WORD PREDICTABILITY AFTER HESITATIONS:
A CORPUS-BASED STUDY

Elizabeth Shriberg   *Andreas Stolcke*

Speech Technology and Research Laboratory
SRI International, Menlo Park, California, USA
ees@speech.sri.com, stolcke@speech.sri.com

ABSTRACT

We ask whether lexical hesitations in spontaneous speech tend to precede words that are difficult to predict. We define predictability in terms of both transition probability and entropy, in the context of an N-gram language model. Results show that transition probability is significantly lower at hesitation transitions, and that this is attributable to both the following word and the word history. In addition, results suggest that fluent transitions in sentences with a hesitation elsewhere are significantly more likely than transitions in fluent sentences to contain out-of-vocabulary words and novel word combinations. Such findings could be used to improve statistical language modeling for spontaneous-speech applications.

1. INTRODUCTION

Early psycholinguistic experiments found that human subjects asked to 'guess' next words in a transcription of spontaneous speech required more guesses for words that had been preceded by a hesitation [1]. These and other studies suggest a correlation between the location of hesitations and the predictability of following words. If such a correlation exists, it has potential implications for language modeling in speech applications.

In previous work we found that "cleaning up" disfluencies reduces perplexity [2]. However, if there is a correlation between hesitations and word predictability, we would lose information if we eliminate disfluencies completely. Thus, in this study we seek to determine whether such a correlation exists, but using a method based on corpus statistics instead of human judgments. We define predictability in two different ways:

1. **Transition probability**: the conditional probability of a word given its history

2. **Entropy**: the uncertainty in predicting the words following a history

Thus, transition probability is a function of both the history and the next word, whereas entropy depends only on the history. By definition, transition probability and entropy are inversely related over an entire sample; however, they are not mutually constrained at any

particular transition. We ask whether transition probability, entropy, or both measures are correlated with the location of hesitations.

2. METHOD

We examined speech from two large corpora of spontaneous speech: the ATIS corpus of human-computer dialog in the air travel planning domain [3], and the Switchboard corpus of human-human telephone conversations on prescribed topics [4]. Past work has shown that these corpora differ significantly in a number of dimensions [5], which lends generalizability to the present analyses. In both corpora, data were segmented into linguistic or "sentence" units.

The hesitations studied were (1) filled pauses ("uh" and "um"), and (2) disfluent repetitions (such as "on the on the"). These two types of disfluencies have the property of not changing the syntax of an utterance, and have similar distributional and acoustic characteristics, across different corpora [5]. All other disfluencies (e.g., false starts, self-repairs) were "cleaned up" (extraneous words removed from the text).

Word histories consisted of one or two words, corresponding to the bigram and trigram language models usually employed in speech recognition. Inclusion of both models allows us to assess whether the length of the history has any effect on results. Bigrams provide the advantage of more reliable probability estimates, while trigrams provide more resolution in modeling the probabilities.

Our hesitant transitions (HES transitions) were of two types: transitions containing filled pauses (FP), and transitions containing disfluent repetitions (REP), defined as follows:

HES transition: either a FP transition or a REP transition.

FP transition: a pair of words (w_k, w_{k+1}) occurring in the text as "w_k FP w_{k+1}" where FP (filled pause) is "uh" or "um". *Example: "the um billboard".*

REP transition: a pair of words (w_k, w_{k+1}) such that there is a sequence of words "$w_1 ... w_k\ w_1 ... w_k\ w_{k+1}$" where "$w_1...w_k$" represents a disfluent repetition. *Example: "on the on the billboard".*

Fluent transitions were also of two types, as defined below. We included the "FH" category (defined below) to test for a nonlocal

effect of hesitations on word predictability—i.e. to test whether the presence of a hesitation affected word predictability elsewhere in the sentence.

FLUENT transition: a pair of words (w_k, w_{k+1}) occurring in the text with no intervening words or hesitations.
Example: "the billboard".

FF transition: a fluent transition in a sentence containing no HES transitions.

FH transition: a fluent transition in a sentence containing one or more HES transitions.

Data from each corpus were divided into a training set and a test set. Training was based only on fluent transitions, that is, on the data after all disfluencies (other than hesitations) were "cleaned up".

We also eliminated from the analyses all 1- and 2-word sentences from the Switchboard set, as these were very frequent, and consisted almost exclusively of backchannels (e.g., "uh-huh") with no hesitations; they would thus have contributed disproportionately to the FF transitions condition.

