Natural Language Engineering **6** (3–4): 205–212. Printed in the United Kingdom

Best practice in spoken language dialogue systems engineering
Introduction to the special issue

JAN VAN KUPPEVELT,
ULRICH HEID and HANS KAMP
Institute for Natural Language Processing (IMS)
University of Stuttgart
Azenbergstrasse 12
70174 Stuttgart, Germany
e-mail: {kuppevelt,uli,hans}@ims.uni-stuttgart.de

(*Received November 2000*)

1 Introduction

This special issue brings together representative views on what has come to be known as "best practice" in the development and evaluation of spoken language dialogue systems (SLDSs). The issue was initiated in the context of the European Esprit project DISC, which ran from June 1997 till February 2000. DISC's main goal was to identify current practice in both the development and the evaluation of SLDSs, in order to arrive at a useful definition and description of best practice. The project has resulted in a collection of guidelines which are intended for different target groups, in particular developers, deployers and customers.[1]

The last few years the interest in SLDSs has increased enormously. At present there is a large number of systems available, many of them for commercial use. Their number is growing rapidly, and so are the variety of their functionalities and the diversity of their application domains. The tasks that advanced systems are able to perform are often more complex, less stereotypical, and are often carried out in the context of several interconnected domains of application. With these advances have come higher expectations of the naturalness and intelligence with which SLDSs fulfill their assignments, and as a consequence the interest in such systems has even grown more, both within academic and commercial circles. As far as natural human-system interaction is concerned, one significant change in SLDS design concerns the interaction between natural language understanding and dialogue management. Here we see a clear tendency towards models that incorporate a substantial amount

[1] DISC partners were: Natural Interactive Systems Laboratory, Odense University, Denmark (coordination); Department of Speech, Music and Hearing (KTH), Stockholm, Sweden; Human-Machine Communication Department, CNRS-LIMSI, Orsay, France; Institute for Natural Language Processing (IMS), University of Stuttgart, Germany; Vocalis Ltd, Cambridge, United Kingdom; DaimlerChrysler Research Center Ulm, Germany; and the ELSNET foundation, Utrecht, The Netherlands.

of discourse semantics and make use of some conception of context-change. This allows for more natural interactions between the system and its human users, due on the one hand to the system's improved ability to compute the intended meaning of the user's input and on the other to the increased sophistication of the strategies it uses for planning its own responses. Such improved capacities are crucial when the system is to leave more of the initiative to the user, instead of keeping the dialogue on a narrowly circumscribed path of largely predictable exchanges. Further, there is a tendency to combine spoken language human-system interaction with other modalities of information exchange and representation (e.g., images and gestures), asking for both modality-specific and modality-integrating syntactic and semantic processing capabilities. All these developments have led to a situation in which there is a great need, shared by developers, deployers and customers alike, for effective guidelines, which will enable them to make accurate and successful design and implementation decisions, in accordance with broad consensus of what must be best practice in this particular engineering domain.

The growing need for best practice guidelines concerning the development and evaluation of SLDSs has three aspects which are fairly closely related, but which it is all the more important to distinguish. The first is the need for *access to the state of the art*. The enormous growth of information about SLDSs of the past years has led to a situation where it has become very difficult to obtain a comprehensive picture of the field. Developers, deployers and customers, who wish to know about the range of options for design, implementation and evaluation of a system feel this need acutely. The very first need here is for a description of current practice that provides a general survey of the spectrum of existing dialogue engineering options that make it possible for the one who needs or wants to develop any particular type of SLDS to get a clear idea of what to look for and where to look for it.

The second aspect is the need for *quality control*. The thicket of different technical designs is making it ever harder for those who want to put an SLDS together to find optimal solutions to their specific engineering problems. What is required here is a best practice definition in terms of effective guidelines which make it possible to select from among the existing design, implementation and evaluation options the ones which are best suited to the particular demands and constraints that come with any given application.

The third aspect is *economic control*. The development of SLDSs tends to be time-consuming and expensive, and it is important to have efficient ways of minimizing both costs and time. In this connection it is of great importance that existing components and design know-how can be reused in new systems which differ in their over-all system task and domain of application from those for which the components were originally built or which led to the acquisition of the know-how. An example are labour-intensive design techniques like the Wizard of Oz method for collecting data about the ways in which end users behave in various possible dialogue situations. The experience that has been gathered in applying these methods now often enables developers to proceed much more efficiently than would have been possible before, and by delving into the vast and growing depository of established results they can sometimes find what they need to know without having to collect

the data themselves. Generally there has been a growing tendency in SLDS design and development to redeploy existing resources – from generic architectures and common toolkits to an array of techniques and components aimed at very specific goals. Current discussions within the SLDS community reveal the importance which it attaches to these possibilities.[2]

Within the domain of SLDS evaluation we can observe a similar concern for "reusability". Here the focus is on common tools of evaluation, which are applicable to a variety of SLDS types and, which – this is a point to which advanced SLDS evaluation models pay particular attention – look at systems within the broader context of the human (or non-human) environment in which they are to perform. Important factors which these evaluation tools take into account are the knowledge, communicative preferences and other "quirks" of the particular user group(s) for which the system is intended. Of special importance is the problem solving task of the user, which leads him to consult the system.

