Grammar-based context-specific statistical language modelling

Rebecca Jonson
Department of Linguistics, Göteborg University & GSLT
Box 200, 40530 Göteborg, Sweden
rj@ling.gu.se

Abstract

This paper shows how we can combine the art of grammar writing with the power of statistics by bootstrapping statistical language models (SLMs) for Dialogue Systems from grammars written using the Grammatical Framework (GF) (Ranta, 2004). Furthermore, to take into account that the probability of a user's dialogue moves is not static during a dialogue we show how the same methodology can be used to generate dialogue move specific SLMs where certain dialogue moves are more probable than others. These models can be used at different points of a dialogue depending on contextual constraints. By using grammar generated SLMs we can improve both recognition and understanding performance considerably over using the original grammar. With dialogue move specific SLMs we would be able to get a further improvement if we had an optimal way of predicting the correct language model.

1 Introduction

Speech recognition (ASR) for dialogue systems is often caught in the trap of the sparse data problem which excludes the possibility of using statistical language models (SLMs). A common approach is to write a grammar for the domain either as a speech recognition grammar (SRG) or as an interpretation grammar which can be compiled into a speech recognition grammar (SRG) using some grammar development platform such as Gemini, Regulus or GF (Rayner et al., 2000; Rayner et al., 2006; Ranta, 2004). The last option will assure that the linguistic coverage of the ASR and interpretation are kept in sync. ASR for commercial dialogue systems has mainly focused on grammar-based approaches despite the fact that SLMs seem to have a better overall performance (Knight et al., 2001; Bangalore and Johnston, 2003). This probably depends on the time-consuming work of collecting corpora for training SLMs compared with the more rapid and straightforward development of SRGs. However, SLMs are more robust for out-of-coverage input, perform better in difficult conditions and seem to work better for naive users as shown in (Knight et al., 2001). SRGs on the other hand are limited in their coverage depending on how well grammar writers succeed in predicting what users may say.

An approach taken in both dialogue systems and dictation applications is to write a grammar for the particular domain and generate an artificial corpus from the grammar to be used as training corpus for SLMs (Galescu et al., 1998; Bangalore and Johnston, 2003; Jonson, 2006). These grammar-based models are not as accurate as the ones built from real data as the estimates are artificial, lacking a realistic distribution. However, as has been shown in (Bangalore and Johnston, 2003; Jonson, 2006) these grammar-based statistical models seem to have a much more robust behaviour than their corresponding grammars which leaves us with a much better starting point in the first development stage in a dialogue system. It is a way of compromising between the ease of grammar writing and the robust-

Proceedings of SPEECHGRAM 2007, pages 25–32,
Prague, Czech Republic, June 2007.

ness of SLMs. With this methodology we can use the knowledge and intuition we have about the domain and include it in our first SLM and get a much 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. In this paper the advantage of this method is shown further by evaluating a different domain in greater detail.

Context-specific models have shown important recognition performance gain (Baggia et al., 1997; Riccardi et al., 1998; Xu and Rudnicky, 2000; Lemon and Gruenstein, 2004) and have usually been of two types: created as state-specific grammars or built from collected data partitioned according to dialogue states. Both methods have their disadvantages. In the first case, we constrain the user heavily which makes them unsuitable for use in a more flexible system such as an information-state based system. This can be solved by having a back-off method but leaves us with extra processing (Lemon and Gruenstein, 2004). In the latter case, we have an even more severe sparse data problem than when creating a general SLM as we need enough data to get a good distribution of data over dialogue states. In an information-state based system where the user is not restricted to only a few dialogue states this problem gets even worse. In addition, why we chose to work with grammar-based SLMs in the first place was because data is seldom available in the first stage of dialogue system development. This leaves us with the requirement of an SLM that although being context-specific does not constrain the user and which assures a minimal coverage of expressions for a certain context. In (Gruenstein et al., 2005) this is accomplished by dynamically populating a class-based SLMs with context-sensitive content words and utterances. In this paper, we will show how we can use the same methodology as in (Jonson, 2006) to create context-specific SLMs from grammars based on dialogue moves that match these criteria.

