Modeling Conversational Speech for Speech Recognition

Marie Meteer and Rukmini Iyer
BBN Systems & Technologies
70 Fawcett St.
Cambridge MA 02138
mmeteer@bbn.com, riyer@bbn.com

Abstract

In language modeling for speech recognition the goal is to constrain the search of the speech recognizer by providing a model which can, given a context, indicate what the next most likely word will be. In this paper, we explore how the addition of information to the text, in particular part of speech and dysfluency annotations, can be used to build more complex language models. In particular, we ask two questions. First, in conversational speech, where there is a less clear notion of "sentence" than in written text, does segmenting the text into linguistically or semantically based units contribute to a better language model than merely segmenting based on broad acoustic information, such as pauses. Second, is the sentence itself a good unit to be modeling, or should we look at smaller units, for example, dividing a sentence into a "given" and "new" portion and segmenting out acknowledgments and replies. To answer these questions, we present a variety of kinds of analysis, from vocabulary distributions to perplexities on language models. The next step will be modeling conversations and incorporating those models into a speech recognizer.

1 Introduction

In language modeling for speech recognition the goal is to constrain the search of the speech recognizer by providing a model which can, given a context, indicate what the next most likely word will be. Currently, the field is predominated by the use of n-grams, in particular bi-grams and tri-grams. One advantage of this type of model is that only the text itself is needed to create the model. In the case where what is being modeled is written text in a style for which millions of words of text are available, such as the WSJ corpus, this kind of modeling is effective. However, this is rarely the case since the ultimate goal of speech recognition is to model extemporaneous spoken language.

The interesting problem in language modeling is how to bring generalizations above the level of the words themselves to the text. One approach is to annotate text, either by hand or semi-automatically, to bring additional information to the text. Another is to develop algorithms that use rules or heuristics that operate over the corpus, again bringing additional information. It is this additional information that can help create generalizations over the text and contribute to a model which can go beyond the training corpus.

In this paper, we describe annotations to the Switchboard corpus which add part of speech and dysfluency markings and our latest work using those annotations to create a level of generalization on top of the annotation to capture the structure of conversational speech. In particular, we ask two questions. First, in conversational speech, where there is a less clear notion of "sentence" than in written text, does segmenting the text into linguistically or semantically based units contribute to a better language model than merely segmenting based on broad acoustic information, such as pauses. Second, is the sentence itself a good unit to be modeling, or should we look at smaller units, for example, dividing a sentence into a "given" and "new" portion. To answer these questions, we present a variety of kinds of analysis, from vocabulary distributions to perplexities on language models.

1.1 The Switchboard Corpus

The Switchboard corpus (Godfrey, et al. 1992) consists of 2 million words of conversational English collected over the phone by having strangers chat with one another about 70 different topics ranging from pets and family life to education and gun control. Each conversation is about five minutes long and the transcription includes information on background noises and where conversants overlapped (talked over one another).

Conversational speech is particularly difficult for speech recognizers, since not only is the speech often very quick and "sloppy" (e.g. with "mgonna" for "I am going to"), but it is also very dysfluent, with many fillers, such as "uh" and "you know", and many restarts, such as "I, uh, well, I went, I went away last summer". It has been surmised that it is the dysfluencies that make the language modeling particularly difficult, but no studies have been able to conclusively show that that is true or explain the problems sufficiently. In fact, at the 1995 Language Modeling Workshop at Johns Hopkins they showed that sentences with dysfluencies do not perform any worse in recognition word error rates than sentences without dysfluencies; they have a slightly lower error rate. Whether this has to do with the overall sentence length or if different kinds of dysfluencies correlate differently with error rate still needs to be explored. Stolcke and Shriberg (1996) have begun work in this area. In particular, they show that some results are different depending on whether the data was segmented linguistically rather than acoustically (see §3).

The Switchboard annotation effort had four parts: part of speech annotation, "sentence" boundaries, dysfluencies, and bracketing. Part of speech and bracketing were done in the UPenn treebank style (and all annotation was done at Upenn by the Linguistic Data Consortium (LDC). The dysfluency and sentence boundary annotation stylebook was begun at BBN by the authors of this paper and completed at UPenn by Ann Taylor. We elaborate on these later two types of annotation in Section Two. The main target audience for this annotation was the 1995 Language Modeling Workshop at Johns Hopkins, which brought together researchers with diverse backgrounds to tackle the problem of language modeling for conversational speech. Some of the work we report on in this paper was begun at the workshop.

The annotation notation was based on the work of Elizabeth Shriberg (1994). We used her work as a starting point for several reasons. First, her notation was quite comprehensive, covering all (and more) of the phenomena we were interested in annotating. Second, she had done much of her work on the Switchboard corpus, annotating 40,000 words, so we knew that most, if not all, the idiosyncrasies of that corpus would be dealt with. Third, others can more readily build on the extensive analysis in Shriberg's thesis if the annotations are not gratuitously different.

1.2 Goals of the work

The ultimate goal of this work is to improve the performance of a speech recognizer. In the process we hope to gain a better understanding of the structure of conversational speech and how styles of speech differ.

Annotating data is an enormously expensive task, so before we select what information to annotate, we need to think about what use we can make of that information. In the case of dysfluency annotation, the sheer number of dysfluencies in conversational speech, plus the fact that dysfluencies do not occur at all in written speech and thus cannot be modeled using that source, are indicators that this annotation can contribute to our language modeling work. Some other questions about the structure of conversational speech that can be addressed with annotated data are as follows:

- Are dysfluencies more likely to occur in some places in the sentence than others?
- Do dysfluencies carry some useful information?
- What is the notion of a sentence or segment in conversational speech?
- Is conversational structure different from formal written structure and if it is, can we exploit that difference in our language models?

The most innovative part of this paper is the use of the information structure of conversational utterances to develop targeted language models for different parts of a sentence. According to many theories of discourse in linguistics (e.g. Clark and Haviland, 1977, or Haliday and Hasan, 1976), sentences have an "information structure" in addition to the syntactic structure that distinguishes between *given* information, which is already shared by the conversational participants, either from shared knowledge or information contained earlier in the dialog, and *new* information that the speaker is conveying to the hearer. In the unmarked case (e.g. sentences not marked with emphatic stress or cleft sentences, which are rare in conversational speech), *given* information tends to occur in the beginning of a sentence where the topic is established, whereas *new* information tends to occur at the end, the comment on the topic. This work is currently in the data analysis phase, where we are developing heuristics to divide a sentence into its *given* and *new* parts, which roughly correspond to the parts of a sentence before and after the verb. While this simple heuristic is far from the complex notion of given and new in the linguistic literature, it has on main advantage: it can be computed automatically given sentence boundaries, part of speech information, and dysfluency information. If the simple notion pans out and we can show a significant improvement in recognition performance, we may want to move to a more complex notion and hand annotate it in the corpus. However, many ideas which are

intuitively appealing are difficult to incorporate in a language model or do not appear to make a difference in overall performance.

