A Linguistically Motivated Analysis of Intonational Phrasing in Text-to-Speech Systems: Revealing Gaps in Syntactic Sensitivity

Charlotte Pouw[1], **Afra Alishahi**[2], **Willem Zuidema**[1]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]Cognitive Science and Artificial Intelligence, Tilburg University
{c.m.pouw,w.h.zuidema}@uva.nl
a.alishahi@tilburguniversity.edu

Abstract

We analyze the syntactic sensitivity of Text-to-Speech (TTS) systems using methods inspired by psycholinguistic research. Specifically, we focus on the generation of intonational phrase boundaries, which can often be predicted by identifying syntactic boundaries within a sentence. We find that TTS systems struggle to accurately generate intonational phrase boundaries in sentences where syntactic boundaries are ambiguous (e.g., garden path sentences or sentences with attachment ambiguity). In these cases, systems need superficial cues such as commas to place boundaries at the correct positions. In contrast, for sentences with simpler syntactic structures, we find that systems *do* incorporate syntactic cues beyond surface markers. Finally, we finetune models on sentences without commas at the syntactic boundary positions, encouraging them to focus on more subtle linguistic cues. Our findings indicate that this leads to more distinct intonation patterns that better reflect the underlying structure.

1 Introduction

Humans use prosody to convey meaning beyond words. Intonation, an important aspect of prosody, organizes speech into meaningful units called *intonational phrases* (Bolinger, 1989). Linguistic theory suggests that in human speech, the positioning of boundaries between these phrases is closely linked to syntactic structure. Some theorists claim that intonational phrasing can directly be derived from syntactic structure (e.g., Steedman, 1991; Cooper, 1980); others argue that the mapping is more complex and there must exist an independent level of intonational structure (e.g., Pierrehumbert, 1980; Selkirk, 1984; Nespor and Vogel, 2007).

Regardless of the theoretical perspective, it is well-established that intonational and syntactic boundaries often overlap. Acoustic markers of intonational boundaries (i.e., pauses, syllable lengthening, and pitch contour changes) are frequently observed at syntactic boundary positions (Klatt, 1975; Cooper, 1976; Ferreira, 1993; Croft, 1995; Watson and Gibson, 2004). Psycholinguistic experiments have also shown that the placement of intonational boundaries influences parsing decisions in speech processing (e.g., Pynte, 1996; Kjelgaard and Speer, 1999; Pauker et al., 2011), and that speakers adjust their intonation to signal the underlying structure of an ambiguous sentence (e.g., Snedeker and Trueswell, 2003; Kraljic and Brennan, 2005; Schafer et al., 2005).

In this paper, we analyze if we can observe a similar link between syntax and intonational phrasing in the behavior of Text-to-Speech (TTS) systems. Such systems have become increasingly capable of mimicking human intonation patterns, but it remains an open question to what extent these patterns are shaped by linguistic structure. We propose to use methods from psycholinguistics to investigate this question, an approach previously used to assess the syntactic sensitivity of text-based language models (e.g., Linzen et al., 2016; Futrell et al., 2019; Ettinger, 2020; Jumelet et al., 2024). This involves the use of controlled stimuli that require a reliance on specific (linguistic) information to elicit specific behavioral responses.

We find that TTS systems incorporate syntactic information when it reliably signals the need for an intonational boundary (i.e., obvious clause boundaries in simple sentence structures), although the duration of intonational boundaries is also modulated by lexical cues. In more complex cases such as garden path sentences and attachment ambiguities, systems need explicit punctuation cues to place intonational boundaries at the correct syntactic positions. In the absence of such cues, TTS systems tend to default to the statistically most likely intonation pattern, which may not align with the underlying structure.

Encouragingly, we also find that with increased exposure to sentences where we have removed ex-

plicit punctuation cues at the intonational boundary positions, TTS systems *can*, to some extent and under some conditions, learn to generate more distinct intonation patterns that better reflect alternative syntactic structures. We hope that these findings contribute to the development of more linguistically informed TTS training and evaluation paradigms. All code is available at our GitHub repository.

2 Psycholinguistic Background

The relationship between intonation and syntax has been explored in various psycholinguistic studies. These studies often use sentences with (temporary) syntactic ambiguity (Cutler et al., 1997), as listeners have to make a decision about the syntactic structure based on controlled evidence (e.g., the position of an intonational boundary). These sentences therefore provide a unique opportunity to study the interplay between intonational boundary placement and syntactic parsing decisions in speech processing.

A key area of research has focused on **garden path** sentences—structures that initially lead the listener to a syntactic interpretation that must later be revised (Bever, 1970). From the extensive literature on the human processing of such sentences, we mention Kjelgaard and Speer (1999), who examined sentences such as *When Roger left the house was dark*, which initially confuses the listener into interpreting *left the house* as a single constituent. They found that an intonational boundary after *left* facilitated processing speed, as it helped to clarify the syntactic structure. However, a boundary after *the house* led to processing difficulty because it interfered with the underlying structure.

A related phenomenon occurs with sentences that exhibit **attachment ambiguity**, where there are two alternative syntactic structures based on the attachment site of a prepositional phrase. Many psycholinguistic studies have revealed details of how humans deal with such ambiguity. For instance, Pynte (1996) showed that, in sentences such as *The spies inform the guards of the conspiracy*, an intonational boundary after *inform* leads to the NP-attachment interpretation (i.e., *of the conspiracy* attaches to *the guards*), whereas a second boundary after *guards* leads to the VP-attachment interpretation (i.e., *of the conspiracy* attaches to *inform*). These findings illustrate how the position of intonational boundaries can guide listeners towards alternative syntactic structures.

