Resource Evaluation for Usable Speech Interfaces: Utilizing Human – Human Dialogues

Pepi Stavropoulou[1,3], Dimitris Spiliotopoulos[2], Georgios Kouroupetroglou[1]

[1]Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilisia, GR-15784, Athens, Greece
E-mail: {pepis, koupe}@di.uoa.gr

[2]Athens Technology Centre
Rizariou 10, Chalandri, GR-15233, Athens, Greece
E-mail: d.spiliotopoulos@atc.gr

[3]Omilia Ltd
Konitsis 3-5, Marousi GR-15125, Greece

Abstract

Human-human spoken dialogues are considered an important tool for effective speech interface design and are often used for stochastic model training in speech based applications. However, the less restricted nature of human-human interaction compared to human-system interaction may undermine the usefulness of such corpora for creating effective and usable interfaces. In this respect, this work examines the differences between corpora collected from human-human interaction and corpora collected from actual system use, in order to formally assess the appropriateness of the former for both the design and implementation of spoken dialogue systems. Comparison results show that there are significant differences with respect to vocabulary, sentence structure and speech recognition success rate among others. Nevertheless, compared to other available tools and techniques, human-human dialogues may still be used as a temporary at least solution for building more effective working systems. Accordingly, ways to better utilize such resources are presented.

Keywords: Spoken Dialog Interfaces, Resources for Language Modeling, Human-Human Dialogue.

1. Introduction

Commercial Spoken Dialogue Systems may range from simple directed dialogue applications involving a primarily menu driven interface, where the user navigates through a menu of options, to open-ended natural language conversational systems. The latter are commonly used for large scale applications, and are targeted for user satisfaction and naturalness, as they allow the user to respond in a more natural, "in his own words" way. Whether it is a directed dialogue or a more elaborate open-ended conversational system, there are three generally acknowledged major steps in the process of building such a system (Cohen et al., 2004):

- Design: Requirements Analysis and High-Level Design, Detailed Design
- Implementation, iterative testing and deployment: Testing and development of system components (Automated Speech Recognition (ASR) Module, Natural Language Understanding (NLU) Module, Dialogue Manager (DM), Language Generation Module, Text to Speech Synthesizer)
- Evaluation

Empirical approaches involving the collection and analysis of "in domain" spoken corpora are commonly employed throughout these steps especially in the case of open ended natural language systems, where such corpora are necessary for design issues as well as the development of statistical language models for ASR, input classifiers for NLU, other machine learning modules for dialogue management, user simulation and so on.

More specifically, during the early design cycle, analysis of such corpora may provide insights in the vocabulary used, the nature of the interaction, user's attitude and mental model of the task in general. They are, therefore, very important for the development of grammars later on – providing the list of slots and meaningful key-words – as well as the design of the dialogue flow and the prompt specification. The latter is particularly important taking into account that the success of a speech based application greatly depends on the correspondence between the "natural" mental model that first time users bring to the interaction and the proposed model afforded by the design of the interface (Norman, 1988; Galitz, 2007; Weinschenk, 2000). Even by simply adjusting the wording of the prompts to conform with user discourse patterns, the ease of use and clarity of the interface may be significantly improved. In short, corpus analysis comprises a significant aspect of user-centred design, and is, thus, of great significance when it comes to creating a usable, user-friendly application.

Furthermore, with regards to the implementation cycle, the utilization of domain specific corpora comprises a "sine qua non" for large scale open-ended conversational systems. Such systems must use stochastic models, for ASR at least, as it is practically impossible to predict in advance and subsequently specify the variation in user input using a handcrafted rule-based recognition grammar. Instead, Statistical Language Models (SLMs) are trained on domain specific corpora, in an attempt to essentially model what the callers are likely to say when interacting with the system. Coupled with robust natural language grammars or machine learning techniques for interpretation, they can lead to successful recognition and interpretation of free style speech allowing for a more efficient and natural interaction and thus enhancing user satisfaction. Moreover, by utilizing machine learning techniques to model user behaviour, the system is better adapted to the user's prior knowledge and experience, which is also particularly significant for creating a more familiar, intuitive, easier to learn and use interface.