As far as SLDS development is concerned, best practice methodology and the general guidelines it has yielded aim at optimizing the performance of systems and system components, given a specification of developmental goals and constraints (for example, overall system performance goals, developmental constraints imposed by developer, customer and user preferences, as well as constraints imposed by costs and available resources in terms of reusability). One purpose of such general guidelines is to make those who wish to acquire or develop a new SLDS more sharply aware of the particular things they want from it and of the constraints imposed by the environment in which it will have to work. Moreover, once these parameters have been defined, the guidelines should point them towards those of the more specific environment-, platform-, and application-specific guidelines which fit their case. The notion of best practice as it is defined by guidelines at this general level is an abstract and delicate one. Best practice in this sense is something that varies; what constitutes best practice in one design, development or evaluation situation won't be the same as what constitutes best practice in another. For instance, systems performing simple tasks can often manage without semantic analyses of the inputs they receive, but when the tasks get more complicated or the repertoire of possible in- and outputs more diverse, semantic analysis may prove indispensable; or, for another example, with growing task complexity it may be necessary to control the input, and thus the course and form of the dialogue, more carefully, something which will require more sophisticated dialogue management components; systems that are meant to work in noisy environments will need special filters to support speech recognition; and so on.

That best practice is variable in this way – that it works like a mapping from functional parameters to parameters of design and development – is an unsurprising banality – this kind of dependence is after all what we find in almost any kind of engineering discipline. Still, it may be a point that needs stressing. But what really matters is to determine exactly what the mapping is like, and how its salient

[2] Cf., e.g., *Proceedings of the First ACL SIGdial Workshop on Discourse and Dialogue*, Hong Kong, October 7-8, 2000 (URL SIGdial: http://www.sigdial.org).

properties are best explained to a broad spectrum of laymen and professionals who find themselves confronted with the problem of getting an SLDS that answers their needs.

In the next section of this introduction we sketch, in the briefest possible terms, the best practice methodology followed and expanded within DISC. In Section 3 follow short descriptions of the contributions to this issue, with a focus on the contributions they make to the issues of best practice.

2 The DISC Best Practice Dialogue Engineering Model

The DISC Best Practice Dialogue Engineering Model (Bernsen 1999, van Kuppevelt and Heid 1999)[3] is a best practice evaluation methodology comprising both system or system component evaluation and the evaluation of the system's (or system component's) life cycle. The DISC best practice model is an empirical, "bottom-up" model which, in contrast to for example EAGLES methodology, bases the evaluation of SLDSs and their components on a description of the current practice.[4] The development of the model consisted in a three-stage process which resulted in three interrelated levels: (i) a descriptive level, (ii) a non-comparative evaluation level, and (iii) a best practice level.[5]

The descriptive part of the model which was developed first, consists in a systematic description of the current practice of SLDSs. A variety of exemplars was used for the analysis of current systems and main components or aspects of these systems. In total 26 analyses of exemplars/aspects were carried out. Data collection was carried out by making use of various information sources, mainly available documentation (both published material and developers' internal documentation), site visits and telephone interviews. The data were analyzed according to two description types: (i) system (component) description by means of a detailed set of questions (called "grid questions" in DISC) concerning the technological choices made in the design and implementation of each system aspect and (ii) system development description by means of a detailed set of "life cycle questions" concerning the procedures and ways in which the development including the evaluation at different steps was carried out. Both descriptions are based on earlier work by (Bernsen, Dybkjaer and Dybkjaer 1998).[6] The analysis of technological options used in SLDSs makes use of

[3] Bernsen N. 0. (1999), *Working Paper on Dialog Management Evaluation*, DISC deliverable D3.10; van Kuppevelt J. and Heid U. (1999), *From a Description of Spoken Language Dialogue Systems to their Evaluation and Best Practice*, DISC Deliverable D3.8b.