In speech production, it has been shown that speakers adjust their intonation to signal the underlying structure of an ambiguous sentence. For example, Snedeker and Trueswell (2003) studied the placement of intonational boundaries in a referential game setting. Speakers had to refer to objects with instructions such as *Tap the frog with the flower*. The attachment site of the PP *with the flower* was ambiguous, as the room contained a frog toy with a flower on its head, as well as a frog and a flower separately. When speakers were aware of the ambiguity, they produced a boundary after *frog* to signal the VP-attachment structure (i.e., when they wanted the addressee to use the flower as an instrument); they did not do this for the NP-attachment scenario (i.e., when they wanted the addressee to tap the frog which had the flower on its head). In other (similar) studies, this pattern has been observed even for speakers who were unaware of the potential ambiguity (Kraljic and Brennan, 2005; Schafer et al., 2005).

Taken together, these studies illustrate how both listeners and speakers use intonational boundaries to interpret and signal syntactic structures. In the present study, we systematically analyze whether and how TTS systems are informed by syntax to determine the placement of intonational boundaries.

3 Text-to-Speech Models

We select three publicly available TTS systems with diverse architectures. We also provide Mean Opinion Scores (MOS) (i.e., human ratings of the naturalness of each system's output speech, on a scale from 1-5) reported for each system, while noting that these scores were not consistently measured, and should therefore only been seen as approximate (Kirkland et al., 2023; Chiang et al., 2023; Le Maguer et al., 2024).

Tacotron2 (Shen et al., 2018) is an LSTM-based encoder-decoder. The bidirectional encoder converts a character sequence into a hidden feature representation, which the decoder (with attention) takes as input to autoregressively predict spectrogram frames. A WaveNet vocoder (Van Den Oord et al., 2016) transforms these spectrogram frames into a waveform. The model was trained on an internal US-English dataset containing 24.6 hours of speech from one female speaker. MOS: 3.52[1]

[1]The original release paper of Tacotron2 reports a MOS of 4.53, but the model scores much lower on LJSpeech.

Speech-T5 (Ao et al., 2022) is a Transformer-based encoder-decoder. The encoder embeds token indices based on which the decoder predicts a log Mel-filterbank. A HiFi-GAN vocoder (Kong et al., 2020) is used to convert the predicted log Mel-filterbank to a waveform. The encoder-decoder is jointly pre-trained on speech and text from audiobooks (960h of spoken language and 400M written sentences from LibriSpeech, Panayotov et al. (2015)). For TTS, the model is fine-tuned on 460 hours from LibriTTS (Zen et al., 2019). MOS: 3.65

Parler-TTS (Lyth and King, 2024) is a decoder-only Transformer. The model autoregressively predicts latent audio tokens given a sequence of prepended text tokens. These audio tokens are then decoded into a waveform using the Descript Audio Codec (DAC) (Kumar et al., 2023). We use Parler-TTS Mini v0.1, which was trained on 10k hours from the English portion of Multilingual LibriSpeech (Pratap et al., 2020) plus 585 hours from LibriTTS-R (Koizumi et al., 2023). MOS: 3.92

4 Experiment 1: Ambiguous Structures

The goal of this experiment is to assess whether TTS systems can correctly analyze the structure of sentences with (temporary) syntactic ambiguity, and place intonational boundaries in the correct positions accordingly. Using controlled stimuli, we analyze which cues are used by the systems to disambiguate these sentences.

4.1 Syntactic Disambiguation

Garden path sentences contain temporary syntactic ambiguity because the syntactic closure point can either appear early or late in the sentence. Consider the following examples:

1. **Early closure:** *When Roger left$_A$ the house was dark.*
2. **Late closure:** *When Roger left the house$_B$ it was dark.*

In the early closure condition, the syntactic boundary occurs at position *A*, while in the late closure condition, the boundary appears later, at position *B*. The word *was* or *it* resolves the ambiguity. We investigate if TTS systems are sensitive to these syntactic cues and place intonational boundaries in the correct positions accordingly.

As a control, we use the same sentences with a comma inserted at the syntactic closure point (i.e., *When Roger left,$_A$ the house was dark* and *When Roger left the house,$_B$ it was dark*). These commas should provide the systems with more explicit, surface-level cues for generating intonational boundaries. Having this control condition allows us to observe a clear "ground-truth" intonation pattern for each underlying structure.

For our stimuli, we used 45 garden path sentences from several psycholinguistic studies (Kjelgaard and Speer, 1999; Pauker et al., 2011), which are listed in Appendix Table 3.

4.2 Semantic Disambiguation

In addition to syntactic cues, semantic information can also be used to resolve syntactic ambiguity. To test whether TTS systems are sensitive to semantic cues, we used sentences with attachment ambiguity containing a semantic bias towards either high (VP) or low (NP) attachment. For example:

1. **High attachment:** *The boy looked at the painting$_A$ with much enthusiasm.*
2. **Low attachment:** *The boy looked at the painting with muted colours.$_B$*

The prepositional phrase *with enthusiasm* is more likely to attach to *looked at*, whereas *with muted colours* is semantically more likely to attach to *the painting*. We analyze if TTS system can distinguish between these structures based on this semantic bias. If so, we would expect an intonational boundary at position *A* to signal the high attachment structure, and no boundary at that position to signal the low attachment structure. Again, we add a control condition with a comma placed at the boundary position, but only for the high attachment cases (e.g., *The boy looked at the painting,$_A$ with much enthusiasm*), since the comma would be unnatural in the low attachment cases (e.g., *The boy looked at the painting, with muted colours$_B$*).

We generated stimuli using the following template: `<Animate Subject> <Verb> <Inanimate Object> with <Inanimate/Animate Property>`. We filled each slot with six different phrases and generated all possible combinations, resulting in a dataset of 1296 sentences with a semantic bias towards low attachment and 1296 with a bias towards high attachment. Examples are listed in Appendix Table 4.

