Considering the subjectivity to rationalise evaluation approaches : the example of "Spoken Dialogue Systems"

Marianne Laurent, Philippe Bretier, Ioannis Kanellos

CONSIDERING THE SUBJECTIVITY TO RATIONALISE EVALUATION APPROACHES: THE EXAMPLE OF SPOKEN DIALOGUE SYSTEMS

Marianne Laurent, Philippe Bretier

Orange Labs, Lannion, France

Ioannis Kanellos

Telecom Bretagne, Brest, France

ABSTRACT

We present in this paper a reading grid aiming at helping the evaluator take into account the subjectivity factor when designing evaluation protocols. Actually, however most contributions to Spoken Dialogue Systems evaluation tend to objectify their approach with rationalising purpose, we believe that subjectivity needs to be considered for valuable evaluations. The first section shows how closely evaluation processes are dependant on their contexts and on the evaluators' perspectives. We then present an anthropocentric framework that establishes the evaluator as a mediator between the consideration of contextual elements and a rationalising corpus of evaluation procedures. We finally anticipate the benefits our framework brings at both individual and community levels.

Index Terms— SDS, Evaluation, Context, Practice

1. INTRODUCTION

The issue of SDS evaluation is tackled with opposite approaches from academia and industry.

On the one hand, the academics look for *one-size-fits-all* evaluation solutions, i.e. shared metrics dedicated to scientific communication within the domain. They need benchmark protocols to evaluate complete Spoken Dialogue Systems (SDS) solutions as well as separate technical modules to be integrated in bench platforms. Moreover, they claim for a portable methodology to process commensurability exercises between SDSs [1]. As a result, they need well-specified protocols being both portable from a system to another and adequate to the validation criteria of targeted research communities. This quest for objectivity relies on the idea that an impartial methodology could provide an external perspective to arbitrate on the qualities of various solutions.

On the other hand, as pointed out by Tim Paek, the industrials focus more on the design of services and on best practices than on evaluation [1]. The various contributors involved in SDS projects tend to accommodate with *ad hoc* evaluation protocols that support the local decision-making processes. Therefore, more than on the protocol, they insist on the definition and monitoring of local Key Performance Indicators (KPIs) to scrutinise the design of service. In line with their business objectives, they identify and correct design weaknesses and optimise the user experience. For example, the Orange Labs' development suite [2] enables to statistically measure the impact of a given prompt, synthesis voice, prosody, etc. on the overall scenario success. Consequently, commensurability is not of great value for them since the design requirements evolve at the same time of the system design itself.

However the sharing of practices and common evaluation protocols is not a priority, the nomadism of these evaluations approaches may lead to counter-productive efforts. As a matter of fact, the isolated evaluators generally develop and maintain custom-made evaluation spreadsheets, defining self-suitable KPIs, based on personal definitions and calculation methods [3]. First, this reduces the opportunity and the ease for collaboration between project stakeholders. Second, the lack of convention for communication of results and comparison of systems' performance may either lead to misunderstanding or to audience disinterest. On the contrary, the implementation of shared practices and standardised indicators may lead to timesaving and improved cooperation.

We conceive both points of view and claim for some efforts to bridge the gap between them. Indeed, it would enable both fields to rationalise their approaches to evaluation, by providing a framework to the nomadic evaluation efforts and considering the subjectivity inherent to each evaluation process at the same time. Just as [1] and [4] advocate for the convergence of academic and commercial approaches of SDS design and evaluation, this article presents a framework to combine the value of a rationalisation of procedures with the consideration of the evaluator subjectivity.

In the second section we underline that, even if some research groups are looking for an objectification of evaluation, subjectivity is an inevitable factor to consider when designing an evaluation process design. The third section presents an anthropocentric framework that considers the evaluator as a mediator betweenthe rationalising corpus of evaluation protocols, its community of practice (CoP) and the evaluation situation. The fourth section explains how such an approach could support the individual activity, as well as the evaluation practices shared in both the communities of practice and the communities of interest.

2. THE EVALUATOR'S POINT(S) OF VIEW

2.1. Evaluation is a subjective process

Along the lifecycle of a project, many are the stakeholders involved [3]; project owners, technical developers, ergonomics, marketing people, customers and end-users are only examples of the various communities of individuals possibly contributing to a SDS project. Of course, all of them are prone to assess their contribution to the overall project by analysing the system from their activity-biased point of view.

