Chapter 1

Natural and multimodal interactivity engineering – directions and needs

Niels Ole Bernsen and Laila Dybkjær
Natural Interactive Systems Laboratory
University of Southern Denmark
Campusvej 55, 5230 Odense M, Denmark

Abstract
This introductory chapter takes a global view of the field of natural and multimodal interactivity engineering and superimposes the contributions of the following 15 chapters onto this view to gauge the current state of the field, its needs and future prospects.

Keywords: Natural and multimodal interactivity engineering

1. Introduction

Chapters 2 through 16 of this book present original contributions to the emerging field of natural and multimodal interactivity engineering. In this chapter, we discuss the nature of the field and present an admittedly incomplete and sketchy chart of the current state of natural and multimodal interactivity engineering, where the field is going, and what are the needs which should be met for the field to advance as effectively and efficiently as possible in the years to come. A simple matrix for the field provides the structure of the chapter itself and helps plot the multi-dimensional contributions to the field made in Chapters 2 through 16.

2. Current characteristics of natural and multimodal interactivity engineering

The first and most prominent characteristic of natural and multimodal interactivity engineering (henceforth NMIE) is that the field is not yet an established field of research and commercial development but, rather, an emerging one in all respects, including applicable theory, experimental results, platforms and development environments, standards (guidelines, de facto standards, official standards), evaluation paradigms, coherence, ultimate scope, enabling technologies for software engineering, general topology of the field itself, “killer applications”, etc.

A second important characteristic of the NMIE field is that its practitioners come from very many different, and often far more established, fields of research and industrial development, such as signal processing, speech technology, computer graphics, computer vision, human-computer interaction, virtual and augmented reality, non-speech sound, haptic devices, telecommunications, computer games, etc. It may be noted that the fact that a field of research has been established over decades in its own right is fully compatible with many if not most of its practitioners being novices in NMIE. It follows that NMIE community formation is an ongoing challenge for all.

Thirdly, the field is expanding very rapidly, primarily, it seems, driven by a shared but perhaps not quite unified vision of the potential of new interactive modalities of information representation and exchange for radically transforming interaction with computer systems, networks, devices, applications, etc. from the GUI (graphical user interface) paradigm into something which will achieve a far deeper and much more intuitive and natural integration of computer systems in people’s lives. The inherent, ultimate vision which the field is gradually unveiling, is, we believe, one in which even

the age-old term “interaction” becomes inadequate for characterising the systems which will emerge. Despite its generality, the notion of interaction still embodies the idea of conscious control of something else, such as a tool. Rather, or so goes the vision, people will often no longer have to *interact* with the system - i.e. consciously and deliberately exchanging information with the system - in order to get things done, the system will get things done by itself based on observation and knowledge of its human companions. Thus, with all its dangers and pitfalls, the inherent, ultimate vision of the NMIE field might be that of *the caring system* with which we will often interact, of course, but which aims to do what we need done whether or not its pursuance of those aims result from traditional human-system interaction.

Fourthly, the field of natural and multimodal interactivity engineering is vast no matter how one looks at it: whether in terms of the new classes of applications envisioned, the theoretical, empirical, developmental, and evaluation research challenges, the community integration needs, the supporting knowledge and craft skills which can help newcomers off to an early start, its future impact, etc. Obviously, the field inherits all or most current trends in today’s world of computing more generally, such as ambient intelligence [Ducatel et al. 2001] which, in fact, incorporates most of the others, including ubiquitous computing, cognitive systems, pervasive computing, higher mobile bandwidth, new agent architectures, powerful networks, new sensors and actuators, new devices, etc.

3. Multimodality and natural interactivity

Conceptually, NMIE combines natural interactive and multimodal systems and components engineering. While both concepts, natural interactivity and multimodality, have a long history, it would seem that they continue to sit somewhat uneasily side by side in the minds of most of us. *Multimodality* is the idea of being able to choose any input/output modality or combination of input/output modalities for optimising interaction with the application at hand, such as speech input for many heads-up, hands-occupied applications, speech and haptic input/output for applications for the blind, etc. A *modality* is a particular way of representing input or output information in some *physical medium*, such as something touchable, light, sound, or the chemistry for producing olfaction and gustation [Bernsen 2002, see also Carbonell and Kieffer]. The physical medium of the speech modalities, for instance, is sound or acoustics but this medium obviously enables the transmission of information in many acoustic modalities other than speech, such as earcons, music, etc. The term multimodality thus refers to any possible combination of elementary or *unimodal* modalities.

Compared to multimodality, the notion of *natural interactivity* appears to be the more focused of the two. This is because natural interactivity comes with a focused vision of the future of interaction with computer systems as well as a relatively well-defined set of modalities required for the vision to become reality. The natural interactivity vision is that of humans communicating with computer systems in the same ways in which humans communicate with one another. Thus, natural interactivity specifically emphasises human-system communication involving the following input/output modalities used in situated human-human communication: speech, gesture, gaze, facial expression, head and body posture, and object manipulation as integral part of the communication (or dialogue). As the objects being manipulated may themselves represent information, such as text and graphics input/output objects, natural interaction subsumes the GUI paradigm. Technologically, the natural interactivity vision is being pursued vigorously by, among others, the emerging research community in embodied conversational agents as illustrated by [Bickmore and Cassell, Cole et al., Granström and House, Heylen et al., Massaro]. An embodied conversational agent may be either virtual or a robot [Sidner and Dzikovska].

A weakness in our current understanding of natural interactivity is that it is not quite clear where to draw the boundary between the natural interactivity modalities and all those other modalities and modality combinations which could potentially be of benefit to human-system interaction. For instance, isn’t pushing a button on the mouse or otherwise, although never used in human-human communication for the simple reason that humans do not have communicative buttons on them, as natural as speaking? If it is, then, perhaps, all or most research on useful multimodal input/output modality combinations is also research into natural interactivity even if the modalities addressed are not being used in human-human communication? In addition to illustrating the need for more and better NMIE theory, the point just made may explain the uneasy conceptual relationship among the

two paradigms of natural interactivity and multimodality. In any case, we have decided to combine the paradigms and address them together as natural and multimodal interactivity engineering.