This study is organized as follows. First, we introduce our methodology for developing SLMs from grammars. Section 3 describes the data collection of test utterances and how we have partitioned the data into different test sets depending on grammar coverage, types of users and types of dialogue moves. In section 4, we show and discuss the results of the different models for different test sets and finally we draw some conclusions from the experiments.

2 Grammar-based SLMs

In (Jonson, 2006) we described how we could generate an SLM from an interpretation grammar written in GF for an MP3 player application and get a much more robust behaviour than by using the original grammar for ASR. In this study, we approach a different domain using a GF grammar written for a dialogue system application called AgendaTalk (Ericsson et al., 2006). It is one of several applications that has been developed in the TALK project (www.talk-project.org) and has been built with the TrindiKit toolkit and the GoDiS dialogue system (Larsson, 2002) as a GoDiS application. It works as a voice interface to a graphical calendar. Apart from evaluating a different domain in a more extensive way to see if the tendency we found in (Jonson, 2006) is consisting over domains, we have driven the methodology a bit further to be able to generate context-specific SLMs that favour certain parts of the grammar, in our case certain dialogue moves. We call these SLMs "dialogue move specific SLMs" (DMSLMs). Both types of models are obtained by generating all possible utterances from a GF grammar, building trigram SLMs from the grammar-based corpus using the SRI language modelling toolkit (Stolcke, 2002) and compiling them into recognition packages. For comparison we have also compiled the GF grammar directly into a Nuance speech recognition grammar using the GF compiler.

2.1 Building a general SLM from grammar-based corpora

The GF grammar written for the calendar domain consists of 500 GF functions (rules) where 220 are domain-specific and 280 inherited from a domain-independent grammar. It exists in two equivalent language versions that share the same GF functions: English and Swedish. We have used GF's facilities to generate a corpus from the Swedish version consisting of all possible meaningful utterances generated by the grammar to a certain depth of the analysis trees in GF's abstract syntax. The grammar is written on the phrase level accepting spoken language utterances such as e.g. “add a booking please”. The resulting corpus consists of 1.7 million utterances and 19 million words with a vocabulary of only 183 words. All utterances in the corpus occur exactly once. However, all grammar rules are not expanded which leaves us with a class-tagged corpus without e.g. all variants of date expressions but with the class `date`. What we get in the end is therefore a class-based SLM that we compile into a recognition package together with a rule-based description of these classes. The SLM has 3 different classes: `time`, `date` and `event` and the domain vocabulary when including all distinct words in these classes make up almost 500 words.

Adding real speech corpora

In (Jonson, 2006) we saw that the use of real corpora in interpolation with our artificial corpus was only valuable as long as the real corpora approximated the language of use. The big news corpus we had available did not give any significant improvement but the transcribed Swedish speech corpus we used was much more helpful. In this study we have therefore once again used the GLSC corpus to improve our word occurrence estimates by interpolating it with our grammar-based SLM. 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). The corpus is composed of about 1,300,000 words and is turn-based which gives it long utterances including e.g. transcribed disfluencies. From this corpus we have built an SLM which we have interpolated with our grammar-based SLM keeping our domain vocabulary. This means we are just considering those n-grams in the GSLC corpus which match the domain vocabulary to hopefully get a more realistic probability distribution for these. We will call this model our `Extended SLM`.

