

# Using Statistical Language Modelling to Identify New Vocabulary in a Grammar-Based Speech Recognition System

*Genevieve Gorrell*

Department of Computer and Information Science  
Linköping University, Sweden

gengo@ida.liu.se

## Abstract

Spoken language recognition meets with difficulties when an unknown word is encountered. In addition to the new word being unrecognisable, its presence impacts on recognition performance on the surrounding words. The possibility is explored here of using a back-off statistical recogniser to allow recognition of out-of-vocabulary words in a grammar-based speech recognition system. This study shows that a statistical language model created from a corpus obtained using a grammar-based system and augmented with minimally-constrained domain-appropriate material allows extraction of words that are out of the vocabulary of the grammar in an unseen corpus with fairly high precision.

## 1. Introduction

As spoken dialogue systems increase in their capabilities, becoming able to support more complex and diverse domains, the pressure is on to provide speech recognition capabilities appropriate to their demands. High performance speech recognition depends on high-quality language modelling. Current work in language modelling for speech recognition in spoken dialogue systems focuses on two main areas. Grammar-based language modelling involves writing a hand-coded complete specification of allowable phrases, which ideally will be based on observation of a corpus, and may or may not be augmented with probabilities derived from a corpus. Statistical language modelling is entirely empirical, and involves compilation of a finite-state machine in which the likelihood of a given word occurring is calculated based on the corpus, possibly given the context of the preceding  $n$  words; two is usual. Whilst the statistical approach has long been preferred in academia, and has been shown to produce good results, industry is currently showing a marked preference for grammar-based language modelling. The reasons for this include that the collection of a sufficiently extensive corpus to allow for the creation of a high-performance statistical language model (SLM) is time-consuming and expensive. Furthermore, grammars can be manipulated online, for example, to include new vocabulary. Where systems need to be created quickly to cover diverse domains and specific coverage, grammar-based language modelling has the edge. It is also very appropriate in a system where users can be expected to become experts in using the system, as in for example command and control systems, such as CommandTalk [1]; naturalness of language use can be sacrificed for fast and accurate recognition. However, an SLM can be more flexible in the range of language and constructions it allows, and being entirely corpus-based, can model more accurately the language it can expect to receive as input. Where the majority of users are novices, as in

for example information systems such as JUPITER [2], this feature becomes very important. Also, in an experimental system, where the domain can be matched to the corpus available and the emphasis is on the interpretation and dialogue management lying behind the recognition, the permissiveness of the SLM becomes very attractive.

The two approaches clearly have different advantages. In terms of actual recognition performance, their behaviour also shows marked differences. Knight et al. [3] compared recognition performance in a statistical and a grammar-based speech recogniser within the same domain. They showed that where users spoke in coverage of the grammar, the grammar-based recogniser had a small performance edge on producing a correct semantic interpretation. But where the user spoke out-of-coverage, the performance of the grammar-based recogniser dropped off much more quickly than the performance of the SLM. Each type of language modelling has its own, different areas of superiority. Such a division of strengths suggests that it may be possible to combine the two approaches in the same system such as to capitalise on their respective advantages. Different approaches have been suggested for how to do this. For example, Gorrell et al. [4] use a back-off statistical recogniser to provide assistance to the user where grammar-based recognition fails. Here, we consider a new suggestion.

Speech recognition of out-of-vocabulary words is a challenging problem. Accurate speech recognition, especially of the speaker-independent variety, such as is used in the majority of spoken dialogue systems, depends on the system having a fairly precise representation of what the user is likely to say. When the user says a word that is out-of-coverage, not only is that word impossible for the system to recognise, but recognition performance on the surrounding words is greatly reduced, as shown in [2]. Isolation of the boundaries of the new word is very difficult to do. Without the ability to identify the problem word, or even to identify that an out-of-vocabulary word was used, recovery strategies are limited. If the word could be identified, however, then the user could be alerted to the problem, and could avoid using that word in the future. The system could also make an attempt to learn the new word, and be able to recognise it in future.