In our work so far, our analysis of the vocabulary and its distribution in these two parts of the sentence shows a significant difference. We are currently building language models from the two parts and determining ways of integrating them that takes advantage of this difference. The next step will be modeling conversations and incorporating those models into a speech recognizer.

We discuss the dysfluency annotation of Switchboard in detail in Section 2. In Section 3, we describe work on linguistic vs. acoustic segmentation and its effect on the language model. In Section 4, we describe the given/new distinction as we have implemented it and our analysis of the Switchboard corpus. Finally, in Section 5, we describe a model of conversational speech that takes advantage of the given/new distinction and how it can be used in a speech recognition system.

2 Annotations of Switchboard

There were three major kinds of annotations done as part of the dysfluency annotation of Switchboard: sentence boundaries, restarts, and non-sentence elements. The assumption underlying the dysfluency annotation was that when it was complete, sentences could be separated, restarts "folded", and non-sentential elements removed and the result would be a reasonably grammatical sentence (that is, grammatical for conversational speech, not necessarily compliant with your third grade English teacher), though some may not be "complete" in that they may be replies or acknowledgments and some may be interrupted by either the speaker herself or the other conversant and never completed. As mentioned earlier, much of the choice of what to annotate and details of the notation are based on the work of Shriberg (1994). The main difference is that our work is not as detailed as Shriberg's, since we were not planning as fine grained analysis, and it covered significantly more data (Shriberg annotated 40,000 words, whereas this effort annotated 1.4 million words). In the next three sections, we describe these types of annotations and provide some examples.

2.1 Sentences

In written text, the definition of a sentence is clear and marked in the text itself by capitalization and punctuation. For conversational speech, the most natural division would appear to be the turn, when one speaker stops speaking and another starts. However, when we look at the data, we see that participants often interrupt and talk over one another, so even separating turns is not so simple. Within a turn a participant may ramble on and on, making the utterance too long for a speech recognizer to handle.

In annotating Switchboard, we choose to divide turns into "sentences" consisting each of a single independent clause. When two independent clauses are connected by a conjunction, they are divided with the conjunctions marked as described in §2.3.4.

Sentence units are followed with "/" indicating a sentence boundary, as shown in example 1. A sentence is considered to begin either at turn beginning, or after completion of a preceding sentence. Any dysfluencies between the end of a previous sentence and beginning of the current one is considered part of the current sentence.

In Example 2, there are essentially two sentences. The first sentence is across a turn by speaker A, namely "we did get the wedge cut out by building some kind of a cradle for it". The other sentence is by speaker B which is "A cradle for it". "You know" at the end of a sentence (Example 3) is considered as a part of the current sentence, as described in more detail in §2.3.1.

Ex 1: A: You interested in woodworking? /

Ex 2: A: we did get the wedge cut out by building some kind of --

B: A cradle for it. /

A: -- a cradle for it. /

Ex 3: B: I painted, about eight different, colors, you know. / the crayons that are sticking up, it will be the headboard -- /

Each sequence of words consisting of only continuers or assessments (expressions such as "uh-huh", "right", "yeah", "oh really") is also coded as a sentence, as in Examples 4 and 5.

Ex 4: Yeah /

Ex 5: Right / Right /

2.1.2 Incomplete sentences

Sentences that do not end normally are treated as incomplete sentences. They are marked with "-/". In some cases the speaker stops a sentence and starts over (in contrast with restarts where just a few words are repeated, as described in §2.2). In other cases, the other participant in the conversation interrupts the speaker and the speaker never finishes the sentence (in contrast with cases such as

example 2 above, where the first speaker finishes the sentence in the next turn after or during the interruption).

Ex 1: B: what I've seen of this kind before is you have the, -/ if you're looking at adding on you have, -/

Ex 2: A: Perhaps things that we didn't think of before and just concentrated on the lawmaking or the results that would be seen in public works or bills that are passed or, et cetera like that -- -/

Ex 3: B: -- it was very unfortunate thing that occurred there / it's, -/

A: Where do you live? /

B: we live in Utah. /

2.2 Restarts

Restarts are considered to have the following form in Shriberg's work and elsewhere. The initial part is the reparandum (RM), which is the part that the speaker is going to repair. The interruption point (IP) markes the end of the reparandum and it is followed by an optional interregnum (IM), which includes editing phases, such as a filled pause or editing terms. Finally, the repair (RR) is what the speaker intends to replace the RM with.

Show me flights from Boston on uh from Denver on Monday

|-------- RM --- | IM |----- RR ---- |

IP

In order to simplify the notation, the restart notation we developed marks only the boundaries of the entire restart (RM to RR) with square brackets and the interruption point with a "+" . Partial words are also not marked specially (though in the transcripts they end in a "-"); they appear directly to the left of the interruption point. In contrast with Shriberg's work, no internal structure of the restart is included (e.g. which words are repeated, substituted or deleted).

Show me flights [from Boston on + {F uh } from Denver on] Monday

A restart is "repaired" by deleting the material between the open bracket and the interruption point (+). (Note that fillers such as "uh" in the above example are deleted as a separate process in cleaning the text. We discuss them in §2.3) Some examples of restarts and repairs are given below. Note in Example 1, it is not always clear how much should appear in the repair. "In the book" could also have been included. However, to try to reduce the variation in annotation, annotators were instructed to keep the repair as short as possible. In Example 2, it can be seen that a restart has been marked across a turn, with the RM and IM in one turn and the RR in the next turn.

Ex 1: A: [it, + the instructions] in the book I had said use a coping saw but there's no coping saw big enough [to, + for] a fourteen inch wide watermelon /

Ex 2: B: [It's, + uh

A: -- pine? /

B: It's,] plywood face I guess.

In the second restart in Example 3, it is not clear what is the RR for the RM "and". In cases where there does not appear to be a suitable replacement for the restart, annotators were instructed to place the "]" as close to the IP as possible. One rule of thumb that can be followed in case of marking restarts without repairs is that they are always at the beginning or in the middle of the sentence, the sentence continues after the restart and the restart usually comprises one to three function words.

