure 2). It is a scene of rolling hills dotted with icons: an e-mail kiosk, a TV satellite dish, a juke box, a person reading in a lawn chair, a newspaper stand, the moon and stars, and the sun. Each of these icons can move the user into another part of the world depending on his needs and wishes at the time. We have experimented with different renditions of physical world imagery or so-called “ecological interfaces”. Ecological interfaces have been shown to improve speed and accuracy of selection over two-dimensional interfaces when users are familiar with them (Ark et al. 1998).

Pointing and selecting it, the kiosk enlarges to fill the screen, bringing the user into another space. A smaller rendition of the rolling hills at the top of the screen points to the original main screen where the user came from. The user can similarly watch TV, read the newspaper or read an online book while lying on their back in bed. The display is projected upward to cover the ceiling above the bed. When reading something or watching TV or a movie, the user no longer has to prop themselves up with their arms or find a comfortable position to sit in. If the user has back or neck problems, this is especially important.

Once the user has finished reading e-mail selecting the hills at the top of the screen returns them to the initial selection screen. It’s time to go to sleep, so the user moves to the moon and stars, a soothing song begins to play and a sunset that gradually darkens to reveal the night sky is projected. The bed can subtly and playfully encourage or persuade a person to go to sleep at an hour that they should by shifting to this mode as well. (Fogg, B., 1998) Selecting the moon presents the outlines of constellations. As the user explores the night sky, the names of the constellations and planets appear. Selecting a planet brings up its path and other information. This is an example of how the system can function in an educational and informational role as well. As the user falls a sleep (their eyes close and they move less), the bed recognizes the hour, and sets sunrise wake up music to accommodate the user’s sleep patterns. The bed has learned how long it’s occupant likes to sleep by

monitoring the use patterns of the alarm clock. Since the bed has access to the user's calendar, it knows the user will not miss any appointments by waking up at eleven o'clock. In the morning, the sun rises on the ceiling, accompanied by morning music. The room is gradually lit up by the sunlight, and the day's schedule is presented for review along with e-mail and newspaper customized to the user's interests and preferences. In this scenarios the user is able to enjoy the activities that they normally enjoy with the media selection assistance from the computer.

Selection of functions on the Media-Bed selection of items on the ceiling was originally accomplished with a Polhemous 6-degree of freedom system in a ball. The position of the ball controlled a ball-shaped cursor on the landscape imagery of the ceiling interface. The ball used a bed based coordinate system to control a cursor on the screen. It was tiresome to hold it in exactly the right position on the bed to activate the functions . The Gyromouse™ did not require the person's hand to go to a specific place in the air or on the bed to use spatial control. The TrackPoint™ in a custom built handle and a TrackPoint keyboard were much easier to use allowing hands to rest on the device. The next step in evaluating the Media Bed interface was to add an eye tracker. The newer Eye-Bed system uses the eye-tracker, positioned in a lamp mounted to the headboard, to control the system.

Through the construction of user model profiles, the Media-Bed and Eye-Bed can learn to suit the user's wishes by understanding what they are interested in seeing, doing, and listening to. The boom box and media presenting applications in the bed do this explicitly. A hiking boot icon when selected kicks the juke box or media player indicating to the user that the system will try to change what media to present. The system changes the current media and updates the model of what to try in the future. It uses artificial intelligence to record actions and reactions of the user to build a model of what kind of information and media will be useful in which situations.

The Eye-Bed version augments the positional syntax of a cursor on a GUI with a language of few simple eye gestures to make an even more interesting interaction scenario. This is done through a paradigm of *relaxed eye tracking*. The Eye-Bed version develops a contextual knowledge of the situation. It uses the "eyes shut" condition to know when a person is asleep or not wanting to see imagery anymore. "Eyes open" to tell the bed that a person need not hear the loud version of the alarm clock, "excessive blinking" or "nervous eyes" to change the station of the radio or TV, and "gazing" into a sparse ecological interface to select interface icons. The eye position itself and the way that a person is looking at something can determine what should be done. If the eye isn't wandering and there is only one nearby object of interest the selection is obvious. Using this multi-modal and contextually aware approach we have enhanced the user interface in the Media bed.

### 3.1.1. Nervous Eyes Want Change

Work with Eye-aRe and the work of many other researchers have shown that it is easy to recognize rapid blinking as a sign of dissatisfaction. In the Eye-Bed we integrate rapid blinking as the syntactic way to say you are

not satisfied with the current interactions. For example, we used rapid blinking to change the channel on the radio and video, in a similar manner to the boot kicking the player. Since this action is similar to the natural way of communicating dissatisfaction, people are able to remember the action and accomplish it with ease.

### 3.1.2. Open Eyes

It is extremely easy to know when an eye is open or closed. Eyes open presumes the person is not asleep and is thus the syntax for telling the bed to activate wake-up imagery of a sunrise and turning off the loud alarm if the time is morning or if the user generally wakes up at that time of day. Likewise if a person is not in bed the wake up alarm is not needed. An eye projected on the ceiling shows the eye open and labels the status "open". This projected eye is part of the feedback to the user that the eye tracking is on and working.

### 3.1.3. Eyes Closed

Missing pupils is the syntax for putting the system into a sleep mode. Of course, a person need not watch TV or other things when they are asleep so it can fadeout these media. The Eye-Bed system puts up a black screen with "ZZZZZZ..." written across it when a person closes their eyes for several seconds.

