# Practical Evaluation of Speech Recognizers for Virtual Human Dialogue Systems

**Xuchen Yao*, Pravin Bhutada†, Kallirroi Georgila‡, Kenji Sagae‡, Ron Artstein‡, David Traum‡**

*Faculty of Arts, University of Groningen
†Department of Computer Science and Center for Computational Learning Systems, Columbia University
‡Institute for Creative Technologies, University of Southern California
{yaoxuchen,pravin.bhutada}@gmail.com, {kgeorgila,sagae,artstein,traum}@ict.usc.edu

**Abstract**

We perform a large-scale evaluation of multiple off-the-shelf speech recognizers across diverse domains for virtual human dialogue systems. Our evaluation is aimed at speech recognition consumers and potential consumers with limited experience with readily available recognizers. We focus on practical factors to determine what levels of performance can be expected from different available recognizers in various projects featuring different types of conversational utterances. Our results show that there is no single recognizer that outperforms all other recognizers in all domains. The performance of each recognizer may vary significantly depending on the domain, the size and perplexity of the corpus, the out-of-vocabulary rate, and whether acoustic and language model adaptation has been used or not. We expect that our evaluation will prove useful to other speech recognition consumers, especially in the dialogue community, and will shed some light on the key problem in spoken dialogue systems of selecting the most suitable available speech recognition system for a particular application, and what impact training will have.

## 1. Introduction

This paper evaluates several publicly available Automatic Speech Recognition (ASR) systems, using data collected from deployed spoken dialogue systems. Since ASR systems are typically tuned to specific applications and the environments they operate in, performance is affected by many factors, among them:

- The domain and vocabulary that the recognizer is expected to handle.
- The acoustic environment in which the recognizer operates.
- The speech recognition engine.
- The procedure for adapting the recognizer to a particular domain.
- The possibility for training on individual speakers, and the amount of available user-specific training data.

Additionally, there is often a trade-off between the quality of the speech recognition output and the time it takes to reach that output; real-time conversational systems may be willing to accept a somewhat degraded output in return for lower latencies.

The evaluation described in this paper was performed by consumers of speech recognition systems, not ASR researchers, and is targeted to other ASR consumers and potential consumers with limited experience with readily available recognizers. We focused on practical factors to determine what levels of performance can be expected from different available recognizers in various projects featuring different types of conversational utterances. While comparative evaluations of speech recognizers are available, e.g. (Dybkjaer et al., 1998; Young and Chase, 1998; Devine et al., 2000; Lamel et al., 2000; Broughton, 2002; Berger et al., 2006), we do not know of any other large-scale evaluations of multiple recognizers across diverse domains in a conversational setting, in particular, for virtual human dialogue systems.

The remainder of the paper describes the data used, the ASR engines, the adaptation procedures, and the results of the comparison.

## 2. Data

We evaluated the speech recognizers on six data sets representing different dialogue domains and types of speaker. All of the data sets consist of spontaneous speech spoken in actual conversations from the target user populations for virtual reality interactions.

### 2.1. Dialogue Systems and Domains

Each of our six domains involves conversation between a human participant and one or more virtual characters, except for the domain of the IOTA system (see Section 2.1.6.), which has conversations between two human participants. The data sets derived from collected interactions of humans and virtual characters include only utterances spoken by human participants, and not by the virtual characters. In the remainder of this subsection, we describe each of the domains used in our evaluation, including brief descriptions of the dialogue systems with which the utterances were collected.

#### 2.1.1. SGT Star

SGT Star (Artstein et al., 2009a) is a question-answering character who talks about careers in the Army. Speech collected in the context of the SGT Star system comes from trained handlers who operated SGT Star at job fairs in 2008, presenting to people attending the event.

Interaction with SGT Star is typically in the form of independent direct questions. The following utterances illustrate questions asked by handlers using speech recognition, and SGT Star's answers:

**Handler** Who are you?

**SGT Star** Who, me? Oh, I'm Sergeant Star! I'm a virtual character. Maybe you've seen me before on the web at GoArmy dot com. But there, I'm stuck looking like this. But now with a brilliant blend of pixels, polygons and programming, coupled with real time animation and spoken word interface, I can do my very best to answer your questions on Army careers. And my lips move when I do it.

**Handler** Is the pay good in the Army?

**SGT Star** The Army is a career and like most jobs, you start at the bottom and work your way up. And when you throw in the allowances for housing, meals, uniforms, medical benefits, we're very competitive with the civilian world.