The numbers of words and sentences in the training and test sets are shown in Table 1.

Table 1: Size of Training and Test Sets

	ATIS train	ATIS test	SWBD train	SWBD test
words	199,621	8511	363,426	778,487
sentences	20,718	827	39,726	68,908

The size of the test set was determined by the number of transitions containing a hesitation; we aimed to roughly equate the size of the fluent-transition test set to the size of the set of transitions containing a hesitation. ATIS has a much lower rate of hesitations than Switchboard; this explains the much smaller test set size for ATIS.

Because we were interested in the role of history, we eliminated from the analysis all transitions that had no or a reduced lexical history; these included sentence-initial hesitations for the bigram language model analysis, and hesitations with only one preceding word in the sentence for the trigram model analysis.

In addition, we wanted to remove any factors that could affect the comparison of transition probabilities across test conditions. One way to make data more homogeneous is to not include N-grams that are not represented in the training data. However, unseen N-grams are quite important to the current study since they represent very low-probability events. We therefore kept counts of the number of unseen N-grams for each condition.

3. RESULTS AND DISCUSSION

3.1. Transition Probability

We first searched for a difference between HES and FLUENT transitions by using transition probability, or the conditional probability of a word given its history:

$$p(w_{k+1}|h_k) = \frac{n(h_k w_{k+1})}{n(h_k)}$$

where h_k is the history preceding word w_{k+1}, and $n(.)$ denotes the observed count of a word sequence in the training set. For bigram analyses $h_k = w_k$; for trigram analyses $h_k = w_{k-1} w_k$.

We examined frequency histograms of transition probabilities for the four different test conditions, for each corpus and for both bigram and trigram language models. Results for the ATIS bigram analysis are shown in Figure 1. As shown, HES transitions had lower overall transition probabilities than either of the two types of FLUENT transitions. For ATIS, the HES transitions were comprised exclusively of FP transitions, since there were not enough non-initial REP transitions to merit analysis in this corpus.
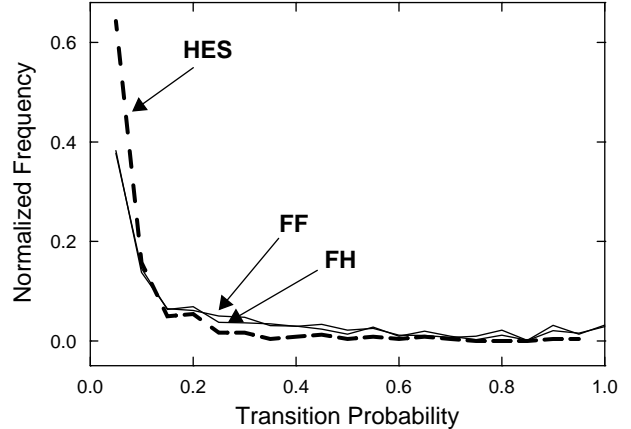
Figure 1: Transition probability distributions for FLUENT (FF and FH) and HES transitions in ATIS bigram analysis. Frequencies are normalized by the total number of transitions.

To test whether HES transitions had significantly lower probabilities than FLUENT transitions, we conducted a Kruskal-Wallis analysis of variance by ranks. Results showed that the FF and FH conditions did not differ significantly from each other, but each had significantly higher probabilities than the HES condition.

We found the same pattern of results for the Switchboard data. For Switchboard, we had in addition the REP condition; results showed no significant difference between FP and REP transitions for this corpus. Furthermore, the same pattern of results was obtained in the trigram version of the analyses, for both corpora. This suggests that the difference between HES and FLUENT transitions is robust to the trade-off between accuracy of probability estimates (bigram) and resolution in modeling conditional word probabilities (tri-

gram). For reasons of simplicity and lack of space we discuss only the bigram results in all further analyses.

3.2. Effect of Following Word

We then asked whether the association between transition probability and HES transitions was attributable to an effect of the following word (w_{k+1}). We investigated the effect of the following word by controlling for entropy. Since it is difficult to control for entropy directly, we chose to equate the distribution of word histories across test conditions. Since entropies are a function of the history, this produces the desired control.