Finally, by NMI 'engineering' we primarily refer to software engineering. It follows that the somewhat innovative expression 'natural and multimodal interactivity engineering' primarily represents the idea of creating a specialised branch of software engineering for the field addressed in this book. It is important to add, however, that NMIE enabling technologies are being developed in fields whose practitioners do not tend to regard themselves as doing software engineering, such as signal processing. For instance, the recently launched European Network of Excellence SIMILAR [http://www.similar.cc] addresses signal processing for natural and multimodal interaction.

4. A matrix for the field

Broadly speaking, the emergence of a new systems field, such as NMIE, takes *understanding* of issues, problems and solutions, knowledge and skills for *building* (or developing) systems, and *evaluation* of any aspect of the process and its results. In the particular case of NMIE, these goals could be expanded as shown in Table 1.1. The table thus aims to specialise, to a modest extent, general software engineering needs for the particular purposes of NMIE.

General	**Generic**	**Specific to NMIE**
Understand it	Future visions	Visions, roadmaps, etc., general and per sub-area
	Applicable theory	Applicable theory for any aspect of NMIE
	Empirical work	Controlled experiments, behavioural studies, simulations, scenario studies, task analysis on roles of, and collaboration among, specific modalities to achieve various benefits
	Coding and analysis	New quality data resources, coding schemes, coding tools, and standards
Build it	Enabling technologies	New basic technologies needed
	More advanced systems	New, more complex, versatile, and capable system aspects
	Make it easy	Re-usable platforms, components, toolkits, architectures, interface languages, standards, etc.
Evaluate it	All aspects of it	Evaluate components, systems, technologies, processes, etc.

Table 1.1. Needs for progress in natural and multimodal interactivity engineering.

Using the Column 3-structure of Table 1.1, Table 1.2 indicates how the 15 chapters in this book contribute to the NMIE field and its needs.

A preliminary conclusion based on Table 1.2 is that, for an emerging field which still has not seen any but the simplest of commercial exploitation yet, the NMIE research being done world-wide today is already pushing the frontiers in many of the directions needed. In Section 14, we will contrast the picture provided by Table 1.2 by a view of future NMIE needs (Table 1.3).

Specific to NMIE	**Contributions**
Visions, roadmaps, etc., general and per sub-area	Full natural interactive systems interacting with people like people interact with each other. 6
Applicable theory for any aspect of NMIE	Arguably, no completely new theory but plenty of applied theory.
Controlled experiments, behavioural studies, simulations, scenario	Effects on communication of animated conversational agents. 2, 10, 11 Spoken input in support of visual search. 3 Gesture and speech for video game playing. 7

studies, task analysis on roles of, and collaboration among, specific modalities to achieve various benefits	Multimodal segmentation of multiple speakers for multi-speaker speech recognition. 8 Animated talking heads for more intelligible and efficient spoken output. 10, 11 Gaze behaviour for more likeable animated interface agents. 2, 11 Audio-visual speech for child language learning. 13 Animated talking heads and agents for child language and reading learning. 6, 13 Tutoring systems 5, 6, 15
New quality data resources, coding schemes, coding tools, and standards	Coding scheme and tool for NMIE systems evaluation. 12 Gesture annotation scheme. Standard for internal representation of NMIE data codings. 12 Coding tool for multilevel NMIE data coding. 12
New basic technologies needed	Interactive robotics: robots controlled multimodally, tutoring and hosting robots. 15 Multi-speaker speech recognition. 8 Machine learning: of language. 9 Machine learning: of dialogue acts assignment. 16 Audio-visual speech synthesis for talking heads. 10, 13
New, more complex, versatile, and capable system aspects	Multilinguality. 14 Ubiquitous (mobile) application. 14 On-line observation-based user modelling for adaptivity. 4, 14 Complex natural interactive dialogue management. 4, 5, 14, 16 Machine learning of language. 9 Machine learning for dialogue modelling. 16 Dialogue patterns. 16
Re-usable platforms, components, toolkits, architectures, interface languages, standards, etc.	Platform for natural interactivity. 6 Re-usable components (many papers). Development toolkit for multimodal dialogue management. 6 Architectures for (multimodal) dialogue management. 5, 6, 14, 15, 16 Multimodal interface language. 14 Tools for developing natural interactive systems. 6 XML for data exchange. 14, 15
Evaluate components, systems, technologies, processes, etc.	Effects on communication of animated conversational agents. 2, 10, 11 Evaluations of talking heads. 10, 11, 13 Evaluation of audio-visual speech synthesis for learning. 10, 13

Table 1.2. How the chapters in this book address current NMIE needs.

5. Modalities investigated

We argued in Section 3 that multimodality includes all possible modalities for the representation and exchange of information among humans and between humans and computer systems, and that natural interactivity includes a rather vaguely defined, large sub-set of those modalities. Within this wide space of unimodal modalities and modality combinations, it may be useful to look at the modalities actually addressed in the following chapters. These are summarised by chapter in the following list. I means input, O means output, and N/A means not discussed in detail, or not relevant.

2. I: speech, gesture (via camera) vs. speech-only. O: embodied conversational agent + images vs. speech-only + images.
3. I: gesture (mouse). O: speech, graphics.
4. I: speech, text, gesture (pointing). O: speech, graphics.

5. I: speech, gesture (pointing). O: speech, text, graphics.
6. Most natural interactive modalities.
7. I: speech, gesture, object manipulation/manipulative gesture. O: video game.
8. I: speech, camera-based graphics. O: N/A.
9. I: speech, keyboard, mouse, pen-based drawing and pointing, camera. O: speech, graphics and text display.
10. O: audio-visual speech synthesis, talking head.
11. I: typed text, O: talking head, gaze.
12. I: gesture. O: N/A.
13. I: mouse, speech, O: audio-visual speech synthesis, talking head, images, text.
14. I: speech, haptic buttons. O: music, speech, text, graphics, tactile rhythm.
15. I: Speech, mouse clicks, (new version includes face and gesture input via camera). O: robot pointing and beat gestures, speech.
16. I: Speech and possibly other modalities, O: Speech and possibly other modalities. Focus is on dialogue modelling so input and output modalities are not discussed in detail.