Stufflebeam [5] defines the evaluation as a process through which one defines, obtains and delivers useful pieces of information to settle between the alternative possible decisions. Likewise, we define an *evaluator* as an individual who, at some point of a SDS project lifecycle, analyses the system's compliance with a set of expectations or compare systems so as to make a decision. This comprises the evaluation planning (choice or design of the protocol), implementation and report. The evaluation is actually a rationalising contribution to the decision process, even if irrational parameters always gain the upper hand. Therefore, it must be considered as an input to its relative decision-making process, not as a context-independent tool. It is part of an encompassing project in which the evaluator is involved. Examples of questions for SDS evaluation may be: Which of these solutions should I deploy considering the users needs? Has this iteration enhanced or worsen the global service quality? Is the application ready to be deployed or does it need further developments? Such questions are of course tainted with the motivations that animate the stakeholder raising the enquiry.

We hence dismiss the idea of an *objective* evaluation. First, the evaluation's definition itself underlines its necessary link to a subject. To evaluate is to assign a value to a given object, which presupposes a value judgement delivered by a subject. Second, the subjectivity is present all along the evaluation process, as when understanding the problem than when building an appropriate response to the latter. The top-down approach of the V-Model process described in 4.1 illustrates how the evaluator's subjectivity influences the interpretation of the situation.

2.2. Evaluation as a goal-driven argumentation discourse

Yarbus' eye-tracking studies [6] show that the eye movements depend on what the observer aims at displaying. When observing a scene, the look is goal-guided, trying to identify a selection of clues relevant to a leading issue (see Fig. 1).

These findings are of prime relevance for us since an evaluation protocol is designed, or selected, by an evaluator we consider as a *situated subject*. Consequently, this design/selection is conditioned by both the evaluator's profile (see 3.1) and the *situation* that establishes the goals for evaluation (see 3.2). To these regards, setting up an evaluation is like building an argumentative discourse: so as to provide convincing clarifications to the decision-makers, an evaluator selects the most relevant points of view considering the decision to be made. Moreover, such an argumentation is adjusted to both self-conviction and the conviction of an audience. This means that the discourse must be both consistent to the position taken by the evaluator (the conclusion he aims at reaching) and adapted to the interlocutors' sensibility (norms for argumentation acceptance). The choice of evaluation criteria (see 4.1) is consequently a salient point since they act as argumentative markers for the discourse.

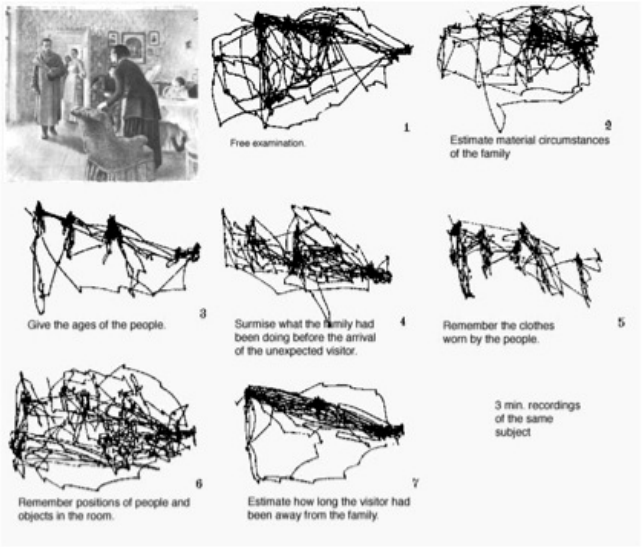

Fig. 1. The task influences the observer behaviour [6]

We explained that the evaluators' perspective is of major influence for the design of evaluation; even so, evaluators are not the only source of subjectivity. The following section presents how the evaluators must cope with both subjective factors linked to the project in which they are involved and the system of values inherited from their CoP.

3. THE EVALUATOR IS A MEDIATOR

This section introduces a reading grid that assimilates the previous reflections. Acting as a thinking model, it consists in an anthropocentric framework where the evaluator assumes the mediation between three dimensions that must be understood both as resources (systems of answers) and as constraints (limits to the evaluator liberty) (see Fig.2).