However, the collection, transcription and annotation of such corpora is a time-consuming process that often requires the existence of an almost complete or deployed system. On the other hand, for some applications, such as phone banking, help desk, customer care and stock trading applications, there may be an abundance of recordings of human to human dialogues immediately available. In addition, the analysis of actual user calls to human agents is already considered to be a significant resource for effective design (McTear 2004; Cohen et al. 2004). Nevertheless, one should be cautious, as users bring different expectations to the dialog, when talking to a computer and not a human, which in turn may undermine the validity and utility of such resources.

In an attempt to formally assess the utility of human-human dialogue resources, this paper examines the differences in the user attitude toward human agents in comparison to their attitude toward the automated voice agent, and discusses the implications of these differences for the design and development cycles and –ultimately – for the development of usable interfaces per se. To do that, a corpus of human-human dialogues is compared to a corpus of human-system dialogues, both corpora pertaining to the same domain of customer care.

In the following section the advantages and disadvantages of human-human dialogues are analyzed in comparison to other commonly applied techniques for corpora collection and task analysis. Next, the experimental setup and the measures for comparing the two types of corpora are presented, results are outlined and major findings are discussed in the light of the theoretical and practical issues set out in the introductory section.

2. Corpora collection tools and design techniques comparison

The first stage in the interface lifecycle is the requirements specification and design phase. Common approaches to the above include application simulations, such as the Wizard of Oz method, where a human simulates the behaviour of the system (Fraser & Gilbret, 1991), testing with limited functionality systems ("System in the Loop" method (McTear, 2004)) and rapid prototyping as part of an iterative design process. While testing for usability and design issues, developers can – at the same time – utilize these methods for collecting corpora to train stochastic models for various system components.

Even though analysis of the corpora collected through these methods may allow for iterative design and more informative decisions on dialogue structure early in the development lifecycle, they face certain drawbacks with regards to the quantity and quality of the collected data. More specifically, a typical test session involves 10-15 participants (Cohen et al., 2004), who are asked to perform specific tasks. As a result, the dialogs collected are usually limited in number - taking into account that for a typical large scale application of a ~2000 word vocabulary a training set of at least ~20000 utterances is required - and also lack the realistic aspect of actual system use. Recruited subjects are not motivated in the same way as real users are, and are often not representative of the end user population. Earlier studies have actually shown that there are differences between usability testing and actual use conditions (Turunen et al., 2006).

The human-human dialogue approach, on the other hand, first of all offers the advantage of having an abundance of recordings immediately available for training and design purposes. Availability of recordings eliminates the need to collect caller utterances from scratch as part of the application's implementation phase. This is particularly important for commercial applications, considering the strict industry time-frames and pressing deadlines. Furthermore, dialogues are collected from real users similarly motivated and representative of the end users of the automated system, allowing, thus, for an early, pre-design even, study of actual user behaviour.

Alternatively, corpora from real users can then only be collected during the pilot phase, which comes late in the development process, and so involves the risk of hard and costly changes due to overlooked early design shortcomings. In addition, having adequate resources for ASR and NLU prior to the pilot phase allows the developers to obtain more reliable results from formative evaluation usability tests, which typically precede pilot testing in the development/evaluation process.

On the downside, human-human dialogues are intrinsically less restricted than human-system dialogues, reflecting distinct conversational situations. Callers behave differently; differences may lie in the vocabulary used, the sentence structure, the tempo, the mental model of the interaction, speaker style (e.g. politeness) etc. Thus, whilst developers may observe actual users, they cannot observe actual user-system interaction. As an additional consequence, no data can be collected for dialog situations that typically come up in human-machine interaction alone.

