Generating statistical language models from interpretation grammars in dialogue systems

Rebecca Jonson
Dept. of Linguistics, Göteborg University and GSLT
rj@ling.gu.se

Abstract

In this paper, we explore statistical language modelling for a speech-enabled MP3 player application by generating a corpus from the interpretation grammar written for the application with the Grammatical Framework (GF) (Ranta, 2004). We create a statistical language model (SLM) directly from our interpretation grammar and compare recognition performance of this model against a speech recognition grammar compiled from the same GF interpretation grammar. The results show a relative Word Error Rate (WER) reduction of 37% for the SLM derived from the interpretation grammar while maintaining a low in-grammar WER comparable to that associated with the speech recognition grammar. From this starting point we try to improve our artificially generated model by interpolating it with different corpora achieving great reduction in perplexity and 8% relative recognition improvement.

1 Introduction

Ideally when building spoken dialogue systems, we would like to use a corpus of transcribed dialogues corresponding to the specific task of the dialogue system, in order to build a statistical language model (SLM). However, it is rarely the case that such a corpus exists in the early stage of the development of a dialogue system. Collecting such a corpus and transcribing it is very time-consuming and delays the building of the actual dialogue system.

An approach taken both in dialogue systems and dictation applications is to first write an interpretation grammar and from that generate an artificial corpus which is used as training corpus for the SLM (Raux *et al*, 2003; Pakhomov *et al*, 2001; Fosler-Lussier & Kuo, 2001). These models obtained from grammars are not as good as the ones built from real data as the estimates are artificial, lacking a real distribution. However, it is a quick way to get a dialogue system working with an SLM. When the system is up and running it is possible to collect real data that can be used to improve the model. We will explore this idea by generating a corpus from an interpretation grammar from one of our applications.

A different approach is to compile the interpretation grammar into a speech recognition grammar as the Gemini and REGULUS compilers do (Rayner *et al*, 2000; Rayner *et al*, 2003). In this way it is assured that the linguistic coverage of the speech recognition and interpretation are kept in sync. Such an approach enables us to interpret all that we can recognize and the other way round. In the European-funded project TALK the Grammatical Framework (Ranta, 2005) has been extended with such a facility that compiles GF grammars into speech recognition grammars in Nuance GSL format (www.nuance.com).

Speech recognition for commercial dialogue systems has focused on grammar-based approaches despite the fact that statistical language models seem to have a better overall performance (Gorrell *et al*, 2002). This probably depends on the time-consuming work of collecting corpora for training SLMs compared with the more rapid and straightforward development of speech recognition grammars. However, SLMs are more robust, can handle out-of-coverage output, perform better in difficult conditions and seem to work bet-

ter for naive users (see (Knight *et al*, 2001)) while speech recognition grammars are limited in their coverage depending on how well grammar writers succeed in predicting what users may say (Huang *et al*, 2001).

Nevertheless, as grammars only output phrases that can be interpreted their output makes the following interpretation task easier than with the unpredictable output from an SLM (especially if the speech recognition grammar has been compiled from the interpretation grammar and these are both in sync). In addition, the grammar-based approach in the experiments reported in (Knight *et al*, 2001) outperforms the SLM approach on semantic error rate on in-coverage data. This has lead to the idea of trying to combine both approaches, as shown in (Rayner & Hockey, 2003). This is also something that we are aiming for.

Domain adaptation of SLMs is another issue in dialogue system recognition which involves reusing a successful language model by adapting it to a new domain i.e. a new application (Janiszek *et al*, 1998). If a large corpus is not available for the specific domain but there is a corpus for a collection of topics we could use this corpus and adapt the resulting SLM to the domain. One may assume that the resulting SLM based on a large corpus with a good mixture of topics should be able to capture at least a part of general language use that does not vary from one domain to another. We will explore this idea by using the Gothenburg Spoken Language Corpus (GSLC) (Allwood, 1999) and a newspaper corpus to adapt these to our MP3 domain.

We will consider several different SLMs based on the corpus generated from the GF interpretation grammar and compare their recognition performance with the baseline: a Speech Recognition Grammar in Nuance format compiled from the same interpretation grammar. Hence, what we could expect from our experiment, by looking at earlier research, is very low word error rate for our speech recognition grammar on in-grammar coverage but a lot worse performance on out-of-grammar coverage. The SLMs we are considering should tackle out-of-grammar utterances better and it will be interesting to see how well these models built from the grammar will perform on in-grammar utterances.