Despite the difficulties, flexible-vocabulary speech recognition is an ongoing area of research. Chung [5] [6] uses subword units to allow recognition of new vocabulary. Her approach is shown to reduce the damage to overall recognition performance that occurs when the user uses an unknown word. Dusan and Flanagan [7] describe a complete system in which new words are acquired and semantics associated with them. The approach they use to recognise the new vocabulary is to back off to a broader-coverage recogniser. So in this way, words out of cov-

erage of the primary recogniser are handled by using an additional recogniser that those words are not out of coverage of.

Multiple layered recognisers can provide a practical solution to the problem of out-of-vocabulary words. By layering grammar-based and statistical language models, the performance advantages of each are able to be capitalised on. The broader, more flexible coverage of the SLM is used where grammar-based recognition fails, but the performance advantage of the grammar-based recogniser, as the primary recogniser, on “easy” utterances is not compromised. To summarise the approach, the grammar-based recogniser is initially applied to the user utterance. If this recogniser fails to produce a high-confidence interpretation, then the statistical recogniser is applied to the utterance. The SLM produces an interpretation of the utterance. In some cases, one or more words in this interpretation will be out of coverage of the grammar-based recogniser. These words are potential new words. Whether this happens often enough and reliably enough to be a useful strategy for recognising additional vocabulary is an empirical question. How likely is it that the user will use a new word and that this word will appear in the statistical language model? Will recognition quality be high enough that the words can be known with any certainty? Here, we attempt to answer these questions.

In section 2, the approach is described in more detail. The language models and test corpus used are described. Section 3 presents the most interesting results. The results are then discussed in section 4.

## 2. Method

The test domain used here is the home control domain. The grammar used is a fairly comprehensive grammar, derived from a unification grammar, that has been used for some time as part of an integrated mature system. This has enabled the collection of an extensive corpus. The SLM used is one created largely from material forming part of this corpus. Some of the corpus was put aside as test data. The Nuance 8 speech recognition system [8] was used here, both for the grammar-based and the statistical recognition. First, the grammar-based recogniser is used to recognise the utterances in the test corpus. The statistical recogniser is then used as a back-off, on those utterances where a good result was not obtained using the grammar. If the statistical recogniser recognised a word that was not in coverage of the grammar, then this word was identified as an out-of-vocabulary word. These words were then compared to the correct answers.

Section 2.1 describes in more detail the grammar used. 2.2 says more about the SLM. The corpus is then briefly described, before 2.4 describes the method in more detail.

### 2.1. Grammar

The grammar used is a home-control domain grammar developed for use in the On Off House system, described in [4]. It is implemented in the Nuance Toolkit Grammar Specification Language, and directly encodes slot-value semantics. It offers coverage of a fairly broad range of language, including commands (“Turn on the heater”, “Turn off the light in the bathroom”), several types of questions (“Is the heater switched on?”; “What is there in the kitchen?”; “Where is the washing machine?”; “Could you tell me which lights are on?”), universal quantification (“Switch off everything in the bathroom”), conjunction (“Are the hall and kitchen lights switched on?”; “Switch off the radio, TV and computer”), ellipsis (“Turn on

the cooker”... “now the microwave”) and pronouns (“Switch off the stereo and the hi-fi”... “switch them on again”). It is written for a domain in which eight rooms and some 20 devices are simulated. It has been tuned over four or five iterations of user testing. It has a total vocabulary size of 166 words. It achieves an overall word error rate of 31% on the test corpus used here.

