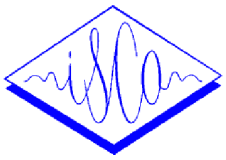

Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

Predicting Intrasentential Pauses: Is Syntactic Structure Useful?

Joshua Tauberer

Department of Linguistics
University of Pennsylvania
tauberer@ling.upenn.edu

Abstract

Syntactic structure plays a role in prosodic phrasing, which in turn guides where speakers may pause during a sentence. To what extent does syntactic structure predict the locations and durations of unfilled, intrasentential pauses in conversational speech? Here we attempt to answer that question by using the word-aligned and syntactically annotated Switchboard corpus [3]. Automated binary classification revealed that while syntactic structure helps above and beyond lexical features such as POS, knowledge of constituent node labels was less useful than knowledge of the starts and ends of constituents. On the other hand, some POS and constituent labels were more predictive than others in predicting word boundary durations. Some labels indicated the presence of a disfluency; others picked out syntactic contexts such as between a noun phrase and a prepositional phrase, which was predictive of a lack of or shorter pause.

10.21437/SpeechProsody.2008-89

1. Introduction

Intrasentential unfilled pauses occur frequently in spontaneous speech, resulting from processing limitations, prosodic structure, and other factors. Understanding where and why pauses occur, and providing a computational model for predicting them, can be useful both in speech recognition applications, which must distinguish a pause within an utterance from silence indicating the completion of an utterance, and in speech synthesis, in which inserting appropriate pauses may improve the comprehensibility and perceived fluency of the speech system. The present paper investigates whether pauses between words within a sentence could be predicted on the basis of lexical-level features and syntactic structure.

This problem has been treated in the past as a binary classification problem, where each word boundary is to be labeled as a pause or a no-pause, based on some features extracted from the surrounding context. [1] used a 6-gram HMM model over part-of-speech (POS) tags with a corpus of BBC Radio speech and achieved 68% recall of pauses at word boundaries with 89% of word boundaries overall classified correctly. (A simple baseline is choosing the most frequent category at each word boundary, and in this case choosing no-pause everywhere would have yielded 80% accuracy.) [7] compared two feature sets: the first included time and various features at the level of the syllables surrounding word boundaries; the second included part-of-speech tags surrounding the word boundaries. Pauses were predicted at roughly the same accuracy by the two models using conditional random fields, but below baseline accuracy.

A connection between intrasentential, unfilled pauses and syntactic structure has been investigated at least since [9], an early work in corpus phonetics. They found that nearly half of all filled and unfilled pauses ocurred within and not at the boundaries of what they considered phrases, which were two-to-four-word NPs and PPs. [4] recorded slow speech and found that the duration of (unfilled) pauses relative to the other pauses in a sentence were highly correlated with a 'complexity index' assigned to each word boundary, which reflected roughly the sizes of the constituents around the boundary. But, they noted that pauses that deviated from the complexity index often occurred in a context where the constituents on either side were of unequal length — in these cases, the pause would occur *within* the larger constituent. Experimentally, [2] found that pauses before clausal complements were fewer and shorter than before clausal modifiers, attributing the difference to processing and complexity.

Syntactic features have also more recently been implicated by means of machine learning as predictors of pauses. [8] compared the feature set of mostly lexical-level features reported in [5] — included POS labels, time from the start of the sentence, the identity of any punctuation at the word boundary, and two basic syntactic features for the position of word boundaries within NPs — to a new syntax-based feature set involving preceding and following constituent sizes and whether the preceding or following phrase was a 'major' phrase or 'SBAR' phrase. The word-level feature set predicted pauses much better than the syntactic feature set (f-scores of 83.0% and 70.8%, respectively[1]), but combined (84.8%) showed a small improvement over lexical-level features alone. The improvement indicates either constituent sizes or labels are predictive of pauses.