### 3.1.4. Stare

Attention is a fundamental communication act. When a person looks at something intently we call it staring. In the Eye-Bed we use dwell time to activate a spatial icon. Eye-aRe demonstrates that staring at a toy dog is an obvious way to make it respond with a bark; staring at a TV is an obvious way to demonstrate interest in the TV show. Therefore staring in the Eye-Bed is used to select and activate media.

### 3.1.5. Gaze

When a person looks around we could say they are gazing. In the Eye-Bed the eye moving around without staying anywhere is interpreted as lack of focus on the bed interface. The system shows the interpretation on the ceiling display eye indicator.

The eye gesture syntax described in this section is small. The simple language of eye states has been enough to drive the entire Eye-Bed demonstration.

## 3.2. Discussion

Typical spatial interfaces use a spatial inclusion syntax. (Selker, T.; Appel, A, 1991). The control moves an indicator or cursor to within the boundaries of a spatial object or icon to associate syntax to it. The eye gesture language is an augmented visual language in which some eye gestures have global consequence while others act as parameters of a selection device just as mouse buttons on a mouse are parameters to the graphical object that the cursor associates it with. The Eye-Bed eye gesture language has made it possible for people to control the entire Media-Bed interface using only their eye gestures.

In using a gesture-based interface it usually becomes difficult to teach and use the gestures. This system's use

of natural eye gestures, which people do anyway, makes using the bed almost as natural as a social interaction. One goal of creating “natural interfaces” is to create interfaces that use the actions that people are familiar with and relate them to actions the system might expect of users. This can be achieved by copying the actions of people. Studying perceptual and physiological actions and capabilities of people is important as well. It has been shown that in many situations people treat computers as they do people (Reeves, B.; Nass, C., 1986). This paper and these uses of eye input demonstrate how the higher order behavioural and social psychological areas can be used as a motivating approach for interface design. By carefully studying these fields exciting taxonomies of natural behaviour can be found. Once found these can become a basis for more natural, social, and gesture-based interaction languages with the computer. Our goal is developing interaction languages that are amalgamations of typical human actions with appropriate computer augmentation to assistance people in what they want to do.

### **3.3. Status**

#### **3.3.1. Media Bed**

The Media-Bed is a Macromedia Director program running on a computer. The Media-Bed with physical inputs has been demonstrated to hundreds of people at the MIT Media Lab; the opening of Media Lab Europe in Dublin, Ireland; and at the AAAI Fall 2000 workshop in Falmouth, MA. We are surprised at how relaxing it is to lie down to demonstrate the night time and wake up scenarios. Within days of it working people were approaching us to form marketing alliances. We have used the media Bed and its display as a place to work and find that it is quite relaxing.

#### **3.3.2. User Model**

All of the selection scenarios are enhanced by the creation of the user model. The simplest user model is that a person whose eyes are closed need not be shown imagery. Currently we consider a person whose eyes are closed to be asleep.

The user models in the radio and TV are the most sophisticated. These models notice what time of day it is, what has been playing and how long a user listens or watches it as a basis for appreciation. If a user likes the music then similar music continues to play. Of course we have found that some people don't like to hear the same music over and over again. Refining the heuristics for this is a current goal. The eye tracking approach has allowed us to simulate nervousness or detect actual nervousness as the way to tell the media generator that it should attempt to find other media to play. If a person is not paying any attention to anything near the media player and has not recently turned it on, these analyses of nervousness most likely are not about the media

#### **3.3.3. Eye-Bed**

An early version of the Eye-Bed was demonstrated on Good Morning April 10 2001 (Shelley, R. 2001). The Eye-Bed is the Media-Bed with another computer running

the eye gesture recognition software. Mike Li wrote a Java version of the Eye-Bed software. It was replaced with a C version written by Jessica Scott that requires much less of the Ethernet communication for its interpretations. The New version has a much better ability to interpret eye gestures. Further, the new version includes the eye indicator on the ceiling bed display.

The Eye-Bed eye gesture based interface has been demonstrated dozens of times at the MIT Media Lab. The ability to control it with less than a minute of instruction amazes everyone. The impressive thing about Eye-Bed is that people enjoy using it and don't need much instruction. The system is so easy to use that we often have visitors demonstrate the eye-gesture based interface to one another. The real value of this interface is the ease with which we can recognize the gestures of eyes closed, open, gaze, stare, blink, and nervous blinking.

The current system has limitations. Text entry has not been satisfactorily resolved. There are good and bad times to use the system. So far the system is designed for a single user and does quite well at integrating the many controls of the previously discussed hospital bed. However the system does not make any accommodation for the social or sexual activities that take place in bed in fact at this point many users think that the current features are too intrusive. They are appalled at the thought of email intruding into their bedroom and literally “hanging over their heads”.

### **3.4. Future Work**

The interface is effective enough for us to sleep with it on and beneficial enough for us to enjoy it when we are awake. The goal of demonstrating the limits of time and fidelity of eye gesture are central to our future work. The integration and evaluation of new eye gestures and other physiologically natural gestures is central to the context aware stance of the research group that this work takes place in. Understanding what social cues are for and how to make them reliable within a graphical interface system continues to be exciting. We will extend the language that we have developed to include other forms of implicit communications such as facial gestures. The question as to whether a serious formal theory will aid in this endeavour stands before us.