### 2.2. Statistical language model

The SLM used is a standard trigram model created from some 4000 utterances collected using the On Off House system. Additionally, 200 utterances were included of a corpus of more free-form utterances collected using a minimal “starter” recogniser and no integration to a system. This data broadens the coverage of the SLM. This recogniser was again a standard trigram model, but one created from a much smaller data set, and designed to give users some minimal feedback without affecting the range of language they were inclined to use. The users in this data collection were also given a graphical representation of the home control scenario appropriate to the grammar, but containing no language that would be likely to influence the users in their choice of words. Four users took part in the data collection. The final SLM achieves an error rate of 25% on the test corpus used here. It has a total vocabulary size of 325 words. The recognisers are described in more detail in [3], where more detailed performance figures are also given, including semantic error rate, on which the grammar outperforms the SLM (40% compared to 44%), constituting a reason why one would want to use the grammar as the primary recogniser in this case, despite a higher word error rate.

### 2.3. Test corpus

The test corpus consisted of 645 utterances, of which 438 were collected using the On Off House system. The remainder were collected during the same free-form data collection described in section 2.2. The test corpus was entirely unseen by the statistical recogniser. It contained a total of 3350 words, of which 449 were out of coverage of the grammar. 101 words were out of coverage of the SLM.

### 2.4. Detailed description

The out-of-vocabulary material in the test corpus was first identified, providing a set of “right answers” that were later used to assess performance. Then, the grammar-based recogniser was applied to the test utterances. Any utterance that was recognised with a confidence threshold of 45 or greater was then removed from consideration. This was done to reflect the situation created in a real system, where a result accepted by the primary recogniser would not be further processed by the system, but instead taken to be correct. Nuance confidence scores give an indication of the reliability that the system attaches to its interpretation. They lie between 0 and 100. The threshold score of 45 is a common choice. It should be noted that this removal of the utterances that the grammar-based recogniser produced a good result for means that the utterances remaining are the ones that are less easy to recognise. They are more likely to contain out-of-vocabulary words, but also to display other features that cause recognition to fail. For example, the sound quality could be poor.

The SLM was then used to produce one interpretation for each of the remaining test utterances, which numbered 134. The SLM result was compared to the list of words that are in coverage of the grammar. A word that appeared in the SLM in-

terpretation but not the grammar lexicon was considered to be a potential new vocabulary item. These items were then compared to the correct answers prepared earlier.

This basic method was repeated with varying utterance-level confidence thresholds both on the grammar recognition and on the SLM recognition. Then, a variety of word-level confidence thresholds were applied on the SLM recognition. If the utterance or word did not produce a result with a score at or over the threshold, then it was discarded from consideration. Alteration of the grammar confidence threshold is somewhat unrealistic, since in a live system, performance with the primary recogniser would be the first priority, so this variation was not focused on. Variation of the confidence threshold on the SLM recognition however constitutes an important area of investigation in this work. Introduction of a threshold has the potential to greatly improve precision without damaging recall to a great extent, since the words excluded will be those that the recogniser is not confident of. Results will be presented for various SLM confidence threshold variations in the next section.

### 3. Results

The results presented in tables 1 to 4 show performance figures for experiments in which the confidence threshold on the grammar-based recogniser was set to 45. The four tables present results for word confidence thresholds of 30, 45, 50 and 60 on the statistical recognition.

Table 1: *Grammar 45 SLM Word Confidence Threshold 30*

True positives	118
False positives	65
Omissions over included words	112
Omissions total	331
Precision	118/183=64%
Recall	118/230=51%
Total recall	118/449=26%

Table 2: *Grammar 45 SLM Word Confidence Threshold 45*

True positives	108
False positives	40
Omissions over included words	112
Omissions total	341
Precision	108/148=73%
Recall	108/230=47%
Total recall	108/449=24%

Table 3: *Grammar 45 SLM Word Confidence Threshold 50*

True positives	102
False positives	33
Omissions over included words	128
Omissions total	347
Precision	102/135=76%
Recall	102/230=44%
Total recall	102/449=23%

“Omissions over included words” is the number of out-of-vocabulary words that the method failed to pick up of those in