Ex 3: B: [I got, + uh it got] delayed for a little bit [and,+] because of work /

2.2.1 Complex Restarts

Multiple restarts are handled as embedded and are repaired from left to right. Some examples of complex restarts are shown below. In Example 2, the left to right annotation has not been strictly observed since "[Ber-, + Bermuda]" appears as a restart with repair within the repair of another restart.

Ex 1: A: to keep an inmate in there [[on a, + on a,] + on a] life sentence /

Ex 2: B: Yeah, / [[they're, + Um you know they're] like Ber-, +

A: Dress shorts. /

B: they're like black corduroy [Ber-, + Bermuda]] shorts.

2.3 Non-Sentence Elements

Non-sentence elements are words or phrases which are inserted in an utterance, disrupting the flow of the sentence. They are simple units with no internal structure and no interruption point. There are five types of non-sentence elements: filled pause {F }, editing term {E }, discourse marker {P }, conjunction {C }, and aside {A }.

2.3.1 Filled Pause

Filled pauses have unrestricted distribution and no semantic content. A few examples of fillers are "uh" "um", "huh". There can also be other filled pauses which are rare such as "eh" and "oops". "oh" can be treated as a filled pause if it appears along with other words for example "oh yeah", "oh really", as in Example 1. Otherwise, "oh" is treated as a regular word unit of language if it appears by itself as a reply, as in Example 3.

Ex 1: B: {F Oh }, yeah. / Uh-huh. /

Ex 2: B: Actually, [I, + {F uh, } I] guess I am /[laughter]. {F um }, it just seems kind of funny that this is a topic of discussion. / {F uh }, I do, {F uh }, some, {F uh }, woodworking myself / [noise]. {F uh }, in fact, I'm in the middle of a project right now making a bed for my son. /

Ex 3: B: Oh /

2.3.2 Explicit Editing Term

Editing terms are usually restricted to occur between the restart and the repair and have some semantic content (e.g. "I mean" "sorry", "excuse me"), as shown in Example 1, through it is possible that editing terms occur outside the RR.

Ex 1: A: {F Oh, } yeah, / {F uh, } the whole thing was small and, [you, + {E I mean, } you] actually put it on [laughter], /

2.3.3 Discourse Marker

Discourse markers have a wider distribution than explicit editing phrases but are unlike filled pauses in that they are lexical items (e.g. "well", "you know", "like"). "You know" is the most frequent discourse marker and is used very frequently by some speakers, as shown in Example 3. There are some other terms such as "so" and "actually" which can also serve as discourse markers, as in Example 2; however, "so" can also be a coordinating conjunction or a subordinating conjunction, as discussed in §2.3.4. In Example 4, it can be observed that the discourse marker is within the RR of a restart.

Ex 1: B: {P Well }, we have a cat who's also about four years old. /

Ex 2: A: he comes back. / {P So } [he, + he's] pretty good about taking to commands /

Ex 3: B: Yeah, / with, {P you know, } me being at home and just having the one income, {P you know, } you don't have, this lot o f extra money [to, + to] do a lot of, {P you know, } extra things. /

Ex 4: B: [We take, + {P you know, } whenever we take] them to Showbiz or -/ they think it's wonderful just to go to McDonalds, /

2.3.4 Coordinating Conjunction

Coordinating conjunctions occur at the inter-sentential level and generally include "and", "but" and "because". In some cases it is possible that two words together constitute a conjunction, for example "and then", as in example 2. Most of the conjunctions that appear between two full clauses are marked as coordinating conjunctions. The rule of the thumb to be followed is "split sentences whenever possible" except when the two sentences, if split, are grammatically incorrect (for example the second sentence in the spilt does not have a subject since it is in the earlier sentence).

Ex 1: A: Yeah, / {C and } we got him when he was about eight weeks old / {C and } he's pretty okay, /

Ex 2: B: {C and then } I painted, {F uh }, about eight different, {F uh }, colors, /

Example 3 is of "so" as a coordinating conjunction. Note that in Example 4, the second "and" is NOT treated as a coordinating conjunction, as the two sentences it conjoins ("I call him" and "he comes back") are both short and both appear to be modified by the initial "if" clause.

Ex 3: B: {P Well }, {F uh }, we just moved recently [laughter] / {C so} now we're in the, {F uh }, Dallas area /

Ex 4: A: he's pretty good. / He stays out of the street / {C and, } {F uh }, if I catch him I call him and he comes back. / {P So } [he, + he's] pretty good about taking to commands /

2.3.5 Asides

This is a category for "asides" that interrupt the flow of the sentence. Interjections are rare and are considered only when the corresponding sequence of words interrupt the fluent flow of the sentence AND the sentence later picks up from where it left. The examples below clearly illustrate this.

Ex 1: B: I, {F uh }, talked about how a lot of the problems they have to [come, + overcome] [to, + {F uh, } {A it's a very complex, {F uh, } situation } to] go into space. /

Ex 2: A: {P So } we built a cradle for it / {C and } [we got th-, + {A once it was turned, } we got] [one s-, + one] cutout on the table saw, on the radial saw, /

3 Linguistic Segmentations

As explained in the previous section, one of the important annotations of the Switchboard corpus involved the issue of sentence boundaries, or segment boundaries. Sentence boundaries are easy to detect in the case of read speech where there is a distinctive pause at the end of the sentence and the sentence is usually grammatically complete (the second also holds true in case of written speech, where in addition a period marks the end of a sentence). However, this is not so in the case of conversational speech as is clear from the examples above. In conversational speech, it is possible to have incomplete sentences, sentences across turns and complex sentences involving restarts and other dysfluencies.

Prior to having annotated data, the segment boundaries for conversational text data were provided in the form of acoustic segmentations. These segmentations were based on pauses, silences, non-speech elements (e.g. laughs and coughs) and turn taking. The differences between the two forms of segmentations can be observed with the example given below[1]:

Acoustic segmentations

> I'm not sure how many active volcanoes there are now and and what the amount of material that they do <s> uh <s> put into the atmosphere <s> I think probably the greatest cause is uh <s> vehicles <s> especially around cities <s>

Linguistic segmentations

> I'm not sure how many active volcanoes there are now and and what the amount of material that they do uh put into the atmosphere <s> I think probably the greatest cause is uh vehicles especially around cities <s>

In the n-gram approach to statistical language modeling, the segment boundary is treated as one of the symbols in the lexical dictionary and modeled similar to other words in the data stream. The segment boundaries provide an additional source of information to the language model and hence it appears intuitively correct to use linguistic segmentations for training language models. The notion of segmentation is also an important issue if we use higher level language models such as phrase structure models or sentence-level mixture models (Iyer, et al. 1994). However, given only a speech signal during recognition with no text cues available for segmentation, there will be an inherent mismatch between the linguistically segmented training data and the acoustically segmented test data.

