Revisiting Parity of Human vs. Machine Conversational Speech Transcription

Courtney Mansfield[1], Sara Ng[1], Gina-Anne Levow[1], Richard A. Wright[1], Mari Ostendorf[2]

[1]Linguistics Department, University of Washington, Seattle, WA, USA
[2]Electrical & Computer Engineering Department, University of Washington, Seattle, WA, USA

{coman8,sbng,levow,rawright,ostendor}@uw.edu

Abstract

A number of studies have compared human and machine transcription, showing that automatic speech recognition (ASR) is approaching human performance in some contexts. Most studies look at differences as measured by the standard speech recognition scoring criterion: word error rate (WER). This study looks at more fine-grained analysis of differences for conversational speech data where systems have reached human parity in terms of average WER, specifically insertions vs. deletions, word category, and word context characterized by linguistic surprisal. In contrast to ASR systems, humans are more likely to miss words than to misrecognize them, and they are much more likely to make errors in transcribing words associated primarily with conversational contexts (fillers, backchannels and discourse cue words). The differences are more pronounced for more informal contexts, i.e. conversations between family members. Although human transcribers may miss these words, conversational partners seem to use them in turntaking and processing disfluencies. Thus, ASR systems may need superhuman transcription performance for spoken language technology to achieve human-level conversation skills.

Index Terms: speech recognition, conversational speech, human parity

1. Introduction

Recent advances in deep learning have led to significant reductions in word error rate (WER) in ASR systems, and recent studies have suggested that ASR systems are approaching human levels of performance. This study provides an analysis of conversational telephone speech transcriptions (CTS) generated by an ASR system which has reached human parity according to WER. It extends previous work by providing a fine-grained comparison of ASR errors and errors by human transcribers on the same dataset.

The performance of ASR is measured with the standard scoring criterion of $WER = (I + D + S)/N_r$, i.e., the number of errors relative to the reference length (N_r), including insertions (I), deletions (D) and substitutions (S) in aligning the hypothesis to the reference. Human performance on speech transcription is used to benchmark ASR systems. However, human performance on transcription varies depending on the speaker, the experience and training of the transcriber, and other factors. An LDC study on transcription of the RT-03 evaluation set (76k tokens) reports human WERs of 4.1-4.5% for careful transcription and 9.6% for quick transcription [1]. Transcription in other domains (e.g. broadcast news, interviews, meetings) is cited at 1.3-6.3%. The 2000 Hub5 CTS evaluation dataset [2] is commonly used to benchmark ASR systems [3, 4, 5]. For this data, studies report WERs of 5.1-5.9% and 6.8-11.3% for human transcribers on the Switchboard (21k tokens) and CallHome (22k tokens) portions of Hub5, respectively [3, 5]. The difference in performance between Switchboard and CallHome in these studies illustrates how conversational style also affects transcription error rates. CallHome, which consists of conversations between family and friends, is more informal and is notably more challenging for both human and ASR transcribers. The current study aims to examine some of these differences. Finally, factors such as tokenization can impact reported WER. Mississippi State University researchers corrected conversations from Switchboard (3M tokens) citing a reduction in WER of 8% [6], but a different alignment of the same transcriptions reported a reduction of 5% [7], with differences purportedly due to transcription conventions.

In order to benchmark ASR systems, a particular set of CTS corpora (e.g. Switchboard) is often used. These corpora are well-studied, are not noisy, and are not representative of current US demographics. Differences in speaking style and sociolinguistic factors have been shown to markedly affect ASR performance. One recent study on transcriptions of structured interviews illustrates ASR performance disparities with regards to race [8]. Comparing several state-of-the-art ASR systems, WER was on average more than 80% higher for Black speakers relative to the WER for white speakers. Another study [9] considers emotional speech, showing that state-of-the-art systems have significantly higher WERs on scripts read with emotional readings vs. emotionally-neutral readings. Our analysis highlights differences in terms of conversational style. We explore the 2000 Hub5 CTS evaluation set which includes conversations between strangers (Switchboard portion) and between familiar friends and family members (CallHome portion). Understanding differences in performance across various speech styles can help ASR to perform more evenly in a variety of contexts.

A number of studies directly compare ASR performance to human transcription performance. One such study [3] examining an ASR system which claimed to reach human parity in 2016, highlights similarities between human and machine recognition performance. They compare the most frequent word types from ASR and human errors. However, error counts ranged from 45 to as few as 4 tokens, and these tokens are also high-frequency words in English. Corpus frequency may account for these findings; in fact, previous work has described a linear relationship between the number of transcription errors and their log frequency [7]. The current study extends the results on this data by providing a more fine-grained analysis of human vs. machine errors considering style differences associated with familiarity. We use word classes normalized by corpus frequency to account for this effect, motivated by earlier linguistic analyses of ASR errors in CTS [10].