Table 4: *Grammar 45 SLM Word Confidence Threshold 60*

True positives	64
False positives	20
Omissions over included words	166
Omissions total	385
Precision	64/84=76%
Recall	64/230=28%
Total recall	64/449=14%

the corpus of utterances on which the grammar-based recogniser did not achieve a confidence of above 45. “Omissions total” is the number of words in the entire corpus that the method did not pick up. Precision is calculated as the number of correct positive answers divided by the total number of words the method produced as potential out-of-vocabulary words, and so the divisor is different in each condition. Recall is the number of correct positive answers divided by the actual number of out-of-vocabulary words in the section of the corpus that the grammar failed to produce a good result on. There were 230 actual out-of-vocabulary words in this section of the corpus, as indicated by the fact that true positives plus omissions equals this figure in each case. True positives plus omissions total adds up to 449 in each case, the total number of out-of-vocabulary words in the entire corpus, and it is this figure that is the divisor in the total recall calculation.

Further results obtained include those produced by varying the utterance-level confidence score on the SLM-based recognition. By using a threshold of 45 a precision of 63% was obtained, with a recall of 50% and a total recall of 26%. Using an utterance threshold of 60 on the SLM recognition produced an increase in precision, to 65% but with a recall of 42% and a total recall of 21%. This variation was considered therefore less successful, since in obtaining a precision comparable with even the lowest produced using word confidence scores, much lower recalls are obtained.

Note that in the results tables presented above, if a word appeared twice in the automatic result and once in the correct list of OOV items it was counted as correct twice. If it appeared once in the automatic result and twice in the correct list it was counted once. If it appeared twice in the correct list and not in the result it was counted as two omissions.

### 4. Discussion

The results show that precisions in the 64-76% range are achievable in identifying vocabulary not present in the primary grammar-based recogniser using this approach. Overall recall is low; in the mid 20-something percent range at best, though factoring in the fact that any utterance that was accepted by the grammar-based recogniser was removed from consideration, the recall figures are a more impressive 28-51%. Increasing the word confidence threshold on the statistical recognition predictably improves precision, but at the expense of recall. The results shown here demonstrate how word confidence level tuning on the SLM-based recognition can be performed to reflect different priorities, with 45 in this case appearing to be a “happy medium” and 50 somewhat more cautious. Increasing the word confidence threshold to 60 appears not to improve precision noticeably.

As is evident from the difference in the number of out-of-vocabulary words present in the total corpus and that in the

subcorpus remaining after the grammar-based recogniser had been used, a significant number of utterances are being accepted by the primary recogniser despite their containing out-of-vocabulary words (449-230=219). The number of OOV words in the section rejected by the grammar is clearly greater than chance would predict (51% of the OOV words in 21% of the utterances). However, that 49% of the OOV words are missed at this stage suggests room for improvement. It is a subject open to investigation whether a better way to detect utterances possibly containing OOV words can be developed. Such a method should however not impact on the performance of the primary recogniser, which is a danger if the interpretation of utterances accepted by the primary recogniser is to be questioned. Ultimately, the aim in speech recognition for spoken dialogue systems is to produce the lowest possible *semantic* error rate, and the presence of an OOV word in an accepted utterance does not necessarily mean that an incorrect semantic result has been accepted by the system.

By tuning using confidence scores, precision can be prioritised over recall. This priority choice reflects what would be appropriate in a system in which a range of strategies are available for handling a poor recognition result. An incorrectly identified new word might lead to a confusing response on the part of the system; however, other strategies can be applied in the situation where the system fails to identify a new word. In any case where a new word is identified, this can be used to assist the user; by making them aware of the word the system cannot deal with, or by attempting to acquire a semantics for the word. Even if out-of-vocabulary words are identified only relatively infrequently, use of this information can only be positive.