To produce a sample of transition probabilities for which the distribution of histories was identical, we used the subset of transitions with histories that occurred in all three test conditions. We then randomly resampled N-gram tokens from each condition to produce identical numbers of samples across conditions for each history. Results for the bigram analyses are shown in Figure 2. Statistical analysis was conducted in the same manner as in the previous analysis.
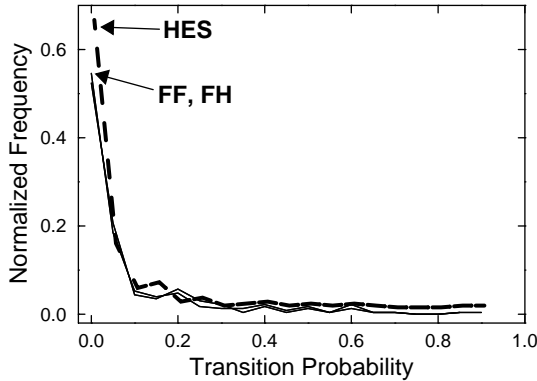
Figure 2: Transition probability distributions for FLUENT and HES transitions in ATIS bigram analysis, using subset of transitions equated for follow-word entropy across conditions.

As expected, the equating-for-history procedure resulted in a smaller overall difference between HES and FLUENT transitions. However, the pattern of results was the same as found previously, with a significant difference between the HES and FLUENT transitions. This result suggests that entropy alone cannot account for the association between HES events and less-predictable word transitions; there is clearly some effect attributable to the identity of the following word.

3.3. Effect of Entropy

As shown in Section 3.2, word history alone could not account for the difference in transition probabilities between the HES and FLUENT conditions. Here we ask whether history plays a role at all in characterizing this difference.

First, we need to make the notion of entropy precise. In information theory, the entropy of a distribution denotes the average information conveyed by a sample from the distribution, and is defined

as the expected negative log probability of an event drawn from the distribution: $H(p) = -\sum_{x} p(x) \log p(x)$. In our case, the distribution in question is given by the conditional probabilities of all words following a certain history. We therefore seek to compute the correlation between the probability that a history will be followed by a hesitation, and that history's follow-word entropy.

Since entropies of empirical distributions are inherently higher for larger samples, i.e., for word histories that occur more often, we equated the sample sizes for all histories obtained on the training corpus. We chose a minimum number of occurrences, eliminated all histories occurring less often than the minimum, and downsampled the training corpus occurrences of the remaining histories to the chosen minimum count. Each history's entropy was then computed by using the relative word frequencies of the downsampled data.

We applied the method described above to bigram histories in the Switchboard corpus. Results are shown in Table 2, where RS denotes the Rank-Spearman correlation coefficient, a measure appropriate for relating values on different scales. We show the correlation obtained for a range of minimum counts, since there is an inherent trade-off between the accuracy of the entropy estimates and the number of histories included in the sample.

Table 2: Correlation Between Entropy and Probability of HES

Min. History Count	No. Histories	RS	p
20	1222	0.13	<0.0001
50	628	0.22	<0.0001
100	377	0.16	<0.0001

As shown, there was a weak, but highly significant positive correlation between likelihood of a hesitation and entropy after a history. The exact value varies with the minimum number of histories used in the sample. These results suggest that while entropy alone cannot account for the lower transition probabilities after HES events (Section 3.2), entropy does contribute to the overall effect.

According to Bayes law, however, the probability of a word, given that a HES follows, is proportional to both (1) the probability of a HES following the word, $p(\text{HES} \mid \text{word})$, and (2) the overall probability of the word, $p(\text{word})$. We showed that (1) is positively correlated with entropy. However, (2) is also typically correlated with entropy in human language, because there tends to be a wide choice of words following high-frequency words (we verified this for our corpora in a separate study). Both (1) and (2) will therefore contribute to the overall correlation between HES locations and high entropies, and we cannot from these results alone determine the relative contribution of each.

3.4. Rate and Location of Unseen N-grams

An interesting result was discovered when we examined the counts for unseen N-grams, i.e., N-grams in the test conditions that were not present in the training data. As mentioned earlier, these N-grams were not included in the comparison of transition probabili-

ties, but rather simply tallied for each test condition. We distinguished four different cases. **OOV1** indicates that the word history is an out-of-vocabulary word; **OOV2** indicates that the following word is out of vocabulary; **OOV12** indicates that both the history and the following word are out of vocabulary; and **Novel Combination** indicates that both the history and the following word are in vocabulary but were never present in sequence.

As expected, the HES condition showed higher rates of unseen N-grams than either of the FLUENT conditions. The result of interest, however, was that FH transitions were significantly more likely to involve unseen N-grams (of all types) than FF transitions. Relative counts for the bigram analyses are shown in Table 3. We show only the set of FH transitions that did not share any words with HES transitions, since the result would be uninteresting if the increase in unseen N-grams in the FH condition were attributable to the same words responsible for low transition probabilities in the HES conditions. This finding held for both corpora, and was statistically significant in a Chi-square test.