Discussions in bed, on the phone, or in person will be augmented by pervasive access to information. The nature of this information will also rely upon user models. For example, a four-year old who wants to know what bears eat, is looking for a different answer than what a college biology major with the same question is looking for. We will continue to explore the integration of health monitoring and feedback systems. Sound sensing and acoustical feedback will be used to monitor sleep apnoea and snoring. The Media-Bed and Eye-Bed has moved into educational areas, starting with astronomy. We will soon move on to other contextually appropriate topics. Especially interesting is the context of looking up such as in auto mechanics, marine biology, meteorology, ornithology, and rainforest canopy sciences. This work will also be extended into the realms of fun, play, and creativity by implementing games and motivational activities.

## 4. Conclusion

The appropriate use of interface techniques should be the focus of the Computer Human Interface field. Unfortunately as industry develops new interface techniques and scenarios designers bring untested ideas into the market. In this paper we attempt to show that a well-understood language of a few eye gestures can simplify the use of the eyes as a control for user interfaces. We further use an ecological interface to simplify teaching control of the user interface. In doing so we create a system that is natural and ease for people to learn, use, and remember. The goal of developing improved user interactions will continue to require us to invent new scenarios and test where and how they can be applied.

## 5. Acknowledgements

We thank the MIT Media Lab for supporting this work; Jesse Pavel for the Polhemus to Director Interface; Kim May and Ian May for the hand held track point.

## 6. References

- Ark, W.; D. Christopher Dryer, Ted Selker, Shumin Zhai. 1998 Representation Matters: The Effect of 3D Objects and a Spatial Metaphor in a Graphical User Interface. People and Computers XIII, Proceedings of HCI'98, (Eds) H. Johnson, N. Lawrence, C. Roast. Springer. 209–219.
- Carpenter, M., 1976. Human Neuroanatomy. Waverly Press. Baltimore, MD.
- Clark, H. H., & Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition*, 22:1-39. Reprinted in P. R. Cohen, J. L. Morgan, & M. E. Pollack (Eds.), 1990. *Intentions in Communication*. Cambridge: MIT Press.
- Flickner, M. 2001. Blue eyes: Suitor [WWW Document]. URL <http://www.almaden.ibm.com/cs/blueeyes/suitor.html> (visited 2001, February 2).
- Fogg, B.J. 1998. Persuasive Computers: Perspectives and Research Directions. Conference on Human Factors in Computing Systems: CHI 1998 Conference Proceedings. 225-233
- Lieberman, H.; Selker, T. 2000. Out Of Context: Computer Systems That Adapt To, and Learn From, Context. IBM Systems Journal 39, Nos. 3&4: 617-632.
- Hinckley, K., Pausch, R., Goble, J. C., Kassell, N. F. April 1994. A Survey of Design Issues in Spatial Input, Proc. ACM UIST'94 Symposium on User Interface Software & Technology. 213-222.
- Li, M.; Selker, T. 9-2001. Eye Pattern Analysis Ocular Computer Input, IVA.
- Maglio, P.P., Barrett, R. Campbell, C.S., and Selker, T. SUITOR: An attentive information system. In *Proceedings of IUI2000* (New Orleans, LA, Jan. 2000), ACM Press.
- Pastoor, S.; Skerjanc, R. 1997. Autostereoscopic user-computer interface with visually controlled interactions. Digest of Technical Papers, SID'97 International Symposium. 277-280.
- Reeves, B.; Nass, C. 1996. The Media Equation. How People Treat Computers, Television, and New Media Like Real People and Places. Stanford, CA: CSLI Publications.
- Selker, T.; Burleson, W. 6-2000. Context-Aware Design and Interaction in Computer Systems. IBM Systems Journal 39, Nos. 3&4.p. 880-891
- Selker, T.; Lockerd, A.; Martinez, J.; Burleson, W. 2001. Eye-aRe: A Glasses-Mounted Eye Motion Detection Interface. Conference Proceedings Conference on Human Factors in Computing Systems, CHI2001. Extended Abstracts. 179-180.
- Selker, T.; Appel, A, 1991. Graphics as Visual Language.. Handbook of Statistics, Vol. 9, Elsevier Science Publishers BV.
- Shelley, R. April 10, 2001. Multimedia Bed. ABC News "Good Morning America."
- Yarbus, A. L. (1967). Eye movements during perception of complex objects, in L. A. Riggs, ed., 'Eye Movements and Vision'. New York: Plenum Press, chapter 7, 171-196.
- Zhai, S., Morimoto, C. & Ihde, S. (1999). Manual and gaze input cascaded (MAGIC) pointing. Proc. ACM CHI '99 Human Factors in Computing Systems Conference, Addison-Wesley/ACM Press, 1999. 15-20.

# PROMISE - A Procedure for Multimodal Interactive System Evaluation

Nicole Beringer<sup>1</sup>, Ute Kartal<sup>1</sup>, Katerina Louka<sup>1</sup>, Florian Schiel<sup>2</sup>, Uli Türk<sup>1</sup>

<sup>1</sup> Institut für Phonetik und Sprachliche Kommunikation,

<sup>2</sup> Bavarian Archive for Speech Signals (BAS)

Schellingstr. 3, D-80799 München, Germany

{beringer,ukartal,kalo,schiel,tuerk}@phonetik.uni-muenchen.de

## Abstract

This paper describes a general framework for evaluating and comparing the performance of multimodal dialogue systems: PROMISE (Procedure for Multimodal Interactive System Evaluation). PROMISE is a possible extension to multimodality of the PARADISE framework ((Walker and al., 1998; Walker et al., 1997) used for the evaluation of spoken dialogue systems), where we aimed to solve the problems of scoring multimodal inputs and outputs, weighting the different recognition modalities and of how to deal with non-directed task definitions and the resulting, potentially uncompleted tasks by the users.