3. Study Description and Experimental Setup

To assess the severity degree of these shortcomings, a corpus of human-human dialogues was compared to a corpus of human-system dialogues. The study was conducted using an open-ended spoken dialog system built for a Customer Care call centre of a Greek Mobile Telephony company. The system performs two major tasks: a) appropriate routing of the client's call to one of approximately 20 dedicated queues, and b) database information retrieval for speech-based self-service modules. In order to assess the validity, usefulness and generalization capability of human-human dialogues, a corpus of 2100 turns (~33100 words) – hereafter referred to as Agent corpus – was collected from existing recordings of phone calls between customers and live agents in the customer care department, prior to the development and deployment of the automated dialog system. The Agent corpus was then compared to a corpus of human-system dialogues – hereafter referred to as System corpus – collected from calls to the automated system. For the creation of the corpora, in both sets of dialogues only the first dialogue turn of the caller was used, where the caller responded to the same initial "How may I help you" question. Furthermore, in order for the corpora to be comparable, it was important that they are of the same size. As the average number of words per turn was significantly higher for the Agent corpus, two versions of the System Corpus were developed, one with the same number of turns as the Agent corpus – hereafter referred to as TSystem corpus – and one with the same number of words – hereafter referred to as WSystem corpus. The Agent corpus was compared against both versions on the basis of the following measures: number of words per turn, part of speech ratio, type/token ratio, term frequency-inverse document frequency, language model perplexity, word error rate, concept error rate and mean ASR confidence. For the last four measures, three distinct statistical language models were trained on each corpus respectively, and tested against a test set of 200 utterances. Apart from that, the same recognition and interpretation resources were employed (e.g. acoustic models, dictionaries, robust interpretation grammars). Other parameters, such as the order of the n-gram model, the discounting and backing-off strategy and the language model scaling factor, were optimized for each model separately. In the following section each measure is presented in detail.

4. Evaluation Measures

Depending on type, the following measures were used to assess differences in style and vocabulary, appropriateness for speech recognition or/and dialog structure and interpretation.

Number of words per dialogue turn: indicate length and complexity of utterances. Note that due to limitations of existing speech recognition and understanding technology all commercial speech platforms introduce specific parameters that define the maximum time users are allowed to talk within a turn, before the system interrupts them. Default values range from 10 to 60 seconds, setting corresponding expectations for the user. In general, too long utterances cause recognition problems, so that it is considered best practice to keep the number of seconds relatively low.

Part of speech distribution: per dialogue turn and total per corpus.

Type/Token ratio and Vocabulary size: The ratio between the distinct words in a text and the total amount of words in a text. It constitutes a measure of lexical density indicating stylistic differences among corpora. Furthermore, the total number of types in an application correlates with the amount of data needed to train the language models for speech recognition; the higher the number, the more data is required.

Term Frequency-Inverse Document Frequency (TF-IDF): Based on the frequency of a term within a corpus and the distribution of terms across corpora, this form is used here as a measure of similarity among corpora providing the list of terms that are useful for discriminating them, i.e. the terms in which they differ. For the purposes of this study the TF-IDF weight is used:

TF-IDF = (C / T) * log(D / DF), where

C = the number of times a word appears in a document

T = the total number of words in the document

D = the total number of documents in a corpus

DF = the number of documents in which a particular word is found.

The term "document" in the above definition refers to the Agent, WSystem or TSystem corpus, while the term "corpus" refers to the {Agent, WSystem} and {Agent, TSystem} corpus set.

Language Model Perplexity: It measures the quality of the language model and correlates with overall speech recognition accuracy; the lower the perplexity, the less confusion in recognition, the better the model. As a general rule, a decrease in perplexity of 10% or more is indicative of better recognition performance.

Word Error Rate (WER): Most common measure for evaluating recognition quality, taking into account false word insertions, deletion and substitutions.

Concept Error Rate (CER): The percentage of interpretation errors (rejections and misinterpretations). Both Word and Concept Error Rate greatly correlate with usability aspects such as user satisfaction, effectiveness and efficiency, and are included in widely spread evaluation schemes (Dybkjær & Bernsen, 2000; Walker et al., 1998). Within the Paradise framework mean concept recognition score is identified as one of the most significant factors for predicting user satisfaction (Kamm & Walker, 1997).

Average ASR Confidence: It practically measures the reliability of the recognition result, and constitutes "one of the most critical components in a practical speech recognition system" (Huang et al, 2001). It has also been used in other studies of differences between spoken corpora collected under diverse conditions (Turunen et al., 2006; Ai et al., 2007)

Lexical analysis measures were computed using the Ellogon Language Engineering Platform (Petasis et al., 2002). Other measures were derived from system log files and appropriate annotations of corpora. The application was built and logs were collected using the DiaManT platform (TR01, 2011).

5. Results

In general, the Agent corpus proved to be more complex, exhibiting significantly longer turns (approximately 3 times longer than the corresponding mean value of the System corpora) larger vocabulary size and higher lexical density (7.8% type/token ratio against 5.4%). Table 1 summarizes the results. Note that, even though the TSystem corpus is characterized by higher type/token ratio, it is not necessarily lexically denser, as it consists of fewer tokens (10500 tokens vs. 33100 of the Agent corpus). Taking into account that the type/token ratio varies with the length of the text, and that longer texts typically exhibit lower density, only the Agent and the

WSystem corpus are safely comparable, as they consist of the same number of tokens.

