Syntactic Information Content in Word Duration and Pause

Junghyun Min **Ethan Wilcox**
Georgetown University
jm3743@georgetown.edu
https://github.com/Aatlantise/prosody-syntax-interface

Introduction. Speakers use prosodic cues to signal syntax, and listeners rely on these cues to disambiguate syntax (Price et al., 1991). However, the way prosody and syntax interact, as well as the magnitude of information that prosody carries about syntax, remain contested. Direct Reference (DR; e.g. Wagner, 2015) claims that prosody is a direct representation of syntactic structure, Indirect Reference (IR; e.g. Selkirk, 2011) claims that prosody "indirectly refers to" syntax while having its own well-formedness constraints, and Prosody-driven Syntax (PDS; e.g. Richards, 2016) claims that syntax follows prosodic structure. The three theories propose different nature of the interaction.

Methods. Working with English, we measure the amount of syntactic information encoded in two prosodic features–**word duration** and **pause length**. Specifically, we investigate (1) How much syntactic information do these prosodic features convey? (2) Do they carry syntactic information beyond the syntactic informatio carried by words alone? (3) How does this information content vary across speech styles? To investigate, we first introduce three variables: S (Syntax) and two variables that contain information about S: P (Prosody) and W (Text). Following Wolf et al. (2023), we define information content about S in P and W as the amount of uncertainty in predicting S that is reduced when P or W is provided. We use entropy H to quantify uncertainty. Thus, (1)–the information P encodes about syntax–becomes $H(S) - H(S \mid P)$. Furthermore, (2)–the information P encodes about S beyond that W already carries–can be written as $H(S \mid W) - H(S \mid P, W)$.

Our measurement requires four estimations: $H(S)$, $H(S \mid P)$, $H(S \mid W)$, and $H(S \mid P, W)$, where H is the expected value of surprisal, which is equivalent to the average sequence-level language modeling loss in a dataset of sequences representing S. We operationalize S, W, and P as follows: S as linearized constituency parses (e.g. `(ROOT (S (NP NN) ...)` that corresponds to a sequence of words W (e.g. "This is an ...") that contains a sequence of word-level prosodic features P measurable in centiseconds (e.g. `[0.2, 0, 1.1, ...]`). For each estimation, we estimate H by measuring the the language modeling loss on a held-out test set on two datasets: LibriTTS (Zen et al., 2019), a corpus of read-out-loud audiobook recordings and CANDOR (Reece et al., 2023), a corpus of naturalistic conversational speech.

All H are estimated as the sequence-level language modeling loss. $H(S)$ can be measured by autoregressively modeling S. $H(S \mid P)$ can be measured by predicting S given corresponding P with a sequence-to-sequence model; $H(S \mid W)$ by predicting S given W; and $H(S \mid P, W)$ by predicting S given both P and W.

	LibriTTS		CANDOR	
Input	$I(S, P)$	$\mid W$	$I(S, P)$	$\mid W$
Text	**44 ± 0.7**	—	**38 ± 0.4**	—
Paus	0.8 ± 1.1	**0.8 ± 0.2**	**1.6 ± 0.7**	-0.3 ± 0.1
Dur	**9.4 ± 1.0**	-0.2 ± 0.3	**1.8 ± 0.8**	-0.9 ± 0.2

Table 1: Syntactic information content in each input feature. Boldfaced indicates statistical significance.

Results. As illustrated in Table 1, pause and duration carry syntactic information content in both datasets, echoing previous work in the syntax-prosody interface. Duration carries significant syntactic information, but this information is also carried by words in both datasets. This **supports IR** as it argues syntactic information is also contained in lexical identity. Finally, prosody contains greater syntactic information in planned speech than in naturalistic speech, **undermining PDS** that predicts greater shared information between syntax and prosodiy in spontaneous speech, where phonology can affect syntax.

References

P. J. Price, M. Ostendorf, S. ShattuckHufnagel, and C. Fong. 1991. The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90(6):2956–2970.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

N. Richards. 2016. *Contiguity Theory*. Linguistic Inquiry Monographs. MIT Press.

Elisabeth Selkirk. 2011. *The Syntax-Phonology Interface*, chapter 14. John Wiley & Sons, Ltd.

Michael Wagner. 2015. Phonological evidence in syntax. *Syntax–theory and analysis: An international handbook*, pages 1154–1198.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Gotlieb Wilcox, and Tamar I. Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

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. *Preprint*, arXiv:1904.02882.