4.3 Measuring Intonational Boundaries

We use the Montreal Forced Aligner (MFA) (McAuliffe et al., 2017) to align the generated

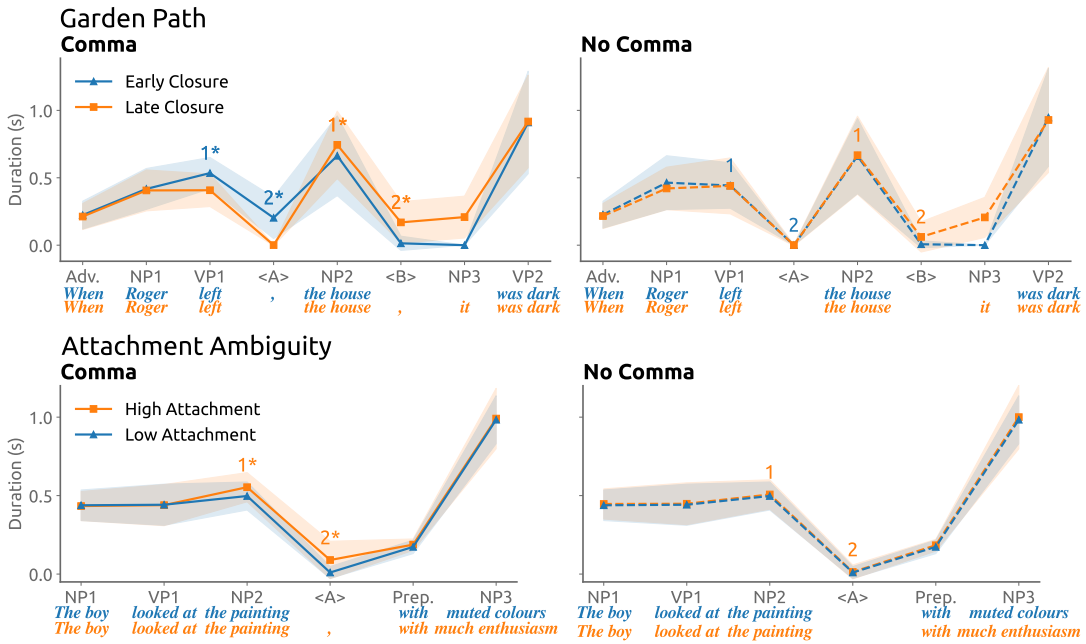

Figure 1: Average durations of sentence regions in garden path sentences (top) and sentences with attachment ambiguity (bottom), generated by Parler-TTS. An intonational boundary consists of *lengthening* at the pre-boundary position (1), and insertion of a *pause* at the syntactic boundary position (2); asterisks indicate the presence of these effects. Example sentences are annotated on the x-axes; shading indicates the standard deviation across sentences.

speech with the input text and measure the duration of each sentence region within the garden path and attachment ambiguity sentences. We identify the presence of intonational boundaries by examining two durational cues: 1) lengthening at the pre-boundary position and 2) the insertion of a pause (i.e., silence, indicated by an unannotated segment by the MFA) at the boundary position. However, we acknowledge that this method has limitations (see Appendix B), as other prosodic cues such as pitch and intensity also contribute to the perception of intonational boundaries.

4.4 Results

Figure 1 shows average durations across sentence regions as generated by Parler-TTS. The results for Tacotron2 and Speech-T5 are highly similar and shown in Appendix Figure 6.

We observe a strong dependence on comma cues: the systems lengthen the pre-boundary position (1) and insert a pause at the syntactic boundary position (2) only in the presence of a comma at position (2). Without comma cues, the systems default to the statistically most likely intonation pattern. For garden path sentences, this means that no intonational boundaries are generated at position <A>, and occasionally, a pause is inserted at position <B>, since late closure sentences are statistically more likely than early closure sentences. For attachment ambiguity, this means that no intonational boundaries are generated, even if it does not align with the semantic bias of the prepositional phrase.

5 Experiment 2: Simple Structures

Our previous experiment indicates that TTS systems struggle to resolve local or global ambiguities in syntactic structure, and are much more dependent on explicit punctuation cues for the generation of intonational boundaries at the correct positions. This is in a sense a human-like effect, as the syntactic structure of garden path and attachment ambiguity sentences is hard to parse, even for humans. It is possible that models correctly incorporate syntactic cues when these are more reliable (i.e., not ambiguous).

In the next experiment, we analyze the role of syntactic cues for intonational boundary placement in simpler sentence structures. We also investigate the role of commas in more detail: are they purely mechanical markers that always trigger a pause, or can TTS systems combine evidence from commas and syntax? To address this, we place commas in syntactically natural and unnatural positions (i.e., aligned with a clause boundary or not), and then compare the strength of the intonational boundaries generated at these points.

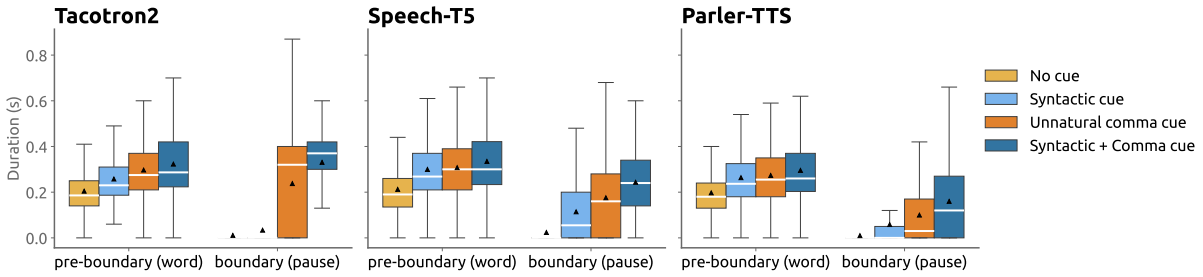

Figure 2: Durations of critical regions (i.e., pre-boundary word and pause at the boundary position), as generated by three TTS systems given different cues: presence or absence of a comma (light vs. dark); measurement of the pause at a syntactic or non-syntactic boundary (blue vs. orange). Black triangles are means, white lines are medians.

5.1 Data

From Simple Wikipedia[2], we select sentences that contain exactly one comma, marking a syntactic boundary.[3] We select boundaries that signal major structural breaks, which typically lead to an audible intonational boundary in spoken language. We use dependency parsing to detect such structural breaks (examples are listed in Appendix Table 5). We create different versions of each sentence, such that the TTS systems have access to different cues for potentially generating an intonational boundary. After synthesizing these sentences, we measure the presence of an intonational boundary at position *A*:

1. **Comma cue + Syntactic cue**: *Most links are blue,$_A$ but they can be any color.*

2. **Syntactic cue**: *Most links are blue$_A$ but they can be any color.*

In (1), the systems can use the comma at position *A* as a cue for generating an intonational boundary. Additionally, they can use the fact that position *A* is a clause boundary. In (2), the systems can only rely on the clause boundary information, since the comma is absent.