#### 2.1.2. Amani

Amani (Artstein et al., 2009b) is a bargaining character used as a prototype for training soldiers to perform tactical questioning. Speech comes from cadets at the U.S. Military Academy in April 2009, who interacted with Amani as a university course exercise on negotiation techniques.
The following utterances illustrate a segment of a dialogue between a cadet and Amani:

**Cadet** Do you know who did the shooting?

**Amani** So you want to talk about the man who did the shooting? I have heard that he is called Saif.

**Cadet** Do you know where he lives?

**Amani** You are asking for a lot of information. If you expect me to tell you that you must offer me secrecy.

**Cadet** I will keep this a secret.

**Amani** You have assured me of secrecy. I believe he hides on the second floor of Assad's shop. I know he is in there but when I go in the shop I don't see him. And I have no idea where he is now.

#### 2.1.3. SGT Blackwell

SGT Blackwell (Leuski et al., 2006b) is a question-answering character who answers general questions about the Army, himself, and his technology. Speech comes from visitors to the Cooper-Hewitt Museum in New York from December 2006 to March 2007, who interacted with SGT Blackwell at his booth as part of the National Design Triennial exhibition (Robinson et al., 2008).
SGT Blackwell is designed to answer independent questions, like SGT Star. However, the questions collected from sessions with SGT Blackwell come from the general public, and not from trained handlers as was the case with the questions collected from sessions with SGT Star. The museum exhibit listed a set of about five sample questions, but visitors were free to ask anything they wanted. The following utterances illustrate a segment of a dialogue between a museum visitor and SGT Blackwell:

**Museum visitor** What is your favorite color?

**SGT Blackwell** I like red, white, and blue.

**Museum visitor** Why do you like red?

**SGT Blackwell** I am not authorized to comment on that.

#### 2.1.4. SASO

The SASO system (Traum et al., 2008) is a negotiation training prototype in which two virtual characters negotiate with a human "trainee" about moving a medical clinic. Speech was collected at the USC Institute for Creative Technologies (ICT) during 2006–2009, mostly from visitors and new hires.
The following utterances illustrate a segment of a dialogue between a trainee and two virtual characters (Doctor Perez and Elder Al Hassan) in the SASO system:

**Trainee** I have orders to move this clinic to a camp near the U.S. base.

**Elder Al Hassan** We have many matters to attend to.

**Trainee** I understand, but it is imperative that we move the clinic out of this area.

**Elder Al Hassan** This town needs a clinic.

**Doctor Perez** We can't take sides.

**Trainee** Would you be willing to move downtown?

**Elder Al Hassan** We would need to improve water access in the downtown area, captain.

**Trainee** We can dig a well for you.

**Doctor Perez** Captain, we need medical supplies in order to run the clinic downtown.

#### 2.1.5. Radiobots

The Radiobots system (Roque et al., 2006) is a training prototype that responds to military calls for artillery fire in a virtual reality urban combat environment. Speech was collected in 2006 at Fort Sill, Oklahoma, during two evaluation sessions from volunteer trainees who performed calls for specific missions (Robinson et al., 2006).
Examples of user and system utterances in this system are shown below:

**Trainee** M T O kilo alpha four rounds target number alpha bravo one out.

**System** Shot over.

**Trainee** Shot out.

**System** Splash over.

#### 2.1.6. IOTA

IOTA is an extension of the Radiobots system. Speech for the IOTA domain was collected in 2008 from training sessions in the virtual reality environment at Fort Sill between a human trainee and a human instructor on a variety of missions, including some that are similar to Radiobots and others that are more complex. Audio was captured over a simulated radio with reduced sampling rate.
Examples of utterances from a complex mission spoken by a trainee and an instructor are shown below:

| | Words | | | Turns | | | Mean Turn Length |
|---|---|---|---|---|---|---|---|
| | TRAIN | TEST | DEV | TRAIN | TEST | DEV | (TEST) |
| Star | 16340 | 2137 | 2051 | 2974 | 400 | 400 | 5.3 |
| Amani | 15553 | 1855 | 1503 | 1479 | 188 | 187 | 9.9 |
| Blackwell | 80901 | 11520 | 11141 | 17755 | 2500 | 2499 | 4.6 |
| SASO | 22703 | 3483 | 2892 | 3601 | 510 | 466 | 6.8 |
| Radiobots | 6841 | 1163 | 1325 | 1082 | 167 | 190 | 7.0 |
| IOTA | 49633 | 5441 | 6552 | 4939 | 650 | 608 | 8.4 |

Table 1: Data used in the evaluation. Mean turn length is measured in words.

