Is that a Good Spoken Language Dialogue System?

Niels Ole Bernsen and Laila Dybkjær

Natural Interactive Systems Laboratory, University of Southern Denmark
Science Park 10, 5230 Odense M, Denmark
{nob, laila}@nis.sdu.dk

Abstract

This paper steps back from the specific issues involved in evaluating spoken language dialogue systems (SLDSs) to discuss instead the meta-questions of what constitutes a good SLDS and how we can decide if system A is better than system B. It is argued that system goodness is not a property which is amenable to scientific study. Rather, what science can do is unravel the numerous factors which, viewed in isolation, contribute to SLDS quality. At some point, even that exercise stops being part of science and becomes an integral part of manufactoring.

1. Introduction

On the market for more than a decade, spoken language dialogue systems (SLDSs) are, finally, proliferating in a large variety of applications and an increasing number of languages. Another good piece of news for at least some of the colleagues working in the field, is that SLDS evaluation has become a respected research topic which is often deemed to be of critical importance to both research and industry. The first piece of slightly less good news in this paper follows from this fact. Had we known (i) what constitutes a good SLDS or, even more ambitiously, (ii) how to decide if SLDS (A) is a better or poorer system than SLDS (B), then SLDS evaluation would hardly have become a new important research issue in the first place. In other words, speaking generally, we don't know the answers to (i) and (ii).

This paper aims to step back from the specific issues involved in evaluating particular SLDSs, asking instead what the questions (i) and (ii) above really mean and which kinds of answers to (i) and (ii) it is meaningful to expect. So, to warn the reader one last time, this is a "meta-paper" on SLDS evaluation. The authors feel that this might be the time to ask meta-questions about SLDS evaluation in order not to risk being trapped in a futile search for the impossible. Meta-papers often run a different risk, though, namely that of being seen as never getting to the important points be they practical, theoretical or otherwise, and tend to arouse suspicion that their authors have little to say on what really matters. In the present case, the authors have been through the full process of evaluating an SLDS research prototype (Bernsen, Dybkjær and Dybkjær 1998); they have in the DISC project been through a painstaking process of reviewing what many strong groups of developers have done to evaluate their SLDSs and components (http://www.-disc2.dk); they have repeatedly tried to express what it takes to evaluate an SLDS dialogue manager and what it takes to evaluate an SLDS from a human factors point of view, see e.g. (Dybkjær and Bernsen 2000); they have also devised and used a template for evaluation of SLDSs and their components (Bernsen and Dybkjær 2000). What they haven't done, is to find all or most of the right answers. Also, to conclude this lengthy confession, they have approached the problem in what may be described as a rather classical academic fashion, i.e. trying first to get a better understanding of some respected core issues and then working their way "outwards" into increasingly messy problems. Still, they did not go all the way out there yet because, when the issues get too messy and the data to work from too few, such as data from the actual installation of a wide range of SLDSs, science tends to throw up its hands in frustration and return to its home ground: the simpler problem, the really comprehensive data etc. In this paper, we would like to contemplate the evaluation problem from the outside. We discuss what "good" means in relation to the particular issue of SLDSs quality, potential implications for SLDSs evaluation, and expectations to the future tasks of a science of evaluation.

2. What does "good" mean?

In general, components are easier to evaluate than the systems they are part of. We are not claiming that it is exactly simple to evaluate speech recognisers or speech synthesisers. Even a "good" speech recogniser or speech synthesiser is subject to the problems inherent to what it means to be "a good X" (see below). Yet it seems clear that evaluating a system consisting of a series of components must be orders of magnitude harder than to evaluate the components themselves one by one. And in all or most cases, the quality of the system is not a simple function of the quality of its components and the way the components have been put together.

Despite the fact that cars have been produced for a century and watches for several centuries, we still can choose among a huge variety of brands and models of cars and watches. It would even seem likely that most people would resent *not* being able to do that. Why haven't we been able to sort out a long time ago which cars are good and which cars are not-so-good? The answer seems to be that most present-day cars for sale are "good" cars. By and large, that's why they are (still) able to compete in the market. Their goodness, as the goodness in "a good spoken dialogue system", is not an objective quality had by some systems and lacked by others. Rather, "good" means something like *good enough for someone right now*. This means that goodness changes over time for a particular individual, that goodness differs from one individual to another, and that the *for* (someone) indicates that goodness is relative to a wealth of contextual factors, such as the economy of the buyer, the priorities of the buyer, or the sheer importance which the buyer attaches to different things. Note that all of this is compatible with there being objective technological goodness, only superior technical quality is not that important to some customers, others cannot afford it, and yet others can but won't.