To investigate the extent to which TTS systems generate intonational boundaries at syntactically unnatural positions, we measure the presence of an intonational boundary at position *B*:

3. **Unnatural comma cue**: *Most links are blue but they can,$_B$ be any color.*

4. **No cue**: *Most links are blue but they can$_B$ be any color.*

In (3), the systems can use the comma as a cue for generating an intonational boundary at position *B* (although it appears at a syntactically unnatural position). In (4), there is no cue that indicates the need for an intonational boundary at position *B*.

5.2 Evaluation

Besides comparing the durations for critical regions (i.e., the (pre-)boundary position) across conditions[4], we compute a **Syntactic Sensitivity Score** for each system. This consists of precision, recall and F1 scores based on the following counts in the sentences without commas: **True Positives** occur when the model generates a pause at a syntactic boundary (position *A*), **False Positives** when it generates a pause at a syntactically unnatural position (position *B*), **False Negatives** when no pause is generated at position *A*, and **True Negatives** when no pause is generated at position *B*.

5.3 Results

Figure 2 shows the durations for the pre-boundary word and boundary pause, depending on condition. We see that all models show a similar pattern: the strongest intonational boundaries are produced in the Syntactic + Comma cue condition. None of the models produce an intonational boundary in the No cue condition. The Syntactic cue and Unnatural comma cue conditions are inbetween, with the comma cue leading to a slightly stronger intonational boundary than the syntactic cue. This indicates that in simple sentence structures, TTS systems do pick up on syntactic cues, but that commas simply provide more direct evidence for intonational boundaries. It also shows that models

[2]https://simple.wikipedia.org

[3]Additional filters were applied: sentences had to be between 7 and 15 words long and free of digits, punctuation (except commas and final periods), and bracketed phrases.

[4]The words preceding the syntactic boundary position *A* and non-boundary position *B* may have different lengths, which could affect the average duration. To account for this, we averaged word duration by syllable count.

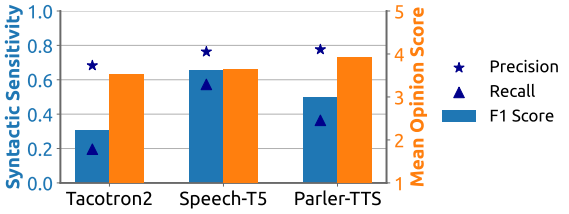

Figure 3: Syntactic Sensitivity versus Mean Opinion Score across TTS systems. The F1 score represents the harmonic mean of a system's precision and recall in generating pauses at syntactic boundaries.

integrate evidence from multiple sources: the combination of a comma and a syntactic cue leads to a stronger intonational boundary than only one cue.

In Figure 3, we compare our Syntactic Sensitivity score with reported MOS for each system. We see that *precision* mirrors the MOS pattern (Tacotron2 < Speech-T5 < Parler-TTS), while Speech-T5 has better *recall* than Parler-TTS. In other words: False Positives (i.e., pauses placed at syntactically unnatural positions) seem to affect human ratings more than False Negatives (i.e., no pauses at syntactic boundaries). This illustrates that our Syntactic Sensitivity score provides complementary insights that MOS does not capture.

6 Interpreting Boundary Placement

In the previous experiments, we used controlled stimuli to analyze how two specific cues influence intonational boundary placement in TTS systems. It could be the case, however, that systems' predictions are modulated by the presence of lexical items associated with pauses (e.g., conjunctions such as *but*, *and*, *or*). To gain insight into these cues, we train regression models with a range of different predictor variables to approximate the intonational boundary placement behavior of each TTS system.

For each TTS system, we train two regression models to predict the following outcome variables for a given position in a sentence: the duration of a pause in that position (**pause duration**), and the duration of the word before that pause (**pre-boundary word duration**) (i.e., the two aspects of an intonational boundary we focus on). We again use the sentences from Simple Wikipedia as input and extract the features listed in Table 1 at the positions marked as *A* or *B* (see Section 5.1).

Implementation Since we have a large number of (correlated) features, we use LASSO (Least Absolute Shrinkage and Selection Operator; Tibshirani 1996). This regularization technique introduces a penalty term that encourages sparsity in the model, allowing only a subset of features to be used in predicting the outcome variable, preventing overfitting and reducing the effect of multicollinearity (when features are highly correlated, LASSO tends to select only one of them). We apply standardization to the numerical features to ensure they have a mean of zero and a standard deviation of one (unit variance). We train the regression models on 80 percent of the data and evaluate on the remaining 20 percent. We use R^2 (explained variance) as our evaluation metric to gauge how well the predicted regression lines fit the data.

Category	Predictor
Punctuation	Comma Presence (1 or 0)
Lexical	Preceding POS tag (one-hot) Following POS tag (one-hot)
Constituency	Is Clause Boundary (1 or 0) Num. Closing Brackets Max. Tree Depth
Dependency	Preceding Token: Is Dep. Head (1 or 0) Preceding Token: Num. Dependents Preceding Token: Depth in Subtree
Length	Preceding Token Length Following Token Length Sentence Length Number of Preceding Tokens
Interaction	Is Clause Boundary * Comma Presence

Table 1: Predictor variables for regression models. Global features are extracted from the entire sentence; the other features are extracted at the boundary positions described in Section 5.1.

6.1 Results

Model	Pause Dur.	Pre-boundary Word Dur.
Parler-TTS	.14	.37
Speech-T5	.30	.44
Tacotron2	.44	.42

Table 2: Explained variance (R^2) of linear regression models for predicting *pause duration* and *pre-boundary word duration* as generated by three different TTS systems. Reported scores are for a held-out test set.