PROMISE is used in the end-to-end-evaluation of the SmartKom project - in which an intelligent computer-user interface that deals with various kinds of oral and physical input is being developed. The aim of SmartKom is to allow a natural form of communication within man-machine interaction.

## 1. Introduction

The aim of this paper is to give an extended framework on dialogue system evaluation for multimodal systems in the end-to-end evaluation of SmartKom.

In the SmartKom project, an intelligent computer-user interface is being developed which deals with various kinds of oral and physical input. The system allows and reacts to gestures as well as mimic and spoken dialogue. Potential benefits of SmartKom include the ease of use and the naturalness of the man-machine interaction which are due to multimodal input and output. However, a very critical obstacle to progress in this area is the lack of a general methodology for evaluating and comparing the performance of the three possible scenarios provided by SmartKom:

- SmartKom Home/Office allows to communicate and to operate machines at home (e.g. TV, workstation, radio),
- SmartKom Public provides public access to public services, and
- SmartKom Mobile

Because of the innovative character of the project, new methods for end-to-end evaluation had to be developed partly through transferring established criteria from the evaluation of spoken dialogue systems (PARADISE), and partly through the definition of new multimodal measures. These criteria have to deal with a fundamental property of multimodal dialogue systems, namely the high variability of the input and output modalities with which the system has to cope. As an example, the system can accept a task solution not only via three different input devices (mimic-camera, voice-recording, gesture-camera), but also many different long-term solution strategies. For example, in order to plan an evening watching TV, a subject using the system may start with a sender (or a time, or an actress, etc.), progress to give a time (or sender, or actress, etc.) and end choosing a single programme (or a series, or

many programmes, or none, etc.). This inherent complexity and the tasks to be completed make it necessary to find an evaluation strategy which measures up to the possibilities of such a flexible system. Earlier evaluation strategies (PARADISE) had to deal with systems with only one input modality, used given solution strategies and thus were easily able to measure the success of a task. The advancements of SmartKom, though, cannot be adequately measured nor evaluated using an evaluation-strategy based on a monomodal system with pre-given solution strategies.

The following section gives an overview of standard problems of dialogue system evaluation which principally can be solved by the PARADISE framework (Walker and al., 1998; Walker et al., 1997). Section three describes how to define a task and extract the attribute value keys out of the description - solving a problem not uniquely belonging to multimodal dialogue evaluation. How we deal with incomplete tasks or tasks that get a very low task success measure due to incooperativity of the user, is described in section four. The scoring of multimodal inputs and outputs can be found in section four. Sections five to six give a detailed description of the status of multiple-to-one input facilities, i.e. the possibility to express the same user intention via multiple input as well as via different input modalities. Section seven defines the approach of PROMISE as a multimodal dialogue evaluation strategy which normalizes over dialogue strategy and dialogue systems. In the last section we sum up some ideas to be implemented in our framework.

## 2. Standard problems of dialogue system evaluation

Of course, multimodal dialogue evaluation has to deal with the same problems spoken dialogue system evaluation has to deal with, namely

- How can we abstract from the system itself, i.e. the different hardware and software components, in order to evaluate across dialogues and scenarios (see above)?

- How can we abstract from different dialogue strategies?

The PARADISE framework (for detailed description please refer to (Walker and al., 1998; Walker et al., 1997)) gives a useful and promising approach of how to compare different spoken dialogue strategies and different spoken dialogue systems via attribute value matrices (AVM), to compute the (normalized) task success measure (provided that a clearly defined task description is given to the user) define several (normalized) quality and quantity measures so-called cost functions, and to weigh their importance for the performance of the system via multiple linear regression dependent on the User Satisfaction value (cumulative function on the questionnaire completed by the subjects). This is not practicable, though, when dealing with a multimodal system like SmartKom.

Unfortunately, in dealing with multimodal systems we find a number of components which do not fit into the PARADISE approach, which are:

- The user is given a rather unprecise task definition, in order to enable a mostly natural interaction of user and system. Therefore there exist no static definitions of the “keys” (a PARADISE term) necessary to compute an AVM. Our solution is to extract different superordinate concepts depending on the task at hand. For example, when planning an evening watching TV, these superordinate concepts - we call them “information bits” - would contain movie title, genre, channel, timeslots, actors etc. Similar to a content-analysis, these “information bits” are carefully selected, categorized and weighted by hand before the tests start. This makes it possible for us to compute, normalize and compare across different tasks and scenarios.
- The number of information bits can vary within one completed task, but they must define a task unambiguously in order to finish it completely and successfully. For example, when a user asks to explicitly view a movie by name, assuming this movie is broadcasted at a set time and in only one channel, the number of information bits necessary to complete the task is one (the name of the movie). Whereas if a user doesn't know exactly what show he wants to watch, the number of information bits necessary to complete the task of planning an evening watching TV must be at least two (for example, time and channel).

In contrast to computing the task success via AVMs like PARADISE, in which case not completed tasks could implicitly influence the results, our information-bit-approach ensures that task success can only be calculated if the task has been completed.

### 3. How to deal with a bad performance due to user incooperativity?

One of the main problems of dialogue systems is an incooperative user. We consider only those users to be truly incooperative, who fall out of the role or purposely misuse the system. As an example, a user reading a book to the

system or using his mobile phone will be classified as an “incooperative” user. These, of course, are not unique to multimodal system evaluation, but can occur in other situations as well. On a first cue, it is impossible to incorporate incooperative users in an evaluation without lowering task success and thus the system performance. To avoid this, there exist the following approaches:

- Only dialogues with cooperative users are evaluated using empirical methods
- Only dialogues which terminate with finished tasks are evaluated.