**Trainee** Roger where do you want hog to look from now that I'm looking at that building, where do you want me to go?

**Instructor** Follow the y to the south.

**Trainee** Okay you mean the y that follows to the southwest?

**Instructor** Affirmative.

**Trainee** Roger contact on that east west road.

**Instructor** From that unit from that intersection go west three units of measure.

### 2.2. Creating Data Sets from Collected Utterances

The utterances collected from user sessions in the domains described above were transcribed manually to create a separate corpus for each of the domains. We selected utterances from each corpus randomly to create training, development and test sets: development and test sets were each slightly over 10% of the total utterances (dialogue turns) in each corpus, and the remaining utterances were assigned to the training set. The sizes of the training, development and test sets for each domain are shown in Table 1. We show set sizes in terms of word (token) count and the number of dialogue turns. In addition, we also show the mean turn length for each domain.

The SGT Blackwell corpus is the largest of our six corpora, with a training set containing over 80,000 words, in almost 18,000 dialogue turns. This is also the corpus with the shortest turns on average, with a mean turn length of 4.6. In comparison, the mean turn length in the Amani corpus is more than twice as long, at 9.9. The smallest of the six corpora is the Radiobots corpus, with a training set under 7,000 words and about 1,000 utterances.

Table 2 shows the vocabulary size and density for the training set in each domain. Size is the number of unique words, or types, in each of the training sets, and density is the total number of tokens divided by the vocabulary size.[1] The table also shows the number of words in the development set that are not in the training set vocabulary, or out-of-vocabulary (OOV) words. Counts are included for OOV types and tokens in the development set. Finally, the table also includes the OOV rate, defined as the OOV token count divided by the total number of tokens in the development set. Vocabulary size and OOV rate are indicative of the difficulty of the recognition task in these specific domains. Tables 1 and 2 suggest, for example, that the amount of data collected in the Amani domain may be inadequate, given the small training set size and high OOV rate. Although the Radiobots corpus is even smaller, its vocabulary size is very small, and its OOV rate low.

| | Vocabulary | | OOV | OOV tokens | |
|---|---|---|---|---|---|
| | size | density | types | N | % |
| Star | 516 | 31.7 | 35 | 37 | 1.80 |
| Amani | 1194 | 13.0 | 58 | 67 | 4.46 |
| Blackwell | 2568 | 31.5 | 128 | 147 | 1.32 |
| SASO | 808 | 28.1 | 38 | 43 | 1.49 |
| Radiobots | 198 | 34.6 | 16 | 18 | 1.36 |
| IOTA | 1878 | 26.4 | 114 | 143 | 2.18 |

Table 2: The vocabulary size and density of the training set for each corpus, the number of unique out-of-vocabulary (OOV) words in each development set, and the total number and rate of out-of-vocabulary words in each development set.

## 3. Methodology

### 3.1. General Steps

The following open source recognizers were used in the evaluation:

Cambridge HTK family: HVite (v3.4.1), HDecode and Julius (v4.1.2).[2]

CMU Sphinx family: Sphinx 4 and Pocket Sphinx (v0.5).[3]

For these recognizers, acoustic models and language models were first trained on the training set (TRAIN). Then the recognizers were tuned on the development set (DEV) and the final result was calculated on the test set (TEST).

### 3.2. Acoustic and Language Models

Acoustic models and language models were trained as follows.

[1] The density figure is not normalized to the size of the corpus; generally, a higher number (indicating lower density) is expected for larger corpora

[2] HTK is available from `http://htk.eng.cam.ac.uk`; Julius `http://julius.sourceforge.jp` is compatible with acoustic and language models trained using HTK so we include it with the HTK family

[3] Both are available from `http://cmusphinx.sourceforge.net`

| | Non Real-time | | | Real-time | |
|---|---|---|---|---|---|
| | HVite | HDecode | Sphinx4 | Julius | PocketSphinx |
| Star | 22 | 20 | 33 | 27 | 25 |
| Amani | 47 | 49 | 42 | 54 | 35 |
| Blackwell | 34 | 31 | 60 | 35 | 49 |
| SASO | 32 | 28 | 32 | 33 | 36 |
| Radiobots | 10 | 11 | — | 17 | 7 |
| IOTA | 66 | 49 | 76 | 61 | 55 |