The vision chapter by [Cole et al.] discusses most of the natural interactive modalities. Combined speech input/output, which, in fact, means spoken dialogue almost throughout, is addressed in about half of the chapters [Bickmore and Cassell, Chai et al., Clark et al., Cole et al., Corradini and Cohen, Dusan and Flanagan, Reithinger et al., Sidner and Dzikovska]. Two thirds of the chapters address gesture input in some form [Bickmore and Cassell, Chai et al., Clark et al., Cole et al., Corradini and Cohen, Darrell et al., Dusan and Flanagan, Martell, Reithinger et al., Sidner and Dzikovska]. Six chapters address output modalities involving talking heads, embodied animated agents, or robots [Bickmore and Cassell, Cole et al., Granström and House, Heylen et al., Massaro, Sidner and Dzikovska]. Three chapters [Darrell et al., Cole et al., Bickmore and Cassell] address computer vision input and Sidner and Dzikovska mention that they are looking into adding vision to their robot system. Dusan and Flanagan also mention that their system has camera-based input. Facial expression of emotion is addressed by [Cole et al., Granström and House]. Despite its richness and key role in natural interactivity, input or output speech prosody is hardly discussed. Granström and House discuss graphical ways of *replacing* missing output speech prosody by facial expression means.

In general, the input and output modalities and their combinations discussed would appear representative of the state-of-the-art in NMIE. The authors make it quite clear how far we are from mastering the very large number of potentially useful unimodal “compounds” theoretically, in input recognition, in output generation, as well as in understanding and generation.

In the following Sections 6 through 13, we briefly review the NMIE contributions in this book, following the structure of Tables 1.1 and 1.2. Throughout, discussion focuses on what is being done in the current state-of-the-art to further the global goals of NMIE, what is not being done, and what still needs to be done. Based on the discussion, Section 14 presents a view of current NMIE research needs following the now familiar, proposed structure of the field.

6. Visions for natural and multimodal interactivity engineering

In their chapter on visions for NMIE, [Cole et al.] highlight several important points. Using natural interactive teaching systems (or tutorial systems) applications for illustration, the paper evidences the important driving role of re-usable platforms and tools, such as the DARPA Communicator system manager or “hub” (http://fofoca.mitre.org), for rapid progress. Moreover, Cole et al. point out the future importance to NMIE of two generic technologies which are needed to extend spoken dialogue systems to full natural interactive systems. These technologies are (i) computer vision for processing camera input, such as face tracking, eye tracking, expression recognition, and gesture recognition, and (ii) computer animation systems. It is only recently that the computer vision community has begun to address issues of natural interactive and multimodal human-system communication, and there is a long way to go before computer vision can parallel speech recognition as a major input medium for NMIE.

On the issue of interdisciplinarity and community convergence, Cole et al. point out that it takes a diverse community of researchers working together to develop perceptive animated interfaces.

Cole et al. aptly illustrate current NMIE initiatives to extend natural and multimodal interaction beyond traditional information systems to new major application areas, such as training and education which has been around for a while already, notably in the US-dominated paradigm of tutoring systems using animated interface agents, but also to edutainment and entertainment. While the GUI, including the current WWW, might be said to have the edutainment potential of a schoolbook or newspaper, NMIE systems have the much more powerful edutainment potential of brilliant teachers, comedians, and exiting human-human games. Some recent, major embodied conversational character system development efforts following these leads are the US Army-sponsored systems for tactical situation control and tactical Arabic training (http://www.isi.edu/isd/carte/proj_tactlang/), and the EU Human Language Technologies-sponsored NICE (Natural Interactive Conversation for Edutainment, http://www.niceproject.com) system for spoken conversation with fairytale author Hans Christian Andersen [Bernsen and Dybkjær 2004].

7. The need for applicable theory

It may be characteristic of the NMIE field at present that our sample of papers does not include a single contribution of a primarily theoretical nature. However, the absence of theoretical papers is *not* characteristic in the sense that the field does not make use of, or even need, applicable theory. On the contrary, a large number of chapters actually do apply existing theory in some form, ranging from empirical generalisations to full-fledged theory of many different kinds. For instance, Bickmore and Cassell test generalisations on the effects on communication of involving embodied conversational agents; Carbonell and Kieffer apply modality theory; Chai et al. apply theories of human-human dialogue to the development of a fined-grained, semantics-based multimodal dialogue interpretation framework; Clark et al. mention support by activity theory of dialogue; Cole et al. mention cognitive theory as underlying their literacy tutor; Massaro applies theories of human learning; and Sidner and Dzikovska draw on conversation and collaboration theory. The interesting points are, rather, we submit, the following. Firstly, it is only natural that NMIE researchers, facing more or less daunting tasks of engineering innovation, initially orient themselves towards applying established theories, and start by proposing empirical generalisations rather than full-fledged theories. Secondly, NMIE theory development is hard to do, slows down what we want to accomplish in engineering, and tends to be regarded with caution not only by funding agencies but also by our fellow researchers because, by adopting the theory, they often have to revise their ways of thinking.

In terms of needs for applicable theory, the NMIE field may be unparalleled in its needs for a large variety of theoretical support. All of a sudden, for instance, when developing embodied, life-like conversational agents, we find ourselves deeply into high-complexity areas, such as human personality, emotions, attitudes, detailed situated non-verbal behaviour, etc. This might be viewed to suggest that many theoretical needs of NMIE can be addressed by adapting, in order to make them applicable and operational, theoretical results from many different disciplines. However, adapting a theory is a theoretical exercise in itself, and it seems likely that we shall need lots of new theory anyway. Some sources for new theory are the emerging generalisations discussed in Section 8 and the much needed new coding schemes discussed in Section 9.

8. Empirical results

By contrast with new theories and fortunately so, the NMIE field is replete with empirical studies of human-human and human-machine natural and multimodal interaction [Dehn and van Mulken 2000]. By their nature, empirical studies are far closer to the process of engineering than is theory development. We build NMIE research systems not only from theory but, perhaps to a far greater extent, from hunches, contextual assumptions, extrapolations from previous experience and untried transfer from different application scenarios, user groups, environments, etc., or even Wizard of Oz studies, which are in themselves a form of empirical study, see, e.g., [Corradini and Cohen, Bernsen et al. 1998]. Having built a prototype system, we are keen to find out how far those hunches, etc. got us. Since empirical testing, evaluation, and assessment are integral parts of software and systems

engineering, all we have to do is to include “assumptions testing” in the empirical evaluation of the implemented system which we would be doing anyway.