The present study sought to determine the extent to which syntactic structure such as constituent size (as above), category labels, and structural relations could each improve a model of the placement of these pauses above the lexical-level features considered previously, such as POS tags. A decision tree classifier and multiple regression were used.

2. Methods

2.1. Corpus

Word-aligned transcripts [6] from the Switchboard Telephone Speech Corpus [3] were combined with Penn Treebank Release 3's [10] syntactic annotations of 650 Switchboard conversations. The word-aligned transcripts marked regions of silence on each channel, which were considered 'pauses' here. Because the syntactic annotations omitted all information about the location of silences in the channels, it was necessary to combine the transcripts with the syntactic parses.

I would like to thank Jiahong Yuan, Tony Kroch, and Catherine Lai for their helpful comments.

[1] Note that these measures cannot be easily compared across studies because of the different relative frequencies of pauses in different corpora, among other factors.

Merging the two annotation sources required special processing. The annotations of the two corpora disagreed significantly on many levels, resulting both from annotation error and different annotating conventions, and as a result the list of word tokens from the transcripts and that from the leaf nodes of the syntactic trees did not match up straightforwardly. An automated method of merging the timestamps, silence annotations, and syntactic structure from the two corpora was developed.

The merged corpus contained 110,504 utterances (sentences, fragments, and interjections). There were 10.6 non-pause word boundaries for every pause boundary, which was relatively high compared to other corpora (8.7:1 in [7], and 4.1:1 in [1]'s corpus of radio speech).

2.2. Causes of pauses

Pauses are the result of a number of independent factors, relating to prosodic structure, production mistakes, processing limitations, etc. In Switchboard, pauses were also the result of conversational effects, such as when a speaker is interrupted by the other conversant, pauses for a moment to hear what was said, but retains the floor and continues with the sentence. A sample of pauses was (impressionistically) classified into one of five categories to determine the proportion of pauses attributable to something outside of grammar competence. This is reported in Table 1. Besides 'interruption', as described above, 'parenthetical' refers to pauses surrounding phrases including "you know", 'uncertainty' refers to fairly clear cases where the speaker was mentally searching for what to say next (some instances of which were searching for a numeric magnitude of something) but involved no disfluency that could have been inferred from the transcript, 'disfluency' refers to pauses adjacent to restarts, and 'other' refers to all other pauses, which could be related to intonational structure or any other covert factor.

It is unclear to what degree the locations of interruptions and uncertainty-type pauses are influenced by the structure of the utterance in which they appear. To the extent that they have a free distribution, they would be mere noise for the present analyses.

Table 1: *Types of Pauses (N=84)*

Type	Percent
Interruption	11
Parenthetical	13
Uncertainty	13
Disfluency	17
Other	46

2.3. Features

At each word boundary, a set of features were recorded, both categorical and continuous. The 'left' and 'right' words below refer to the word preceding and following the boundary. The lexical-level features investigated were:

- The time of the end of the left word from the start of the sentence, in seconds and relative to the length of the sentence.
- The part of speech of the left word, the right word, and a combined feature of both.

The syntactic features were:

- The label of the largest constituent ending at the boundary, similarly for beginning at the boundary, and a combined feature of both. At any word boundary any number of constituents may be ending (or beginning). For instance, in Figure 1, in which asterisks denote pauses, at the word boundary containing the first pause five constituents are ending: a NN, NP, VP, S-NOM, and a PP. In this case, the word boundary is categorized as following the PP only, the largest/highest constituent.
- The size of the preceding and following largest constituent, in terms of number of words and duration in seconds.
- The depth of the boundary, measured as the larger of the number of close-brackets or open-brackets at the word boundary.
- A feature encoding the syntactic relation between the surrounding constituents, collapsing a few configurations into head-complement, head-adjunct, specifier/adjunct-head, specifier-phrase, adjunct-phrase, phrase-adjunct, or other structure. The structural relations were determined based on 12 hand-crafted rules. Word boundaries with an interjection or disfluency on either side were classified in a separate category.