Performance The performance of the regression models is displayed in Table 2. We see that our predictor variables generally explain more variance in the *pre-boundary word duration* data compared to the *pause duration* data, which makes sense given that we use explicit features of the pre-boundary

word (e.g., its length). We also see that pause duration is more predictable for Tacotron2 than for the other two systems. The behavior of Parler-TTS is least predictable, indicating that this model relies on other features than the ones we included in our regression models, or on more complex interactions between those features.

Feature Importance Figure 4 shows the top 10 selected predictors for *pause duration* for each of the TTS systems, together with their regression coefficients. We see that *comma presence* is the strongest predictor for all three TTS systems, verifying their strong reliance on punctuation cues. For Parler-TTS and Speech-T5, *is clause boundary* is also an important predictor.[5] We also see that specific lexical items are selected, e.g., words with the POS tag SCONJ or CCONJ. Depending on the model, different length-related features are also selected: *sentence length* for Parler-TTS and Speech-T5, *preceding/following token length* for Speech-T5, and *num. preceding tokens* for Tacotron2.

Overall, the analysis confirms that punctuation plays a major role in determining the duration of intonational boundaries in TTS systems. It also demonstrates that specific lexical items and length-related features influence pause duration. This reliance on surface cues is particularly evident in the LSTM-based system Tacotron2, while the Transformer-based systems Parler-TTS and Speech-T5 also seem to incorporate some syntactic information.

7 Changing the Training Distribution

While TTS systems may see plenty of examples of simple syntactic structures with obvious clause boundaries, garden path sentences are likely underrepresented in their training data. Sentences with attachment ambiguity may occur more frequently. However, even for such sentences, the intonation patterns we aim to capture (where high attachment introduces an intonational boundary and low attachment does not) may still be rare in the training data. As discussed in Section 2, speakers use distinct intonation patterns to disambiguate high and low attachment in conversational settings, helping to convey the intended meaning. In non-conversational speech, this distinction is less frequently observed.

[5]We verified that *is clause boundary* was a predictor by itself by running LASSO on different subsets of sentences: with/without commas, and with/without predictive lexical items (e.g., conjunctions). In all cases, *is clause boundary* was still selected as an important predictor.

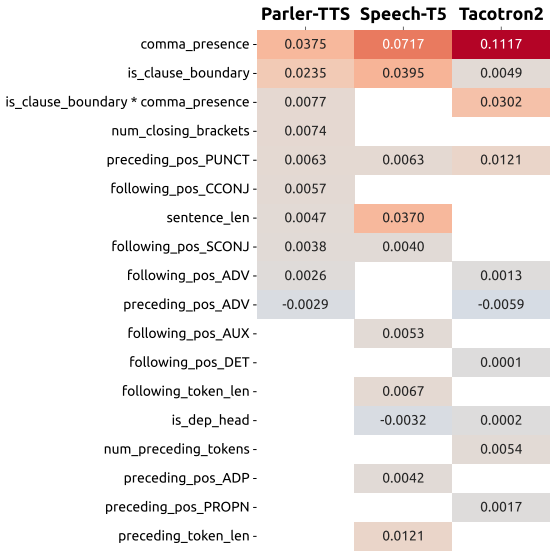

Figure 4: Coefficients of LASSO-selected predictor variables for pause durations of TTS systems.

Consequently, TTS systems trained on audiobooks may not have sufficient exposure to the nuanced intonation patterns associated with the different syntactic structures.

7.1 Training data analysis

Out of the three TTS systems we investigated, Parler-TTS was trained on the largest amount of data. To check if it missed important evidence for high and low attachment structures, we selected a subset of the MLS corpus that Parler-TTS was trained on (5000 examples, ~12000 sentences) and counted the occurrences of pauses, commas, and frequent prepositions[6], as well as the overlap between them. The detailed results are shown in Appendix Figure 7. While we cannot directly determine how often the model encountered high or low attachment structures, we observe that prepositions *without* a preceding pause (aligning with low attachment) appeared almost 5 times more frequently than those *with* a preceding pause (aligning with high attachment). This imbalance may explain why the model struggles to generate distinct intonation patterns for the two structures.

7.2 Altering the training distribution

We hypothesize that a greater balance in the occurrence of high and low attachment structures in the training data will enable the model to generate more varied intonation patterns that better reflect the underlying structure. To test this hypothesis,

[6]*of, to, in, for, with, as, at, on, by, for*.

we conducted two finetuning experiments aimed at rebalancing the data. These experiments are not meant to directly improve the performance of Parler-TTS, but merely to diagnose the role of (lack of) exposure to certain structures.

Finetuning on sampled data For the first experiment, we selected all sentences from the Jenny corpus[7] containing a preposition preceded by a pause (~5000 sentences, ~6 hours of speech). To ensure that the model would not be able to rely on commas as a cue for generating intonational boundaries, we removed all commas from the transcriptions. Our hope was that showing the model more examples of *general* PPs preceded by a pause would lead to more varied intonation patterns for sentences with an *ambiguous* PP.

Finetuning on synthetic data For the second experiment, we created a synthetic dataset to provide the model with more explicit examples of high and low attachment. Using the template described in Section 4.2, we generated 2500 sentences with a semantic bias towards high attachment, and 2500 sentences with a bias towards low attachment (resulting in ~6 hours of speech). We synthesized these sentences using Tacotron2, inserting commas at positions that would correspond to intended pauses (e.g., before the preposition *with* in high attachment cases). We again removed these commas from the text to ensure that the model could not rely on punctuation, but instead learn to use the semantic bias of the sentences to predict the presence of a pause.

Evaluation We created an evaluation set consisting of sentences containing function words that could be interpreted in two different ways, with one interpretation requiring a pause before the word (e.g., *The boy looked at the painting <pause> with genuine interest*) and the other not (e.g., *The boy looked at the painting with muted colors*). These function words include *with* (our primary example for high and low attachment), but also *as*, *for*, and *to*, as shown in Table 6 in the Appendix. We created 30 sentences per category and sampled them three times from the models (using three different random seeds). We then measured the pause duration before the critical function word across the resulting 90 data points.