The comparative ease of doing empirical studies as part of the normal business of engineering sometimes tempts us to think that analysis and reporting of empirical experimentation is easy to do and interpret. When this happens, we get published empirical results stating, for instance, that animated speaking interface agents are liked by users who, almost invariably, are university students conveniently grabbed in the corridors. This blatant over-generalisation is then countered by other findings showing that users prefer, e.g., speech-only communication. Given the importance of solid empirical investigation for NMIE progress, it is important to emphasise that proper empirical evaluation and reporting is hard to do, requiring meticulous description of setup, dependent and independent variables, instructions given to subjects, etc., as well as painstaking analysis to avoid over-generalisation and other forms of misleading presentation of the findings. When proper experimental and reporting practice is followed, we encounter another characteristic of empirical studies in the NMIE field. It is that most controlled experimental setups include such a multitude of independent variables that the results are unlikely to generalise much. This point is comprehensively argued and illustrated for the general case of multimodal and natural interactive systems which include speech in [Bernsen 2002]. Still, as we tend to work on the basis of only slightly fortified hunches anyway, the results could often serve to inspire fellow researchers to follow them up. Thus, best-practice empirical studies are of major importance in guiding NMIE progress.

The empirical chapters in this book illustrate well the points made above except for the one on misleading presentation of findings. One cluster of findings demonstrate the potential of audio-visual speech output by animated talking heads for child language learning [Massaro] and, more generally, for improving intelligibility and efficiency of human-machine communication, including the substitution of facial animation for the, still-missing, prosody in current speech synthesis systems [Granström and House]. In counter-point, so to speak, Darrell et al. convincingly demonstrate the advantage of audio-visual *input* for tackling an important next step in speech technology, i.e. the recognition of multi-speaker spoken input. Jointly, the three chapters do a magnificent job of justifying the need for natural and multimodal (audio-visual) interaction independently of any psychological or social-psychological argument in favour of employing animated conversational agents.

A key question seems to be: for which purpose(s), other than harvesting the benefits of using audio-visual speech input/output described above, do we need to accompany spoken human-computer dialogue with more or less elaborate animated conversational interface agents [Dehn and van Mulken 2000]? By contrast with spoken output, animated interface agents occupy valuable screen real estate, do not necessarily add information of importance to the users of large classes of applications, and may distract the user from the task at hand. Whilst a concise and comprehensive answer to this question is still pending, it seems, Bickmore and Cassell go a long way towards explaining that the introduction of life-like animated interface agents into human-computer spoken dialogue is a tough and demanding proposition. As soon as an agent appears on the display, users tend to switch expectations from talking to a machine to talking to a human. By comparison, the finding in [Heylen et al.] that users tend to appreciate an animated cartoon agent more if it shows a minimum of human-like gaze behaviour might speak in favour of preferring cartoon-style agents over life-like animated agents because the former do not run the risk of facing our full expectations to human conversational behaviour. Cole et al. point out that, so far, studies of the effectiveness of animated agents have failed to reveal any significant improvement in user performance. However, they also stress the importance of continued system development to provide testbeds for further research and improvement of agents, identify missing knowledge, and assess the benefit of perceptive animated interfaces in learning. It may be added that user performance improvement is not a relevant criterion for all NMIE systems. In particular, the criterion does not obviously apply to entertainment systems.

On the multimodal side of the natural interactivity/multimodality semi-divide, several papers address issues of modality collaboration, i.e. how the use of modality combinations could facilitate, or even enable, human-computer interaction tasks that could not be done easily, if at all, using unimodal interaction. Carbonell and Kieffer report on how combined speech and graphics output can facilitate display search, and Corradini and Cohen show how the optional use of different input modalities can improve interaction in a particular virtual environment.

Using scenario-based use case analysis, i.e., a borderline case of empirical investigation which is normally undertaken in the very early stages of systems development, Sidner and Dzikovska reach out towards real conversational interaction. Bickmore and Cassell call their animated agent a conversational interface agent, Chai et al. claim that their system engages users in intelligent multimodal conversation, and Clark et al. work on conversational interaction. Cole et al. also talk about conversational agents but make clear that there is still a long way to go before spoken dialogue systems have human-like conversational skills. Similarly, Wilks et al. point out that the usual approaches to dialogue modelling produce systems which have far less than optimal conversational capability. We would like to expand on the use of the term 'conversational'. Despite the fact that the notion of embodied conversational agents seems to have reached canonical status in the NMIE community, real conversational systems are virtually non-existent today. Possibly because most NMIE practitioners come from research fields other than spoken dialogue systems, we sense a certain inflation in the use of the notion of conversation. A 'conversational system' tends to mean a system using speech input-output, even if it only understands spoken commands in a, say, 50 words vocabulary. Some require that, for the spoken dialogue to qualify as conversational, the dialogue should be mixed-initiative. However, mixed-initiative spoken dialogue, or even the ability to add a couple of small talk phrases during interaction, is not the same as conversational dialogue. Mixed initiative represents an important step beyond command-and-control dialogue, use of designer-designed keywords, system-directed dialogue, and user-directed dialogue. Still, today's mixed-initiative dialogue is, almost entirely, *task-oriented* spoken dialogue. Apart from the discussion of conversational dialogue in the chapters mentioned above, and the arguments for going beyond finite-state representations of dialogue structure in [Clark et al.] and a related discussion in [Wilks et al.], no real conversational dialogue is envisioned in this book. In general, however, human conversation is *not* task-oriented but, rather, meanders rhapsodically among different domains of discourse and rarely seeks to accomplish particular tasks. Since natural interactivity requires conversational dialogue, it would seem preferable to reserve the term 'conversation' for describing the real conversational spoken systems of the future, whether unimodal or multimodal. To develop these, we may not need new basic theory of conversation because there is already a proliferation of theories available. However, we are likely to need new theory at the design and implementation levels for how to manage the complexity of spoken conversation which goes far beyond that of task-oriented dialogue [Bernsen and Dybkjær 2004].

9. Coding natural interactive and multimodal data

It is perhaps not surprising that we are not very capable of predicting what people will do, or how they will behave, when interacting with computer systems using new modality combinations and possibly also new interactive devices. More surprising, however, is the fact that we are often just as ignorant when trying to predict natural interactive behaviours which we have the opportunity to observe every day in ourselves and others, such as: which kinds of gestures, if any, do people perform when they are listening to someone else speaking? This example illustrates that, to understand the ways in which people communicate with one another as well as the ways in which people communicate with the far more limited, current NMI systems, we need extensive studies of behavioural data. The study of data on natural and multimodal interaction is becoming a major research area full of potential for new discoveries.