3.1. The community of practice

As mentioned above, SDS projects involve various families of practitioners prone to evaluate the relative solution. A given evaluator is necessarily biased by its belonging CoP that acts as a normative system. First, the community predisposes the evaluator with a repertory of practices, methodologies and

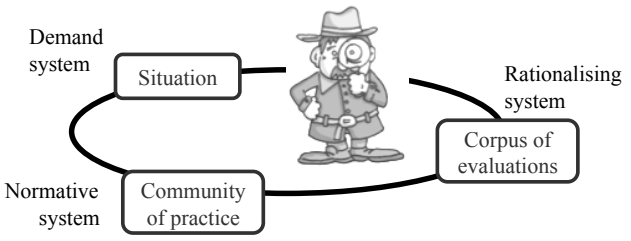

Fig. 2. The evaluation ecosystem

points of view for analysis. And, second, it promotes interpretation grids and a *system of values* that bias the way situations are apprehended. Lorino [7] links this phenomenon with the notion of *genre*, seen as a common repertory of gestures, wordings, tacit meanings, to which the actors of a given job family refer to. However these actors may adapt the common practice according to their personal *style*, they remain under the influence of their community. This CoP influences both the evaluator's understanding of the situation and its consequent behaviour as regard evaluation.

3.2. The situation acts as an operationalisable context

The situation conditions the goals that motivate the evaluation. Its characterisation is consequently of prime importance since it orients the design, choice and the potential parameterisation of the evaluation to be carried out. We devise the situation as a practical version of the context, the latter being itself a too general concept to be usable. Consequently, we empirically restrain its definition to three factors: the moment in a SDS project lifecycle, the scope of the evaluation and the resources/constraints applied to the evaluation.

3.2.1. Moment within a SDS project

We assume that one of the most significant parameters in the evaluation situation is the stage of the project during which the evaluation is carried out. Actually, it strongly infers evaluation goals and influences the designed protocols.

Three major types of evaluation [8] may be distinguished. First, in the early stages of a project, the *prognostic evaluations* guide orientation and admission decisions. Examples are return on investment, feasibility and distance-to-target evaluations. Second, the *formative evaluations* support the intern regulation within the course of the project. For example, they may help spotting design issues or measuring improvements between two versions of a solution in the frame of iterative development. And third, the *certifying evaluations* are involved, at the last stages of a project, to validate the service with respect to the specifications or predefined tolerance thresholds.

These examples show how the stakes and the protocols may vary between evaluations. Additionally, data capture and analysis differ too. For example, testing a prototype with a panel of test users and testing a live solution with real customers supply different types of data.

3.2.2. Scope of the evaluation

The decision-making context strongly conditions the evaluation scope. First, evaluation may focus on either a technical module of the solution (glass-box evaluation) or a global view on the solution (black-box evaluation). Second, it may address various fields of interest. For example, to measure the impact of the vocal recognition performance on the global usability of the system requires appreciating both the vocal recognition and the ergonomics points of view. Therefore, before building or choosing an evaluation methodology, evaluators must stand back on their practice particularities considering other possible methodologies.

3.2.3. Resources and constraints

As mentioned above, resources and constraints applied to the evaluation strongly depend on the phase in the SDS project. They may also result from the development and financial contexts of the project. Wizard-of-Oz (WoZ) simulations, for example, are known as costly, laborious and time-consuming experimental set-ups, and thus will not be applicable in every context. Similarly, the evaluations based on automatically computed log files require specific tools able to compile and parse them.

3.3. The corpus of evaluations

The existing corpus of evaluation contributions may act as a source of inspiration for evaluators. This is also a rationalising system that enables to gain in transparency and reflexive knowledge on the various evaluation procedures. So as to facilitate the acquisition of such knowledge of specialist on the corpus, we proposed a meta-model to classify evaluations (see Fig.3). It is articulated around three levels of analysis, so as to obtain an individualised knowledge on given evaluation contributions. The first level suggests four categories based on the type of data captured for evaluation: performance measures, predictions of user perception, user perception and *ad-hoc* mixed evaluations. The second one focuses on the models and calculation procedures used to transform the row data into meaningful KPIs for a couple evaluator/situation. The third one distinguishes the points of view selected for the analysis. This systematic taxonomy enables to describe the particularities of each evaluation, considering the methodologies as well as the critical viewpoints.

It provides the evaluator with thinking categories, which helps for making their practice explicit and sharable. Yet, it does not claim for being exhaustive in the categories it suggests and allows for the framework extension. It (i) facilitates

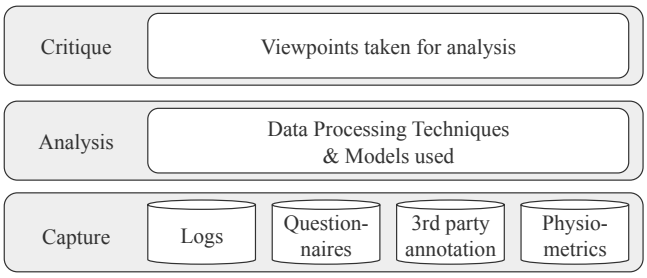

Fig. 3. A three-layer definition of evaluation

the communication and the sharing of results within the domain and (ii) permits to position the solution within a map of existing methodologies. However, it does not take any contextual factor into account.