[7] https://github.com/dioco-group/jenny-tts-dataset

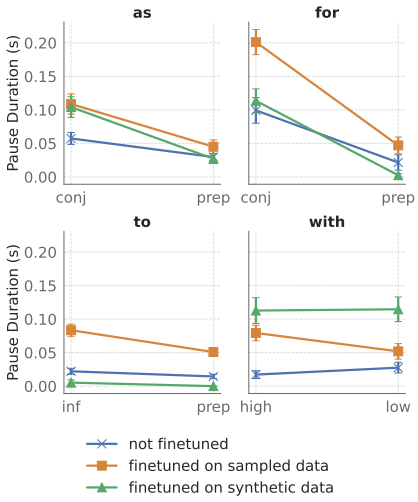

Figure 5: Average pause duration before the function words *as, for, to, with* (each used in two different ways, e.g., as a preposition vs. conjunction) for three versions of Parler-TTS. Error bars indicate the standard error.

7.3 Results

Figure 5 shows that the model finetuned on sampled data (orange lines) generates longer pauses than the non-finetuned model (blue lines). Interestingly, this increase is more pronounced in contexts where a pause is expected, i.e., before *for* and *as* when used as a conjunction, before *to* when used as an infinitive, and before *with* in the high attachment case. This suggests that training the model on a more balanced data distribution leads to more distinct intonation patterns that reflect different syntactic structures.

In contrast, the model finetuned on synthetic data did not learn to distinguish between high and low attachment based on semantic cues, as the pause duration before *with* remains the same in both cases (although it did increase compared to the non-finetuned model). These results indicate that, even with more exposure, TTS systems cannot disambiguate syntactic structure based on semantic cues. However, this observation requires further investigation, particularly regarding the role of natural versus synthetic speech and the amount of data necessary for robust results.

8 Conclusion

We evaluated the syntactic sensitivity of TTS systems by analyzing their intonation patterns generated for controlled stimuli. We find that systems can identify obvious clause boundaries in simple sentences but struggle with more complex, locally

or globally ambiguous structures. We also investigated the role of (lack of) exposure to such structures, and show that systems can generate more syntax-aligned intonation patterns if provided with appropriate evidence.

Future work should study a broader range of phenomena to better understand the types of linguistic associations captured by TTS systems. One potential direction would be to develop a resource similar to BLiMP (Warstadt et al., 2020) for TTS, which could serve as a more comprehensive framework for evaluating their syntactic sensitivity. Additionally, *structural probing* (Hewitt and Manning, 2019; Shen et al., 2023) could offer a more detailed look at the internal representation of syntax in TTS systems, complementing our behavioral measures.

9 Acknowledgements

We are grateful to Grzegorz Chrupała, Jaap Jumelet and Tom Lentz for reviewing early versions of this paper. This research is funded by the Netherlands Organisation for Scientific Research (NWO) through NWA-ORC grant NWA.1292.19.399 for *InDeep*.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738.

Thomas G Bever. 1970. The cognitive basis for linguistic structures. cognition and the development of language, ed. by John R. Hayes, 279-362.

Dwight Bolinger. 1989. Intonation and its uses: Melody in grammar and discourse. *Edward Arnold*.

Sara Bögels, Herbert Schriefers, Wietske Vonk, Dorothee J. Chwilla, and Roel Kerkhofs. 2010. The Interplay between Prosody and Syntax in Sentence Processing: The Case of Subject- and Object-control Verbs. *Journal of Cognitive Neuroscience*, 22(5):1036–1053.

Cheng-Han Chiang, Wei-Ping Huang, and Hung-yi Lee. 2023. Why we should report the details in subjective evaluation of tts more rigorously. In *Proc. Interspeech 2023*, pages 5551–5555.

W Cooper. 1980. Syntax and speech.

William E. Cooper. 1976. Syntactic control of timing in speech production: a study of complement clauses. *Journal of Phonetics*, 4(2):151–171.

William Croft. 1995. Intonation units and grammatical structure. *Linguistics*, 33(5):839–882.

Anne Cutler, Delphine Dahan, and Wilma Van Donselaar. 1997. Prosody in the comprehension of spoken language: A literature review. *Language and speech*, 40(2):141–201.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Fernanda Ferreira. 1993. Creation of prosody during sentence production. *Psychological review*, 100(2):233.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 14727–14742. Association for Computational Linguistics (ACL).

Ambika Kirkland, Shivam Mehta, Harm Lameris, Gustav Eje Henter, Eva Székely, and Joakim Gustafson. 2023. Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. In *12th Speech Synthesis Workshop (SSW) 2023*.

Margaret M. Kjelgaard and Shari R. Speer. 1999. Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language*, 40(2):153–194.

Dennis H. Klatt. 1975. Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, 3(3):129–140.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023. LibriTTS-R: A restored multi-speaker Text-to-Speech corpus.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative Adversarial Networks for efficient and high fidelity speech synthesis. *CoRR*, abs/2010.05646.

Tanya Kraljic and Susan E. Brennan. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology*, 50(2):194–231.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved RVQGAN.

Sébastien Le Maguer, Simon King, and Naomi Harte. 2024. The limits of the Mean Opinion Score for speech synthesis evaluation. *Computer Speech & Language*, 84:101577.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Dan Lyth and Simon King. 2024. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech 2017*, pages 498–502.

Marina Nespor and Irene Vogel. 2007. *Prosodic Phonology: With a new foreword*, volume 28. Walter de Gruyter.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Efrat Pauker, Inbal Itzhak, Shari R. Baum, and Karsten Steinhauer. 2011. Effects of Cooperating and Conflicting Prosody in Spoken English Garden Path Sentences: ERP Evidence for the Boundary Deletion Hypothesis. *Journal of Cognitive Neuroscience*, 23(10):2731–2751.

Janet Breckenridge Pierrehumbert. 1980. *The phonology and phonetics of English intonation*. Ph.D. thesis, Massachusetts Institute of Technology.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

Joel Pynte. 1996. Prosodic breaks and attachment decisions in sentence parsing. *Language and cognitive processes*, 11(1-2):165–192.