To achieve stable and useful results on the behaviours involved in natural and multimodal interaction, we need, first, *high quality data.* Available natural interactive and multimodal data resources world-wide is reported in [Knudsen et al. 2002b]. First guidelines on how to handle, i.e., create, document, etc., natural interactive and multimodal data resources are presented in [Knudsen et al. 2003]. Second, we need *coding schemes* for all relevant classes of behavioural phenomena involved in natural and multimodal interaction. A report on available natural interactive and multimodal coding schemes world-wide is [Knudsen et al 2002a]. The report shows the need for a large variety of new NMIE coding schemes. First guidelines on how to handle, i.e., create, document, etc., natural interactive and multimodal coding schemes are proposed in [Dybkjær et al. 2003]. Thirdly, as data coding by-hand is costly and time-consuming, we need general-purpose *coding tools* which can facilitate the coding and analysis of all or most aspects of natural and multimodal interactive behaviour. A first report on

available natural interactive and multimodal coding tools world-wide is [Dybkjær et al. 2001], see also the Eurospeech 2003 update at [http://nite.nis.sdu.dk/eurospeech/tutorialslides]. The report shows that there is no general-purpose coding tool available yet for coding and analysing all or most aspects of natural and multimodal interactive behaviour.

A number of chapters make use of, or refer to, data resources for NMIE, but none of them take a more general view on data resource issues. One chapter addresses NMIE needs for new coding schemes. Martell presents a new, kinematically-based gesture annotation scheme for capturing the kinematic information in gestures from videos of speakers. Linking the urgent issue of new, more powerful coding tools with the equally important issue of standardisation, Martell proposes a standard for the internal representation of NMIE codings.

10. Improving enabling technologies

An enabling technology is often developed over a long time by some separate community, such as by the speech recognition community from the 1950s to the late 1980s. Having matured to the point at which practical applications become possible, the technology transforms into an omnipresent tool, as is the case with speech recognition technology today. NMIE needs a large number of enabling technologies and these are currently at very different stages of maturity. Several enabling technologies, some of which are at an early stage and some of which are finding their way into useful applications, are presented in this book in the context of application to NMIE problems, including robot interaction and agent technology, multi-speaker interaction and recognition, machine learning, and talking face technology.

Sidner and Dzikovska focus on robot interaction in the general domain of “hosting”, i.e., where a virtual or physical agent provides guidance, education, or entertainment based on collaborative goals negotiation and subsequent action. A great deal of work remains to be done before robot interaction becomes natural in any approximate sense of the term. For instance, the robot’s spoken dialogue capabilities must be strongly improved and so must its embodied appearance and global communicative behaviours. In fact, Sidner and Dzikovska make some of the same conclusions as Bickmore and Cassell, namely that agents need to become far more human-like in all or most respects before they are really appreciated by humans. Cole et al. envision that it will be possible to build lifelike characters in a near future that interact with people much like people interact with each other despite the research advances required to realise this vision and despite sparse evidence that animated agents can improve human-computer interaction.

Darrell et al. address the problem in multi-speaker interaction of knowing who is addressing the computer when. Their approach is to use a microphone array combined with computer vision to find out who is talking to the computer. In-car application developers are faced with the problem of not only deciding when the driver is speaking as opposed to one of the passengers, but also when the driver is addressing the system rather than a passenger. Some applications use a push-to-activate button to partly overcome the latter problem.

Developers of spoken dialogue applications must cope with problems resulting from vocabulary and grammar limitations and from difficulties in enabling much of the flexibility and functionality inherent in human-human communication. Despite having carried out systematic testing, the developer often finds that words are missing when a new user addresses the application. Dusan and Flanagan propose machine learning as a way to overcome part of this problem. Using machine learning, the system can learn new words and grammars taught to it by the user in a well-defined way. Wilks et al. address machine learning – or transformation-based learning - in the context of assigning dialogue acts as part of an approach to improved dialogue modelling. As another part of their approach, Wilks et al. consider the use of dialogue action frames, i.e., a set of stereotypical dialogue patterns which perhaps may be learned from corpus data, as a means for flexibly switching back and forth between topics during a dialogue.

Granström and House and Massaro describe the gain in intelligibility that can be obtained by combining speech synthesis with a talking face. There is still much work to do both on synthesis and face articulation. For most languages, speech synthesis is still not very natural to listen to and if one wants to develop a particular voice to fit a certain animated character, this is not immediately possible

with today's technology. With respect to face articulation, faces need to become much more natural in terms of, e.g., gaze, eyebrow movements, lip and mouth movements, and head movements, as this seems to influence users' perception of the interaction [Granström and House, Heylen et al.].

An important enabling technology for NMIE which is not mentioned in this book is audio-visual speech recognition. It is not only the human perception of speech which is improved by visual cues. The same seems to be true for computers. Thus, audio-visual speech recognition may help improve speech recognition. Another important enabling technology is prosody recognition. Work on prosody recognition has been going on for decades at what appears to be a rather slow pace, and we are still far from being able to harness the technology for NMIE purposes. Many cues in the speech input signal will continue to be lost as long as recognisers cannot cope with prosody, which propagates important losses in the naturalness of the system's dialogue behaviour. Computer vision and multi-party speech recognition are two other enabling technologies for NMIE which need further progress.

11. Building more advanced systems

Enabling technologies for NMIE are often component technologies, and their description, including state of the art, current research challenges, and unsolved problems, can normally be made in a relatively systematic and focused manner. It is far more difficult to systematically describe the complexity of the constant push in research and industry towards exploring and exploiting new NMIE application types and new application domains, addressing new user populations, increasing the capabilities of systems in familiar domains of application, exploring known technologies with new kinds of devices, etc. In general, the picture is one of pushing present boundaries in most directions. However, we do seem to spot surprisingly underdeveloped areas in research and development, i.e. areas in which, e.g., the enabling technologies seem to be in place and the user (or consumer) interest is strong but where little is happening nevertheless. During the past few years, a core trend in NMIE has been to combine different modalities in order to build more complex, versatile and capable systems, getting closer to natural interactivity than is possible with only a single modality. This trend is reflected in several chapters in this book.