2.2 Dialogue move specific SLMs

SLMs capture the lexical context statistics in a specific language use. However, the statistical distribution in a dialogue is not static but varies by boosting and lowering probabilities for different words and expressions depending on contextual appropriateness. It is not only words and expressions that vary their distribution but on a semantic level different conceptual messages will be more or less probable as a user utterance at different points of the dialogue. This means that certain dialogue moves will have a higher degree of expectancy at a specific point of the dialogue. To capture this phenomenon, we want to build models that raise the probability of certain dialogue moves in certain contexts by giving a higher probability for utterances expressing these dialogue moves. These are models where utterances corresponding to a certain dialogue move are more salient (e.g. a model where all ways of answering yes or no are more plausible than other utterances). Such a model will account for the fact that the expectation of dialogue moves a user will perform varies in a dialogue and thereby their statistics. We can obtain this by using a version of the grammar-based corpus where the dialogue moves for each utterance are generated which allows us to partition the corpus in different ways based on dialogue moves. We can then take out part of the corpus e.g. all utterances corresponding to a certain dialogue move, create an SLM and interpolate it with the general grammar-based SLM. In this way, we get SLMs where certain dialogue moves are more probable than others and where minimally all possible expressions for these, which the grammar describes, are covered. By interpolating with the general SLM we put no hard constraints on the expected dialogue move so the user can in fact say anything at any point in the dialogue despite the raised expectancy for certain dialogue moves. We just boost the expected probability of certain dialogue moves and their possible expressions. By using contextual constraints in the information state we could then predict which model to use and switch SLMs on the fly so that we obtain a recognizer that takes account of expected user input.

2.2.1 Partitioning the training data by dialogue moves

In GoDiS, dialogue moves are activity related and exist in seven different types: `request` moves, `answer` moves, `ask` moves (i.e. questions), yes and no (`yn`) answers, `greet` moves, `quit` moves and feedback and sequencing moves which are called `ICM:s` (Larsson, 2002). We have chosen to focus on the first four of these dialogue move types to build up our DMSLMs. We have used GF to generate a corpus with all possible dialogue moves and their combinations with their corresponding expressions. From this corpus we have extracted all utterances that can be interpreted as an `answer` move or a sequence of answer moves, all expressions for specification of a `request` (in GoDiS what type of action to perform e.g. deleting a booking), all ways of expressing questions in our grammars (i.e. `ask` moves) and all possible `yn` answers. This leaves us with four new sets of training data.

The decision to partition the data in this way was based on the distribution of dialogue moves in our data where the moves we focus on are the most common ones and the most critical for achievement of the dialogue tasks. As these dialogue moves are abstract and domain-independent it would be possible to use a domain-independent prediction of these dialogue moves and thereby the language models although the structure of the SLMs would be different in different domains.

2.2.2 Building dialogue move specific SLMs

For each set of dialogue move specific training data we created an SLM that only captures ways of expressing a specific dialogue move. However, we are looking for less constrained models which just alter the probability of certain dialogue moves. By interpolating the SLMs built on dialogue move specific corpora with the general grammar-based SLM we achieve models with contextual probabilities but that generalize to avoid constraining the user input.

The interpolation of these models was carried out with the SRILM toolkit based on equation 1. The optimal lambda weight was estimated to 0.85 for all models with the SRILM toolkit using held-out data.

$$P_{dmslm}(W) = \lambda P_{movespec}(W) + (1 - \lambda)P_{general}(W) \quad (1)$$

We ended up with four new SLMs, so called DMSLMs, in which either the probability of `answer`, `ask`, `request` or `yn` moves were boosted.

3 Test Data

The collection of test data was carried out by having people interacting with the AgendaTalk system using the grammar-based SLM. The test group included both naive users with no experience of the system whatsoever and users that had previous experience with the system to varying extents. We have classified the latter group as expert users although the expertise varies considerably. All users were given a printed copy of a calendar month with scheduled bookings and some question marks and were assigned the task of altering the voice-based calendar so that the graphical calendar would look the same as the printed copy except for the question marks which they were to find values for by querying the system. This would mean that they would have to add, delete and alter bookings as well as find out information about their schedule e.g. the time of an event. The tasks could be carried out in any order and there were many different ways to complete the schedule.