AJ Schafer, SR Speer, and P Warren. 2005. Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. *Approaches to studying world situated Language Use: Psycholinguistic, linguistic and computational perspectives on bridging the product and action tradition*.

E Selkirk. 1984. Phonology and syntax. *The relation between sound and structure*.

Gaofei Shen, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2023. Wave to Syntax: Probing spoken language models for syntax. In *Proc. INTERSPEECH 2023*, pages 1259–1263.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. Natural TTS synthesis by conditioning WaveNet on Mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE.

Kim EA Silverman, Mary E Beckman, John F Pitrelli, Mari Ostendorf, Colin W Wightman, Patti Price, Janet B Pierrehumbert, and Julia Hirschberg. 1992. ToBI: A standard for labeling English prosody. In *ICSLP*, volume 2, pages 867–870.

Jesse Snedeker and John Trueswell. 2003. Using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, 48(1):103–130.

Mark Steedman. 1991. Structure and intonation. *Language*, 67(2):260–296.

Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, et al. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Duane Watson and Edward Gibson. 2004. The relationship between intonational phrasing and syntactic structure in language production. *Language and cognitive processes*, 19(6):713–755.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from LibriSpeech for Text-to-Speech. *Interspeech 2019*.

A Appendix Figures and Tables

Figure 6, Figure 7 and Table 3, Table 4, Table 5, Table 6 are shown on the next pages.

B Limitations

This study has several limitations that should be acknowledged. First, we analyzed intonational boundaries based duration measures only. While pause duration and word lengthening are well-established proxies for intonational boundaries, other prosodic features (e.g., pitch contour and intensity) also contribute significantly to their perception. Although previous research suggests that duration measures alone can reliably indicate the presence of an intonational boundary, and that pitch and intensity are less consistent across speakers and contexts (Kjelgaard and Speer, 1999; Bögels et al., 2010; Pauker et al., 2011), incorporating these additional prosodic cues would allow us to better characterize intonational structure (as generated by TTS systems).

Second, we did not consider different levels of boundary strength, a distinction made by the Tones and Break Indices (ToBI) framework (Silverman et al., 1992). Future work could benefit from adopting this gradation to more fully capture the complexity of intonational phrasing.

Third, the Parler-TTS model supports conditioning on voice characteristics specified through natural language descriptions. However, in this study, we only used a single voice description to synthesize our stimuli. It remains an open question how varying these voice characteristics might influence the resulting intonation patterns.

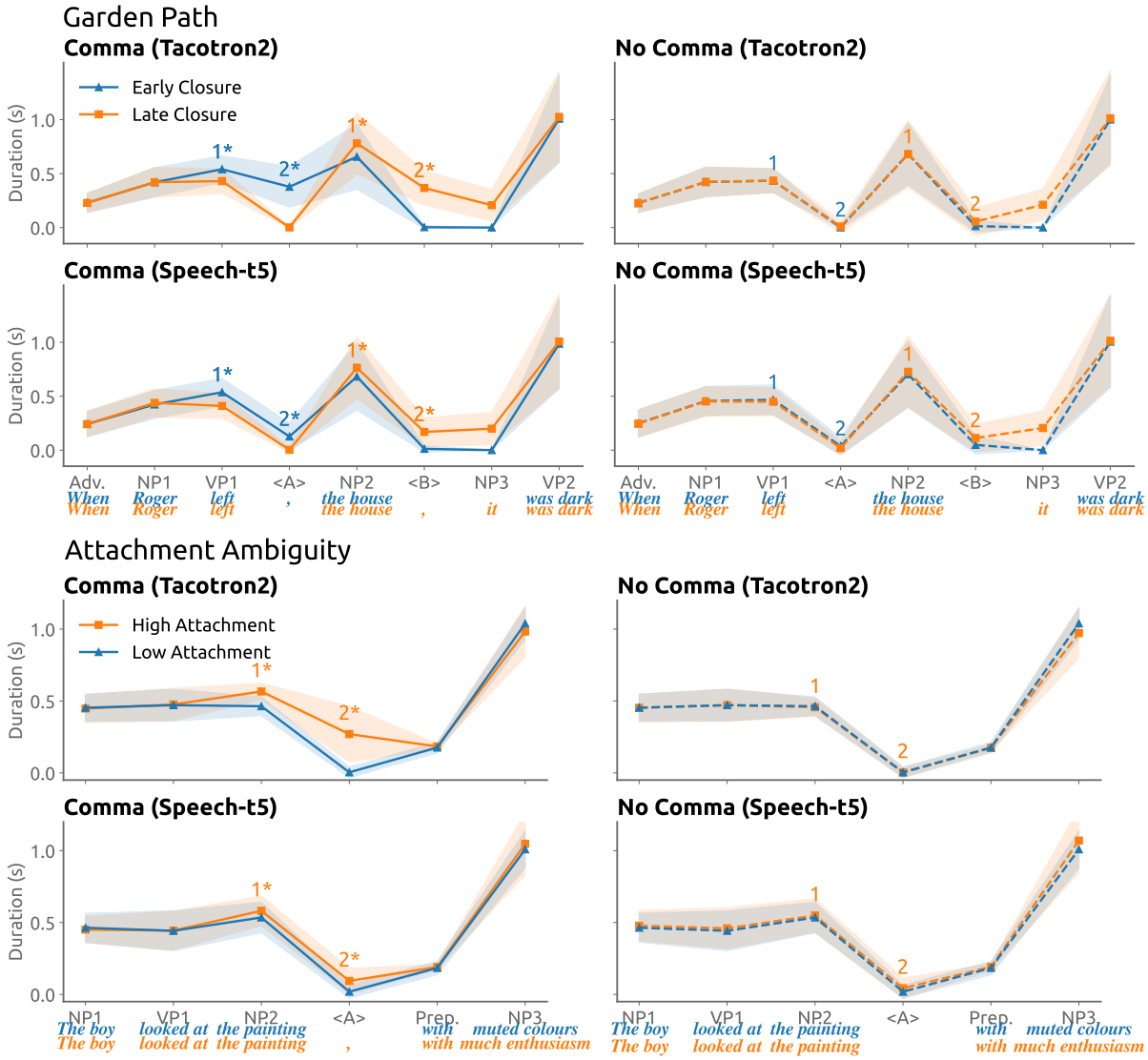

Figure 6: Average durations of sentence regions in garden path sentences (top) and sentences with attachment ambiguity (bottom), generated by Tacotron2 and Speech-T5. An intonational boundary consists of *lengthening* at the pre-boundary position (1), and insertion of a *pause* at the syntactic boundary position (2); asterisks indicate the presence of these effects. Example sentences are annotated on the x-axes; shading indicates the standard deviation across sentences.