Part of the NMIE paradigm is that systems must be available whenever and wherever convenient and useful, making ubiquitous computing an important application domain. Mobile devices, such as mobile phones, PDAs, and portable computers of any (portable) size have become popular and are rapidly gaining functionality. However, the interface and interaction affordances of small devices require careful consideration. Reithinger et al. present some of those considerations in the context of providing access to large amounts of data about music.

It can be difficult for users to know how to interact with new NMIE applications. Although not always very successful in practice, the classical GUI system has the opportunity to present its affordances in static graphics (including text) before the user chooses how to interact. A speech-only system, by contrast, cannot do that because of the dynamic and transitory nature of acoustic modalities. NMIE systems, in other words, pose radically new demands on how to support the user prior to, and during, interaction. Addressing this problem, several chapters mention user modelling or repositories of user preferences built on the basis of interactions with a system [Chai et al., Reithinger et al.]. User modelling may be done via information acquired either off-line or on-line during interaction. In the latter case, the user model may be updated on the fly or only between interactions. In any case, the idea is to enable more natural interaction based on the system's knowledge about the user's preferences, habits, etc. Machine learning, although another example of less-than-expected pace of development during the past 10 years, has great potential for increasing interaction support. In an advanced application of machine learning, Dusan and Flanagan propose to increase the system's vocabulary and grammar by letting users teach the system new words and their meaning and use. Wilks et al. use machine learning as part of an approach to more advanced dialogue modelling.

Increasingly advanced systems require increasingly complex dialogue management, cf. [Chai et al., Clark et al., Wilks et al.]. As discussed in Section 8, it is an exaggeration to call most existing spoken or spoken-cum-animated dialogue system 'conversational'. Real conversational dialogue is hinted at in [Sidner and Dzikovska, Bickmore and Cassell, Chai et al, Clark et al.] and envisioned by [Cole et al.]. Like life-likeness of animated interface agents, conversational dialogue is among the key challenges in achieving the NMIE vision.

Multilinguality of systems is an important goal which is not merely one of adding speech and language processing for different languages. Research also needs to overcome unsolved issues, such as language recognition, user modelling of the user's preferred language, the enormous challenge of recognising cross-language pronunciation variants, distributed speech recognition for limited-power devices, etc., which are beyond the scope of this book. Multilingual applications are addressed in [Reithinger et al.]. In their case, the application is running on a handheld device.

Multi-speaker input speech is mentioned by [Chai et al., Darrell et al.]. For good reason, recognition of multi-speaker input has become a lively research topic. We need solutions in order to, e.g., build meeting minute-takers, separate the focal speaker's input from that of other speakers, exploit the huge potential of spoken multi-user applications, etc.

12. Building systems easily

Due to the complexity of multimodal natural interaction it is becoming dramatically important to be able to build systems as easily as possible. It seems likely that no single research lab or development team in industry, even including giants such as Microsoft, is able to master all of the enabling technologies required for broad-scale NMIE progress. To advance efficiently, everybody needs access to those system components, and their built-in know-how, which are not in development focus. This implies strong attention to issues, such as re-usable platforms, components and architectures, development toolkits, interface languages, data formats, and standardisation.

Cole et al. mention various research tools in support of developing perceptive animated interfaces, including the DARPA Communicator hub (http://fofoca.mitre.org/) which supports a modular plug-and-play approach, the CU Conversational Agent Toolkit for developing advanced dialogue systems, and CU Animate which supports the development and rendering of 3D animated characters. Clark et al. have used the Open Agent Architecture (OAA, http://www.ai.sri.com/~oaa/), a framework for integrating heterogeneous software agents in a distributed environment. What OAA and other architectural frameworks, such as CORBA (http://www.corba.org/), aim to do is provide a means for modularisation, synchronous and asynchronous communication, well-defined inter-module communication via some interface language, such as IDL (CORBA) or ICL (OAA), and the possibility of implementation in a distributed environment.

XML (Extensible Markup Language) is a simple, flexible text format derived from SGML (ISO 8879) which is becoming popular as, among other things, a message exchange format, cf. [Reithinger et al., Sidner and Dzikovska]. Using XML for wrapping inter-module messages is one way to overcome the problem of different programming languages used for implementing different modules. XML is becoming a standard for data exchange. It is one of the activities in which W3C (the World Wide Web Consortium) (http://www.w3.org/) is involved.

Some chapters express a need for reusable components. Many of the applications described include off-the-shelf software, including components developed in other projects. This is particularly true for mature enabling technologies, such as speech recognition and synthesis components. As regards multimodal dialogue management, there is an expressed need for reuse in, e.g., [Clark et al.] who discuss a reusable dialogue management architecture in support of multimodal interaction.

In conclusion, there are architectures, platforms, and software components available which facilitate the building of new NMIE applications, and standards are underway for certain aspects. There is still much work to be done on standardisation, new and better platforms, and improvement of component software. In addition, we need, in particular, more and better toolkits in support of system development and a better understanding of those components which cannot be bought off-the-shelf and which are typically difficult to reuse, such as dialogue managers. Advancements such as these are likely to require significant corpus work. Corpora with tools and annotation schemes as described by [Martell] are exactly what is needed in this context.

13. Evaluation

Software systems and components evaluation is a broad area, ranging from technical evaluation over usability evaluation to customer evaluation. Customer evaluation has never been a key issue in research but has rather tended to be left to the marketing departments of companies. Technical

evaluation and usability evaluation, including evaluation of functionality from both perspectives, are, on the other hand, considered important research issues. The chapters show a clear trend towards focusing on usability evaluation and comparative performance evaluation.

Comparative performance evaluation objectively compares users' performance on different systems with respect to, e.g., how well they understand speech-only versus speech combined with a talking face or with an embodied animated agent [Granström and House, Massaro, Bickmore and Cassell]. The usability issues evaluated all relate to users' perception of a particular system and include parameters, such as life-likeness, credibility, reliability, efficiency, personality, ease of use, and understanding quality [Heylen et al., Bickmore and Cassell].

It is hardly surprising that performance evaluation and usability issues are considered key topics today. We know little about what happens when we move towards increasingly multimodal and natural interactive systems, both as regards how these new systems will perform compared to alternative solutions and how the systems will be received and perceived by their users. We only know that a technically optimal system is not sufficient to guarantee user satisfaction.