[4] The EAGLES evaluation model can be characterized as a theoretical model (cf., e.g., King M. and Maegaard B. (1998), Issues in Natural Language Systems Evaluation, *Proceedings of the First International Conference on Linguistic Resources and Evaluation*, Granada, pp. 225-230). It stipulates a set of seven evaluation criteria, called quality characteristics. In order to obtain a specific evaluation model applying to a particular kind of language engineering product, the seven evaluation criteria must be extended with subcriteria on various levels. The choice, organization and relative importance of these subcriteria define the evaluation model for the class of language engineering products under consideration.

[5] Detailed information on the DISC Best Practice Dialogue Engineering Model and its results can be found at the DISC best practice website (URL: http://www.disc2.dk).

[6] Bernsen N. O., Dybkjaer H. and Dybkjaer L. (1998), *Designing Interactive Speech Systems: From First Ideas to User Testing*, Springer Verlag, Berlin/Heidelberg.

questionnaires for each of the system aspects described in their Speech Interaction Model. The life cycle questionnaire has grown out of the Software Engineering Life Cycle Model by the same authors. The first stage in the development of the model resulted in a description of the current practice in terms of current issues in the design, implementation and evaluation of SLDSs, as well as in a description of related dialogue engineering options.

The second, non-comparative evaluation level which takes the results of the descriptive part as input is two-fold: (i) the determination of best practice evaluation criteria based on current practice, and (ii) the evaluation of design, technological and evaluation options related to a given issue in terms of (constraints-dependent) advantages and limitations ("pros" and "cons"). Best practice evaluation criteria were defined by evaluation questions; these help assess the adequacy of the technological and system development solutions identified thus far. A key element of the DISC best practice model is the possibility of constrained evaluation: a given technological or design option is evaluated against the background of possible development objectives and constraints on the development process.

Results of the second level, in turn, form the input of the third, best practice level. This level contains the best practice methodology. It is defined in terms of a comparison of language engineering options on the basis of the pros and cons which were assigned to them on the second, non-comparative evaluation level. The result is an evaluation scale representing an ordering on the set of technological options under discussion.

3 Best practice in this special issue

The special issue of NLE presents nine articles on best practice in SLDS development and evaluation. Though some of them deal with certain matters which do not concern best practice directly, all have been selected because of the contribution they make to SLDS best practice issues. In fact, we, the guest editors, believe that as regards such issues, the present collection is representative of current views and discussions within the field of SLDS development and evaluation.

The special issue opens with the article *An architecture for a generic dialogue shell* by Allen, Byron, Dzikoska, Ferguson, Galescu and Stent. The paper which directly contributes to best practice anticipates future needs in the performance of dialogue systems. It proposes an architecture for a generic dialogue shell for unrestricted natural conversation in terms of a reduction of limitations on user's options in the interaction, within the large domain of practical dialogues that concentrate on the accomplishment of an objective or the performance of a task. On the basis of their experience in developing dialogue systems in different domains, the authors believe that the development of a generic dialogue shell for practical dialogues is feasible, both technologically and commercially. It is hypothesized that (i) dialogues belonging to this category do not require the full conversational competence needed for general human conversation, and (ii) most of the complexity in natural language understanding and dialogue management is domain-independent.

The second article *User-guided system development in interactive spoken language*

education by Atwell, Baldo, Bisiani, Bonaventura, Herron, Howarth, Menzel, Morton, Pezzotta, Schmidt and Souter contributes to best practice in an expanding subdomain of spoken language dialogue systems, namely second language learning systems. The article presents specific results of the European project ISLE (Interactive Spoken Language Education). While the paper does not so much discuss technical issues and options, it focuses on the improvement of the performance of second language tutorial systems in terms of their design. Based on consumer and market research, it provides an extensive specification of user requirements for this growing class of systems.

Central to the article *Usability issues in spoken dialogue systems* by Dybkjaer and Bernsen is SLDS usability best practice. Following the principles of the DISC best practice methodology, the authors show the relevance of a systematic understanding of factors which optimize SLDS usability. The space of usability is divided into eleven issues, on the basis of which they propose evaluation criteria for SLDS usability. Their target group are developers, who can make use of these criteria during requirement specification, design, development and evaluation of SLDSs and their components. The authors draw attention to limitations of current usability practice.

The next article *Speech technology on trial: experiences from the August system* by Gustafson and Bell forms an example of SLDS description. The system described is the experimental Swedish spoken dialogue system "August". The authors had placed August, an animated agent, in the streets of Stockholm, and they collected the dialogues of inexperienced users. The analysis of this material leads to new insights with respect to SLDS design.

