HOW FAR DO SPEAKERS BACK UP IN REPAIRS? A QUANTITATIVE MODEL

Elizabeth Shriberg *Andreas Stolcke*

SRI International, Menlo Park, CA 94025
http://www.speech.sri.com

ABSTRACT

Speakers frequently retrace one or more words when continuing after a break in fluency. Syntactic principles constrain the points from which speakers retrace; however syntactic principles do not provide predictions about the relative usage of different allowable retrace points. Such predictions are useful for automatic processing of repairs in speech technology, particularly if they use information readily available to a speech recognizer. We propose a quantitative model that predicts the overall distribution of retrace lengths in a large corpus of spontaneous speech, based only on word position. The model has two components: (1) a constant, position-independent probability for extending a retrace by one more word; and (2) a position-dependent probability to "skip" to the beginning of the sentence. Results have implications for modeling repairs in speech applications and constrain explanatory models in psycholinguistics.

1. INTRODUCTION

When speakers resume after a disfluency, they often retrace back one or more words before continuing, producing simple repetitions as well as repeated words in repairs. A question important to modeling repairs in both psycholinguistics and in speech technology is: *when speakers retrace, what predicts how far back they go?* Previous accounts of retracing in linguistics and related fields have illuminated syntactic constraints on retracing—namely that speakers retrace to points that correspond to the onsets of syntactic phrase boundaries, and which can produce a well-formed syntactic coordination between the original utterance and the continuation [3]. The syntactic phrasing accounts match native speaker judgments about what constitutes a "bad" retrace point. However, they do not predict which of many possible remaining retrace points are chosen. In English as well as other right-branching languages, many locations in an utterance correspond to the onsets of constituents, and a large subset of these correspond to points that produce a well-formed coordination—including retracing back to the sentence onset. For example in the following case, all possible previous words constitute viable retrace points:

At the end of the road – ((((at) the) end) of) the block

Our goal in this study was to explore whether overall corpus statistics on the length of retracings could be predicted using information readily available to a speech recognizer. We focus on word position for this purpose, since retracing is inherently constrained by this factor and since information on words can be easily modeled in speech systems.

2. METHOD

2.1. Data

Data consisted of transcripts from the Switchboard corpus of human-human dialog over the telephone [2], distributed by the Linguistics Data Consortium (LDC). We used a subset of 1115 conversations (roughly 1.4M words, 350 different speakers) that had been marked for sentence boundaries and for disfluencies by the LDC, as described in [5]. Word correspondences within disfluencies were not marked, but retraced words could be detected automatically with high accuracy by aligning reparandum and repair regions via dynamic programming. We recorded all instances of simple repeats, as well as cases of retracing before changed words in other repairs. This resulted in a set of 30,524 disfluencies containing one or more retraced words.

2.2. Measures

We characterize retracing in terms of two measures, the number of retraced words (retracing length, k) and the position relative to the start of the utterance at which the retraced word sequence ends, (retracing position, m) as illustrated in Figure 1.

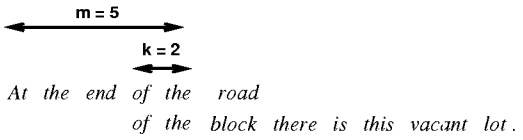

Figure 1: Word-Based Measures

In conducting the analyses, we found two general principles to be true empirically:

(1) Retrace points do not occur within words.
(2) Retracing does not cross sentence boundaries.

We found no cases of retracings involving the second word in contractions. For examples, cases like "I'll – 'll " did not occur.

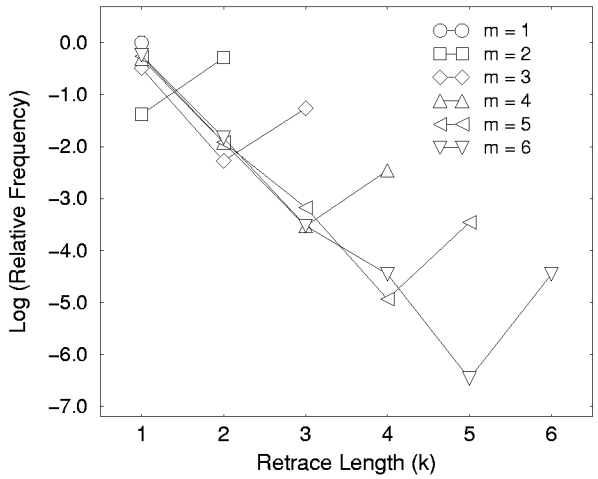

Figure 2: Distribution of Retrace Lengths by Position

This is not surprising given the phonotactic and prosodic characteristics of contracted forms. We found empirically that results overall are considerably cleaner for the analyses herein, if contractions are counted as a single word (while their noncontracted counterparts are counted as two). We also found that speakers did not back up across sentence boundaries in retracing, and any made-up examples sound quite odd. This constraint is predicted by the coordination rule [3], but notably not simply by a phrase-onset constraint without some notion of maximal projection. Because speakers do not back up across sentence boundaries, the maximum value of k for any particular token is equal to m.