Both approaches will be used in conjunction, so that a clearly defined set of data can be evaluated. When deciding to follow the first idea, uncooperative users as defined above, are of course interesting for other than purely empirical reasons. Evaluating the data generated by these incooperative users, in the above sense, in order to improve the system for future development, though, is not part of our aim to judge the quality of the present state of the system SmartKom.

### 4. How to score multimodal inputs or outputs?

In contrast to interactive unimodal spoken dialogue systems, which are based on many component technologies like speech recognition, text-to-speech, natural language understanding, natural language generation and database query languages, multimodal dialogue systems consist of several such technologies which are functionally similar to each other and therefore could interfere with each other. To make this clear, just imagine the similar functions of ASR and Gesture Recognition: while interacting with a multimodal man-machine interactive system like SmartKom users have the possibility to say what information they want to have and to simultaneously give the same, an additional, or a more specific input by an “interactional gesture” (Steininger et al., 2001). There are several possible problem solving strategies for the system namely:

- First match: the information which was recognized first is taken for further system processing, regardless of the recognition method. This would of course not help in multimodal processing.
- “Mean” match: the system takes the information which is common to both of the recognition modules. This could be called multimodal verification.
- Additional match: take all the information given by several recognizers for further system processing. This would be the best solution, if we assume all recognizers to be highly accurate, which leads us to the next problem:

### 5. How to weight the several multimodal components of recognition systems?

How can we estimate the accuracy of different recognizers? For example, in talking about speech recognition,

we have to deal with a very complicated pattern match, whereas gesture recognition has a limited set of recognizable gestures which can be found in a given coordinate plane.

It should be clear, that

- the gesture recognizer will be more accurate than the ASR system but
- the ASR system must get a higher weight than the gesture recognition when evaluating the system, since the complexity of the gesture recognition is much lower than the complexity of the ASR system.
- Apart from the problems of how to weight the different multimodal system components in an end-to-end evaluation of a multimodal system, there is also the problem of synchrony:
- Are multimodal inputs synchronous or linear within the evaluation? Are inputs from different modalities synchronous, i.e. are they describing the same user intention, although they may not be synchronous in time? Or does the system have to cope with different inputs?

## 6. PROMISE - A Procedure for Multimodal Interactive System Evaluation

In the last sections we have identified the most characteristic problems which show the need for an extended framework for multimodal dialogue system evaluation. We already gave some examples of possible problem solving strategies. Within this section we will specify these ideas and present the current version of PROMISE. Given the normalized performance function of PARADISE

$$\text{performance} = \alpha \mathcal{N}(\kappa) - \sum_{i=1}^n \omega_i \mathcal{N}(c_i)$$

with  $\alpha$  the weight for task success<sup>1</sup>  $\kappa$ , the assumed Gaussian cost functions<sup>2</sup>  $c_i$  weighted by  $\omega_i$ , and  $\mathcal{N}$  z-score normalization function. Weights are defined via a linear multiple regression over  $\kappa$  respectively the costs and the cumulative sum of the user satisfaction scores (see usability questionnaire (Walker and al., 1998)). PROMISE now splits this function in two parts in the way that the formula is reduced to normalized cost functions first. Instead of a multiple linear regression between the free cost variables and the dependent user satisfaction score, PROMISE searches correlations via Pearson correlation between User-Satisfaction - Cost pairs. This means that objective measurable costs will be addressed in the questionnaire to be answered by each user.

Tables 1 and 2 give an overview of the costs we defined in SmartKom, some of them equivalent to the PARADISE costs, some of them extended to deal with multimodality or to weed out user incooperativity.

<sup>1</sup>defined as the successful completion of a duty

<sup>2</sup>either mean or cumulative sum of one cost category (quantity and quality measures); differing from system to system

| Quality measures          |  |
|---------------------------|--|
| system-cooperativity      | measure of accepting misleading input  |
| semantics                 | no. of multiple input possible misunderstandings of input/output<br>semantical correctness of input/output   |
| helps                     | no. of offered help for the actual interaction situation   |
| recognition               | speech<br>facial expression<br>gestures  |
| transaction success       | no. of completed sub-tasks   |
| diagnostic error messages | percentage of error prompts  |
| dialogue complexity       | task complexity (needed information bits for one task)<br>input complexity (used information bits)<br>dialogue manager complexity<br>(presentation of results) |
| ways of interaction       | gestures/graphics vs. speech<br>n-way communication (several modalities possible at the same time?)  |
| synchrony                 | graphical and speech output  |
| user/system turns         | mixed initiative<br>dialogue management<br>incremental compatibility   |

Table 1: Quality measures for the SmartKom evaluation

Apart from measuring the quality of dialogue systems in general, like dialogue management (system directed, user directed or mixed initiative), elapsed time of input and output, task completion, mean response time, word count and turn count, we defined measures referring to the problems in section four and five above. Multiple inputs and outputs are scored via the “semantics” cost, to be precised the number of multiple input and the possible misunderstandings of input/output (due to multimodality).

The multimodal components of recognition systems are partly weighted via Pearson correlation using the corresponding user satisfaction scores for the recognition costs defined above. Comparing the accuracies of the different recognition systems for defining a cross-recognizer weight, we calculate “ways of interaction” and “helps”. The latter defines the quality and quantity of dynamic help offered by the system in situations where the emotional status of the user changes.