Also included were:

- The conditional log-probability of the right word given the left word, based on bigram frequencies from the entire merged corpus.
- The pointwise mutual information of the bigram, $\log_2$(P(bigram)/(P(left)*P(right))), a symmetric measure of association between the two words. The measure increases as the association between the words increases.

Various combinations of the features above were used with the C4.5 [11] decision tree binary classifier to assess precision and recall on the presence or absence of a pause at each word boundary. The classifier was used in two ways. In the first way, it was trained on roughly 150,000 word boundary examples from around 35,000 sentences and tested on a smaller independent set, drawn from a random sample of utterances from the merged corpus.

In the second way the classifier was used, no-pause word boundaries were thrown out from the training and testing sets to achieve a 50/50 balance between pause and no-pause word boundaries. The training set was roughly 25,000 word boundaries, and the testing set 6,000.

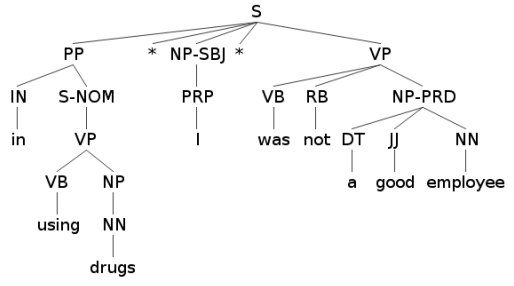

Figure 1: *A sentence from the PTB/Switchboard corpus with silence tokens added, denoted by asterisks.*

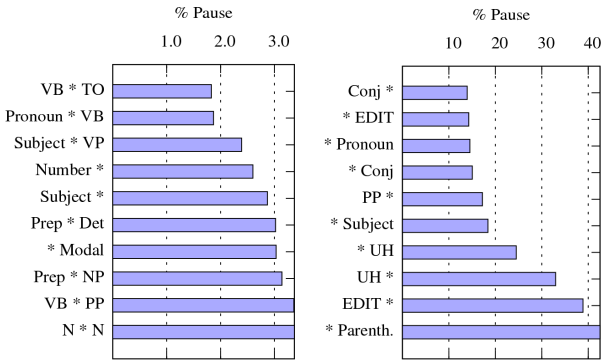

Figure 2: *The ten POS labels and syntactic categories with the fewest occurrences of pauses (left) and with the greatest occurrences (right), out of the 100 most frequently occurring contexts in the corpus. The asterisk indicates the potential location of the pause. 'Prep' refers to both prepositions and complementizers. 'EDIT' refers to a restart disfluency. 'Parenth.' refers to parentheticals like 'you know'.*

A multiple linear regression was also used to predict the duration of each word boundary (i.e. zero for no pause, otherwise the duration of the pause) in order to compare the relative predictive power of each feature. The independent variables were the features. The categorical features were broken down into binary dummy variables for each possible value, 0 if not present and 1 if present. All of the features were included, except those correlated with other features at $r > .96$ to avoid problems of multicollinearity. 91,225 word boundaries were used.

3. Results

3.1. Pauses in context

Figure 2 shows the percent of pauses that occur along with some of the 100 most frequent feature values in the corpus. At the low end (left) are the position between a subject and verb phrase and between a preposition and a determiner (i.e. between a preposition and its complement). At the high end (right) are surrounding conjunctions, surrounding restart disfluencies and filled pauses, and preceding subjects and parentheticals. Because some of these contexts are correlated with each other, a multiple regression analysis is more revealing.

3.2. Decision tree analysis

On the full training/testing sets, the decision tree classifier achieved an accuracy marginally above the baseline. Baseline accuracy was 91.60% by classifying all word boundaries as no-pause. Starting with all features and performing leave-one-out repeatedly ended up with the best feature set (in terms of accuracy) comprising all of the features except those related to part of speech labels and the feature for the duration of the following constituent. This feature set achieved an accuracy of 91.78% (f-score: 22%).