3. A QUANTITATIVE MODEL

Since m determines the maximum retrace length k for each token, we plotted our tokens separately by m values. Figure 2 shows the relative frequency of each retrace length k in our corpus, in the log frequency domain. Lines connect points for each m value, and since the measure on the ordinate is relative frequency, points for each m sum to 1.0. Only data for $m = 1,\ldots,6$ are shown, but higher m values show the same pattern. The case of $m = 1$ is degenerate; it includes only one point, at 0 on the log plot (since all retracings from a retrace point that is one word into an utterance must be one word long).

A first observation is that for each m value, there is an apparent exponential decay in frequency with increasing values of k. The only exception is that the final "hooks" for each trend ($k = m$) are consistently higher than expected based on the previous points. We will come back the question of final points later. Ignoring the final hooks for the moment, it is also apparent that the lines for different m values are parallel, suggesting that the rate of decay in retrace length is about the same, regardless of where in the sentence the retrace was started from.

3.1. Retrace length decay rate

The simplest model we can propose for the decay rate for $k < m$ is the following: assume that at any position m, there is a uniform probability, p, which corresponds to the probability that the speaker will retrace back one additional word. Under this simple assumption we can model the probability of observing a retrace-length of exactly k words by a geometric distribution:

$$\text{Prob}(k) = (1-p)p^{k-1}$$

Since the world of cases we are considering all contain at least one retraced word, the first time this probability is applied is after the first word has been retraced; thus the exponent is $k-1$. We can think of p as the "pass" probability, or the probability that the speaker will retrace back one more word. Note that p is independent of how many words were retraced already (lines for each m are straight in the log plot); and p is independent of where in the sentence retracing started (lines for different m values have the same slope).

If p is the probability of retracing one more word, then $1-p$ can be thought of as a "stop" probability, or the probability that the speaker will stop adding words to the retraced string (i.e., that he will back up no further than the current location). The resulting "stopping" points are locations that speakers choose as the left edge of their retraced string. Although we do not know what they correspond to in this study, they are the same points that other accounts have explained as onsets of syntactic or prosodic phrases. Thus we may want to view the model in terms of some type of phrase rather than as probabilities of stopping applied at each word. We will define a construct, r–phrase (retracing-model-phrase), for this purpose; an r–phrase is the shortest distance between two stopping points. Note that the r–phrases so defined differ from syntactic phrases because r–phrases are purely concatenative; the end of a previous r–phrase is the beginning of the next, so there is no nesting or overlap of r–phrases. Since under the present model, we would have an r–phrase boundary between any two words with a probability of $1-p$, the average length of an r–phrase is $\frac{1}{1-p}$ words.

3.2. Skip parameter

We next focus on accounting for the final hooks. One possible explanation is that they result from applying the geometric distribution to the finite range of observed k values, accumulating the excess probability from the tail of the distribution at $k = m$. To explore whether the truncation explanation could account for the observed final hooks, we plotted predicted versus actual values of k individually for each position. Predicted values were computed by tying p to a single value for all positions, obtaining the values for each k from the geometric distribution determined by p, and then for each position, subtracting the total probability over all k from 1 to obtain the probability mass associated with tail of the distribution for that position. This excess mass was then added to the value at $k = m$, producing the quantity p^m. Results showed that the predicted values at $k = m$ consistently underestimated actual values, even after adding the excess probability mass. Therefore the truncation explanation alone cannot account for the hook: people skip back to the start of the sentence more often than can be predicted by a truncated distribution.

To account for this extra "skip" probability, we introduced an additional parameter, p_s. This factor expresses the additional tendency of speakers to retrace back to sentence beginning—after accounting for the probability that they will end up there by suc-