This study is organized as follows. Section 2 introduces the domain for which we are doing language modelling and the corpora we have at our disposal. Section 3 will describe the different SLMs we have generated. Section 4 describes the evaluation of these and the results. Finally, we review the main conclusions of the work and discuss future work.

2 Description of Corpora

2.1 The MP3 corpus

The domain that we are considering in this paper is the domain of an MP3 player application. The talking MP3 player, DJGoDiS, is one of several applications that are under development in the TALK project. It has been built with the TrindiKit toolkit (Larsson *et al*, 2002) and the GoDiS dialogue system (Larsson, 2002) as a GoDiS application and works as a voice interface to a graphical MP3 player. The user can among other things change settings, choose stations or songs to play and create playlists. The current version of DJGoDiS works in both English and Swedish.

The interpretation and generation grammars are written with the GF grammar formalism. GF is being further developed in the project to adapt it to the use in spoken dialogue systems. This adaptation includes the facility of generating Nuance recognition grammars from the interpretation grammar and the possibility of generating corpora from the grammars. The interpretation grammar for the domain, written in GF, translates user utterances to dialogue moves and thereby holds all possible interpretations of user utterances (Ljunglöf *et al*, 2005). We used GF's facilities to generate a corpus in Swedish consisting of all possible meaningful utterances generated by the grammar to a certain depth of the analysis trees in GF's abstract syntax as explained in (Weilhammer *et al*, 2006). As the current grammar is under development it is not complete and some linguistic structures are missing. The grammar is written on the phrase level accepting spoken language utterances such as e.g. "next, please".

The corpus of possible user utterances resulted in around 320 000 user utterances (about 3 million words) corresponding to a vocabulary of only 301 words. The database of songs and artists in this first version of the application is limited to 60 Swedish songs, 60 Swedish artists, 3 albums and 3 radio stations. The vocabulary may seem small if you consider the number of songs and artists included, but the small size is due to a huge

overlap of words in songs and artists as pronouns (such as *Jag* (*I*) and *Du* (*You*)) and articles (such as *Det* (*The*)) are very common. This corpus is very domain specific as it includes many artist names, songs and radio stations that often consist of rare words. It is also very repetitive covering all combinations of songs and artists in utterances such as "I want to listen to Mamma Mia with Abba". All utterances in the corpus occur exactly once.

2.2 The GSLC corpus

The Gothenburg Spoken Language (GSLC) corpus consists of transcribed Swedish spoken language from different social activities such as auctions, phone calls, meetings, lectures and task-oriented dialogue (Allwood, 1999). To be able to use the GSLC corpus for language modelling it was pre-processed to remove annotations and all non-alphabetic characters. The final GSLC corpus consisted of a corpus of about 1,300,000 words with a vocabulary of almost 50,000 words.

2.3 The newspaper corpus

We have also used a corpus consisting of a collection of Swedish newspaper texts of 397 million words.[1] Additionally, we have created a subcorpus of the newspaper corpus by extracting only the sentences including domain related words. With domain related words we mean typical words for an MP3 domain such as "music", "mp3-player", "song" etc. This domain vocabulary was handcrafted. The domain-adapted newspaper corpus, obtained by selecting sentences where these words occurred, consisted of about 15 million words i.e. 4% of the larger corpus.

2.4 The Test Corpus

To collect a test set we asked students to describe how they would address a speech-enabled MP3 player by writing Nuance grammars that would cover the domain and its functionality. Another group of students evaluated these grammars by recording utterances they thought they would say to an MP3 player. One of the Nuance grammars was used to create a development test set by generating a corpus of 1500 utterances from it. The corpus generated from another grammar written by some other students was used as evaluation test set. Added to the evaluation test set were the transcriptions of the recordings made by the third group of students that evaluated both grammars. This resulted in a evaluation test set of 1700 utterances.

The recording test set was made up partly of the students' recordings. Additional recordings were carried out by letting people at our lab record randomly chosen utterances from the evaluation test set. We also had a demo running for a short time to collect user interactions at a demo session. The final test set included 500 recorded utterances from 26 persons. This test set has been used to compare recognition performance between the different models under consideration.