<sup>3</sup>(Oppermann et al., 2001)

| Quantity measures     |   |
|-----------------------|---|
| barge-in              | no. of user and system overlap by means of backchannelling, negation of output, further information   |
| Cancels               | planned system interrupts due to barge-in   |
| off-talk <sup>3</sup> | no. of non-system directed user utterances  |
| elapsed time          | duration of input of the facial expression<br>duration of gestural input<br>duration of speech input<br>duration of ASR<br>duration of gesture recognition<br>mean system response time<br>mean user response time<br>task completion<br>duration of the dialogue |
| rejections            | error frequency of input which require a repetition by the user   |
| timeout               | error rate of output<br>error rate of input   |
| user/system turns     | no. of turns<br>no. of spoken words<br>no. of produced gestures<br>percentage of appropriate/inappropriate system directive<br>diagnostic utterances<br>percentage of explicit recovery answers   |

Table 2: Quantity measures for the SmartKom evaluation

The second step is to define another way to calculate the task success.

In PARADISE a set of defined static “keys” was used to measure task success via an attribute value matrix. Since “information bits” (see section two) are used, it makes no sense to calculate an AVM. As described above, these information bits can vary from situation to situation. A successful task is given, if the task was completed according to the necessary number of information bits. A task fails, if it has not been successfully completed. Therefore, we define task success in PROMISE as follows:

$\tau_j = +1$  : task success;  $\tau_j = -1$  : task failure; where  $j$  is the index of the corresponding tests.

For each test the corresponding user satisfaction values are Pearson correlated with  $\tau_j$ .

The system performance results in the following formula:

$$\text{performance} = \alpha \bar{\tau} - \sum_{i=1}^n \omega_i \mathcal{N}(c_i)$$

with  $\alpha$  being the Pearson correlation between  $\tau_j$ (task success) and the corresponding user satisfaction values,  $\bar{\tau}$  the mean value of all  $\tau_j$  with  $j$  index of tests,

$n$  the number of different cost functions,

$c_i$  the assumed Gaussian cost functions - consistently either mean or cumulative sum of one cost category  $i$  (measured over all tests)

weighted by  $\omega_i$  - the Pearson correlation between cost function  $c_i$  and defined associated user satisfaction values, and the z-score normalization function  $\mathcal{N}(c_i) = \frac{c_i - \bar{c}_i}{\sigma_{c_i}}$ , where  $\sigma_{c_i}$  as variance of  $c_i$ .

## 7. Conclusion and future work

Our aim was to roughly define an extended evaluation framework for a multimodal dialogue system evaluation which can deal with multimodal dialogue processing. However, there are still some unresolved or solely unsatisfactorily solved problems dealing with the timing of input in multimodal systems. We are currently specifying different approaches in order to satisfactorily solve the remaining problems which we hope to present at the LREC post-conference workshop on “Multimodal Resources and Multimodal Systems Evaluation” in June.

## 8. Acknowledgements

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the SmartKom project (01IL905E/6).

## 9. References

- D. Oppermann, F. Schiel, S. Steininger, and N. Beringer. 2001. Off-talk - a problem for human-machine-interaction. *Proc. of EUROSPEECH 2001, Scandinavia, Aalborg*.
- S. Steininger, B. Lindemann, and T. Paetzold. 2001. Labeling of gestures in smartkom - the coding system. *Springer Gesture Workshop 2001, London (to appear)*, LNAI 2298.
- M.A. Walker and al. 1998. Evaluating spoken dialogue agents with paradise: Two case studies. *Computer Speech and Language*, 12.
- M.A. Walker, D.J. Litman, C.A. Kamm, and A. Abella. 1997. Paradise: A framework for evaluation spoken dialogue agents. *Annual Meeting of the Association of Computational Linguistics. ACL*.

# MUMIN

## A Nordic Network for MultiModal Interfaces

Patrizia Paggio\*, Kristiina Jokinen\*\* and Arne Jönsson\*\*\*

\* Center for Sprogteknologi, Copenhagen

email: [patrizia@cst.dk](mailto:patrizia@cst.dk)

\*\* University of Art and Design, Helsinki

email: [Kristiina.Jokinen@uia.fi](mailto:Kristiina.Jokinen@uia.fi)

\*\*\* University of Linköping

email: [arnjo@ida.liu.se](mailto:arnjo@ida.liu.se)

### Abstract

This paper reports on a recent initiative undertaken under the language technology programme of the Nordic Research Training Academy (NORFA) to create a network of Nordic research institutes working with multimodal interfaces. In the paper we present the objectives of the network and give an overview of multimodal research and resources in the Nordic countries.

**Keywords:** multimodal integration, cognitive and usability studies, multimodal dialogue, multimodal research and resources in the Nordic countries

## 1. Motivation and objectives

In the Nordic countries as elsewhere, both the research and the industrial communities are showing a growing interest in multimodal interfaces. Therefore, there is a need to channel the efforts of individual organisations and countries in joint activities to establish a common research agenda and to define relevant standards and generic methodologies. The aim of the Nordic Network for MultiModal Interfaces (MUMIN) under the NORFA programme, is to stimulate Nordic research in this area and increase its visibility in the international research community. MUMIN, which was established in January 2002 and has funds for two years of activity, has the following goals:

- ? encouraging joint activities in building generic models and architectures as well as defining standards for the integration and development of multimodal interaction;
- ? encouraging investigations on the use of multimodality in various practical applications;
- ? providing a forum for sharing resources and results, and by encouraging network participants to make their research results available via the network's web site;
- ? organising PhD courses and research workshops on issues related to multimodal interaction;
- ? creating a network of contacts and a pool of shared knowledge that can be taken advantage of for the

definition of collaborative research projects and for product development.