Two chapters address how the intelligibility of what is being said can be increased through visual articulation [Granström and House, Massaro]. Granström and House have used a talking head in several applications, including tourist information, real estate (apartment) search, aid for the hearing impaired, education, and infotainment. Evaluation shows a significant gain in intelligibility for the hearing impaired. Eyebrow and head movement enhance perception of emphasis and syllable prominence. Over-articulation may be useful as well when there are special needs for intelligibility. The findings of [Massaro] support these promising conclusions. His focus is on applications to the hard-of-hearing, children with autism, and child language learning more generally. Granström and House also address the increase in efficiency of communication/interaction produced by using an animated talking head. Probably, naturalness is a key point here. This is suggested by [Heylen et al.] who made controlled experiments on the effects of different eye gaze behaviours of a cartoon-like talking face on the quality of human-agent dialogues. The most human-like agent gaze behaviour led to higher appreciation of the agent and more efficient task performance.

Bickmore and Cassell evaluate the effects on communication of an embodied conversational real-estate agent versus an over-the-phone version of the same system, cf. also [Cassell et al. 2000]. In each condition, two variations of the system was available. One would be fully task-oriented while the second version would include some small-talk options. In general, users liked the system better in the phone condition. In the phone condition, subjects appreciated the small-talk while, in the embodied condition, subjects wanted to get down to business. The implication is that agent embodiment has strong effects on the interlocutors. Users tend to compare their animated interlocutors with humans rather than machines. To work with users, animated agents need considerably more naturalness and personally attractive features communicated non-verbally. This imposes a tall research agenda on both speech and non-verbal output, requiring conversational abilities both verbally and non-verbally.

Jointly, the chapters on evaluation demonstrate a broad need for performance evaluation, comparative as well as non-comparative, that can inform us on the possible benefits and shortcomings of new natural interactive and multimodal systems. The chapters show a similar need for usability evaluation that can help us find out how users perceive these new systems, and a need for finding ways in which usability and user satisfaction might be correlated with technical aspects in order for the former to be derived from the latter.

The chapters do not address technical evaluation apart from comparative performance evaluation of certain parameters. Several ongoing research projects have as part of their agenda to look into evaluation methods for various aspects of natural interactive and multimodal dialogue systems, e.g.: INSPIRE, Infotainment management with speech interaction via remote-microphones and telephone interfaces, 2002-2004, (http://www.inspire-project.org), looks at usability and acceptability evaluation; MIAMM, Multidimensional Information Access using Multiple Modalities, 2001-2004, (http://www.loria.fr/projets/MIAMM), [Reithinger et al.] looks at evaluation methods and protocols for multimodal interaction; and NICE, Natural Interactive Communication for Edutainment, 2002-2005, (http://www.niceproject.com), looks at new ways of evaluating natural human-system interaction. It would seem timely to establish a high-profile, community-wide project which could address best practice in evaluation for natural and multimodal interactivity engineering at a global

level, perhaps along lines similar to those of the DISC project (http://www.disc2.dk) which focused on spoken dialogue systems engineering, and including aspects of usability evaluation inspired by the approach taken in the DARPA Communicator project which also addressed spoken dialogue systems.

14. Future needs of natural and multimodal interactivity engineering

Based on Sections 6 through 13, Table 1.3 presents conclusions on important research needs and working technologies for NMIE. NMIE is a huge area, and Table 1.3 is no doubt inadequate in several ways. It is rather global or coarse-grained, inviting expansion of every entry to provide more detail and stressing the need to collect or create sub-area-specific roadmaps. In the speech area, for instance, the ELSNET (European Language and Speech Network) roadmap provides far more detail than does Table 1.3 [Bernsen 2001, 2002]. Secondly, Table 1.3 is inevitably partial due to the authors' limited experience in NMIE. Their experience includes aspects of spoken dialogue systems, animated interface agent interaction, on-line user modelling, NMIE data resources, coding schemes, and coding tools, gesture input, systems and component evaluation, modality theory, and technology forecasting, brainstorming and roadmapping, but little hands-on experience with, e.g., computer vision.

Specific to NMIE	**Current NMIE needs**
Visions, roadmaps, etc., general and per sub-area	We need to collect, integrate, and iterate detailed NMIE roadmaps, timeline them, and do those which are missing.
Applicable theory for any aspect of NMIE	Encourage and support development of more applicable theory on NMIE interaction, systems, components, and evaluation.
Controlled experiments, behavioural studies, simulations, scenario studies, task analysis on: roles of, and collaboration among, specific modalities to achieve various benefits	More of the same: As many empirical results as we can get. Investigation of new modality combinations for new users, new environments, new applications, etc. Arguably, increase awareness of best practice in conducting, analysing and reporting empirical findings.
New quality data resources, coding schemes, coding tools, and standards	More high-quality NMIE data resources: well-documented, re-usable, easy to find on the web, free for research purposes, based on standards. More consolidated NMIE coding schemes, created and documented according to standards. General-purpose coding tool(s) for multilevel, cross-level, and cross-modality NMIE data coding. Stronger awareness of the area's importance.
New basic technologies needed	Advances in computer vision in order to track, identify, recognise, and interpret users and their communicative behaviours, including lip movements for audio-visual speech, facial expression, gesture, body posture, object manipulation, the physical environment, emotions, personality, etc. Prosody recognition and synthesis. Multi-speaker speech recognition, audio-visual speech recognition. Machine learning for adaptation, language learning, etc.
New, more complex, versatile, and capable system aspects	Applications for small mobile devices. Situation awareness. Easy production of flexible human-like graphical interface agent behaviours. Real conversational spoken and multimodal dialogue systems. Edutainment and entertainment applications.
Re-usable platforms, components, toolkits, architectures, interface	Easy transfer of all relevant NMIE progress to web applications. Freeware, open source, and other versatile plug-and-play platforms for

languages, standards, etc.	NMIE. More re-usable components, component interface standardisation. Freeware, open source, and other development toolkits for NMIE. Extension of existing architectures to full NMIE capability, including, e.g., stable generic models for input fusion and output fission.
Evaluate components, systems, technologies, processes, etc.	More knowledge on the usability of different modality combinations. More knowledge on the parameters behind user satisfaction to enable better prediction of user satisfaction. Better methods for usability evaluation. Technical evaluation parameters for natural interactive and multimodal systems.