Thus the segmentation experiments tried to answer three important issues:

- Does a mismatch in training/testing segmentation hurt language model performance (perplexity and word error rate)?
- Is there any information in segment boundaries?
- If no boundary information is available during testing, can we hypothesize this information using a language model trained with segmented training data?

3.1 Experimental Setup

In order to analyze the above issues, we first obtained our baseline training and testing data. Since the linguistic segmentations are available for only two thirds of the Switchboard data, we decided to use the corresponding two thirds of the acoustically segmented training data for our comparative experiments. The test set was obtained from the Switchboard lattices which served as the baseline for the 1995 Language Modeling Workshop at Johns Hopkins. The test set was acoustically segmented. A corresponding linguistically segmented test set was also made available[2].

3.1.1 Recognition Paradigm

We used the N-best rescoring formalism for recognition experiments with the test data (Ostendorf, et al. 1991). The HTK toolkit was used to generate the top 2500 hypotheses for each segment. A weighted combination of scores from different knowledge sources (HTK acoustic model scores, number of words, different language model scores, etc.) was then used to re-rank the hypotheses. The top ranking hypothesis was then considered as the recognized output.

3.2 Mismatch in Training and Test Data Segmentations

We trained three trigram language models: two using acoustic segmentations and linguistic segmentations respectively and a third model trained on data with no segment boundaries. The models used the Good-Turing back-off for smoothing unseen n-gram estimates (Katz 1987). These models were then used to compute perplexity on the different versions of the test data. The trigram perplexity numbers are shown in Table 1.

[1] <s> represents the sentence/segment boundary.

[2] Unfortunately, the two test sets did not match completely in terms of the number of words since the lattice test set had been hand corrected after the initial transcription to account for some transcription errors. Hence, there is a difference of about two hundred words between the acoustically and linguistically segmented test sets.

Test	Training		
	acoustic-seg	ling-seg	no-seg
acoustic-seg	105	111	
ling-seg	89	78	
seg removed	163	174	130

Table 1: Trigram perplexity measurements on LM95 SWBD dev. test set

As indicated in Table 1, mismatch between training and testing segmentation hurts perplexity. The best perplexity numbers are obtained under matched conditions. Though the results for the linguistically segmented test set (78) are significantly better than the corresponding matched case for the acoustic segmentations (105), we cannot conclusively state that this is due to better segmentation since we have not controlled for the length of the different segments.

3.3 Hypothesizing Segment Boundaries

A second perplexity experiment that we conducted tried to test whether we can hypothesize segmentations, given that we have no boundaries in the test set. Our segment-hypothesizing algorithm[3] assumed that at any word, we have two paths possible,

- A transition to the next word.
- A transition to the next word through a segment boundary.

The algorithm was approximate in that we did not keep track of all possible segmentations. Instead at every point, we picked the most likely path as the history for the next word.

As in the first experiment, we trained two language models on linguistic segmentations and acoustic segmentations respectively. Henceforth, these models are referred to as the **ling-seg** and **acoustic-seg** models. Both models try to hypothesize the segment boundaries while computing the likelihood of the no-segmentation test set.

Test	Training	
	acoustic-seg	ling-seg
no seg	163	127

Table 2: Trigram perplexity on LM95 SWBD dev. test set hypothesizing segment boundaries

[3] This is work done in collaboration with Roni Rosenfeld at the 1995 Johns Hopkins Workshop on Language Modeling.

The perplexity results of Table 2 indicate that the ling-seg model does better than the acoustic-seg model for hypothesizing segment boundaries. Thus, we can gain a significant amount of boundary information by this simple scheme of hypothesizing segmentations.

3.4 Recognition Experiments

There were a couple of experimental constraints to analyze the aforementioned issues in terms of recognition word error rate.

- We were constrained to use the lattices that had been provided to the workshop. Since these lattices were built on acoustic segments, the models had to deal with implicit acoustic segment boundaries. The context from the previous lattice was not provided for the current lattice.
- We tried to alleviate this problem by trying to provide the context for the current lattice by selecting the most likely pair of words from the previous lattice using pair occurrence frequency. One problem with this approch is that since the standard Switchboard WER is about 50%, about 73% of the time we were providing incorrect context using these lattices.
- We used our segment hypothesizing scheme for scoring an N-best list corresponding to these lattices (N=2500). While the initial context was provided for the N-best lists, we had to throw away the final segment boundary. This led to a degradation in performance.

Model	WER (%)	
	acoustic boundaries	hypothesizing boundaries
acoustic seg	50.46	51.85
ling seg	50.88	51.72

Table 3: N-best rescoring performance (N=2500) measurements on LM95 SWBD dev. test set

As shown in Table 3, the mismatch between the training and test segmentations degrades performance by half a point, from 50.46% to 50.88%. Throwing out the end segment boundary from the N-best lists degrades performance by slightly more than an absolute 1%. Also, the ling-seg model does slightly better at hypothesizing segment boundaries than the acoustic-seg model.

3.5 Conclusions

Our experiments indicated the following:

- Mismatch in segmentation hurts language model performance, both in terms of perplexity as well as in terms of recognition word error rate.
- There is information in the knowledge of segment boundaries that should be incorporated in our language models.
- If no segment boundaries are known during testing, it is better to hypothesize segment boundaries using a model trained on linguistic segments than one based on acoustic segments.

The notion of linguistic segmentation is important in language modeling because it provides information that is used in many higher order language models, for example the "given-new" model described in the next section, phrase structure language models, or sentence-level mixture models. However, this information cannot easily be derived from the acoustic signal. In this section, we have described a simple technique of hypothesizing segmentations using an n-gram language model trained on annotated data. We plan to run some controlled perplexity and recognition experiments in the future to use this information in our recognition system.