The article *Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI)* by Hone and Graham aims at the development of a structured and statistically valid tool for the subjective evaluation of speech interfaces, to be used in addition to objective evaluation techniques. The intended tool will contribute to best practice user-oriented design and evaluation of speech systems. The tool makes use of six main factors in user's perception of speech systems, obtained statistically by applying exploratory factor analysis to the results of a questionnaire which was given to users of four different speech systems.

A global overview of best practice in speech recognition is presented in *Towards best practice in the development and evaluation of speech recognition components of a spoken language dialogue system* by Lamel, Minker and Paroubek. The paper which is intended to provide a summary of the state of the art in speech recognition follows and exemplifies the DISC Best Practice Dialogue Engineering Model, starting with a presentation of key issues in current speech recognition practice. For each issue, the article discusses different speech technology options and evaluates them in terms of their advantages and limitations. The paper discusses speech recognition grid issues (see above) concerning relevant aspects of the speech recognizers themselves, speech recognition life cycle issues referring to crucial steps in the development of speech recognizers, and speech recognition evaluation issues concentrating on fundamental aspects in the evaluation of speech recognition components.

The article *Information state and dialogue management in the TRINDI dialogue*

move engine toolkit by Larsson and Traum discusses the Trindi dialogue move engine toolkit (TrindiKit) developed within the context of the European TRINDI project. TrindiKit provides the basic architecture and equipment for building a specific type of dialogue managers, which implement a dynamic theory of dialogue in terms of information state update. Central to TrindiKit are the notions of information state and dialogue move engine. The main function of the latter is dialogue control in terms of updating information states. The contribution of TrindiKit to best practice is diverse, for example, it has the potential to contribute to quality control in the development of dialogue managers, to improve economic or cost control by enabling relatively easy and rapid development of these components, and to enhance comparison and evaluation of the components themselves. Currently, the toolkit is used in the development of a number of systems.

The article *Object-oriented modelling of spoken language dialogue systems* by O'Neill and McTear discusses an object-oriented approach to dialogue management. The main focus of this approach is dialogue modelling in terms of (i) a distinction between high-level, generic dialogue functionalities (e.g. turn taking, confirmation and feedback) and lower-level, specialized functionalities (e.g. request for specific domain information), and (ii) an identification the generic-specific relationships and interactions involved. The major contribution of this paper to best practice is its principled, object-oriented account of how to combine generic dialogue capabilities with domain-specific processing, thereby contributing to the issues of the maintainability and extensibility of SLDSs.

The last article of this special issue is *Towards developing general models of usability with PARADISE* by Walker, Kamm and Litman. The main goal of the research described in this paper is the development of a performance model of SLDS usability that is both predictive and generalizable. The article presents a multivariate linear regression methodology for evaluating SLDSs. Experiments were carried out with three different SLDSs developed at AT&T (ANNIE, ELVIS and TOOT), testing how well the evaluation model generalizes from training data to test data, under different experimental conditions and different user populations.

Acknowledgements

The guest editors would like to thank all members of the review committee for their careful and detailed review reports. The review committee consisted, besides ten members of the DISC consortium, of one member of the DISC Advisory Panel, three members of the NLE editorial board and eleven external referees: James Allen (University of Rochester, USA), Harald Aust (Philips Speech Processing Aachen, Germany), Niels Ole Bernsen (Odense University, Denmark), Peter Bosch (University of Osnabrück, Germany), Phil Cohen (Oregon Graduate Institute of Science and Technology, USA), Robin Cooper (University of Göteborg, Sweden), Morena Danieli (CSELT, Italy), Laila Dybkjaer (Odense University, Denmark), James Glass (MIT, USA), Ulrich Heid (University of Stuttgart, Germany), Paul Heisterkamp (DaimlerChrysler Research Center Ulm, Germany), Julia Hirschberg (ATT Labs Research, USA), Eduard Hovy (University of Southern California,

USA), Stephen Isard (University of Edinburgh, UK), Hans Kamp (University of Stuttgart, Germany), Inger Karlsson (KTH Stockholm, Sweden), Lauri Karttunen (Rank Xerox Research Centre, France), Jan van Kuppevelt (University of Stuttgart, Germany), Lori Lamel (CNRS-LIMSI, France), Patrick Paroubeck (CNRS-LIMSI, France), Karen Sparck Jones (Cambridge University, UK), Simon Thornton (Vocalis Ltd, Cambridge, UK), David Traum (University of Southern California, USA), Marilyn Walker (ATT Labs Research, USA), Yorick Wilks (University of Sheffield, UK), Wolfgang Wokurek (University of Stuttgart, Germany).