Table 1.3. Current research needs for natural and multimodal interactivity engineering.

References

Papers

Bernsen, N. O. (Ed.): Speech-related Technologies. Where will the Field go in 10 Years? ELSNET brainstorming document v.4, March 2001, www.elsnet.org/roadmap.html. Also in Proceedings of the Machine Translation Roadmap Workshop (TMI-2002), Keihanna, Japan, 2002. Utrecht: The European Language and Speech Network (ELSNET), 2002, 29-47.

Bernsen, N. O.: Multimodality in Language and Speech Systems - From Theory to Design Support Tool. In Granström, B. (Ed.): Multimodality in Language and Speech Systems. Dordrecht: Kluwer Academic Publishers 2002.

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

Bernsen, N. O. and Dybkjær, L.: Domain-oriented Conversation with H.C. Andersen. In André, E., Dybkjær, L., Heisterkamp, P. and Minker, W. (Eds.): Affective Dialogue Systems. Proceedings of the Irsee Tutorial and Research Workshop on Affective Dialogue Systems, LNCS 3068, Springer Verlag, 2004.

Bickmore, T. and Cassell, J.: Social Dialogue with Embodied Conversational Agents.

Carbonell, N. and Kieffer, S.: Do Oral Messages Help Visual Search?

Cassell, J., Sullivan, J., Prevost, S., and Churchill, E. (Eds.): Embodied conversational agents. Cambridge, MS: MIT Press 2000.

Chai, J. Y., Pan, S. and Zhou, M. X.: MIND: A Context-based Multimodal Interpretation Framework for Conversational Systems.

Clark, B., Lemon, O., Gruenstein, A., Bratt, E. O., Fry, J., Peters, S., Pon-Barry, H., Schultz, K., Thomsen-Gray, Z. and Treeratpituk, P.: A General Purpose Architecture for Intelligent Tutoring Systems.

Cole, R., van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W. and Yan, J.: Perceptive Animated Interfaces: First Steps towards a New Paradigm for Human-Computer Interaction.

Corradini, A. and Cohen, P. R.: On the Relationships Among Speech, Gestures, and Object Manipulation in Virtual Environments: Initial Evidence.

Darrell, T., Fisher, J., Wilson, K. and Siracusa, M.: Geometric and Statistical Approaches to Audiovisual Segmentation.

Dehn, D. and van Mulken, S.: The Impact of Animated Interface Agents: A Review of Empirical Research. Int. Journal of Human-Computer Studies 52, 2000, 1-22.

Ducatel, K., Bogdanowicz, M., Scapolo, F., Leijten, J. and Burgelman, J.-C.: Scenarios for Ambient Intelligence in 2010. Draft Final Report Version 2. IPTS, Seville, Spain, 2001.

Dusan, S. and Flanagan, J.: Adaptive Human-Computer Dialogue.

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 (International Standards for Language Engineering) Report D11.1, January 2001, isle.nis.sdu.dk

Dybkjær, L., Bernsen, N. O., Knudsen, M. W., Llisterri, J., Machuca, M., Martin, J.-C., Pelachaud, C., Riera, M. and Wittenburg, P.: Guidelines for the Creation of NIMM (Natural Interactivity and Multimodality) Annotation Schemes. ISLE (International Standards for Language Engineering) Report D9.2, February 2003, isle.nis.sdu.dk

Granström, B. and House, D.: Effective Interaction with Talking Animated Agents in Dialogue Systems.

Heylen, D., van Es, I., Nijholt, A. and van Dijk, B.: Controlling the Gaze of Conversational Agents.

Knudsen, M. W., Bernsen, N. O., Dybkjær, L., Hansen, T., Mapelli, V., Martin, J.-C., Paulsson, N., Pelachaud, C., and Wittenburg, P.: Guidelines for the Creation of NIMM (Natural Interactivity and Multimodality) Data Resources. ISLE (International Standards for Language Engineering) Report D8.2, February 2003, isle.nis.sdu.dk

Knudsen, M. W., Martin, J.-C., Dybkjær, L., Ayuso, M. J. M, N., Bernsen, N. O., Carletta, J., Kita, S., Heid, U., Llisterri, J., Pelachaud, C., Poggi, I., Reithinger, N., van ElsWijk, G. and Wittenburg, P.: Survey of Multimodal Annotation Schemes and Best Practice. ISLE (International Standards for Language Engineering) Report D9.1, 2002a, isle.nis.sdu.dk

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 (Natural Interactivity and Multimodality) Data Resources, Current and Future User Profiles, Markets and User Needs for NIMM Resources. ISLE (International Standards for Language Engineering) Report D8.1, 2002b, isle.nis.sdu.dk

Martell, C.: FORM: An Extensible, Kinematically-based Gesture Annotation Scheme.

Massaro, D. W.: The Psychology and Technology of Talking Heads: Applications in Language Learning.

Reithinger, N., Fedeler, D., Kumar, A., Lauer, C., Pecourt, E. and Romary, L.: MIAMM – A Multi-Modal Dialogue System Using Haptics.

Sidner, C. L. and Dzikovska, M.: A First Experiment in Engagement for Human-Robot Interaction in Hosting Activities.

Wilks, Y., Webb, N., Setzer, A., Hepple, M. and Catizone, R.: Machine Learning Approaches to Human Dialogue Modelling.

Websites

CORBA: http://www.corba.org/

DARPA Communicator project: http://fofoca.mitre.org

DARPA Tactical Language Training Project: http://www.isi.edu/isd/carte/proj_tactlang/

DISC project: www.disc2.dk

ELSNET roadmaps: www.elsnet.org/roadmap.html

Eurospeech tutorial on natural interactivity and multimodality coding tools: http://nite.nis.sdu.dk/eurospeech/tutorialslides

INSPIRE project: http://www.inspire-project.org

ISLE project: http://isle.nis.sdu.dk

MIAMM project: http://www.loria.fr/projets/MIAMM

NICE project: http://www.niceproject.com

Open Agent Architecture: http://www.ai.sri.com/~oaa/

SIMILAR: http://www.similar.cc

W3C, World Wide Web Consortium: http://www.w3.org/