	Tokens/Turn Mean	Tokens /Turn SD	Type/Token Ratio	Number of Types (Vocab. Size)
Agent Corpus	15.77	9.22	7.8%	2587
TSystem Corpus	5.00	4.58	9.8%	1032
WSystem Corpus	4.85	4.45	5.4%	1781

Table 1. Lexical analysis results

In addition, the --IDF measure (Figures 1 and 2) revealed a significant percentage of highly weighted terms differentiating the Agent corpus from the rest (~70% and 59% when compared to the TSystem and WSystem corpus respectively). Several of these terms – e.g. “obviously”, “firstly”, “therefore”, “last year” – are not identified as key-words and are not important for succesful call routing. On the other hand, most terms in the System corpora are included in the Agent corpus; 74% and 60% of the total number of types in the TSystem and WSystem respectively are assigned a zero value.

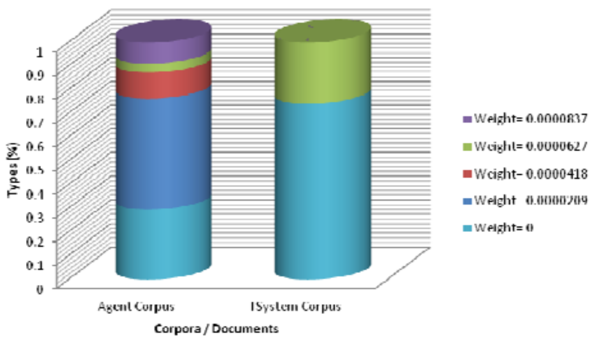

Fig. 1 TF-IDF measure - Agent and TSystem corpus compared: Distribution of weights among word types in each corpus

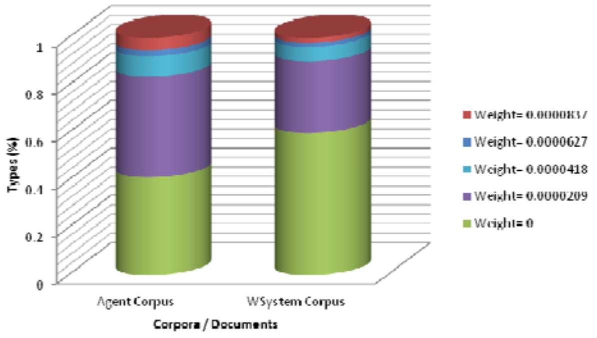

Fig. 2 TF-IDF measure - Agent and WSystem corpus compared: Distribution of weights among word types in each corpus

Table 2 shows the Part of Speech distribution for the most important tag subsets. The most striking difference is the high percentage of nouns of the System corpora compared to the Agent corpus. In contrast, the percentage of verbs is higher for the Agent corpus. Note that noun phrases, as opposed to verb phrases, are commonly used for brevity, conciseness and compactness of expression, as proven by the wide use of noun phrases and nominalizations in captions and titles. Given that nouns are prototypically linked to reference as opposed to predication, and that the discourse topic is commonly a nominal (Lyons, 1977), system users seem to point out what they want to talk about rather than what they want to say or do about it. Also, the percentage of conjunctions is higher for the Agent corpus, indicating subordination and hence more complex sentence structure.

	Nouns	Verbs	Conjunctions	Adjectives	Other
Agent Corpus	28.41%	13.18%	14.22%	6.52%	37.68%
TSystem Corpus	36.68%	8.16%	9.97%	7.07%	38.12%
WSystem Corpus	39.69%	7.83%	9.50%	6.82%	36.16%

Table 2. Part of Speech Distribution per Corpus

Finally, Figure 3 summarizes the results of the evaluation of the language models. The language models trained on the System corpus perform better having an over 100% lower perplexity. Accordingly, they result in a 9-14 and 11-14 percentage point decrease in WER and CER respectively. Furthermore, recognition confidence is slightly higher for these models. Also, note that they result in faster processing times, ranging from 0.19 – 0.20 sec per turn for the System models as opposed to 0.25 sec for the Agent models. As all other recognition parameters that may influence speed, such as pruning, were kept the same, the increased time cost in the case of the Agent model is most likely due to the size and complexity of the model.