The data collection gave us a recording test set of 1000 recorded utterances from 15 persons (all native, 8 female, 7 male). This unrestricted test set was used to compare recognition performance between the different models under consideration. We also partitioned the test set in various ways to explore different features. The test set was parsed to get all in-coverage utterances that the original GF grammar covers to create an in-coverage test set from these. In addition, we partitioned the data by users with a test set with the naive user utterances and another test set from the expert users. In this way we could explore how our models performed under different conditions. Different dialogue system applications will have a different distribution of users. Some systems will always have a large number of naive or less experienced users who will use more out-of-coverage utterances and more out-of-vocabulary (OOV) words whereas users of other applications will have the opportunity to obtain considerable experience which will allow them to adapt to the system, in particular to its grammar and vocabulary.

The recordings for the unrestricted test set have an OOV rate of 6% when using our domain vocabulary. The naive test set makes up 529 of these recordings with an OOV rate of 8% whereas the expert test set of 471 recordings has a lower OOV rate of 4%. The in-coverage test set consists of 626 utterances leaving us with an in-coverage rate of 62.6% for the unrestricted test set. This shows the need for a more robust way of recognition and interpretation if we expect to expose the system to less experienced users.

For the evaluation of the DMSLMs we have partitioned the test data by dialogue moves. The utterances corresponding with the four dialogue moves chosen for our DMSLMs were divided into four test sets. The utterances left were used to create a fifth test set where none of our four DMSLMs would apply but where we would need to use the general model. If we look at the distribution of the test data considering dialogue moves we find that 75.4% of the test data falls into our four dialogue move categories and that only 24.6% of the data would require the general model. This part of the test data includes dialogue moves such as greetings, quit moves and dialogue move sequences with combinations of different moves. The most common dialogue move in our data is an `answer` move or a sequence of `answer` moves resulting in common utterances such as: "a meeting on friday" as answer to system questions such as "what booking do you want to add?".

4 Experimental Results

To evaluate the recognition performance of our different types of models we ran several experiments on the different test sets. We report results on word error rate (WER), sentence error rate (SER) and also on a semantic level by reporting what we call dialogue move error rate (DMER). The dialogue move error rate was obtained by parsing the recognized utterances and comparing these to a parsed version of the transcriptions, calculating the rate of correctly parsed dialogue moves. The calculation was done in the same way as calculation of concept error rate (CER) proposed by (Boros et al., 1996) where the degree of correctly recognized concepts is considered. In our case this means the degree of correctly recognized dialogue moves. For parsing we have used a phrase-spotting grammar written in Prolog that pattern matches phrases to dialogue moves. Using the original GF interpretation grammar for parsing would have restricted us to the coverage of the grammar which is not an optimal choice together with SLMs. Ideally, we would like to use a robust version of GF to be able to use the original GF grammar both for parsing and SLM generation and by that assure the same linguistic coverage. Attempts to do this have been carried out in the TALK project for the MP3 domain by training a dialogue move tagger on the same type of corpus that was used for the DMSLMs where dialogue moves occur together with their corresponding utterances. Other methods of relaxing the constraints of the GF parser are also under consideration. Meanwhile, we are using a simple robust phrase spotting parser. We have investigated both how our grammar-based SLMs perform in comparison to our grammar under different conditions to see how recognition and understanding performance varies as well as how our DMSLMs perform in comparison to the general grammar-based SLM. The results are reported in the following sections. All models have the same domain vocabulary and the OOV figures presented earlier thereby apply for all of them.

4.1 Grammar-based SLMs vs. grammars

Table 1 shows the results for our different language models on our unrestricted test set of 1000 utterances as well as for the part of this test set which is in-coverage. As expected they all perform much better on the in-coverage test set with the lowest WER obtained with our grammar. On the unrestricted test set we can see an important reduction of both WER (26% and 38% relative improvement) and DMER (24% and 40% relative improvement) for the SLMs in comparison to the grammar which indicates the robustness of these to new user input.