The balanced training and testing sets had the same number of pause and non-pause word boundaries so that the baseline accuracy was 50%. With the same leave-one-out procedure, the feature set that maximized accuracy included all features except the time features, the following POS label, and the number of words in the preceding constituent. This feature set had an accuracy of 78.3% (f-score: 78.5%). However, the 12 features in this set only marginally improved on the set of just the duration of the preceding constituent and the number of words in the following constituent: accuracy 78.0%, f-score 78.2%. (No single feature added to this set increases its accuracy.)

Table 2 reports the performance of the classifier on various other sets of features. With the POS features alone (B), the f-score was 66%. This was marginally improved on by adding the conditional probability (A), to 68% (D), but not the mutual information features (E). As in [8], adding constituent size and word boundary time to the POS features improved classification, here by 10 percentage points to 76% (C).

Including all syntactic information (G) yielded an f-score of 71%. This was worse than roughly the features used in [8] (C), and the syntactic features could not be improved on much by adding the lexical features (H).

Table 2: *Precision, recall, accuracy, and f-score for different feature sets on the balanced corpus*

	Features Included	Prec.	Rec.	Acc.	F
A	Conditional Probability	.54	.86	.57	.66
B	POS	.67	.65	.67	.66
C	POS, Const. Size, Time	.74	.77	.76	.76
D	POS, Conditional Prob.	.69	.68	.69	.68
E	D + Mutual Info.	.70	.65	.69	.67
F	Syn. Category, Relation	.74	.63	.71	.68
G	F + Constit. Size & Depth	.74	.67	.73	.71
H	E + G	.72	.71	.72	.72
I	Left Dur./Right # Wds	.76	.81	.78	.78

3.3. Regression analysis

Table 3 reports the regression coefficients for the continuous-valued features significant at $p < .001$. Figure 3 shows the regression coefficients for all of the remaining categorical features significant at the same level (except some features related to disfluencies, which were left out for brevity). Pauses had a median duration of .31 s, with the middle 50 percent between .17 and .48 s.

Following [8], the sizes of the preceding and following constituent, in number of words and duration, were considered. Confirming previous results, we found the duration of the constituent before has a much larger effect on pause duration than the constituent after (the coefficients were 10 times smaller and were not significant at this level). Roughly, for every second spent uttering a constituent, the expected pause duration after the constituent increases by .1 s. Second, although the number of words in and the duration of an utterance are highly correlated (r=.89), duration is a better predictor (if only slightly) of pause duration.

Both conditional probability and mutual information had a significant and negative coefficient, which was expected. The more likely the following word given the previous, or the more associated the words, the shorter the expected pause.

To take several examples, the contexts before conjunctions and between noun phrases and prepositional phrases (ignoring whether the PP attaches to the NP or higher) tended to inhibit pauses. The Specifier-Phrase syntactic relation, which collapsed the contexts of between a subject and verb phrase (i.e. Spec,IP) and a extracted wh constituent and the rest of the clause (i.e. Spec,CP), had one of the strongest negative co-

Table 3: *Regression coefficients predicting duration of pause*

Feature	Coefficient
Time — Relative	-.036
Time — Absolute	.0028
Conditional Probability	-.0039
Mutual Information	-.00067
Preceding Constituent — Words	-.025
Preceding Constituent — Duration	.098
Depth	.012

efficients.

On the other hand, the contexts before gerunds and after verb phrases, non-subject noun phrases, and adverbial subordinate clauses tended to induce a pause or a longer pause.

The context between a verb and its complement has been studied often. For a NP complement, the regression coefficient was negative and significant at $p < .01$. For clausal and adjectival complements, except quotations, the coefficients were not significant. The coefficient for the context preceding adverbial subordinate clauses was significant at $p < .05$, but was negative, thus failing to verify the results in [2], which found more pauses before less-embedded clauses.