In the remainder of the paper, we detail the transcription data and extraction of features (Section 2). We compare features (e.g. word category, frequency, and surprisal) related to human and machine recognition errors, and consider differences in conversational styles (Section 3).

Table 1: *Comparison of WER (%) from Switchboard (SWBD) and CallHome (CH) portions.*

| | ASR | | Human | |
	[3]	This Work	[3]	This Work
SWBD	6.1	6.9	6.2	7.0
CH	13.6	12.9	13.3	14.0

2. Data

We use reference data from the NIST 2000 Hub5 CTS English evaluation dataset [11]. The Hub5 evaluations include 20 conversations each from Switchboard [12] and CallHome [13, 14]. Callhome and Switchboard participants are native English speakers from various dialect regions in the US. The same human and ASR transcripts have been used in previous work [3]. Human transcripts are generated with Microsoft's production transcription pipeline. The audio is transcribed and error corrected in a second pass. ASR transcripts are produced with Microsoft's DNN-HMM speech recognition system, as described in [15]. The acoustic training data uses 2000 hours of CTS; language models use CTS data, LDC Broadcast news data, and data from the University of Washington conversational Web corpus.

2.1. Data cleaning

The transcriptions underwent cleaning and normalization before scoring. Our normalization closely follows the 2000 NIST evaluation plan.[1] Hyphenated words in the transcripts are expanded to multi-word forms (e.g. 'mother-in-law' as 'mother in law'). Some abbreviations are normalized to better match the reference set by including whitespace (e.g. 'dj' to 'd j', 'phd' to 'p h d'). Stylistic differences are normalized, including reductions ('gonna' vs. 'going to') and compound words ('everyday' vs. 'every day'). Acknowledgements and backchannels are grouped in a category '%bcack', and hesitations in '%hes', according to the 2000 NIST evaluation plan. Fragments can be ignored or included in a transcription (e.g. 't-' hypothesized 'the' or 'those') without penalty. Normalization scripts have been made available to the public.[2]

The WER for the normalized data are presented in Table 1 alongside the WER initially reported in [3]. We were unable to replicate the WER from prior work, as their data cleaning process was not detailed. Differences likely result from choices about normalization. One possible difference is our choice to not tokenize contractions, but treat them as a single (and psycholinguistically real) unit of the lexicon. This leads to a difference in the word counts of the reference. The reference data in [3] contain 21.6K and 21.4K tokens from CallHome and Switchboard respectively. Our reference contains 20.2K tokens each for CallHome and Switchboard. To compare the human and machine WERs, we run standard significance tests with the NIST scoring toolkit (SCTK).[3] Wilcoxon tests and Matched Pairs Sentence Segment Word Error tests do not show significant differences between the WERs of humans and ASR, as in [3]. A Sign test at the utterance level finds differences favoring the ASR system.

2.2. Word category

The word category for each token is found by automatically tagging tokens and mapping each tag to a broad word class. Three categories are used: function (closed-class) words, content (open-class) words, and conversational words (e.g. filled pauses, backchannels, and discourse cues).[4] For tagging, a classifier from the Neural Sequence Labeling Toolkit [16] is trained on human-annotated POS tags from the Switchboard section of the Penn Treebank [17]. The tagger has an accuracy of 96.4% on held-out test data from Switchboard. Because the class of some POS tags can be ambiguous (e.g. adverbs 'sometime' vs. 'happily'), a list of function word tokens is also used to supplement the mapping of POS tags.

2.3. Language model context scoring

The **log unigram probability** (word frequency) of each token is calculated to examine possible differences between the distributions of the human and ASR errors. Word frequencies are estimated with a unigram model trained on 17k tokens from the Fisher (CTS) corpus [18].

To consider how word context and the predictability of errors compares across human and machine transcripts, we measure the difference in **linguistic surprisal**. Linguistic surprisal aims to directly capture sentence processing difficulty in an incremental fashion [19, 20] and is associated with empirical metrics such as reading time [21, 22]. Surprisal is the log-inverse of the conditional probability of a word w_i given history $c(w_i)$:

$$H(w_i) = -\log\ p(w_i|c(w_i)) \tag{1}$$

We compute surprisal over sequences of errors from the hypothesis and corresponding reference text. Error sequences are bracketed by error-free forward and backward contexts to ensure the conditioning events in hypothesis and reference are comparable. The surprisal of sequences is normalized by length. Then, the difference between the hypothesis and reference surprisal is computed. This is expressed as:

$$H(w_1 w_2 ... w_n) = -\frac{1}{n}\sum_{i=1}^{n}\log_2 p(w_i|c(w_i)) \tag{2a}$$

$$\Delta H = H(w_1^r ... w_{n_r}^r) - H(w_1^h ... w_{n_h}^h) \tag{2b}$$

where $p(w_i|c(w_i))$ is the probability given by a language model, and w_i^r and w_i^h are the ith tokens of the reference and hypothesis sequences. A positive surprisal difference indicates that the sequence in the reference transcription is more surprising than the error sequence.