It will also be an important objective of the network to support investigation of the use of multimodal interaction in non-expert environments, and the accessibility of disabled people to IT technology. The network will thus contribute to the Nordic countries' social objectives and help them advance in their vision of building democratic and equal societies for everyone. This also conforms to the EU objectives of creating a user-friendly information society, with accessibility of IT benefits and services for all citizens.

A whole range of research issues, some of which have already received attention from the institutes engaged in the network, are relevant to this overall objective, and constitute topics of interest around which the network's activities will be organised.

## 2. Multimodal integration

A central issue is that of multimodal integration, which will be approached from different perspectives. A promising approach put forward by several researchers is that of using techniques known from NLP. A similar distinction to that made in NLP between grammar rules and parsing algorithms can be made between a multimodal grammar and an algorithm for applying the grammar to input from multiple modalities. By upholding this separation of process and data, the process of merging inputs from different modalities can be made more general, as the representation becomes media-independent. Furthermore, defining algorithms for modality integration independent of the specific

modalities used in a particular application, increases the chances that components of the system can be extended and reused. For example in the Danish research project Staging, Center for Sprogteknologi (CST) has developed a multimodal dialogue interface to a virtual environment (Paggio *et al.* 2000) where speech, keyboard and gestural inputs are merged by a feature-based parser. Relevant to this issue is also the work carried out by the Speech and Multimedia Communication (SMC) group from Center for PersonKommunikation, Aalborg, which also has extensive research and teaching experience in the area of multimodality, complemented with expertise in speech and image processing (Larsen & Brøndsted 2001).

Another promising approach to modality integration represented in MUMIN is the use of different machine learning techniques, in particular neural networks. As has been the case with many other application domains, also for multimodal integration hybrid systems mixing rule-based approaches with machine learning algorithms, may well provide the most interesting results. Although rule-based methods in general work reasonably well, it is a well-known problem that an explicit specification of the steps, i.e. rules that are required to control the processing of the input, is a difficult task, and when the domain becomes more complex, the rules become more complex too. Often the correlation between input and output is difficult to specify. This is the case e.g. with multimodal interfaces, and thus approaches which are both robust and able to adapt to new inputs are needed. Expertise in this domain is brought to the network by the Media Lab at the University of Art and Design in Helsinki (UIAH), and especially its Soft-Computing Interfaces Group which is devoted to designing adaptive interfaces and developing tools for human-machine interaction, relying on nature-like emergent knowledge that arises from subsymbolic, unsupervised processes of self-organizing nature (see e.g. Jokinen *et al.* 2001).

In a similar way as the rule-based integration of modalities can be enhanced using machine learning techniques, results obtained through pure probabilistic analysis methods may well be boosted by the addition of symbolic rules. An example relevant to multimodal interfaces are the algorithms for character and word prediction used in connection with eye-tracking, where the system tries to guess what the user is "typing with the eye". Although the performance of the probabilistic approaches implemented in current systems is promising, language technology techniques seem to constitute a valuable add-on. This is an issue that the IT University of Copenhagen is working on.

### 3. Neurocognitive basis and usability studies

To fully exploit multimodality in various interfaces, it is important to know how the neurocognitive mechanisms support multimodal and multisensory integration. In comparison to that devoted to single sensory systems, there has been very little research on the integration mechanisms of information received via different senses. However, the research group of Cognitive Science and Technology at the Helsinki University of Technology is

using various methods to uncover the neurocognitive principles of multisensory integration with the purpose of developing mathematical models of this integration. The group is also developing a Finnish artificial person – a talking and gesturing audiovisual head model – which will serve as a well-controlled stimulus for neurocognitive studies (Sams *et al.* 1998).

A related issue is that of the impact of multimodality on the users. Relevant questions are which input and output modalities should be used for which task, which are the best combinations, and how different modalities are used by or for users with different degrees of expertise. There are in general two ways to use multimodal input: to react directly to the user's intentional input and to observe the user's unconscious use of certain modalities (e.g. eye-gaze). The former method is based on direct control and has been used in earlier conversational interface prototypes. Observing the user and understanding their intentions and mental states has not been extensively studied, and would add valuable information to the design of multimodal systems. The Computer Human Interaction group at the University of Tampere (TAUCHI) will bring to the network its extensive expertise in the design and use of innovative user interfaces and in usability testing, as well as the agent-based development platform Jaspis (Turunen and Hakulinen 2000). The Natural Interactive Systems Laboratory (NISLab) at the University of Southern Denmark, has also made pioneering contributions to the theoretical understanding of unimodal input/output modalities, and of the multiple conditions which determine the usability of individual modalities and their combinations (Bernsen 2001).

## 4. Multimodality and dialogue

A third issue, which encompasses a great deal of research work carried out by several of the network members, is that of multimodal dialogue. In this respect, it is interesting to note that the growing interest in multimodal interaction is opening a new perspective to Nordic research on dialogues, which is already acknowledged internationally. Several institutes in the Nordic countries have in fact contributed substantially to dialogue research, and developed dialogue models as well as implemented dialogue systems.