The recording test set is just an approximation to the real task and conditions as the students only capture how they think they would act in an MP3 task. Their actual interaction in a real dialogue situation may differ considerably so ideally, we would want more recordings from dialogue system interactions which at the moment constitutes only a fifth of the test set. However, until we can collect more recordings we will have to rely on this approximation.

In addition to the recorded evaluation test set, a second set of recordings was created covering only in-grammar utterances by randomly generating a test set of 300 utterances from the GF grammar. These were recorded by 8 persons. This test set was used to contrast with a comparison of in-grammar recognition performance.

3 Language modelling

To generate the different trigram language models we used the SRI language modelling toolkit (Stolcke, 2002) with Good-Turing discounting.

The first model was generated directly from the MP3 corpus we got from the GF grammar. This simple SLM (named `MP3GFLM`) has the same vocabulary as the Nuance Grammar and models the same language as the GF grammar. This model was chosen to see if we could increase flexibility and robustness in such a simple way while maintaining in-grammar performance.

We also created two other simple SLMs: a class-based one (with the classes `Song`, `Artist` and `Radiostation`) and a model based on a variant of the MP3 corpus where the utterances in which songs and artists co-occur would only match real artist-song pairs (i.e. including some music knowledge in the model).

These three SLMs were the three basic MP3

[1] This corpus was made available by Leif Grönqvist, Dept. of Linguistics, Göteborg University

models considered although we only report the results for the MP3GFLM in this article (the class-based model gave a slightly worse result and the a other slightly better result).

In addition to this we used our general corpora to produce three different models: GSLCLM from the GSLC corpus, NewsLM from the newspaper corpus and DomNewsLM from the domain adapted newspaper Corpus.

3.1 Interpolating the GSLC corpus and the MP3 corpus

A technique used in language modelling to combine different SLMs is linear interpolation (Jelinek & Mercer, 1980). This is often used when the domain corpus is too small and a bigger corpus is available. There have been many attempts at combining domain corpora with news corpora, as this has been the biggest type of corpus available and this has given slightly better models (Janiszek *et al*, 1998; Rosenfeld, 2000a). Linear interpolation has also been used when building state dependent models by combining the state models with a general domain model (Xu & Rudnicky, 2000; Solsona *et al*, 2002).

Rosenfeld (Rosenfeld, 2000a) argues that a little more domain corpus is always better than a lot more training data outside the domain. Many of these interpolation experiments have been carried out by adding news text, i.e. written language. In this experiment we are going to interpolate our domain model (MP3GFLM) with a spoken language corpus, the GSLC, to see if this improves perplexity and recognition rates. As the MP3 corpus is generated from a grammar without probabilities this is hopefully a way to obtain better and more realistic estimates on words and word sequences. Ideally, what we would like to capture from the GSLC corpus is language that is invariant from domain to domain. However, Rosenfeld (Rosenfeld, 2000b) is quite pessimistic about this, arguing that this is not possible with today's interpolation methods. The GSLC corpus is also quite small.

The interpolation was carried out with the SRILM toolkit[2] based on equation 1.

$$MixGSLCMP3GF = \lambda * MP3GFLM + (1 - \lambda) * GSLCLM \quad (1)$$

The optimal lambda weight was estimated to 0.65 with the SRILM toolkit using the development test set.

[2]http://www.speech.sri.com/projects/srilm, as of 2005.

3.2 Interpolating the newspaper corpus and the MP3 corpus

We also created two models in the same way as above by interpolating the two variants of the news corpus with our simplest model.

$$MixNewsMP3GF = \lambda * MP3GFLM + (1 - \lambda) * NewsLM \quad (2)$$

$$MixDomNewsMP3GF = \lambda * MP3GFLM + (1 - \lambda) * DomNewsLM \quad (3)$$

In addition to these models we created a model where we interpolated both the GSLC model and the domain adapted newspaper model with MP3GFLM. This model was named TripleLM.

3.2.1 Choice of vocabulary

The resulting mixed models have a huge vocabulary as the GSLC corpus and the newspaper corpus include thousands of words. This is not a convenient size for recognition as it will affect accuracy and speed. Therefore we tried to find an optimal vocabulary combining the small MP3 vocabulary of around 300 words with a smaller part of the GSLC vocabulary and the newspaper vocabulary.