The conditional probabilities are generated with a Gated Recurrent Unit Network (GRU) [23] using PyTorch [24]. The model consists of 2 hidden layers, 256-dimensional word embeddings, 128-dimensional hidden layers and has a 0.2 dropout rate [25]. The model is trained on 14K tokens from Fisher with a count threshold of 10. Training data includes disfluencies. The model shows a perplexity of 84.0 on Switchboard and 107.3 on CallHome.

3. Analysis of recognition errors

3.1. Composition of errors

We first consider the distribution of insertions, deletions, and substitutions in the transcriptions. Wilcoxon signed-rank tests

---

[1] https://mig.nist.gov/MIG_Website/tests/ctr/2000/h5_2000_v1.3.html
[2] https://github.com/cmansfield8/human-parity
[3] https://github.com/usnistgov/SCTK

[4] A detailed description of the word category mapping is available at https://github.com/cmansfield8/human-parity.
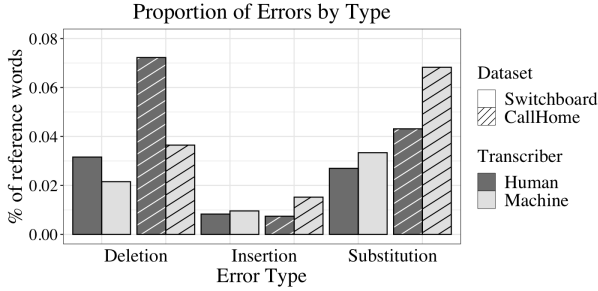
Figure 1: *Proportion of errors of each type as a total of the reference tokens.*

at the utterance level consider differences in counts of each type of error. In both Switchboard and CallHome, there are significant (p<0.0001) differences between the number of deletions and substitutions made by humans and machines. Humans are more likely to miss (delete) tokens, while ASR is more likely to misrecognize (substitute) tokens. ASR inserts significantly (p<0.0001) more tokens than human transcribers in CallHome only. Figure 1 shows insertions, deletions, and substitutions as a proportion of all errors. There are notable differences in the spreads between CallHome and Switchboard. Differences in errors are greater in magnitude in the CallHome corpus portion compared to Switchboard. For instance, there are 49% fewer ASR than human deletions in CallHome compared to 31% in Switchboard.

The more even balance of insertions/deletions in machine errors is likely an artifact of tuning the insertion/deletion trade-off to minimize WER. Humans, however, skew towards missing rather than inserting tokens. Given noisy input, it is possible that human transcribers choose to conserve energy and evade a transcription. Interactions between word category will shed more light on this topic in the following sections.

3.2. Word category

Figures 2 and 3 present the proportion of errors in each word category relative to their frequencies in the reference. In general, the largest differences between humans and ASR are seen in the CallHome data, with particularly large differences associated with conversational words.

Function words are greatest in terms of absolute number of errors, as seen in Table 2. However, function words typically have the lowest error rates relative to their overall frequency in the corpus. The discrepancies between rates of function word insertion, deletion, and substitution between humans and machines are generally smaller than that of content and conversational words. The largest differences are seen in deletions, where ASR is 40% and 23% less likely to delete function words in CallHome and Switchboard, respectively.

Following the general pattern of insertions and deletions, humans are less likely to insert and misrecognize but more likely to delete content words. The CallHome portion of the data shows particularly large differences in terms of content word errors, with ASR being 199% more likely to insert and 80% more likely to misrecognize such words compared to human transcribers.

While conversational word errors are fewest in absolute terms, they are the most likely errors relative to corpus frequency. Humans miss 8% of conversational words in Switchboard and over 13% in CallHome. Conversational words also

Table 2: *Counts of the lexical categories for sequences associated with errors in the hypothesis and reference. The 'Other' category consists of mixed lexical categories other than function-content.*

| | Hypothesis | | Reference | |
	Human	ASR	Human	ASR
All function	582	699	963	981
All content	315	417	427	482
All conversational	164	275	395	336
Function-content	47	78	183	101
Other	6	20	57	26

Table 3: *Wilcoxon signed-rank test of conversational-type error counts for paired human and machine transcripts. A positive stat (Z-value) reflects higher machine error counts.*