Table 3: Percentage of Unseen N-grams in Fluent Transitions

	OOV1	OOV2	OOV12	Novel Comb.
ATIS, FF	0.24	0.22	no data	2.94
ATIS, FH	0.34	0.26	no data	5.66
SWBD, FF	1.11	1.48	0.05	17.79
SWBD, FH	1.62	1.86	0.07	20.74

Further analysis revealed the relevance of location in characterizing the increase in unseen sequences. Figure 3 shows the rate of unseen sequences for FH transitions before and after the first HES in a sentence for Switchboard. Since magnitudes of the different unseen N-gram types vary, we have normalized each rate by the rate of unseen N-grams of the particular type found in the FF condition (fluent transitions in fluent sentences). As shown, the increased rate for all types of unseen N-grams occurred after the first hesitation in a sentence.
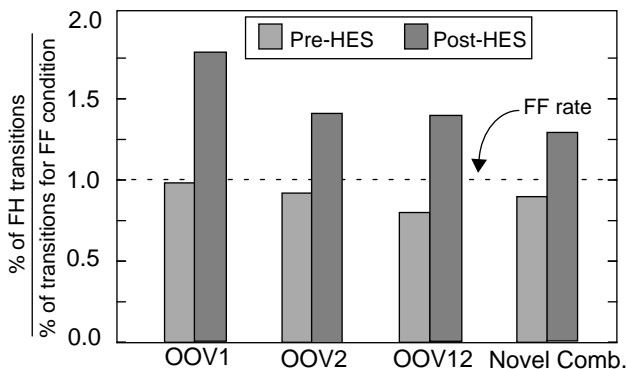
Figure 3: Rate of unseen FH N-grams by type, before/after first HES in sentence. Values are normalized by rate in FF condition.

These results suggest a nonlocal correlation between HES events and unseen N-grams. We cannot, however, infer a direct causal effect of HES. For example, results from other studies [6] show that average transition probability for words early in a sentence is higher than that for later words, presumably because earlier words convey "given" as opposed to "new" information. This is consistent with our finding that unseen N-gram rates are higher after the first HES in a sentence, since words after HES transitions inherently occur later in the sentence.

The increase in unseen N-grams in the FH condition, however, is unexpected given earlier results. We found in Section 3.1 that transition probabilities for the FF and FH conditions were statistically indistinguishable. Yet here we found a significantly higher rate of unseen N-grams in the FH condition. These results must be reconciled in future work, since we typically expect a correlation between the rate of unseen N-grams and transition probabilities.

4. SUMMARY AND CONCLUSIONS

We found both local and nonlocal correlations between HES events and word predictability in spontaneous speech. HES transitions had lower transition probability than FLUENT transitions, and both the following word and the word history were found to contribute to this difference. We also found that FLUENT transitions in hesitant sentences were more likely than FLUENT transitions in fluent sentences to contain unseen N-grams. For language modeling, we conclude that the simple cleanup model proposed in [2] needs to be extended. Specifically, the presence of a hesitation should increase the probability of unlikely following words, and decrease the probability of likely following words. Furthermore, the nonlocal effects found suggest that hesitations should modify the language model beyond the immediate vicinity of the hesitation. For example, the probability of unseen N-grams in a sentence should be increased after detection of a hesitation. Future work should investigate the extent to which modeling these correlations, (as well as others which may obtain for alternate hesitation forms) can improve performance in spontaneous-speech applications.

5. References

1. Goldman-Eisler, F. Psycholinguistics: *Experiments in Spontaneous Speech*. New York: Academic Press, 1968.

2. Stolcke, A. and Shriberg, E.E. "Statistical language modeling for speech disfluencies," *Proc. ICASSP*, 405-408, 1996.

3. MADCOW. "Multi-site data collection for a spoken language corpus," *Proc. Fifth DARPA Speech and Natural Language Workshop*, 7-14, Morgan Kaufmann, 1992.

4. Godfrey, J.J., Holliman, E.C. and McDaniel, J. "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. ICASSP*, 517-520, 1992.

5. Shriberg, E.E. Preliminaries to a theory of speech disfluencies. Unpublished Ph.D. thesis, University of California at Berkeley, 1994.

6. Meteer, M. and Iyer, R. "Modeling conversational speech for speech recognition," *Proc. of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, 1996.