The Department of Linguistics at the University of Göteborg has extensive experience in corpus collection and dialogue management. They have developed tools for spoken language analysis and coding which can be applied to the collection and analysis of multimodal dialogues, thus providing empirical basis and insight for research on multimodal interaction: how different modalities are used in human-human communication (Allwood, 2001). NISLab has a strong background in dialogue management, dialogue systems evaluation, and spoken dialogue corpus coding from a number of EU projects. NISLab is currently addressing best practice in the development and evaluation of natural interactivity systems and components; surveying data resources, coding schemes and coding tools for natural interactivity;

and building a general-purpose coding tool for natural interactive communicative behaviour. The natural language processing research group (NLPLab) at the University of Linköping has for almost two decades conducted research on dialogue systems and now has a platform for the development of multimodal dialogue systems for various applications to be developed further towards an open source code repository (Degerstedt & Jönsson, 2001). Current focus is on integrating dialogue systems with intelligent document processing techniques in order to develop multimodal dialogue systems that can retrieve information from unstructured documents, where the request requires that the user, in a dialogue with the system, specifies their information needs (Merkel & Jönsson, 2001). Finally, the Centre for Speech Technology at the Royal Institute of Technology in Stockholm (KTH) has developed several multimodal dialogue systems with the motivation of studying speech

technology as part of complete systems and the interaction between the different modules that are included in such systems. Their first system, Waxholm, was a multimodal system exploring an animated agent (Carlsson & Granström, 1996). Current work and interests involve research on multimodal output using animations and also to some extent multimodal input using both speech and pointing (Gustafson *et al.* 2000).

## 5. Participating groups

The groups participating in MUMIN are shown in Table 1 below. Currently, three Nordic countries are represented, but the network is interested in welcoming participants from other Nordic countries, as well as in cooperation with non-Nordic countries.

| Denmark   | Finland  | Sweden   |
|---|--|--|
| Center for Sprogteknologi, Copenhagen                         | University of Art and Design Helsinki , Media Lab, Helsinki  | Linköpings Universitet, Institutionen för datavetenskap, Linköping |
| Syddansk Universitet , NISLab , Odense                        | University of Tampere, Department of Computer and Information Sciences, Tampere Unit for Computer-Human Interaction (TAUCHI), Tammerfors | KTH, CTT, Centre for Speech Technology, Stockholm                  |
| Aalborg Universitet , Center for PersonKommunikation, Aalborg | Helsinki University of Technology , Laboratory of Computational Engineering, HUT Espoo   | Göteborgs Universitet, Institutionen för lingvistik, Göteborg      |
| IT-Højskolen, Eye Gaze Research Team, Copenhagen              |  |  |

Table 1: The groups participating in MUMIN

## 6. Conclusion

The MUMIN network is expected to play an important strategic role in the establishment of a common research agenda for Nordic researchers working with multimodal interfaces, but also to relate its activities to those of the international community, and to contribute to the general progress in the area. Therefore, it is highly relevant for MUMIN to participate in this workshop, and to provide a Nordic contribution to the discussion of a multimodal roadmap.

## 7. References

Allwood, J. (2001) Dialog Coding - Function and Grammar, Göteborg Coding Schemas, *Gothenburg*

- Papers in Theoretical Linguistics, GPTL 85.* Göteborg University, Department of Linguistics.
- Bernsen, N. O. (2001) Multimodality in language and speech systems - from theory to design support tool. Chapter to appear in Granström, B. (Ed.): *Multimodality in Language and Speech Systems.* Dordrecht: Kluwer Academic Publishers.
- Carlson, R., Granström, B. (1996) The WAXHOLM spoken dialogue system. *Acta universitatis Carolinae philologica* 1, pp. 39-52.
- Degerstedt, L. and Jönsson, A. (2001) A Method for Iterative Implementation of Dialogue Management , *IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems*, Seattle.
- Gustafson, J, Bell, L, Beskow, J, Boye, J, Carlson, R, Edlund, J, Granström, B, House, D & Wirén M (2000) AdApt - a multimodal conversational dialogue system

- in an apartment domain, In *Proc of ICSLP 2000*, Beijing, 2:134-137
- Jokinen, K., Hurtig, T., Hynnä, K., Kanto, K., Kaipainen, M., and Kerminen, A. (2001) Dialogue Act classification and self-organising maps. In *Proceedings of the Neural Networks and Natural Language Processing Workshop*, Tokyo, Japan.
- Larsen, L.B. and Brøndsted, T. (2001) A Multi Modal Pool Trainer. In *Proceedings of the International Workshop on Information Presentation and Natural Multimodal Dialogue IPNMD-2001*. Verona, Italy 14-15 December 2001, pp. 107-111.
- Merkel, M. and Jönsson, A. (2001) Towards multimodal public informations systems, *Proceedings of 13th Nordic Conference on Computational Linguistics, NoDaLiDa '01*, Uppsala, Sweden.
- Paggio P., Jongejan B. and Madsen C.B. (2000) Unification-based multimodal analysis in a 3D virtual world: the Staging project. In *Proceedings of the CELE-Twente Workshop on Language Technology: Interacting Agents*, pp. 71-82.
- Sams, M., Manninen, P., Surakka, V., Helin, P. and Kättö, R. (1998) Effects of word meaning and sentence context on the integration of audiovisual speech. *Speech Communication*, 26, 75-87.
- Turunen, M. and Hakulinen J. (2000) JASPIS - a Framework for Multilingual Adaptive Speech Applications. In *Proceedings of the 6<sup>th</sup> International Conference of Spoken Language Processing. ICSPL 2000*.