We used the the CMU toolkit (Clarkson & Rosenfeld, 1997) to obtain the most frequent words of the GSLC corpus and the News Corpus. We then merged these vocabularies with the small MP3 vocabulary. It should be noted that the overlap between the most frequent GSLC words and the MP3 vocabulary was quite low (73 words for the smallest vocabulary) showing the peculiarity of the MP3 domain. We also added the vocabulary used for extracting domain data to this mixed vocabulary. This merging of vocabularies resulted in a vocabulary of 1153 words. The vocabulary for the MP3GFLM and the MP3NuanceGr is the small MP3 vocabulary.

4 Evaluation and Results

4.1 Perplexity measures

The 8 SLMs (all using the vocabulary of 1153 words) were evaluated by measuring perplexity with the tools SRI provides on the evaluation test set of 1700 utterances.

In Table 1 we can see a dramatic perplexity reduction with the mixed models compared to the simplest of our models the MP3GFLM. Surprisingly, the GSLCLM models the test set better than

Table 1: *Perplexity for the different SLMs.*

LM	Perplexity
MP3GFLM	587
GSLCLM	350
NewsLM	386
DomNewsLM	395
MixGSLCMP3GF	65
MixNewsMP3GF	78
MixDomNewsMP3GF	88
TripleLM	64

the `MP3GFLM` which indicates that our MP3 grammar is too restricted and differs considerably from the students' grammars.

Lower perplexity does not necessarily mean lower word error rates and the relation between these two measures is not very clear. One of the reasons that language model complexity does not measure the recognition task complexity is that language models do not take into account acoustic confusability (Huang *et al*, 2001; Jelinek, 1997). According to Rosenfeld (Rosenfeld, 2000a), a perplexity reduction of 5% is usually practically not significant, 10-20% is noteworthy and a perplexity reduction of 30% or more is quite significant. The above results of the mixed models could then mean an improvement in word error rate over the baseline model `MP3GFLM`. This has been tested and is reported in the next section. In addition, we want to test if we can reduce word error rate using our simple SLM opposed to the Nuance grammar (`MP3NuanceGr`) which is our recognition baseline.

4.2 Recognition rates

The 8 SLMs under consideration were converted with the SRILM toolkit into a format that Nuance accepts and then compiled into recognition packages. These were evaluated with Nuance's batch recognition program on the recorded evaluation test set of 500 utterances (26 speakers). Table 2 presents word error rates (WER) and in parenthesis N-Best (N=10) WER for the models under consideration and for the Nuance Grammar.

As seen, our simple SLM, `MP3GFLM`, improves recognition performance considerably compared with the Nuance grammar baseline (`MP3NuanceGr`) showing a much more robust behaviour to the data. Remember that these two models have the same vocabulary and are both derived from the same GF interpretation grammar. However the flexibility of the SLM gives a relative improvement of 37% over the Nuance grammar. The models giving the best results are the models interpolated with the GSLC corpus and the domain news corpus in different ways which at best gives a relative reduction in WER of 8% in comparison with `MP3GFLM` and 43% compared with the baseline. It is interesting to see that the simple way we used to create a domain specific newspaper corpus gives a model that better fits our data than the original much larger newspaper corpus.

Table 2: *Word error rates(WER) for the recording test set*

LM	WER(NBest)
MP3GFLM	37.11 (29.48)
GSLCLM	83.04 (71.51)
NewsLM	61.62 (49.53)
DomNewsLM	45.03 (31.58)
MixGSLCMP3GF	34.58 (22.68)
MixNewsMP3GF	38.00 (27.37)
MixDomNewsMP3GF	34.07 (22.07)
TripleLM	33.97 (22.02)
MP3NuanceGr	59.37 (53.19)

4.3 In-grammar recognition rates

To contrast the word error rate performance with in-grammar utterances i.e. utterances that the original GF interpretation grammar covers, we carried out a second evaluation with the in-grammar recordings. We also used Nuance's parsing tool to extract the utterances that were in-grammar from the recorded evaluation test set. These few recordings (5%) were added to the in-grammar test set. The results of the second recognition experiment are reported in Table 3.