3.4. Implementation in an anthropocentric solution

However the design of evaluation is guided by recommendations [9] [10] and influenced by both contextual factors and existing solutions, it is not a tractable task. As an alternative, we adopt an anthropocentric approach that provides the evaluators with a pivotal role in such ecosystems[1] composed of the CoP, the situation and the corpus of evaluations. They are mediators between a rationalising structure (the evaluation description) and less rational factors that are the interpretation of the decision process and the community footprint on their tacit repertory of practices.

We now work on a software implementation of these ideas: a *Multi Point Of vieWs Evaluation Refine Studio* (*MPOWERS*). This system aims the rationalisation - not the standardisation - of evaluation. By rationalisation, we refer to the definition of common norms for the description of processes, common thinking models and vocabulary, for evaluators to make their procedures explicit. Our multi-profile evaluation platform facilitates the design, from a unique corpus of parameters, of personalised evaluations adapted to the particular contexts.

Similarly to our framework for SDS evaluations taxonomy (see Fig. 3), our application will be based on three layers: (i) evaluation data extracted from log files and users and third-party questionnaires, (ii) KPIs aggregated from this row data and (iii) evaluations compiled from these indicators. It stands on the assumption that every evaluation may be modelled as a combination of KPIs, the weight given to each indicator depending on both the system of value supported by the evaluator's community and the situation of evaluation. For a start, the ITU-T Rec. P.Sup24 [12] provides definitions and calculations for sets of common indicators. Our model aims at being dynamic and enriched as new needs arise.

[1]We use here the term *ecosystem* to refer to a dynamical and complex association of interdependent elements that are constantly interacting with each other.

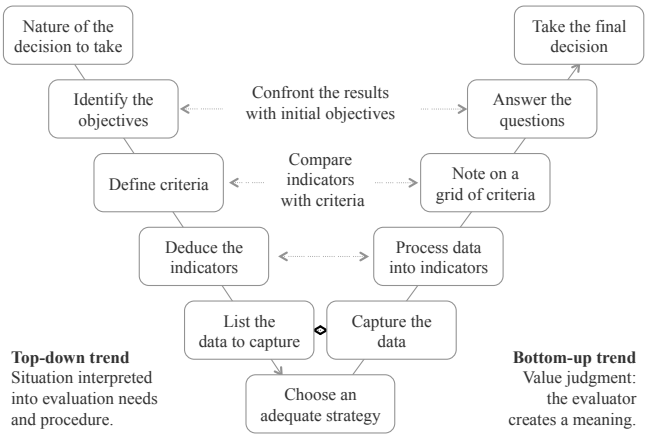

Fig. 4. V-Model evaluation design process. Top-Down process inspired from [13]

By offloading the evaluator user from low value-added and computable tasks (data collection, cleaning, transformation and aggregation), it enables isolated evaluators to concentrate on the creative aspects of evaluation and the ongoing decision-making. Making the methodology explicit thanks to a prescribed framework, they gain on reusability over projects. Serial refinements over evaluation campaigns and new projects should lead to methodological enhancement.

The application is still under development but the following section anticipates a twofold relevance. At the individual level, it should provide the evaluator-users with the required expertise to go beyond their step-by-step laborious design process. It should also encourage the collaborative work inside both CoPs and communities of interest.

4. GAIN EXPERTISE TO GAIN AGILITY

4.1. Evaluation design as a problem solving process

As suggested in the introduction, designing a SDS evaluation mostly comes down to a problem-solving oriented process. The evaluators generally build their evaluation stage-by-stage from their understanding of the problem. Such an approach can be seen as a V-Model development process. As pictured in figure 4, it starts with a top-down approach, during which evaluators interpret the problem deduced from the *situation*. Then they translates the problem into objectives, criteria, KPIs and data to be captured [14]:

- *Objectives*: An evaluation is processed as to prepare a decision. Its objectives are thus constrained by the need to deliver appropriate answers to the decision-makers. The first step therefore consists in interpreting the situation into a set of objectives.
- *Criteria*. Once the objectives identified, evaluators translate them into evaluation criteria. A criterion is a

quality expected in the evaluated object, a viewpoint through which the evaluator examines the object. [15] proposes an interesting taxonomy of *Quality of Service* and *Quality of Experience* aspects. They may be seen as value scales, as for example, the measure of an application's efficiency or effectiveness, its usability or even the more pragmatic performance / cost ratio. Thanks to making the situation and the evaluator profile explicit, the choice of criteria gains in transparency.