4. Information Structure in Conversational Speech

It is well known in the linguistic literature that sentences are not uniform from beginning to end in the kinds of words or structures used. Sentences have a "given/new" or "topic/comment" structure which is especially pronounced in conversational speech. According to discourse theories in linguistics, *given* information tends to occur in the beginning of a sentence where the topic is established, whereas *new* information tends to occur at the end, the comment on the topic.[4]

We are looking at ways of taking advantage of this structure in the language model. The first stage of the work is to devise a method of dividing sentences into these two parts. Next, treating the *before* and *after* portions of the sentences as separate corpora, we look at the distribution of the vocabulary and the distribution of other phenomena, such as restarts. We also build language models using these two corpora and test perplexity both within and across corpora. The final step, which we are currently in the process of, is to find a way to integrate these models and use them within the speech recognition system to see if these more focused models can actually improve recognition performance.

[4] This tendency is overridden in marked syntactic structures, such as cleft sentences ("It was Suzie who had the book last"). These structures are relatively rare in conversational speech.

4.1 Dividing the sentence

In order to divide sentences into their *given* and *new* portions, we devised a simple heuristic which determines a pivot point for each sentence. The underlying assumption is that the dividing line is the verb, or more particularly, the first verb that carries content (disregarding "weak" verbs such as "is", "have", "seems"). The heuristic finds the first strong verb and if there is none, then the last weak verb, and places a pivot point either before the strong verb or after the weak one. Sentences that have no pivot (i.e. that have no verb) are put into one of two classes, those that are considered complete (such as "Yeah" and "OK") and those that are incomplete, that is interrupted by either the speaker or the other conversant (i.e. "Well, I, I, uh."). The no pivot complete set is very similar the "Back Channel" model developed by Mark Liberman at the LM95 Summer Language Modeling Workshop (Jelinek 1995). Liberman separated back channel responses from information-bearing utterances and created a separate language model. Initial experiments shows no overall improvement in word error rate, however, the model was able to identify both previously identified and new backchannel utterances in the test data.

For the purposes of this paper, we will refer to the *given* and *new* parts as *before* and *after* (meaning before and after the pivot), and "NPC" for no pivot complete and "NPI" for no pivot incomplete sentences. The following shows an example dialog and Table 4 shows the corresponding division into the four categories:

A.1: Okay. I think the first thing they said, I have written this down so it would, is it p-, do you think it's possible to have honesty in government or an honest government?

B.2: Okay. You're asking what my opinion about,

A.3: #Yeah.#

B.4: #whether it's# possible [laughter] to have honesty in government. Well, I suspect that it is possible. Uh, I think it probably is more likely if you have a small government unit where everybody knows everybody.

A.5: Right. That's a good point.

B.6: But, uh, other than that I think maybe it just depends on how you define honesty.

A.7: That's an int-, you know, that's interesting.

S	NPC	NPI	BEFORE	AFTER
1	Okay		I think the first thing they	said
2			I have	written this down
3		so it would,		
4			is it p-, do you think it's possible to have	honesty in government?
5	Yeah			
6	Okay		You're	asking what my opinion about : #whether it's# possible [laughter] to have honesty in government
7			Well, I	suspect that it is possible.
8			Uh, I think it probably is more likely if you have a small government unit where everybody knows	everybody
9	Right		That's	a good point.
10			But, uh, other than that I think maybe it just	depends on how you define honesty
11			That's an int-, you know, that's	interesting

Table 4: Dialog divided into subparts

Note that this heuristic doesn't always make the correct division. In sentence 8, "is" is the main verb, however, the algorithm prefers to find a strong verb, so it keeps going until it finds "know", which is actually part of a relative clause. A more complex algorithm that finds the main verb group and uses the last verb in the verb group rather than the last verb in the sentence would remedy this. However, our goal here is to first determine whether in fact this division is useful in the language model. As long as errors such as this are in the minority, we can evaluate the method and then go back and refine it if it proves useful.

In order to do the classification, we relied on three kinds of annotations that were available for the switchboard corpus: sentence boundaries, part of speech, and dysfluency annotation. The dysfluency markings are needed since the pivot point is restricted from being inside of a restart. The following shows the first two turns in the above discourse with both of these annotations[5]:

SpeakerA1/SYM ./.Okay/UH ./. E_S I/PRP think/VBP the/DT first/JJ thing/NN they/PRP said/VBD ,/, N_S I/PRP have/VBP written/VBN this/DT down/RP E_S {C so/RB } it/PRP would/MD ,/, N_S [is/VBZ it/PRP p-/XX ./, + do/VBP you/PRP think/VB it/PRP 's/BES possible/JJ] to/TO have/VB honesty/NN in/IN government/NN or/CC an/DT honest/JJ government/NN ?/. E_S

SpeakerB2/SYM ./. Okay/UH ./. E_S You/PRP 're/VBP asking/VBG what/WP my/PRP$ opinion/NN about/IN ,/, whether/IN it/PRP 's/BES possible/JJ to/TO have/VB honesty/NN in/IN government/NN ./. E_S

Table 5 shows the breakdown of the data into the four divisions, before the pivot (*before*), after the pivot (*after*), complete sentences with no pivot (" NPC"), incomplete sentences with no pivot (" NPI").

	Before	After	NPC	NPI	Total
Total words	272,485	298,460	33,954	9364	614,263
Number of segments	50,741	47,814	25,483	4112	80,336
Avg. segment length	5.37	6.24	1.33	2.772	7.65

Table 5: Switchboard Corpus Divided by Pivot Point

[5] In this version of the annotation, complete sentences are marked E_S and incomplete sentences are marked N_S, rather than / and -/ as described in §2. This is to avoid confusion with the / which delimits words and their part of speech.

Freq. Rank	Word	Before	After	NPC	NPI	Total
1	i	57872	7876	263	1608	67619
2	and	32482	14135	836	2382	49835
3	the	14851	27073	802	809	43535
4	that	22592	17261	301	465	40619
5	you	22774	15896	684	967	40321
6	it	23143	14429	117	548	38237
7	to	13745	19389	102	52	33288
8	a	8354	24024	556	158	33092
9	uh	15480	11734	1666	1798	30678
10	's	24845	4414	71	31	29361
11	of	9244	15615	488	306	25653
12	know	12387	9062	329	696	22474
13	yeah	156	241	20813	10	21220
14	they	14709	4613	48	376	19746
15	do	15284	3252	6	5	18547

Table 6: Totals for 15 Most Frequent words

It is interesting to note that the size of the *before* and *after* corpora are very similar. Note that this is not necessarily because the algorithm is dividing the sentences into two equal portions, as we can see in the example above. Some sentences have a rather long introduction with restarts, as in sentence 4 and 11, whereas others have just a single word and a long *after* portion, as in sentence 6.