So, when we carry out research to produce knowledge about which SLDSs are good, or are better or poorer than other SLDSs, are we trying to identify the Ferraris and Mercedes among SLDSs? These may be too expensive for most customers who will actually choose other products for their business, but still, everybody knows what is ultimately the best product in some sense. That sense, however, is not the "general goodness" discussed above. Rather, it seems to be something like "average technical goodness" which is a function of component quality, design quality and manufacturing quality most of which factors, we submit, are amenable to objective validation. The point is that few users go for top average technical goodness. Together, Ferrari and Mercedes probably sell less than 5% of the cars sold world-wide.

An important reason for the mismatch between technical quality and sales probably is the imponderable issue of taste. The technology itself may be deemed by all to be the best there is but isn't it awful nevertheless! Its (the car's) design exudes the wrong role model; its (the SLDS's) speech output is clear and intelligible but condescending in its tone of voice; its dialogue management makes me feel stupid even if it enables me to get the job done; its mechanically consistent phrasing is exasperating; etc. And don't expect that this will change with more natural interactive technology. More natural interaction only means that, increasingly, SLDSs will be like people. And people, you know - that fellow irritates me even if he knows his stuff, I cannot say why right now but I am certainly not going to buy such a system. Tomorrow I might, though, because taste, like its sibling, fashion, changes in incomprehensible ways. Even our perception of people changes over time, sometimes for the better and sometimes for the worse, when we get to know them more intimately.

Or consider PCs. At NISLab we have a technician who likes to build our PCs himself from inscrutably selected components. Nobody objects too much to his choices most of the time because they haven't themselves formed any strong opinions based on experience or on their favourite PC magazine. When they object, it's mostly because we need the machine right now and cannot wait for the technician to build the thing. In that case, component choices are made from the provider's limited list of compatible components instead, again in partly inscrutable ways. Put simply, life is just too short for subjecting a shopping list of PC components to quasi-scientific scrutiny. If the specimens of a particular PC brand become too annoying, we just make sure not to buy that brand again. Maybe the future market of SLDSs will be comparable to the present PC market: with many different brands, alternative components ("Standard", "Silver", "Gold" and so on) to select from when composing the system, different look-and-feel of different brands, different price levels, different marketing efforts, and relatively low prices overall compared to other kinds of software and hardware. Arguably, that is how the market is shaping up at the moment. The main difference from current PCs may be that many future SLDSs will retain an aspect of the custom made, just as when you buy a new kitchen for your house. But even with kitchens, most customers stick to standard modifications of the standard models. Those who want to argue that future SLDSs will be any different, please stand up!

3. Some implications

Due to the good *for* someone –principle in Section (2), SLDSs evaluation decomposes into (i) technical evaluation of the system and its components in order to assess component quality, design quality and manufacturing quality, (ii) end-user usability evaluation of the system, and (iii) customer evaluation of the system and its components. Although (i)-(iii) are not completely dissociated, a technically excellent system integrating excellent components may have poor end-user usability whilst a technically secondary system may score highly in terms of end-user satisfaction. And the customer may prefer yet a third system for reasons of, say, cost, platform compatibility and a good service deal, all which have little to do with technical perfection or end-user satisfaction, at least as long as the customer expects that the end-users will go along with the choice. As long as this is the case, nothing prevents the customer from profoundly disliking the system when using it as an end-user. The customer, in other words, who may or may not be identical to the end-user, tends to trade off all sorts of properties and criteria against each other to get a good SLDS.

The question to which we would very much like to know the answer is: how far does it make sense to go in aiming to lay down principles for what is a good SLDS? In what follows, we would like to propose what we consider to be some implications of the discussion above.

The first implication is that *subjective evaluation will remain subjective evaluation.* Science will never be able to identify "the best" system for the simple reason that the best system does not exist and never will. Like cars, SLDSs will proliferate to the point where there is - or if there actually isn't, there probably soon will be - an SLDS for almost everybody's preference. Trying to predict goodness from carefully selected questions to be answered on five-point scales is a wild goose chase. These questions never ask if the end-user would buy the system followed by a "Please sign on the dotted line". And if they did, the user would counter by asking: which systems could I choose from?

The second implication is that *customer evaluation will remain unpredictably contextual.* It will also be subjective, to be sure, remember that the customer may or may not be identical to the end-user. When different from the end-users, customers are typically organisations interested in deploying some system in order to provide a set of services for people, sometimes in competition with other service providers. Basically, customers are looking for a good deal. If they believe they are getting a good deal, by definition they also believe that they get a good system. And as we know, there is no end to the constituents of a good deal. "Can you deliver in six weeks max.?" is just one such constituent.