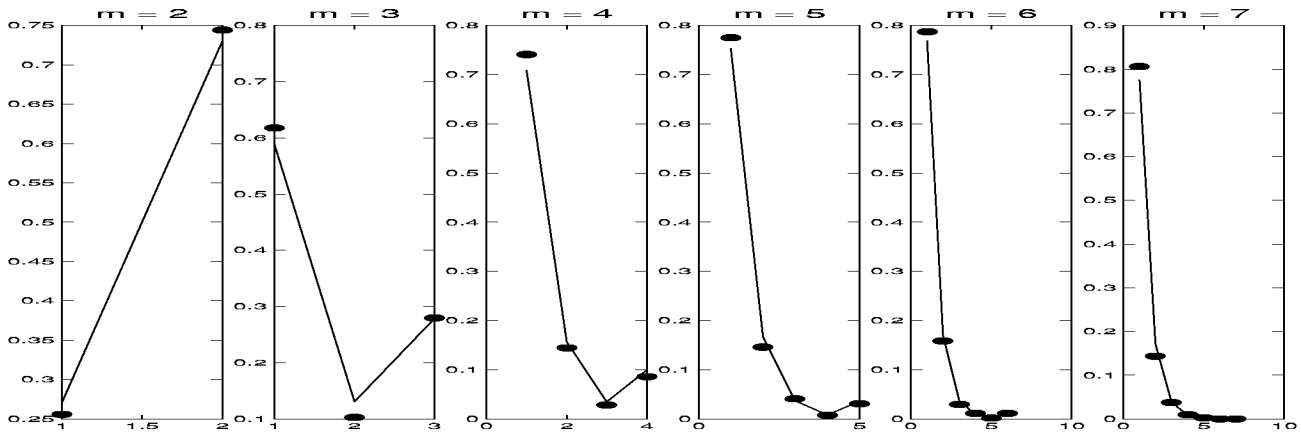

Figure 3: Fits for the frequency of retrace length by position for $m = 2$ through 7, based on the model in equations 1 and 2. Parameters were optimized for all m values simultaneously. Frequencies are plotted in the linear domain; points indicate observed values.

cessively retracing back one word at a time. The overall model is therefore expressed in two parts, one for $k < m$ and the other for $k = m$:

$$\text{Prob}(k) = \begin{cases} (1-p)p^{k-1}(1-p_s) & \text{for } k < m \\ p^m(1-p_s) + p_s & \text{for } k = m \end{cases} \quad (1)$$

When this model was fit with a single value for p_s over all positions however, fitted values deviated systematically from the observed data. Predictions undershot the final hooks at low values of m, and overshot the hooks at high values of m. Fits for p_s individually for each value of m revealed that the skip factor depends strongly on position. For example, when a speaker is only two words into a sentence, the *added* probability of skipping back to the start—after accounting for the probability from the overall decay rate model—is nearly 70%. However by the time the speaker is four words into an utterance, the skip factor has dropped to under 5%. By plotting points for the range of position values, we found that the relationship between position and skip factor is very well fit by an exponential decay:

$$p_s = Ae^{-Bm} \quad (2)$$

where A is a scaling factor that controls the overall rate of skipping, and B is the decay in skip rate associated with position. We can now propose a unified model for retracing length. The overall model, which accounts for all positions m, has three parameters—one to describe the retracing length decay rate and two to describe the probability of skipping to the beginning of the sentence:

p	probability of retracing one more word
A	p_s parameter, controls overall rate of skipping
B	p_s parameter, controls decay of skip rate with m

Using the model in equation 2, we solved simultaneously for the values of the three parameters that minimized the squared prediction errors in the linear domain (results are similar for minimization of prediction error in the log domain). Optimizations were performed using a single value for each parameter for the set of all m values. The resulting fits are shown in Figure 3, in linear rather than log frequency to provide an idea of actual values. As can be seen, the model provides a close fit to the observed data. Note that the upward trend for $m = 2$ is attributable to the final hook, which as explained earlier has a high value for low values of m.

3.3. Part-of-speech simulation experiments

Our model fits the overall distribution well, but this does not mean that it makes the right predictions in any particular sentence. The simplest case for which our model is a good fit may not necessarily be the correct one. Such a case corresponds to no effect of any factors other than position on retracing. We tested this null hypothesis by conducting the following Monte Carlo experiments.

We took the set of actually occurring retracings and divided them into subsets, conditioned on both m and sentence length (since m itself is constrained by sentence length). For each retracing, we recorded the part of speech (POS)[1] of the first word in the retraced string that the speaker actually produced. Next, we produced for each sentence a *simulated retracing* at the indicated position m by randomly drawing a retrace length from the empirical distribution of k values associated with the m value, broken down by sentence length. Note that this method uses the observed data (not our model predictions) to obtain the simulated k value for each token, and thus avoids potential circularity. Also importantly, this method controls for the actual trouble points, since we change only the retrace length in each case. If the only factor governing retrace lengths is position, we expect that the POS distribution for the simulation should roughly match that for the real retracings.

Results, which could not be fully presented here due to space constraints, show that for the majority of POS types the simulation produced a close match to the empirical values. However the null hypothesis notably underpredicted values for prepositions, which were the most frequent location to which speakers retrace. The model also had a high value for these cases, but not quite

[1] We used POS information from the Penn Treebank database of hand-corrected, machine-assisted natural language annotations [4].

high enough—indicating that speakers appear to have some type of preference for retracing to preposition boundaries. The model also overpredicted rates for verbs; speakers appear to have a bias against restarting from the onset of a verb phrase that is not accounted for by position effects alone.