Table 3: Word error rates on the various DEV sets (best results achieved after tuning the parameters).

| | Non Real-time | | Real-time | |
|---|---|---|---|---|
| | HVite | HDecode | Julius | PocketSphinx |
| Star | 33 | 32 | 36 | 33 |
| Amani | 56 | 65 | 50 | 38 |
| Blackwell | 32* | 42 | 32 | 53 |
| SASO | 33 | 29 | 33 | 30 |
| Radiobots | 15 | 12 | 14 | 10 |
| IOTA | 57 | 39 | 42 | 47 |

Table 4: Word error rates on the various TEST sets. Note that the result of HVite on Blackwell is based on only 10% of the data set. To facilitate comparisons the WER of HDecode and Julius on the same portion of Blackwell was 46% and 36% respectively.

**HTK family:** The three decoders used the same acoustic and language models. We used two sets of these models: in one set both models were trained only on TRAIN so they highly fit a specific data set; in the other set both models were adapted with the Wall Street Journal (WSJ) training corpus (Vertanen, 2006)[4]. The training procedure follows Young et al. (2006).

**Sphinx family:** A language model was built from TRAIN of each data set with the CMU SLM toolkit (Clarkson and Rosenfeld, 1997)[5], while the acoustic models were adapted with the WSJ corpus using CMU's SphinxTrain tool[6]. We used the WSJ acoustic models distributed by CMU.

The CMU pronouncing dictionary v0.7a (Weide, 2008) was used as the main dictionary for both of the HTK and Sphinx family. We used trigrams throughout our experiments with the Sphinx family of recognizers. On the other hand, both bigrams and trigrams were used with the HTK family of recognizers (except for HVite, which supports only bigrams).

### 3.3. Evaluation Method

Our main evaluation metric was word error rate (WER). WER was calculated by the *HResults* program of HTK. It can be formulated as:

$$\text{WER} = \frac{\text{Substitutions} + \text{Deletions} + \text{Insertions}}{\text{Length of target string}}$$

Additionally, we note whether the recognition was real-time or not. A real-time recognizer can finish recognizing a segment of speech in a time interval no greater than the length of the speech.

We also measure perplexity as an indication of the complexity of each corpus. Perplexity is a common way of evaluating language models with respect to some text. Perplexity (equation 1) is derived from cross-entropy (equation 2):

$$PP = 2^{H(T)} \tag{1}$$

$$H(T) = -\tfrac{1}{W_T} log_2 P(T) \tag{2}$$

where $P(T)$ is the probability that the language model assigns to text $T$ and $W_T$ is the number of tokens (words) in text $T$. A perplexity of value $N$ means that at each point in the recognition path the recognizer has to choose among $N$ words on average. Thus the lower the perplexity the easier the speech recognition task.

## 4. Results and Discussion

### 4.1. Results Overview

Tables 3 and 4 show the performance of the various recognizers on the different data sets. Table 3 shows the results for each of the recognizers on the DEV set. In cases where multiple language models were trained for one engine, we took the best performing one. More details on individual language model performance for the HTK family are provided in Section 4.2. below. Table 4 shows the performance of recognizers on the TEST set, which had not been examined during model selection and tuning. Several conclusions can be drawn from the tables. First, there are a lot of errors in many domains. This underscores the point that ASR for conversational speech is still a challenging task

[4]Available from `http://www.keithv.com/software`

[5]Available from `http://www.speech.cs.cmu.edu/SLM_info.html`

[6]Available from `http://cmusphinx.sourceforge.net/html/download.php#SphinxTrain`

| | HVite | | HDecode | | | | Julius | | |
|---|---|---|---|---|---|---|---|---|---|
| | Adapted | Unadapted | Adapted | | Unadapted | | Adapted | Unadapted | |
| | bigram | bigram | bigram | trigram | bigram | trigram | bigram | bigram | trigram |
| Star | 22 | 23 | 21 | 20 | 23 | 22 | 27 | 26 | 27 |
| Amani | 56 | 47 | 52 | 50 | 51 | 49 | 54 | 56 | 62 |
| Blackwell | 34 | 44 | 34 | 31 | 47 | 45 | 35 | 46 | 53 |
| SASO | 32 | 32 | 29 | 28 | 36 | 34 | 33 | 38 | 44 |
| Radiobots | 14 | 10 | 13 | 13 | 12 | 11 | 17 | 18 | 18 |
| IOTA | 71 | 66 | 49 | 49 | 66 | 66 | 61 | 69 | 82 |