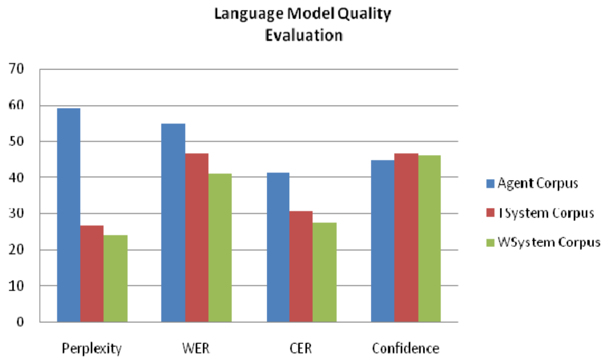

Fig. 3. Evaluation of the language models with respect to perplexity, word error rate (WER), concept error rate (CER) and confidence

6. Discussion and Conclusions

Linguistic comparison of the three corpora indicates that there are significant differences in user's style and attitude towards automated agents compared to human agents. In the first case, users tend to be briefer and less complex. This tendency is reflected upon the shorter sized turns, the lower token/type ratio (i.e. smaller vocabulary size), the limited use of subordination, and the extended use of noun phrases as opposed to verb phrases.

Results of the language model evaluation show that these differences have a significant impact on both recognition and interpretation accuracy and general system performance, corroborating the need for "same style" language resources.

With regards to design considerations, comparison results verify the generally acknowledged importance of setting correct expectations and not misleading the users into believing that they interact with a human as opposed to a machine. Furthermore, it should be noted that certain dialogue situations common in human-machine interaction do not typically come up in human-human interaction. In this telephone based application for example a most common request for "speaking to an agent" does not – as expected – come up in the Agent Corpus (based on the TF-IDF measure the word "agent" ranks among the highest weighted terms), and therefore nor does the handling of such a request, which is particularly important for automated systems in which direct routing to an agent is not an option. Moreover, system users tend to be vaguer in their requests often creating the need for systematic disambiguation. Such situations can only be observed in actual human-system interaction conditions. Therefore the designer needs to be effectively "proactive" or rely on other design resources.

Nevertheless, human-human dialogues are still a useful tool considering the advantages they offer for early observation of actual user behavior, and providing resources for training purposes. It is often the case that such corpora are the only way to obtain resources for developing adequately performing stochastic models for ASR or NLU, and testing with a working system prior to launch. This is particularly true for open-ended systems for which rule-based grammars are not truly an option, as they cannot capture the variability of natural language responses. Therefore, human-human interaction corpora could form a "baseline" solution before the launch of the pilot or the production system. After launch, a "better targeted" corpus should be collected to replace the initial recordings. Furthermore, to avoid investing significant time and resources in transcribing a corpus that will soon be replaced, developers have the option of transcribing a smaller number of recordings and combining them with a generic Statistical Language Model for the ASR module.

It should also be noted that the transcription of human-human dialogues is a more time consuming and costly procedure compared to the transcription of human-system dialogues, as the former are more complicated and difficult to segment on an utterance or dialogue turn basis. In this line of thought and in order to better utilize the human-human recordings, too long or complicated utterances could be rejected during transcription, and long turns could be broken down into smaller utterances. Criteria such as intonation (e.g. final F0 lowering, pausing), syntax (sentence or clause boundary) meaning completeness and dialogue act type (e.g. back-channel, acknowledgement contributions on part of the agent such as "okay" or "uh huh", which the caller does not seem to take into account, are ignored) can be used for consistent utterance segmentation (cf. Heeman and Allen (1994), Nakajima and Allen (1993), Gross et al. (1993)). Hand crafted utterances may also be added to ensure that all the words in the application vocabulary are included in the corpus and can thus be recognized. Based on the results for POS distribution in the system corpora, noun phrases and nominalizations should be preferred, when adding utterances by hand process; however, generalization of this rule to languages other than the language studied here (Greek) should be subject to further investigation.