For the remaining POSs, fits were close to observed values; thus it is conceivable that a position model alone could explain much of the POS distribution for retracing. Determiners are a case in point. It is often noted that speakers retrace to determiner boundaries, and indeed speakers did retrace more often to determiners than to many other POS types. However the simulation produced about the same rate for determiners as was found in the actual data, indicating that the prevalence of determiners could be simply attributed to a correlation with the location of m (for which we controlled the simulation, see above)—rather than a tendency to retrace back to this particular POS. A likely explanation is simply that speakers tend to stop before noun phrases [1]. Further detailed study is needed to understand the relationship between position and syntax in retracing, but these results suggest that both positional and syntactic effects play important roles.

4. DISCUSSION

We found that the probability that a speaker retraces back one more word is uniform; it does not depend on how far they are into an utterance, nor on how many words they have already retraced. Based on recent work on retracing in simple repeats [1], one possible interpretation is that speakers retrace to the onset of the constituent that they are having trouble formulating. Under this view, a longer retracing could indicate trouble beginning further back, on the larger constituent. What remains to be explained however, is why there is a uniform relationship between the probability of encountering trouble N words back and encountering it $N+1$ words back, for the overall data set. Another possible explanation, not mutually exclusive, is that the exponential decay for retracing one additional word is associated with a *temporal* factor: the more time the speaker needs before continuing, the further back the speaker will retrace, since the extra words buy more time (if the speaker is optimizing speaking rather than silent time). Such a factor should not be ruled out, since hesitation pauses also show an exponential decay.

The fitted value for p in the model came out to 0.22, corresponding to an average r–phrase length of 1.3 words. Thus it predicts that in Switchboard there will be acceptable points available for initiating a retraced string roughly every 1.3 words. An interesting question is whether this value is close to the value one would obtain based on the distribution of syntactic phrase boundaries.

The parameter values that determine the skip probability are not directly interpretable. However, it is noteworthy that two values are needed. Since A adjusts the overall scaling of the rate and is independent of position, one reasonable hypothesis is that it represents a stylistic factor associated with certain speakers or registers. The parameter B, on the other hand, *is* tied to position, and therefore may reflect cognitive aspects involved in sentence processing over time. It is also possible that B reflects some qualitative variable associated with position, for example, given versus new information. Since given information tends to occur early in Switchboard sentences, perhaps speakers favor full retracings in these locations because they prefer to retrace given information. Further analysis of the word content could examine this hypothesis. It seems somewhat unlikely, however, that a qualitative factor such as given versus new information would show up as a term in the exponent; a processing explanation seems more consistent with a constant decay applying at each word.

5. CONCLUSION

We found that the overall distribution of retrace lengths in a large corpus can be fit well by a model that uses only information on word position. The model has two components: a constant, position-independent probability for extending a retrace by one more word (p), and a position-dependent probability of skipping back to the start of the sentence (p_s). The skip parameter decays exponentially with position, requiring two added parameters.

For automatic speech processing, since the model uses only word information, results can be directly applied to aid automatic repair detection and correction when syntactic information is unavailable. If there is good evidence for at least one matched word, but the retrace length is in question, then overall probabilities for the retrace lengths can be applied in evaluating hypotheses. If the sentence beginning is known or hypothesized with high confidence, then the skip probabilities can also be taken into account when evaluating hypotheses for different retrace lengths.

For psycholinguistics, it is clear that a generative or explanatory model at the level of individual sentences requires a more detailed linguistic investigation. As was found in the simulation experiment, the position model alone is not sufficient for predicting the probabilities of retrace lengths for any particular sentence; some information on POS or other variables correlated with grammatical phrasing is also needed. Nevertheless, results provide a constraint on explanatory models—they must be able to reproduce the same overall distributions that are effectively estimated by the simple model proposed here.

Acknowledgments

This research was supported by NSF Grants IRI-9314967 and IRI-9619921. The views herein are those of the authors and should not be interpreted as representing official policies of the funding agency.

6. REFERENCES

1. H. Clark and T. Wasow. Repeating words in spontaneous speech. *Cognitive Psychology*, 1998. To appear.
2. J. J. Godfrey, E. C. Holliman, and McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 517–520, San Francisco, 1992.
3. W. J. M. Levelt. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, 1989.
4. M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
5. M. Meteer et al. Dysfluency annotation stylebook for the Switchboard corpus. Distributed by the Linguistic Data Consortium, 1995.