- *KPIs*. A criterion defines a conceptual ideal for the evaluated application. To determine whether a criterion is satisfied, the evaluator considers a set of various indicators of performance. They correspond to real facts or representations, expressed on the evaluated objects (see [12] for a definition of commonly used parameters from which KPIs can be defined).

- *Strategy*. Last, a strategy has to be defined as regard the data capture (automatic log, user questionnaires, third-party annotation, etc.) and their compilation into appropriate KPIs.

Following this work of analysis and translation into methodological traits, evaluators arrange the evaluation protocol following a compositional approach. This is a symmetric phase to the previous top-down one. They gather the data, prepare them into exploitable KPIs, and compare them with the pre-defined criteria framework. Finally, evaluators provide the decision-makers (possibly themselves) with the results, the evaluation being a input sub-process to a decision making process.

Such a problem solving approach corresponds to what Rasmussen [16] positions within the *knowledge based behaviours* in his *Skill-Rule-Knowledge model* (see Fig. 5). Indeed, the field lacks a systematic hindsight on the ecosystems of evaluation contributions' to be able to venture shortcuts in this stage-by-stage design process. Therefore, since no routine or pre-defined rule is available, evaluators have to improvise from their personal *specialised* knowledge on the SDS domain, evaluation protocols and KPIs. We believe that the evaluator would strongly benefit from climbing the SRK model ladder by gaining in expertise.

4.2. Toward an expert problem resolution process

The reading grid we introduce will enable a shift from an inexperienced arduous SDS evaluation design to a documented and process-supported one. Enabling the evaluator to reach the rule-based behaviour level, this grid will support him in processing pattern matching between the evaluation situation recognised by the evaluator and an appropriate adequate methodology to implement. Moreover, this will make meta-evaluation easier thanks to the acquisition of an expertise on the existing evaluation paradigms along with their relative ecosystems. As a consequence, the evaluator will be able to:

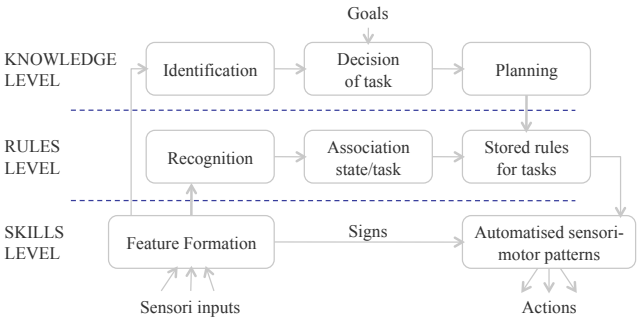

Fig. 5. SRK Model by Rasmussen [16]

- detect more easily changes in evaluation situations, thanks to the disposition of tools to make them explicit,

- dynamically measure the adequacy between an evaluation protocol and the relative situation,

- and, if required, adapt the methodology to the situation.

The decision-making process will then come down to the recognition of a prototypical situation to which an appropriate evaluation may be associated. It requires the capability from evaluators to both recognise these situations and to associate them with types of evaluations. They thus require a frame to recognise situations and a documented and critical description of the evaluation corpus. This includes knowledge on the protocols themselves, but also on their implementation context. In the end, evaluators will adjust the chosen evaluation to local specificities (hierarchy of criteria for example).

4.3. An impact on the collaborative work

The reading grid we propose go beyond the individual impact on the evaluator. It also targets the cooperative work inside *Communities of Practice* (CoPs) on the one hand, and *Communities of Interest* (CoIs) on the other hand.

CoPs *"consist of practitioners who work as a community in a certain domain undertaking similar work (although within each community there are individuals with special expertise, such as power users and local developers")* [17]. Our framework may support the collaborative work inside these communities in terms of methodology sharing and negotiation of common protocols for evaluation. This could lead to the emergence of an explicit (no longer tacit) knowledge of existing norms within communities. As a consequence, a meta-analysis would permit to identify regularities within the practices of the evaluators, and thus enable a mapping of CoPs based on observed practices.