Table 3: *WER on the in-grammar test set*

LM	WER (NBest)
MP3GFLM	4.95 (2.04)
GSLCLM	78.07 (64.15)
NewsLM	48.03 (36.64)
DomNewsLM	26.34 (15.25)
MixGSLCMP3GF	14.23 (6,29)
MixNewsMP3GF	18.63 (10.22)
MixDomNewsMP3GF	15.57 (6.13)
TripleLM	15.17 (6.05)
MP3NuanceGr	3.69 (1.49)

The in-grammar results reveal an increase in WER for all the SLMs in comparison to the baseline `MP3NuanceGr`. However, the simplest model (`MP3GFLM`), modelling the language of the grammar, do not show any greater reduction in recognition performance.

4.4 Discussion of results

The word error rates obtained for the best models show a relative improvement over the Nuance grammar of 40%. The most interesting result is that the simplest of our models, modelling the same language as the Nuance grammar, gives such an important gain in performance that it lowers the WER with 22%. We used the Chi-square test of significance to statistically compare the results with the results of the Nuance grammar showing that the differences of WER of the models in comparison with the baseline are all significant on the p=0.05 significance level. However, the Chi-square test points out that the difference of WER for in-grammar utterances of the Nuance model and the `MP3GFLM` is significant on the p=0.05 level. This means that all the statistical language models significantly outperform the baseline i.e. the Nuance Grammar `MP3NuanceGr` on the evaluation test set (being mostly out-of-coverage) and that the `MP3GFLM` outperforms the baseline overall as the difference of WER in the in-grammar test is significant but very small.

However, as the reader may have noticed, the word error rates are quite high, which is partly due to a totally independent test set with out-of-vocabulary words (9% OOV for the `MP3GFLM`) indicating that domain language grammar writing is very subjective. The students have captured a quite different language for the same domain and functionality. This shows the risk of a hand-tailored domain grammar and the difficulty of predicting what users may say. In addition, a fair test of the model would be to measure concept error rate or more specifically dialogue move error rate (i.e. both 'yes' and 'yeah' correspond to the same dialogue move `answer(yes)`). A closer look at the `MP3GFLM` results give a hint that in many cases the transcription reference and the recognition hypothesis hold the same semantic content in the domain (e.g. confusing the Swedish prepositions 'i' (into) and 'till' (to) which are both used when referring to the playlist). It was manually estimated that 53% of the recognition hypotheses could be considered as correct in this way opposed to the 65% Sentence Error Rate (SER) that the automatic evaluation gave. This implies that the evaluation carried out is not strictly fair considering the possible task improvement. However, a fair automatic evaluation of dialogue move error rate will be possible only when we have a way to do semantic decoding that is not entirely dependent on the GF grammar rules.

The N-Best results indicate that it could be worth putting effort on re-ranking the N-Best lists as both WER and SER of the N-Best candidates are considerably lower. This could ideally give us a reduction in SER of 10% and, considering dialogue move error rate, perhaps even more. More or less advanced post-process methods have been used to analyze and decide on the best choice from the N-Best list. Several different re-ranking methods have been proposed that show how recognition rates can be improved by letting external processes do the top N ranking and not the recognizer (Chotimongkol & Rudnicky, 2001; van Noord *et al.*, 1997). However, the way that seems most appealing is how (Gabsdil & Lemon, 2004) and (Hacioglu & Ward, 2001) re-rank N-Best lists based on dialogue context achieving a considerable improvement in recognition performance. We are considering basing our re-ranking on the information held in the dialogue information state, knowledge of what is going on in the graphical interface and on dialogue moves in the list that seem appropriate to the context. In this way we can take advantage of what the dialogue system knows about the current situation.

5 Concluding remarks and future work

A first observation is that the SLMs give us a much more robust recognition, as expected. Our best SLMs, i.e. the mixed models, give a 43% relative improvement over the baseline i.e. the Nuance grammar compiled from the GF interpretation grammar. However, this also implies a falling off in in-grammar performance. It is interesting that the SLM that only models the grammar (`MP3GFLM`), although being more robust and giving a significant reduction in WER rate, does not degrade its in-grammar performance to a great extent. This simple model seems promising for use in a first version of the system with the possibility of improving it when logs from system interactions have been collected. In addition, the vocabu-

lary of this model is in sync with our GF interpretation grammar. The results seem comparable with those obtained by (Bangalore & Johnston, 2004) using random generation to produce an SLM from an interpretation grammar.