4.2 Vocabulary distributions

It is clear from the definitions of the *given* vs. *new* parts of the sentence, that the vocabularies in the corpora resulting from the division will have different distributions. *given* information will be expressed with a larger number of pronouns whereas the *new* portion will have more complex descriptive noun phrases, and thus a wider ranging vocabulary. Within the verb group, weak (and more common) verbs will appear in the *given* portion, whereas strong verbs that carry content will appear in the *new* portion. But rather than relying on these intuitions, we apply a more careful analysis of the data to determine more closely what the differences are.

4.2.1 Comparing most frequent words

The most frequent words in the corpus divide rather sharply across the data sets. For example, in Table 6, which shows the counts for the top 15 words, pronouns such as "I" and "it" (rows 1 and 6) are much more frequent in the *before* set, since pronouns are used generally to refer to participants in the conversation or things already mentioned, whereas articles such as "the" and "a" (rows 3 and 8) are much more frequent in the *after* part of the sentence, since they are more frequently used in full noun phrases describing new entities. "Yeah" (row 13) occurs almost exclusively in the NPC set, which is comprised mainly of replies.

Table 7 plots the 50 most frequent words in the corpus, showing their *before* and *after* raw totals. Note that while the values cross, they are rarely the same for the same word. This reinforces our intuition that the use of function words (typically the most common words) in the two parts of the sentences are quite different[6].

6 Note that only *before* and *after* sets are plotted, so series 2 and 3, the lower two lines, do not necessarily sum to series 1, the upper line representing totals in the corpus as a whole.

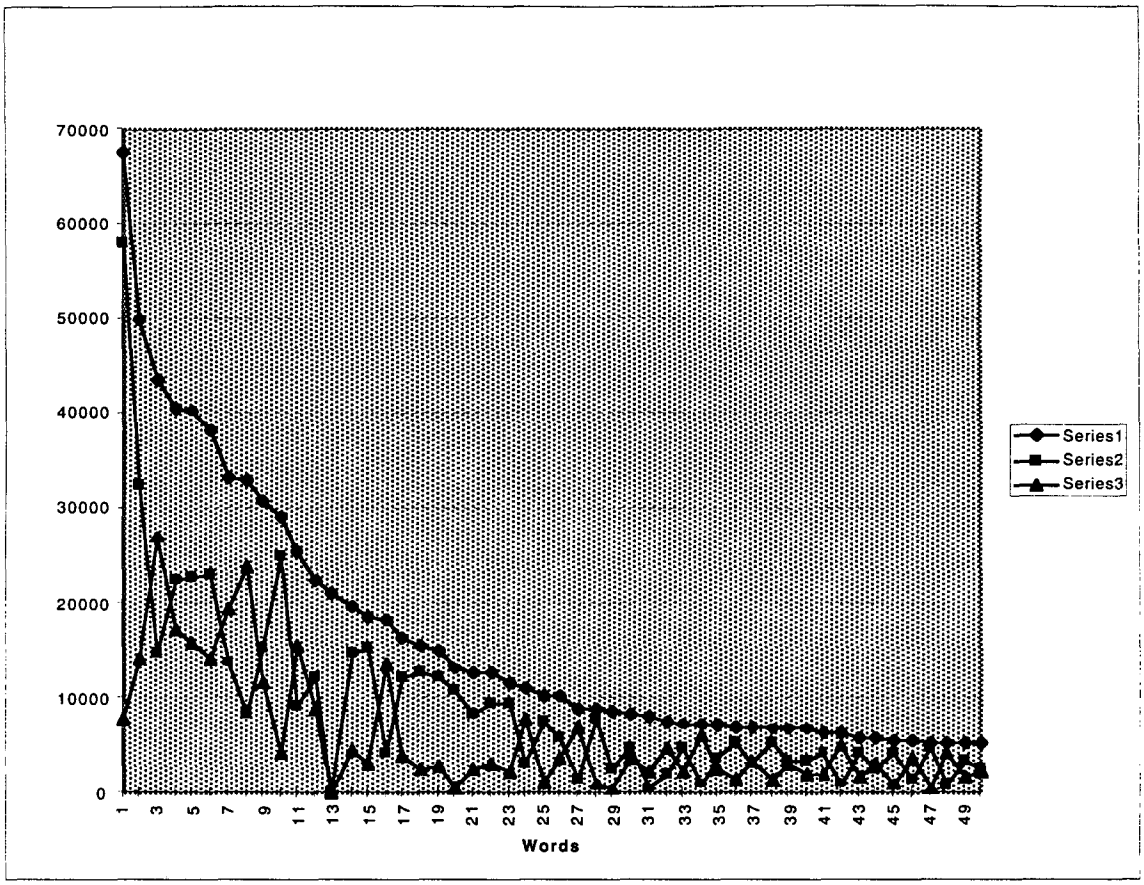

Table 7: *Before* and *After* Totals for 50 Most Frequent words
Series1: Totals for entire corpus
Series2: Counts for *before* pivot sentence parts
Series3: Counts for *after* pivot sentence parts

4.2.2 Differences in vocabulary

We also looked at the differences in the vocabularies of the two parts. Table 8 shows the frequencies of words that appear in one part of the sentence but do not appear in the other in this corpus. The raw totals are quite different, with 7292 words appearing in the *after* portion and not appearing in the *before* portion, while only 1028 appear in *after* and not in *before*. Also, note that less than half on one percent of the words in the *before* are uniquely in that part, where as 7% of the total words in the *after* part never occurred in *before*.

Row 1 shows the words that only occurred once in the corpus (which had to occur in just one side or the other). We see that the tail of new words is much longer in the *after* set than the *before*.

Frequencies	In Before, not in After	In After, not in Before
1	877	5328
2	114	1938
3-5	40	1202
6-10	1	381
11-50	0	217
51-100	0	24
100+ (max 214)	0	6
Total (different words)	1028	9944
Total (instances)	1254	30131
% of corpus	.046%	7%

Table 8: Vocabulary Differences in *Before* Pivot and *After* Pivot Sentence Parts

taken	222	forget	93	notice	75	sold	62
teach	185	expect	90	bother	74	continue	60
suppose	133	quit	89	moving	74	became	59
played	131	follow	85	died	73	mentioned	55
caught	129	wondering	80	selling	70	prefer	54
rid	112	helping	79	choose	68	drove	52
telling	105	born	76	considered	67	depending	52
write	97	stopped	75	covered	62	trust	51
happening	94	notice	75	staying	62	picking	51

Table 9: Words occurring more than 50 times in the *After* Pivot sentence parts and not at all in the *Before* Pivot parts

Table 9 shows the actual words and counts for those words occurring over fifty times in the *after* corpus and not at all in the *before* part. Note that they are all forms of verbs.