CoIs *"bring together stakeholders from different CoPs to solve a particular (design) problem of common concern* [17]. Our reading grid could be used to understand the complementarities between the different groups of actors inside these

temporary communities. Our framework advertises for shared knowledge (corpus of KPIs for example) and enhanced cooperation between actors, instead of the potential conflicts and misunderstandings entailed by the divergence between backgrounds, practices and languages.

Instead of CoI, Lorino [7] suggests the idea of *"Community of Inquiry"* to describe the group of individuals whose cooperation converge around a *"conjoint activity"*, an *"activity that is not characterized by similar practices but their heterogeneous complementarities"*. In our approach, the emergence of such communities is an essential stake for the overall organisation. It helps to render explicit, and thus sharable and modifiable, the collective activity that would otherwise be seen as a constraint by the isolated stakeholders. However the forming of such communities cannot be decided, it may be encouraged by implementing solution similar to the one we present.

5. CONCLUSION

We realised this work considering the evaluation issue for spoken dialogue systems, however it is extensible to any other human-machine domains. The framework we presented aims at fostering the rationalisation of the evaluation protocols design while taking into account the situation and the situated evaluators' subjectivity. On the one hand, it helps to make some elements of the evaluation context explicit to the evaluator. It clarifies the situation in terms of decision-making process, audience and constraints, and the community of practice through its value systems. On the other hand rationalising the evaluation may stimulate the sharing of practices across the communities. This convergence gives transparency to the implemented evaluation approaches. It enables improved communication, cooperation, productivity and comprehension across stakeholders, which for them are requisite to stand back from each others' practices and to allow real collaboration.

6. REFERENCES

[1] Tim Paek, "Toward evaluation that leads to best practices: reconciling dialog evaluation in research and industry," in *Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, New York, 2007, pp. 40–47, Association for Computational Linguistics, Rochester.

[2] Romain Laroche, Ghislain Putois, Philippe Bretier, and Bernadette Bouchon-Meunier, "Hybridisation of Expertise and Reinforcement Learning in Dialogue Systems," in *Interspeech*, 2009.

[3] Marianne Laurent and Philippe Bretier, "Ad-hoc evaluations along the lifecycle of industrial spoken dialogue systems: heading to harmonisation?," in *The seventh international conference on Language Resources and Evaluation (LREC)*, Malta, 2010.

[4] Roberto Pieraccini and Juan Huerta, "Where do we go from here? research and commercial spoken dialog systems.," in *Proceeedings of the 6th SIGDdial Workshop on Discourse and Dialogue*, 2005, pp. 1–10.

[5] Daniel. L. Stufflebeam, *L'évaluation en éducation et la prise de décision*, Ottawa, 1980.

[6] Alfred. L. Yarbus, *Eye Movements and Vision*, Plenum (Originally in Russian 1962), New York, 1967.

[7] Philippe Lorino, "Communities of inquiry and knowledge creation in organizations: the process model in management," *Annals of Telecommunications*, vol. 62, no. 7/8 - Knowledge Management: knowledge networks, pp. 753–771, 2007.

[8] Gérard Guingouain, *Psychologie sociale et évaluation*, Dunod, 1999.

[9] ITU-T Rec. P.851, "Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems," 2003.

[10] Sebastian Möller, *Quality of telephone-based spoken dialogue systems*, Springer-Verlag New York Inc., 2005.

[11] ITU-T Rec. P.Sup24, "Parameters describing the interaction with spoken dialogue systems," 2005.

[12] Jean-Marie De Ketele and Xavier Roegiers, *Méthodologie Du Recueil D'informations*, De Boeck Université, Bruxelles, 1993.

[13] François-Marie Gerard, "L'indispensable subjectivité de l'évaluation," *Antipodes*, vol. 156, pp. 26–34, 2002.

[14] Sebastian Möller, Klaus-peter Engelbrecht, Christine Kuhnel, Ina Wechsung, and Benjamin Weiss, "A taxonomy of quality of service and quality of experience of multimodal human-machine interaction," in *International Workshop on Quality of Multimedia Experience, QoMEx 2009*, 2009, pp. 7–12.

[15] Jens Rasmussen, "Skills, rules, knowledge; signals, signs, and symbols, and other distinctions in human performance models.," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 13, pp. 257–266, 1983.

[16] Gerhard Fischer, "Communities of Interest: Learning through the Interaction of Multiple Knowledge Systems," in *IRIS'24*, Ulvik, 2001, pp. 1–13.