Although interpolating our MP3 model with the GSLC corpus and the newspaper corpora gave a large perplexity reduction it did not have as much impact on WER as expected even though it gave a significant improvement. It seems from the tests that the quality of the data is more important than the quantity. This makes extraction of domain data from larger corpora an important issue and increases the interest of generating artificial corpora.

As the approach of using SLMs in our dialogue systems seems promising and could improve recognition performance considerably we are planning to apply the experiment to other applications that are under development in TALK when the corresponding GF application grammars are finished. In this way we hope to find out if there is a tendency in the performance gain of a statistical language model vs its correspondent speech recognition grammar. If so, we have found a good way of compromising between the ease of grammar writing and the robustness of SLMs in the first stage of dialogue system development. In this way we can use the knowledge and intuition we have about the domain and include it in our first SLM and get a more robust behaviour than with a grammar. From this starting point we can then collect more data with our first prototype of the system to improve our SLM.

We have also started to look at dialogue move specific statistical language models (DM-SLMs) by using GF to generate all utterances that are specific to certain dialogue moves from our interpretation grammar. In this way we can produce models that are sensitive to the context but also, by interpolating these more restricted models with the general GF SLM, do not restrict what the users can say but take into account that certain utterances should be more probable in a specific dialogue context. Context-sensitive models and specifically grammars for different contexts have been explored earlier (Baggia *et al*, 1997; Wright *et al*, 1999; Lemon, 2004) but generating such language models artificially from an interpretation grammar by choosing which moves to combine seems to be a new direction. Our first experiments seem promising but the dialogue move specific test sets are too small to draw any conclusions. We hope to report more on this in the near future.

Acknowledgements

I am grateful to Steve Young, Robin Cooper and the EACL reviewers for comments on previous versions of this paper. I would also like to thank Aarne Ranta, Peter Ljunglöf, Karl Weilhammer and David Hjelm for help with GF and data collection and finally Nuance Communications Inc. for making available the speech recognition software used in this work. This work was supported in part by the TALK project (FP6-IST 507802, http://www.talk-project.org/).

References

Allwood, J. 1999. The Swedish Spoken Language Corpus at Göteborg University. In *Fonetik 99*, Gothenburg Papers in Theoretical Linguistics 81. Dept. of Linguistics, University of Göteborg.

Baggia P., Danieli M., Gerbino E., Moisa L. M., and Popovici C. 1997. Contextual Information and Specific Language Models for Spoken Language Understanding. In *Proceedings of SPECOM'97*, Cluj-Napoca, Romania, pp. 51–56.

Bangalore S. and Johnston M. 2004. Balancing Data-Driven And Rule-Based Approaches in the Context of a Multimodal Conversational System. In *Proceedings of Human Language Technology conference*. HLT-NAACL 2004.

Chotimongkol A. and Rudnicky A.I. 2001. N-best Speech Hypotheses Reordering Using Linear Regression. In *Proceedings of Eurospeech 2001*. Aalborg, Denmark, pp. 1829–1832.

Clarkson P.R. and Rosenfeld R. 1997. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of Eurospeech*.

Fosler-Lussier E. and Kuo H.-K. J. 2001. Using Semantic Class Information for Rapid Development of Language Models within ASR Dialogue Systems. In *Proceedings of ICASSP-2001*, Salt Lake City, Utah.

Gabsdil M. and Lemon O. 2004. Combining Acoustic and Pragmatic Features to Predict Recognition Performance in Spoken Dialogue Systems. In *Proceedings of ACL*, Barcelona.

Gorrell G., Lewin I. and Rayner M. 2002. Adding Intelligent Help to Mixed Initiative Spoken Dialogue Systems. In *Proceedings of ICSLP-2002*.

Hacioglu K. and Ward W. 2001. Dialog-context dependent language modeling combining n-grams and stochastic context-free grammars. In *Proceedings of ICASSP-2001*, Salt Lake City, Utah.

Huang X., Acero A., Hon H-W. 2001. *Spoken Language Processing: A guide to theory, algorithm and system development.* Prentice Hall.