In table 2 we can see how the performance of all our models are better for the expert users with a relative word error rate reduction from 25% to 32% in comparison to the results for the naive test set. The same pattern is seen on the semantic level with important reduction in DMER. The result is expected as the expert users have greater knowledge of the language of the system. This is consistent with the results reported in (Knight et al., 2001). It is also reflected in the OOV figures discussed earlier where the naive users seem to have used many more unknown words than the expert users.

This shows that the models perform very differently depending on the types of users and how much they hold to the coverage of the grammar. Our grammar-based SLM gives us a much more robust behaviour which is good when we expect less experienced users. However, we can see that we get a degradation in in-coverage performance which would be critical if we are to use the model in a system where we expect that the users will achieve cer-

Table 1: *Results on unrestricted vs in-coverage test set*

Model	Unrestricted			In-coverage		
	WER	SER	DMER	WER	SER	DMER
Grammar	39.0%	47.6%	43.2%	10.7%	16.3%	10.3%
Grammar-based SLM	29.0%	39.7%	33.0%	14.8%	18.3%	13.7%
Extended SLM	24.0%	35.2%	25.8%	11.5%	15.8%	10.4%

Table 2: *Results on naive vs expert users*

Model	Naive users			Expert users		
	WER	SER	DMER	WER	SER	DMER
Grammar	46.6%	50.3%	54.7%	31.7%	44.4%	33.2%
Grammar-based SLM	34.4%	42.9%	41.3%	23.8%	35.9%	25.8%
Extended SLM	27.6%	38.2%	29.5%	20.7%	31.8%	22.7%

tain proficiency. The `Extended SLM` seem to perform well in all situations and if we look at DMER there is no significant difference in performance between this model and our grammar when it comes to in-coverage input. In most systems we will probably have a range of users with different amounts of experience and even experienced users will fail to follow the grammar in spontaneous speech. This points towards the advisability of using an SLM as it is more robust and if it does not degrade too much on in-coverage user input like the `Extended SLM` it would be an optimal choice.

From the results it seems that we have found a correlation between the DMER and WER in our system which indicates that if we manage to lower WER we will also achieve better understanding performance with our simple robust parser. This is good news as it means that we will not only capture more words with our SLMs but also more of the message the user is trying to convey in the sense of capturing more dialogue moves. This will definitely result into a better dialogue system performance overall. Interestingly, we have been able to obtain this just by converting our grammar into an SLM.

4.2 Dialogue move specific SLMs vs General SLMs

We have evaluated our DMSLMs on test sets for each model which include only utterances that correspond to the dialogue moves in the model. It should be mentioned that the test sets may include utterances not covered by the original GF grammar e.g. a different wording for the same move. The results for each DMSLM on its specific test set and the performance of the grammar-based SLM and the Extended SLM are reported in tables 3, 4, 5 and 6.

Table 3: *Ask Move SLM*

Model	WER	SER	DMER
Grammar-based SLM	39.2%	68.4%	51.8%
Ask DMSLM	31.8%	68.9%	48.7%
Extended SLM	30.1%	58.0%	44.6%

Table 4: *Answer Move SLM*

Model	WER	SER	DMER
Grammar-based SLM	17.3%	22.0%	16.3%
Answer DMSLM	15.7%	20.1%	14.1%
Extended SLM	18.2%	22.0%	16.7%

Table 5: *Request Move SLM*

Model	WER	SER	DMER
Grammar-based SLM	29.1%	44.3%	27.0%
Request DMSLM	17.0%	36.1%	14.7%
Extended SLM	26.3%	42.6%	22.1%

Apart from these four dialogue moves our test data includes a lot of different dialogue moves and dialogue move combinations that we have not considered. As we have no specific model for these we would need to use a general model in these cases. This means that apart from predicting the four dialogue moves we have considered we would also need to predict when none of these are expected and use the general model for these situations. In table 7 we can see how our general models perform on the rest of the test set. This shows that they seem to handle this part of the test data quite well.