It is possible that performance could be improved through the use of an SLM with broader coverage. Increasing the coverage of the SLM increases the number of OOV words it can potentially recognise, but this trades off against the actual recognition performance of the SLM, so care would need to be taken. It would be interesting to investigate the extent to which the coverage of the SLM could be increased without impacting too heavily on precision, not least because it is through the difference in coverage between the grammar and the SLM that the approach justifies itself; if the coverage of the SLM is not much greater than that of the grammar, then the obvious question is, why not just extend the grammar to cover all the words in the SLM? Reasons for not doing this where the coverage difference is significant include that the grammar-based recogniser will perform better if the coverage is minimised, and that extending the coverage of the SLM is much less labour-intensive than making changes to the grammar, and possibly the system.

The extent to which the result shown here will generalise to other systems is a subject worthy of some attention. The language models used here are restricted-domain command-and-control recognisers, and might be expected to compare well to recognisers in other imperative systems of equivalent coverage. The method applied was designed to reflect the circumstances in which the approach might typically be applied. The SLM used was created primarily from a corpus collected using the grammar. This reflects the nature of corpora typically available to developers working with a grammar-based system. The grammar confidence threshold of 45 on the primary recognition reflects a restriction typical in a system where the first priority is the performance of the primary recogniser. Use of a widely-used commercial recogniser further assists in the creation of a representative environment.

It remains an open question however to what extent the results will generalise to systems with broader coverage and

different foci. A system with broader coverage might lead to less restricted utterances on the part of the user, increasing the demand for and very likely the success of a method such as this. Furthermore, different domains might vary in the extent to which the user is inspired to use new vocabulary. Use of a question-based, as opposed to imperative, system might also produce differing results. Varying results might also be expected to be obtained in systems written for languages other than English.

## 5. Conclusion

Possibilities created by the combination of grammar-based and robust approaches to speech recognition are further extended here with a demonstration that a back-off statistical language model can be used to recognise words out of vocabulary of the primary grammar-based recogniser. The approach identifies new vocabulary with fairly high precision. Precision can be further increased at the expense of recall. Future directions for the work include application of the approach to other systems and domains in order to demonstrate its generalisability.

## 6. Acknowledgements

The author would like to acknowledge the assistance of Nuance Communications Inc. in making available the speech recognition software used in this work, Fluency Voice Technology (UK) for allowing use of their corpus and Matthew Purver for discussion of the idea.

## 7. References

- [1] Moore, R., Dowding, J., Bratt, H., Gawron, J. M., Gorf, Y. and Cheyer, A., "CommandTalk: A Spoken-Language Interface for Battlefield Simulations", Fifth Conference on Applied Natural Language Processing, Washington, D.C., April 1997.
- [2] Zue, V., Seneff, S., Glass, J., Polifroni, J., Pao, C., Hazen, T. J. and Hetherington, L., "JUPITER: A Telephone-Based Conversational Interface for Weather Information", In IEEE Trans. Speech and Audio Processing, 8(1), pp.85-96, 2000.
- [3] Knight, S., Gorrell, G., Rayner, M., Milward, D., Koeling, R. and Lewin, I., "Comparing Grammar-Based and Robust Approaches to Speech Understanding: A Case Study", Proceedings of Eurospeech 2001.
- [4] Gorrell, G., Lewin, I. and Rayner, M., "Adding Intelligent Help to Mixed Initiative Spoken Dialogue Systems", Proceedings of ICSLP 2002.
- [5] Chung, G., "Automatically Incorporating Unknown Words in Jupiter", Proc. 6th International Conference on Spoken Language Processing, Beijing, China October 2000.
- [6] Chung, G., "A Three-stage Solution for Flexible Vocabulary Speech Understanding", Proc. 6th International Conference on Spoken Language Processing, Beijing, China October 2000.
- [7] Dusan, S. and Flanagan, J., "Adaptive Dialog Based upon Multimodal Language Acquisition", Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces, Pittsburgh, PA, USA, Oct. 2002
- [8] Nuance Communications Inc., [www.nuance.com](http://www.nuance.com)