Alternatively, in some cases and when possible, simple mockup systems constitute a far better solution for corpus collection and design. Depending on the application at hand, they can be ideal for early observation of actual system-user interaction and can further provide high quality corpora for training. To take an example, for the call routing application at hand a simple interface was built aiming to collect the initial user request for a particular customer care service (Stavropoulou et al., 2011). No actual interpretation was attempted, a "How may I help you" prompt was played to the caller and after responding the caller was directly routed to an existing DTMF system. Due to its simplicity – dialogue structure and interpretation-wise – the mock-up was easy and fast to develop and the heavy call load of the call centre allowed for the rapid collection of the required amount of utterances. The collected corpus served as the basis for user centred design, and provided developers with the opportunity to get the feel of the task for usability considerations. Furthermore, the collected corpora were used as resources for the ASR and NLU components of the production system.

In conclusion, human-human dialogues display important differences compared to dialogues collected from actual system use. Differences are severe enough to suggest that human-human dialogues should serve as "bootstrap" corpora alone, even though we would expect the performance of statistical models based on such corpora to significantly improve when more training data are added. Still, developers should take into account the extra effort involved in the transcription and normalization of such resources. Overall, human-human dialogues can still comprise a useful tool, especially when no other options such as mock-ups or rapid prototyping are available, and especially with regards to the formulation of the initial evaluation requirements and high-level design of the application.

7. Acknowledgements

The described work was partially funded by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270919 (SUMAT).

We would also like to thank Georgios Petasis, the author of the Ellogon Language Engineering Platform, for his valuable contribution on the extraction of the corpora statistics for this work.

8. References

Ai, H., Raux, A., Bohus, D., Eskenazi, M., & Litman, D. (2007). Comparing spoken dialog corpora collected with recruited subjects versus real users. In Proc. of the 8th SIGdial workshop on Discourse and Dialogue, pp. 124–131.

Cohen, M., Giancola, J.P. and Balogh, J. (2004). Voice User Interface Design, Addison-Wesley.

Dybkjær, L., & Bernsen, N. O. (2000). Usability Issues in Spoken Language Dialogue Systems. Natural Language Engineering, 6(3-4), 243-272.

Fraser, J., and Gilbret, G. (1991). Simulating speech systems. Computer, Speech and Language 5, 81-99.

Galitz, W.O. (2007). The Essential Guide to User Interface Design, Wiley Publishing, Inc.

Gross, D., Allen, J. and Traum, D. (1993) The Trains dialogues. Trains Technical Note. Department of Computer Science. University of Rochester

Heeman, P.A, and Allen, J.F. (1994). *Dialogue Transcription Tools*. TRAINS Technical Note 94-1. August 1994. University of Rochester.

Huang, X., Acero, A. and Hon, H.W. (2001). Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall PTR.

Kamm, C.A. and Walker, M.A. (1997). Design and Evaluation of Spoken Dialogue Systems. In Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barbara (CA), 14–17.

Lyons, John. 1977. Semantics, vol. 2. Cambridge: Cambridge University Press.

McTear, M. F. (2004). Towards the Conversational User Interface. Springer Verlag.

Nakajima, S. & Allen, J.F. (1993) A study on prosody and discourse structure in cooperative dialogues. Phonetica 50:197-210.

Norman, D. (1988). The Design of Everyday Things. New York: Doubleday/Currency.

Petasis, G., Karkaletsis, V., Paliouras, G., Androutsopoulos I. and Spyropoulos, C. (2002). Ellogon: A New Text Engineering Platform. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands, Spain, vol. I, pp. 72 – 78, May 2002.

Stavropoulou, P., Spiliotopoulos, D., Kouroupetroglou, G. (2011): Design and Development of an Automated Voice Agent: Theory and Practice Brought Together. In D. Perez-Marin and I. Pascual-Nieto (Eds), Conversational Agents and Natural Language Interaction: Techniques and Effective Practices, 2011, Information Science Reference Press (IGI Global), Pennsylvania, USA.

TR01 (2011): DiaManT Platform technical report, http://www.omilia.com/index.php?option=com_content&view=article&id=170&Itemid=294&lang=en

Turunen, M., Hakulinen, J., and Kainulainen, A. (2006). Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences, Interspeech 2006.

Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1998). Evaluating spoken dialogue agents with PARADISE: Two case studies. Computer Speech and Language, 12(3), 317-347.

Weinschenk, S. and Barker, D.T. (2000). Designing effective speech interfaces. John Wiley & Sons, Inc