Table 6: *YN Move SLM*

Model	WER	SER	DMER
Grammar-based SLM	37.3%	27.3%	22.7%
YN DMSLM	21.5%	16.5%	11.9%
Extended SLM	25.0%	18.2%	12.5%

Table 7: *General SLM on rest of test data*

Model	WER	SER	DMER
Grammar-based SLM	22.2%	42.7%	31.7%
Extended SLM	19.6%	39.8%	26.0%

We can see that the gain we get in recognition performance varies for the different models and that relative improvement in WER goes from 9% for the `answer` model to 42% for our DMSLMs on appropriate test sets. We can see that our models have most problems with `ask` moves and `yn` answers. In the case of `ask` moves this seems to be because our GF grammar is missing a lot of syntactic constructions of question expressions. This would then explain why the Extended SLM gets a much better figure here. The GSLC corpus does capture more of this expressive variation of questions. In other words we seem to have failed to capture and predict the linguistic usage with our hand-tailored grammar. In the case of `yn` answers the result reveals that our grammar-based SLM does not have a realistic distribution of these expressions at all. This seems to be something the GSLC corpus contribute, considering the good results for the `Extended SLM`. However, we can see that we can achieve the same effect by boosting the probability of yes and no answers in our DMSLM.

If we look at the overall achievement in recognition performance, using our DMSLMs when appropriate and in other cases the general SLM, the average WER of 22% (27% DMER) is considerably lower than when using the general model for the same test data (29% WER, 33% DMER). If we had an optimal method for predicting what language model to use we would be able to decrease WER by 24% relative. If we chose to use the `Extended SLM` in the cases our DMSLMs do not cover we could get an even greater reduction.

We have also tested how well our DMSLMs perform on the general test set (i.e. all 1000 utterances) to see how bad the performance would be if we chose the wrong model. In table 8 we can see that this approach yields an average WER of 30% which is a minimal degradation in comparison to the general grammar-based SLM. On the contrary, some of our models actually perform better than our general grammar-based SLM or very similarly. This implies that there is no substantial risk on recognition performance if our prediction model would fail. This means that we could obtain very good results with important recognition improvement even with an imperfect prediction accuracy. We have a relative improvement of 24% to gain with only a minimal loss.

Table 8: *DMSLMs on general test set*

Model	WER	SER
Answer DMSLM	34.7%	55.6%
Ask DMSLM	28.2%	46.2%
Request DMSLM	26.5%	43.2%
YN DMSLM	29.8%	44.0%

5 Concluding remarks

Our experimental results show that grammar-based SLMs give an important reduction in both WER and DMER in accordance with the results in (Jonson, 2006). We reach a relative improvement of 26% and a further 17% if we interpolate our grammar-based SLM with real speech data. The correlation of the DMER and the WER in our results indicates that the improved recognition performance will also propagate to the understanding performance of our system.

Context-specific language models (statistical and rule-based) have shown important recognition performance gain in earlier work (Baggia et al., 1997; Xu and Rudnicky, 2000; Lemon and Gruenstein, 2004; Gruenstein et al., 2005) and this study reaffirms that taking into account statistical language variation during a dialogue will give us more accurate recognition. The method we use here has the advantage that we can build statistical context-specific models even when no data is available, assuring a minimal coverage and by interpolation with a general model do not constrain the user input unduly.

The language model switch will be triggered by changing a variable in our information state: the predicted dialogue move. However, to be able to choose which language model suits the current information state best we need a way to predict dialogue moves. The prediction model could either be rule-based or data based. Our first experimental tests with machine learning for dialogue move prediction seems promising and we hope to report on these soon. Optimally, we want a prediction model that we can use in different GoDiS domains to be able to generate new DMSLMs from our domain-specific GF grammar for the dialogue moves we have considered here.