Table 5: Comparison of WER for the HTK family considering adaptation on DEV (using both bigrams and trigrams).

and further work is needed on ASR performance and NLU and dialogue techniques to cope with high error rates, e.g. (Leuski et al., 2006a). Second, there are large differences in the recognition rates for the different domains. This underscores the need for further domain typology for virtual humans. Some of these differences may be an artifact of the size of the collected data set, but other aspects concern the domain itself, e.g. size of turns, size of vocabulary, how specialized the vocabulary is, density, perplexity and OOV rate. Virtual human designers may need to pay attention to how people will want to talk in a given domain and the implications for ASR performance. Third, no one recognizer dominates on all data sets, e.g. Julius works best on Blackwell, but is significantly worse than Pocket Sphinx on Radiobots and Amani. The upshot is that training for specific domains is important, and choice of recognizer may again depend on aspects of the domain.

### 4.2. Adaptation Affects Performance

For the HTK family, we did experiments to evaluate bigram vs. trigram language models and whether adapting with both WSJ acoustic and language models helps improve WER.

Table 5 shows the comparison. Again, no one technique dominates. HVite is better with adaptation for Blackwell, but worse for IOTA and Amani. HDecode does best with adapted trigrams for most domains, but unadapted trigrams are best for Radiobots, while trigrams perform at least as well as bigrams for all domains. Adaptation also is optimal for Julius, while bigrams perform better than trigrams for most domains.

Due to the fact that the WSJ corpus is much larger than our data sets, the final adapted models are also much enlarged. This brings a decrease in decoding speed because the search space is widened. The consequences of bigger search space could be two-fold. On one hand, enriched models could compensate for data sparsity and thus lower WER. This appears to be the case for the Blackwell domain, where enriched models cause a drop of more than 10 percentage points in WER for all three decoders, and may also be the cause for the lower drop in WER for the SASO domain. On the other hand, if a data set covers only a closed domain and uses small-size vocabulary, then the additional hypotheses of the enriched models make it more difficult to find the correct interpretation. This may explain the increase in WER with adapted models for the Radiobots domain (for HVite and HDecode).

### 4.3. Perplexity Affects Performance

Table 6 presents perplexity results using both unadapted bigrams and unadapted trigrams on the TRAIN, DEV and TEST data sets of each domain. As expected perplexity is lower on TRAIN since this data set was used for training the language models. Also, trigrams lead to lower perplexities than bigrams. The perplexity on IOTA is very high, especially on DEV, which explains the high WER. On the other hand, the perplexity is low for Radiobots, which explains the low WER for this domain. To calculate perplexity we used the SRI SLM toolkit (Stolcke, 2002)[7].

## 5. Conclusion

We performed an evaluation of multiple off-the-shelf speech recognizers across diverse domains for virtual human dialogue systems. Our evaluation is targeted to ASR consumers and potential consumers with limited experience with readily available recognizers. Our results show that there is no single recognizer that outperforms all other recognizers in all domains.

We expect that our evaluation will prove useful to other ASR consumers, especially in the dialogue community, and will shed some light on the key problem in spoken dialogue systems of selecting the most suitable available ASR system for a particular application, and what impact training will have.

In future work we intend to incorporate these recognizers into our system architectures, so that we can test the effect of each ASR engine on the overall user experience while he/she interacts with the dialogue system. We also intend to work towards developing a regression model that will help us predict which ASR system will perform best based on the characteristics of the domain.

## 6. Acknowledgments

The work described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position of the United States Government, and no official endorsement should be inferred.

The first author acknowledges financial support from the Erasmus Mundus Program through scholarships for the European Masters Program in Language and Communication Technologies (LCT).

[7]Available from `http://www.speech.sri.com/projects/srilm/`

| | TRAIN | | DEV | | TEST | |
|---|---|---|---|---|---|---|
| | bigram | trigram | bigram | trigram | bigram | trigram |
| Star | 7.1 | 4.8 | 10.7 | 7.9 | 13.2 | 10.4 |
| Amani | 26.7 | 17.6 | 47.1 | 39 | 52.9 | 45 |
| Blackwell | 10.8 | 8.1 | 11.5 | 8.9 | 12.3 | 9.9 |
| SASO | 10.1 | 9.9 | 15 | 13.3 | 17.8 | 15.9 |
| Radiobots | 4.8 | 3.7 | 5.5 | 4.7 | 4.9 | 4.2 |
| IOTA | 34.3 | 24.3 | 60.4 | 53 | 34.1 | 27.7 |

Table 6: Perplexity for the HTK family on TRAIN, DEV and TEST data sets (using both unadapted bigrams and unadapted trigrams).