Table 3: Garden path stimuli for Experiment 1. Sentences were presented in two forms: **early closure** (without the word in brackets) and **late closure** (with the word in brackets).

Stimulus
Whenever John walks the dogs (*cats*) are chasing him.
Because John studied the material (*it*) is clearer now.
When Whitesnake plays the music (*it*) is loud.
When Tim presents the lectures (*they*) are interesting.
When the original cast performs the plays (*they*) are funny.
When Madonna sings the song (*it*) is a hit.
Whenever John swims the channel (*it*) is choppy.
When Roger left the house (*it*) was dark.
Whenever Frank performs the show (*it*) is fantastic.
Because Mike phoned his mother (*she*) is relieved.
When the clock strikes the hour (*it*) is midnight.
Whenever the guard checks the door (*it*) is locked.
If Laura folds the towels (*they*) are neat.
If George programs the computer (*it*) is sure to crash.
If Charles babysits the children (*they*) are happy.
When the maid cleans the rooms (*they*) are immaculate.
Before Jack deals the cards (*they*) are shuffled.
While the boy read books (*televisions*) were stolen.
When the dog bites cats (*mice*) run away.
When the man batted balls (*players*) covered the field.
While the man parked cars (*bikes*) were waiting.
After the puppy licked kids (*parents*) were laughing.
Because snakes eat mice (*toads*) hide.
When a bear approaches people (*dogs*) come running.
After the chef cooked cake (*coffee*) was served.
While the artist painted clouds (*stars*) were appearing.
As the cat climbed trees (*leaves*) were falling.
As John hunted the frightened deer (*it*) escaped through the woods.
When Anne visited the British relatives (*they*) were moving to London.
When Rita washed her favorite sweater (*it*) was torn to shreds.
When Joan left her old boyfriend (*he*) stalked her for two months.
While the assistant measured the delicate fabric (*it*) ripped and frayed.
When Greg returned the new car (*it*) was operating smoothly.
Because Cecelia baked the delicious homemade bread (*it*) was served at breakfast.
Even when Todd cleaned the small kitchen (*it*) smelled like old garbage.
Because Grandma knitted wool sweaters (*they*) would appear under the Christmas tree.
Because Maria read the financial news (*it*) was always at her fingertips.
As Sam pounded the thin metal (*it*) ripped and broke into pieces.
When Sonya painted the kitchen walls (*they*) were covered into obvious drops.
As Lia typed the eviction notice (*it*) was cancelled.
When Tina supervised the night crew (*it*) was very efficient.
As Gary watched the drunken workmen (*they*) stumbled off the platform.
When the sheriff patrolled the whole area (*it*) was very safe.
When the musician conducted the symphony orchestra (*it*) was at its peak.
When Molly sang the drinking songs (*they*) sounded like opera.

Table 4: Examples of attachment ambiguity stimuli for Experiment 1. Two prepositional phrases were constructed for each stimulus, the former creating a semantic bias towards high (VP) attachment, the latter creating a semantic bias towards low (NP) attachment.

Stimulus
The boy looked at the painting *with much enthusiasm* / *with muted colors*.
The woman described the table *with much enthusiasm* / *with the smooth surface*.
The man bought the vase *with much happiness* / *with red dots*.
The girl found the chair *with much ease* / *with blue stripes*.
The artist inspected the house *with much interest* / *with wooden details*.

Table 5: Example sentences and counts of selected dependency labels, taken from the Simple Wikipedia Corpus.

Dependency Label	**Example**	**Count**
Conjunction (conj)	Most links are blue, but they can be any color.	420
Adverbial clause modifier (advcl)	Unless the cache is cleared, the link will always stay dark blue.	161
Relative clause modifier (relcl)	Animals are eukaryotes with many cells, which have no rigid cell walls.	49
Appositional modifier (appos)	Almost all animals have neurons, a signalling system.	47
Clausal complement (ccomp)	In Thailand, stingray leather is used in wallets and belts.	67
Open clausal complement (xcomp)	Genes say to the cell what to do, like a language.	70

Category	**Pause**	**Example**
as (preposition)	no	She was hired as the new manager of the team.
as (conjunction)	yes	She left early as she had an important meeting to attend.
for (preposition)	no	The child picked up the toy for his friend who had dropped it.
for (conjunction)	yes	The child picked up the toy for he wanted to play with it.
to (preposition)	no	The man gave the book to his sister who wanted it.
to (infinitive)	yes	The man read the book to learn more about history.
with (preposition, high attach.)	yes	The boy looked at the painting with genuine interest.
with (preposition, low attach.)	no	The boy looked at the painting with muted colors.

Table 6: Example sentences for our evaluation set for the fine-tuning experiments: each function word can be used in two different ways, one of which is associated with a pause.

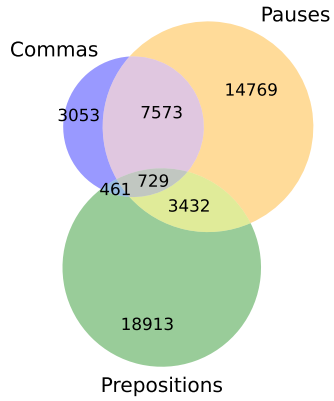

Figure 7: Counts of frequent prepositions, commas and pauses, as well as their overlap, in a subset of the training data of Parler-TTS.