Our experiments show that we could achieve an overall reduction in WER of 46% and 40% in DMER if we were able to choose our best suited SLM instead of our compiled GF grammar. Naturally, we would have to take into account dialogue move prediction accuracy to get a more realistic figure. However, our experiments also show that the effect on performance if we failed to use the correct model would not be too harmful. This means we have much more to gain than to lose even if the dialogue move prediction is not perfect. This makes this approach a very interesting option in dialogue system development.

Acknowledgment

I would like to thank Nuance Communications Inc. for making available the speech recognition software used in this work.

References

Jens Allwood. 1999. The Swedish spoken language corpus at Göteborg University. In *Proceedings of Fonetik'99: The Swedish Phonetics Conference*.

Paolo Baggia, Morena Danieli, Elisabetta Gerbino, Loreta Moisa, and Cosmin Popovici. 1997. Contextual information and specific language models for spoken language understanding. In *Proceedings of SPECOM*.

Srinivas Bangalore and Michael Johnston. 2003. Balancing data-driven and rule-based approaches in the context of a multimodal conversational system. In *Proceedings of the ASRU Conference*.

M. Boros, W. Eckert, F. Gallwitz, G. Görz, G. Hanrieder, and H. Niemann. 1996. Towards understanding spontaneous speech: Word accuracy vs. concept accuracy. In *Proceedings of ICSLP*, Philadelphia, PA.

S. Ericsson, G. Amores, B. Bringert, H. Burden, A. Forslund, D. Hjelm, R. Jonson, S. Larsson, P. Ljunglöf, P. Manchon, D. Milward, G. Perez, and M. Sandin. 2006. Software illustrating a unified approach to multimodality and multilinguality in the inhome domain. Deliverable D1.6, TALK Project.

Lucian Galescu, Eric Ringger, and James Allen. 1998. Rapid language model development for new task domains. In *In Proceedings of the ELRA First International Conference on Language Resources and Evaluation (LREC)*.

Alexander Gruenstein, Chao Wang, and Stephanie Seneff. 2005. Context-sensitive statistical language modeling. In *Proceedings of Interspeech*.

Rebecca Jonson. 2006. Generating statistical language models from interpretation grammars in dialogue systems. In *Proceedings of EACL*, Trento, Italy.

S. Knight, G. Gorrell, M. Rayner, D. Milward, R. Koeling, and I. Lewin. 2001. Comparing grammar-based and robust approaches to speech understanding: A case study. In *Proceedings of Eurospeech*.

Staffan Larsson. 2002. *Issue-based Dialogue Management*. Ph.D. thesis, Göteborg University.

Oliver Lemon and Alexander Gruenstein. 2004. Multithreaded context for robust conversational interfaces: Context-sensitive speech recognition and interpretation of corrective fragments. *ACM Trans. Comput.-Hum. Interact.*, 11(3):241–267.

Aarne Ranta. 2004. Grammatical framework. a type-theoretical grammar formalism. *The Journal of Functional Programming*, 14(2):145–189.

Manny Rayner, Beth Ann Hockey, Frankie James, Elizabeth Owen Bratt, Sharon Goldwater, and Jean Mark Gawron. 2000. Compiling language models from a linguistically motivated unification grammar. In *Proceedings of the COLING*.

Manny Rayner, Beth Ann Hockey, and Pierrette Bouillon. 2006. *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. CSLI Publications.

Guiseppe Riccardi, Alexandros Potamianos, and Shrikanth Narayanan. 1998. Language model adaptation for spoken language systems. In *Proceedings of the ICSLP*, Australia.

Andreas Stolcke. 2002. SRILM - An extensible language modeling toolkit. In *Proceedings of ICSLP*, Denver, Colorado.

Wei Xu and Alex Rudnicky. 2000. Language modeling for dialog system. In *Proceedings of ICSLP 2000*, Beijing, China.