## 7. References


R. Artstein, S. Gandhe, J. Gerten, A. Leuski, and D. Traum. 2009a. Semi-formal evaluation of conversational characters. In Orna Grumberg, Michael Kaminski, Shmuel Katz, and Shuly Wintner, editors, *Languages: From Formal to Natural. Essays Dedicated to Nissim Francez on the Occasion of His 65th Birthday*, volume 5533 of *Lecture Notes in Computer Science*, pages 22–35. Springer, Berlin.

R. Artstein, S. Gandhe, M. Rushforth, and D. Traum. 2009b. Viability of a simple dialogue act scheme for a tactical questioning dialogue system. In *DiaHolmia 2009: Proc. of 13th Workshop on the Semantics and Pragmatics of Dialogue*, Stockholm, Sweden.

S. Berger, Z.A. Sloane, and J. Yang. 2006. Competitive evaluation of commercially available speech recognizers in multiple languages. In *Proc. of Fifth International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy.

M. Broughton. 2002. Measuring the accuracy of commercial automated speech recognition systems during conversational speech. In *Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges*, Melbourne, Australia.

P. Clarkson and R. Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proc. of Eurospeech*, Rhodes, Greece.

E.G. Devine, S.A. Gaehde, and A.C. Curtis. 2000. Comparative evaluation of three continuous speech recognition software packages in the generation of medical reports. *The Journal of American Medical Informatics Association*, 7(5):462–468.

L. Dybkjaer, N.O. Bernsen, R. Carlson, L. Chase, N. Dahlbaeck, K. Failenschmid, U. Heid, P. Heisterkamp, A. Joenson, H. Kamp, I. Karlsson, J. v. Kuppevelt, L. Lamel, P. Paroubek, and D. Williams. 1998. The DISC approach to spoken language dialog systems development and evaluation. In *Proc. of First International Conference on Language Resources and Evaluation (LREC)*, pages 185–189, Granada, Spain.

L. Lamel, W. Minker, and P. Paroubek. 2000. Towards best practice in the development and evaluation of speech recognition components of a spoken language dialog system. *Natural Language Engineering*, 6(3–4):305–322.

A. Leuski, B. Kennedy, R. Patel, and D. Traum. 2006a. Asking questions to limited domain virtual characters: How good does speech recognition have to be? In *25th Army Science Conference*, Orlando, Florida, USA.

A. Leuski, R. Patel, D. Traum, and B. Kennedy. 2006b. Building effective question answering characters. In *Proc. of 7th SIGdial Workshop on Discourse and Dialogue*, pages 18–27, Sydney, Australia.

S.M. Robinson, A. Roque, A. Vaswani, D. Traum, C. Hernandez, and B. Millspaugh. 2006. Evaluation of a spoken dialogue system for virtual reality call for fire training. In *25th Army Science Conference*, Orlando, Florida, USA.

S. Robinson, D. Traum, M. Ittycheriah, and J. Henderer. 2008. What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *Proc. of Sixth International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

A. Roque, A. Leuski, V. Rangarajan, S. Robinson, A. Vaswani, S. Narayanan, and D. Traum. 2006. Radiobot-CFF: A spoken dialogue system for military training. In *Proc. of Interspeech*, Pittsburgh, Pennsylvania, USA.

A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, Denver, Colorado, USA.

D.R. Traum, S. Marsella, J.Gratch, J.Lee, and A. Hartholt. 2008. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Proc. of 8th International Conference on Intelligent Virtual Agents (IVA)*, Tokyo, Japan.

K. Vertanen. 2006. Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments. Technical report, Cavendish Laboratory, University of Cambridge.

R.L. Weide. 2008. The CMU pronouncing dictionary.

S.J. Young and L.L. Chase. 1998. Speech recognition evaluation: a review of the U.S. CSR and LVCSR programmes. *Computer Speech and Language*, 12(4):263–279.

S.J. Young, G. Evermann, M.J.F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P.C. Woodland. 2006. *The HTK Book, version 3.4*. Cambridge University Engineering Department, Cambridge, UK.