Janiszek D., De Mori R., Bechet F. 1998. Data Augmentation And Language Model Adaptation. University of Avignon 84911 Avignon Cedex 9 - France.

Jelinek, F. and Mercer, R. 1980. Interpolated Estimation of Markov Source Parameters from Sparse Data. In *Pattern Recognition in Practice*. E. S. Gelsema and L. N. Kanal, North Holland, Amsterdam.

Jelinek, F. 1997. *Statistical Methods for Speech Recognition.* MIT Press.

Knight S., Gorrell G., Rayner M., Milward D., Koeling R. and Lewin I. 2001. Comparing Grammar-Based and Robust Approaches to Speech Understanding: A Case Study. In *Proceedings of Eurospeech 2001*.

Larsson S. 2002. *Issue-based Dialogue Management.* PhD Thesis, Göteborg University.

Larsson S., Berman A., Grönqvist L., Kronlid, F. 2002. TRINDIKIT 3.0 Manual. D6.4, Siridus Project, Göteborg University.

Lemon O. 2004. Context-sensitive speech recognition in ISU dialogue systems: results for the grammar switching approach. In *Proceedings of CATALOG*, 8th Workshop on the Semantics and Pragmatics of Dialogue, Barcelona.

Ljunglöf P., Bringert B., Cooper R., Forslund A-C., Hjelm D., Jonson R., Larsson S. and Ranta A. 2005. The TALK Grammar Library: an Integration of GF with TrindiKit. Deliverable 1.1, TALK project.

Nuance Communications. http://www.nuance.com, as of May 2005.

Pakhomov SV., Schonwetter M., Bachenko, J. 2001. Generating Training Data for Medical Dictations. In *Proceedings NAACL-2001.*

Ranta A. 2004. Grammatical Framework. A Type-Theoretical Grammar Formalism. In *The Journal of Functional Programming.*, Vol. 14, No. 2, pp. 145–189.

Ranta A. Grammatical Framework Homepage http://www.cs.chalmers.se/ãarne/GF, as of May 2005.

Raux A., Langner B., Black A. and Eskenazi M. 2003. LET'S GO: Improving Spoken Dialog Systems for the Elderly and Non-natives. In *Proceedings of Eurospeech 2003.* Geneva, Switzerland.

Rayner M., Hockey B.A., James F., Owen Bratt E., Goldwater S., Gawron J.M. 2000. Compiling Language Models from a Linguistically Motivated Unification Grammar. In *Proceedings of COLING-2000*.

Rayner M., Hockey B.A., Dowding J. 2003. An Open-Source Environment for Compiling Typed Unification Grammars into Speech Recognisers. In *Proceedings of EACL*, pp. 223–226.

Rayner M. and Hockey B.A. 2003. Transparent combination of rule-based and data-driven approaches in speech understanding. In *Proceedings of EACL.*

Rosenfeld R. 2000. Two decades of statistical language modeling: Where do we go from here? In *Proceedings of IEEE*:88(8).

Rosenfeld R. 2000. Incorporating Linguistic Structure into Statistical Language Models. In *Philosophical Transactions* of the Royal Society of London A, 358.

Solsona R., Fosler-Lussier E., Kuo H.J., Potamianos A. and Zitouni I. 2002. Adaptive Language Models for Spoken Dialogue Systems. In *Proceedings of ICASSP-2002*, Orlando, Florida, USA.

Stolcke A. 2002. SRILM – An Extensible Language Modeling Toolkit. In *Proceedings of ICSLP-2002*, Vol. 2, pp. 901–904, Denver.

van Noord G., Bouma G., Koeling R. and Nederhof, M. 1999. Robust Grammatical Analysis for Spoken Dialogue Systems. In *Journal of Natural Language Engineering*, 5(1), pp. 45–93.

Wright H., Poesio M. and Isard S. 1999. Using high level dialogue information for dialogue act recognition using prosodic features. In *DIAPRO-1999*, pp. 139–143.

Weilhammer K., Jonson R., Ranta A, Young Steve. 2006. SLM generation in the Grammatical Framework. Deliverable 1.3, TALK project.

Xu W. and Rudnicky A. 2000. Language modeling for dialog system? In *Proceedings of ICSLP-2000*, Beijing, China. Paper B1-06.