4.3 Distribution of dysfluencies

Another major difference between the two parts is the kinds of dysfluencies that occur, as shown in Table 10. The *before* part has significantly more non-sentence elements (as described in §2.3)–three times as many. Most of these are conjunctions, which tend to start sentences. There are twice as many discourse dysfluencies, but if you look at just the use of "you know" the totals are nearly the same. There are also approximately the same number of filler words, such as "um" and "uh". A similar result was reported by Shriberg (1994) who showed that the rate of dysfluencies was much higher in sentence initial vs. sentence medial position.

4.3 Perplexity

Since the ultimate goal of the work here is to build a language model, one significant indicator of the uniformity of a corpus is to run perplexity experiments on the data. Perplexity gives a rough indication of the average number of alternatives in the grammar based on the computation of the entropy. This is assuming that the training and the test

	Before	After	NPC	NPI	Total
Total Non-Sentence Elements	89470	26352	9464	9775	57270
Discourse	17796	99445	1110	1646	11940
you know	6433	6538	215	602	
well/uh	7456	646	688	597	
like	1096	1272	53	68	
Conjunctions	46280	380	896	4771	18277
and	25518	209	355	2032	
but	10493	43	312	1300	
so/rb	5485	22	148	798	
Fillers	21693	14035	5787	2163	16548
Editing Terms	2806	534	97	206	1392
Asides	79	121	0	0	69
Restarts					
Begins	32335	13771	655	952	21614
Pivots	32102	13989	660	943	21608
Ends	25884	20160	653	942	21598
Total Words	-	-	-	-	-

Table 10: Distribution of Dysfluencies and Restarts

Training	No. words in training	Test Set		
		Full sentence	Before	After
Full Sentences	1.4M	70	48	226
Before	631K	–	**34**	702
After	693K	–	108	**125**

Table 11: Perplexity experiments on full sentences, *before* and *after* subparts

are well matched. Perplexity can be used to tell how similar or different two corpora are by training on one and testing on the other. It is actually this way of using perplexity that provides the information in this case, since we have not yet developed a complete model of the full sentence that takes into account the *given* and *new* portions.

We trained three bigram language models, one of full sentences, one on just the *before* parts and one on just the *after* parts. The goal here is to find whether this division yields more precise models that are a better fit to the data. Of course, because the data is split into *before* and *after* parts, those two models are less well trained than the full sentences. Nonetheless, the results are interesting and promising, as shown in Table 11.

Note that the lowest perplexity is on the match of the before pivot training and test and the highest perplexity is the *before* model tested on the *after* model. The *after* model itself is much more robust, with roughly equal perplexities when tested with *before* and *after*. The model trained on the full sentences has a much lower perplexity on the *before* parts and a much higher perplexity on the *after* parts. The best combined numbers comes from a match of *before* on *before* and *after* on *after* (in bold in Table 11). This reinforces our intuitions that these subparts of the corpus are quite distinct. Our next challenge is how to combine these models in a way that maximizes the separate models without being penalized by having to choose the pivot point.

5. Language Modeling for Recognition

The major challenge of this work is to take what we have learned from data analysis and perplexity experiments and make use of them in a language model for speech recognition. In Section 5.1 we describe a general approach which develops a conversational model of the different parts of a turn and how they interact. In Section 5.2, we describe some of the other issues and approaches in using this work to improve recognition performance.

5.1 Conversational models

A more general approach is to develop a higher level model of a conversations. We could model a turn with a finite state machine such as that shown in Figure 1. Each state is a model trained on the subpart of the corpus.

Using the same algorithm that divided the corpus for the analysis described above, we created a corpus of segment sequences as shown below. We first show the dialog and then the segment sequences, each segment corresponding to a box in the above FSA.

A1 Okay . E_S
I think the first thing they said , N_S
I have written this down E_S
{C so } it would , N_S
[is it p-, + do you think it 's possible] to have honesty in government or an honest government? E_S

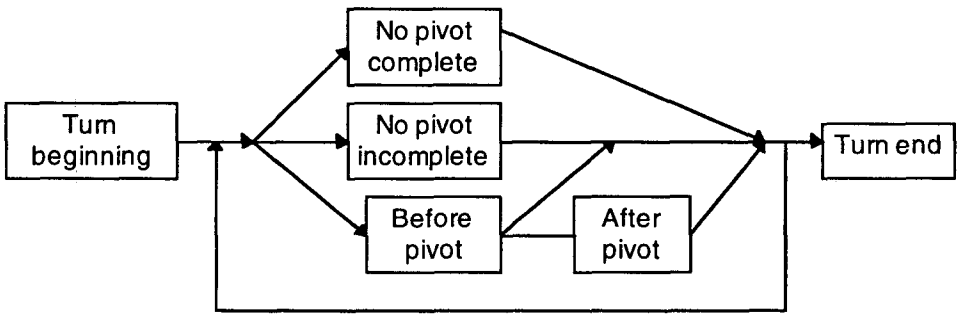

Figure 1: Finite state model of Conversations

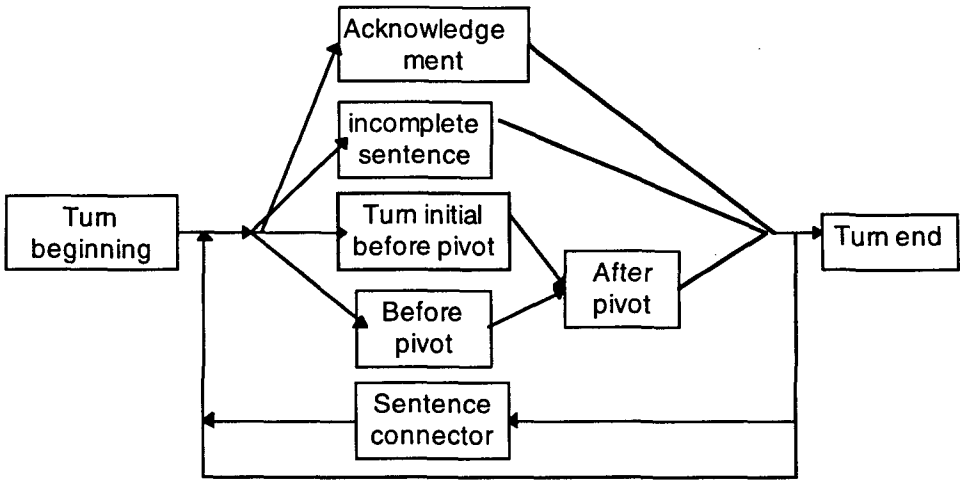

Figure 2: More complex conversational model

B2 Okay. E_S
You're asking what my opinion about, whether it's possible to have honesty in government. E_S
A3 Yeah. E_S