4. Conclusion

We considered above a number of features at word boundaries to model the location and duration of intrasentential pauses with the aim of establishing the extent to which syntactic structure, especially category labels and structural relations, plays a role. A model based on binary classification revealed that syntactic features yield an improvement over lexical (POS and conditional probability) features alone. However, it was not category labels or structural relations that contributed most to the improvement, but rather the duration of the preceding constituent, an observation previously made by [8] and probably the underlying explanation for the results of [4] related to a 'complexity index'.

On the other hand, a look into the regression coefficients revealed that some POS and syntactic category labels were better predictors than others of word boundary durations, and those predictive features that were not indicators of disfluencies suggest that some effects of syntax on prosodic structure are showing through.

5. References

[1] Black, A.W.; Taylor, P. 1997. Assigning phrase breaks from part-of-speech sequences. Proceedings of Eurospeech.

[2] Gayraud, F.; Martinie, B. 2007. Does structural complexity necessarily imply processing difficulty? *Journal of Psycholinguistic Research*, Springer 'online first'.

[3] Godfrey, J.J.; Holliman, E. 1997. *Switchboard-1 Release 2*. Linguistic Data Consortium, Philadelphia.

[4] Grosjean, F.; Grosjean, L.; Lane, H. 1979. The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, 11, 58-81.

[5] Hirschberg, J.; Prieto, P. 1996. Training intonational phrasing rules automatically for English and Spanish text-to-speech. *Speech Communication*, 18, 281-290.

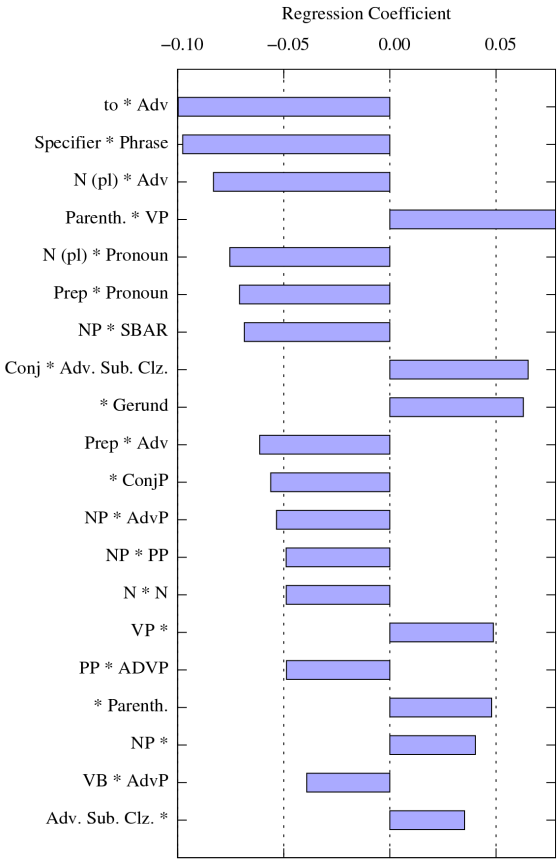

Figure 3: *Regression coefficients for POS and phrasal category features significant to $p < .001$. See the abbreviations in Figure 2 and the explanation in the text.*

[6] ICSI Transcriptions of Switchboard-1 Telephone Speech Corpus. 2001.

[7] Keri, V.; Pammi, S.C.; Prahallad, K. 2007. Pause prediction from lexical and syntactic information. *Proceedings of the International Conference on Natural Language Processing (ICON).*

[8] Koehn, P.; Abney, S.; Hirschberg, J.; Collins, M. 2000. Improving intonation phrasing with syntactic information. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, 3, 1289-1290.

[9] Maclay, H.; Osgood, C.E. 1959. Hesitation phenomena in spontaneous English speech. *Word*, 15, 19-44.

[10] Marcus, M.P.; Santorini, B.; Marcinkiewicz, M.A.; Taylor, A. 1999. Treebank-3. Linguistic Data Consortium, Philadelphia.

[11] Quinlan, J.R. 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.