The third implication is that *we will know in due course which systems are generally the best systems, technologically speaking.* These systems are always real-time, they (almost) never crash, they are serviced well, they help get the task done under almost any circumstance etc. Many users may not like them that much but still ... What we don't seem to know right now is whether this level of quasi-perfection will come to characterise virtually all SLDSs which will be on the market in, say, ten years or whether time will help identify the SLDS Ferraris and Rolexes. For what it is worth, one guess could be that

some of the basic components, such as speech recognisers and synthesisers, will come to be provided by a very small number of world-wide suppliers. The race certainly has started with acquisitions, friendly clubs which we, the researchers, are invited to join etc. If the guess comes true, we won't have much choice among basic components. And the basic components we will be getting, will probably soon stop incorporating ideal state-of-the-art technical perfection. They will just become facts of life just like the Windows platforms have been for some time: they get better or different in some sense with each new release but which sense is not entirely clear. Plug-and-play platforms for component integration, task-independent dialogue managers and the like, may or may not end up becoming proprietary. They might become open source instead. One reason why the open source concept could have a chance in these areas is that SLDS technology has the potential of becoming part of most future applications in some role. In this situation, strong component suppliers could do worse than supporting the open source concept. They will have a hard time delivering to all kinds of need for SLDS technology anyway.

In the, perhaps unlikely, situation that there will exist Mercedes and Rolexes of the SLDS trade in the future, chances are that these will only be sold to comparatively few customers/end-users. NASA or ESA might have some custom-made for the new international space station, for instance. But situations in which money is not an issue and technological perfection is all, will probably remain rare exceptions.

The fourth implication is that *quality of components, design, and manufactoring does matter to some customers and that these customers often have a decisive say in which system to purchase.* In other words, objective component, design and manufactoring evaluation does have an impact, at least until, as regards components, we just have to buy what we are offered because all of the alternative component technologies which used to be around have been acquired. Even if technical quality is far from being the whole story, it provides customers and end-users with a basis from which to add personal, contextual and other preferences which together form their criteria of choice.

4. What can be done by science

SLDSs are becoming increasingly complex and their complexity will continue to grow not least through their integration into the multimodal systems of the future. The systems will be bought and sold subject to so many imponderables due to situation, context, taste, "life is too short for that" etc., that any project aiming to identify the best systems is futile as a scientific enterprise. Rather, what science can do and should continue to be doing, is to identify *quality factors.* A quality factor is some property which, if present or if present to a certain degree in an SLDS, and when viewed in isolation adds quality to the SLDS as a whole. Quality factors cannot be added up, firstly because we shall never find the algorithm for doing so, and secondly because many of the factors will remain qualitative (reflecting experts' opinion) or downright subjective rather than quantitative.

The types of quality factors, it would seem, may be categorised as pertaining to either:

- the technology of a particular component;
- the technology of the system as a whole;
- the way it was designed and manufactored;
- the way end-users perceive particular properties of the system when using it;
- the way customers who are not end-users perceive particular properties of the system.

The quality factors which could be generated from the above list are likely to be in the order of a hundred if not more than that. In the DISC project, we came close to identifying as many quality factors (http://www.disc2.dk). Therefore, the mere combinatoric complexity of interrelating several quality factors prohibits a systematic study of their interrelationships. As SLDSs grow in complexity and multimodal integration, the number of quality factors is set to increase even further. Needless to say, such numbers are far too high for anyone but a developer team to care about them all, or about all those which are relevant to a particular application. The public, the end-users, and the customers will be at the mercy of the sales people, the computer magazines and the possible future omnipotent component suppliers. Which properties the sales people and the magazines will end up focusing on, is anybody's guess. Some guesses could be the existence of a global service network, energy conservation, or environmental preservation, all of which would leave science with little to do, it would appear. To be sure, the systems will all be able to recognise speech and conduct spoken dialogue, of course, else they would not be in the market in the first place.

5. Conclusion

In this paper, we have expressed doubts about the holistic idea that spoken dialogue systems evaluation research should aim at eventually defining what constitutes a good system, which kind of system creates the most user satisfaction etc. We have argued that system goodness is not a property which is amenable to scientific study aiming to identify *the* good system. Rather, what science can do is unravel the numerous factors which, viewed in isolation, contribute to SLDS quality. And at some point, even that exercise will stop being part of science and become an integral part of manufactoring. We are curious to know when, i.e. how soon, that will happen.

6. References

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

Bernsen, N. O. and Dybkjær, L., 2000. A Methodology for Evaluating Spoken Language Dialogue Systems and Their Components. *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000), Athens.

DISC: http://www.disc2.dk

Dybkjær, L. and Bernsen, N. O., 2000. Usability Issues in Spoken Language Dialogue Systems. To appear in *Natural Language Engineering*.