A1 (TB NPC BEF AFT BEF AFT NPI BEF AFT TE)

B2 (TB NPC BEF AFT TE)

A3 (TB NPC TE)

From this corpus, we can create a bigram language model of the transitions. Table 12 shows the transition probabilities between the states. It was trained on about a quarter of the overall data, containing about 133,000 tokens and a vocabulary size of 6 (TB, BEF, AFT, NPC, NPI, TE). As shown in the table below, turns tend to begin (TB) with either a *before* pivot segment (BEF) or an NPC (generally an acknowledgment or reply). *Before* pivots are virtually always followed by *after* segments (AFT). NPI (incomplete segments that were interrupted before the verb) are most likely to be followed by the end of a turn (TE).

	BEF	AFT	NPC	NPI	TE
TB	.41	.001	.44	.04	.11[7]
BEF	0	.999	0	0	0
AFT	.44	0	.01	.03	.51
NPC	.26	.02	.09	.02	.62
NPI	.33	.004	.03	.03	.60

Table 12: Bigram Transition Probabilities for Conversational Model

[7] A transition from TB (turn beginning) to TE (turn end) occurs when there was some non-speech, such as a laugh or cough, but no words in the turn.

The divisions used in the above model are fairly gross. We could develop a more precise model taking into account other differences in the turn. For example in the network shown in Figure 2, we distinguish between turn initial beginnings of sentences and others. We also separate out sentence connectors and optionally include them between sentences.

5.2 Future Work

This work can be extended to improve language modeling and thus recognition performance on the Switchboard corpus. However, to translate our work on the given/new information structure of conversational segments to language modeling, there are some important details. Some of these issues and possible first-cut approaches are as outlined below.

- *Representation of the split:* The split between the given and new parts of a segment can be represented as a lexical entity and treated as part of the data stream as given below. This is similar to the segment boundary representation in n-gram language models.

<s>Uh we were <end-given> <begin-new> thinking mini van for a while </s>

Another approach is to develop a smooth transition from a given model to a new model while computing the likelihood of a segment, as described above. The advantage of the second approach is that there will be more detailed context provided for the beginning of the *new* part of the segment and hence the n-gram estimates will be sharper.

- *Smoothing/Robust parameter estimation:* The given and new models will be trained on subsets of the training data. This fragmentation of the training data can lead to sparse data problems while estimating the language model parameters. We will need to explore robust parameter estimation techniques to smooth our model estimates. One approach would be to smooth the *given* and *new* models with a general "full sentence" model, using n-gram mixtures,

$$P(w_i \mid w_{i-1}) = \lambda_{M_k} P_{M_k}(w_i \mid w_{i-1}) + (1 - \lambda_{M_k}) P_{gen}(w_i \mid w_{i-1})$$

where the subscripts M_k represents *given* or *new* model and *gen* represents the general model trained on full sentences. λ's represent the interpolation weight and can be estimated on held-out data.

- *Hypothesizing the given/new split:* Currently, segments are split based on heuristic rules. We can foresee building upon this scheme during training by developing an iterative pivot-hypothesizing algorithm. In the first iteration, the training sentences can be split using the heuristic rules described earlier. *Given* and *new* language n-gram models are estimated using this data. In the subsequent iterations, the pivot is selected to maximize the likelihood of the sentences as estimated by the *given* and *new* language models. A similar scheme can be used for hypothesizing the pivot on the test data.

6. Conclusions

We have shown a path from annotation, through data analysis, to an implemented language model for speech recognition. Each part of this path is very important. The annotation provides a springboard for a wide range of different research efforts, essentially enabling the work that would be impossible without that effort. We used a simple automatic algorithm for dividing sentences, but based those divisions on both the text and the manual annotations. If we find this approach is a high payoff, then we may want to experiment with hand correcting some of the divisions to see if greater accuracy improves our results. However, it is important to show clear progress before investing the time, since manual annotation is very expensive. The linguistic analysis of the results can provide indications of the effectiveness of our algorithm and point to how best to use the results. In our work on creating new language models based on our analysis, we have only scratched the surface in ways of combining subcorpora and building an overall conversational model. We hope to complete recognition experiments in the next couple months that show the contribution of our generalizations over the text on word error rate.

References

Clark & Haviland (1977) Comprehension and the Given-new Contract" in Freedle (ed.) *Discourse Production and Comprehension*, Ablex Publishing Corporation, New Jersey. 1977.

Group on "Phrase-structure in Language Modeling", Language Modeling Workshop, Johns Hopkins, 1995.

M.A.K. Haliday and R. Hasan, *Cohesion in English*, Longman, London, 1976.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. "Switchboard: Telephone speech corpus for research and development". In *Proceedings of IEEE Conference on Acoustics, Speech, and Signal Processing*. vol. 1 pp. 517-520. San Francisco, March 1992.

R. Iyer, M. Ostendorf, R. Rohlicek, "An Improved Language Model Using a Mixture of Markov Components", Proceedings of the ARPA Workshop on Human Language Technology, pp.82-87, March 1994.

F. Jelinek, et. al 1995. Report on the LM95 Summer Language Modeling Workshop, to appear. See also http://cspjhu.ece.jhu.edu/lm95_workshop.html

S. M. Katz, "Estimation of Probabilities from Sparse Data for the LM Component of a Speech Recognizer", IEEE Transactions on Acoust., Speech, and Signal Proc., Vol. ASSP-35, Number 3, pp. 400-401, March 1987.

M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz and J. R. Rohlicek, "Integration of Diverse Recognition Methodologies Through Reevaluation of N-Best Sentence Hypotheses", Proc. ARPA Workshop on Speech and Natural Language, pp. 83-87, February 1991.

E. Shriberg, Preliminaries to a Theory of Speech Dysfluencies, Ph.D. thesis, Department of Psychology, University of California, Berkeley, CA, 1994.

A. Stolcke and E. Shriberg, "Statistical Language Modeling for Speech Dysfluencies" Proc. ICASSP-96, May 7-10, Atlanta, GA. 1996