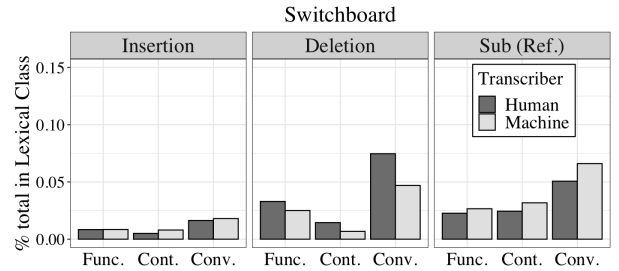| | Switchboard | | CallHome | |
	Stat	P-value	Stat	P-value
Insertion	0.49	NS	1.44	NS
Deletion	-4.95	p < 0.0001	-9.56	p < 0.0001
Sub (hyp.)	3.43	p < 0.001	7.66	p < 0.0001
Sub (ref.)	2.49	p < 0.05	4.44	p < 0.0001

Figure 2: *Proportion of errors as a total of each lexical class in the Switchboard portion of the data. Word categories include function words (func.), content words (cont.), and conversational words (conv.).*

result in the largest discrepancies between humans and ASR, especially in CallHome. ASR misses 65% fewer and misrecognizes 81% more conversational words than human transcribers. In Table 3, a Wilcoxon signed-rank test examines the counts of conversational word errors paired at the utterance-level. While rates of insertion are similar, there are significantly fewer deletions and more substitutions made by ASR.

Table 2 shows counts of word categories for sequences of errors. Sequences represent one or multiple neighboring tokens related to errors. Hypothesis sequences include insertions and the hypothesis of a misrecognized word, while reference sequences include words which were missed and misrecognized. While most sequences can be represented by a single word category, some span multiple categories. Function/content words are the most common 'mixed' category.

3.3. Language model probabilities

The log unigram probabilities of error tokens from human and machine transcripts are compared to the distribution of all ref-
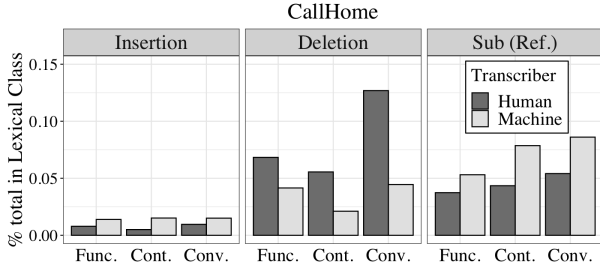
Figure 3: *Proportion of errors as a total of each lexical class in the CallHome portion of the data.*
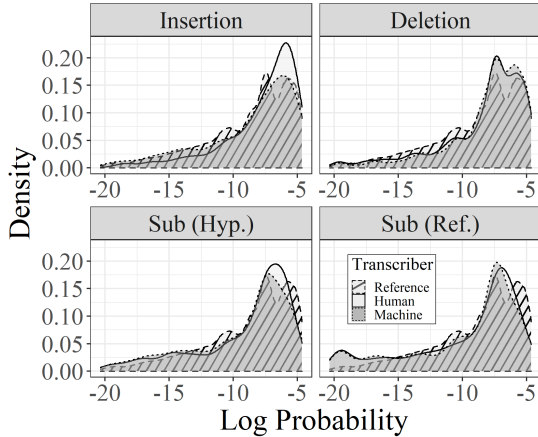

Figure 4: *Log unigram probabilities of words in errors made by humans and machines, vs unigram probabilities of all reference tokens.*
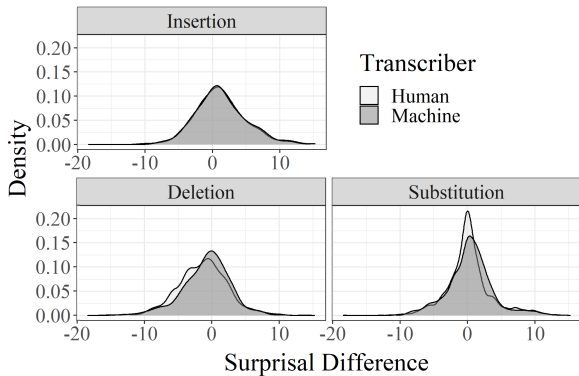

Figure 5: *Surprisal difference of error sequences from humans and ASR. Positive surprisal difference indicates reference is more surprising than hypothesis.*

erence tokens in Figure 4.[5] There is a notable difference between human and machine insertion errors. Human insertions are particularly dense between -5 and -7. This corresponds to very high-frequency function words (e.g. 'a' and 'the') that lead to more grammatical utterances when inserted. ASR insertions have a longer tail of low-frequency words. These words typ-

ically correspond to areas of misrecognition where tokens are both inserted and substituted (e.g. the hypothesis 'jazz exercise' for reference 'jazzercise'). The distribution of deletions is similar between human and machine transcripts. In substitutions, humans are more likely to hypothesize low-probability words.

Differences between the surprisal of hypothesis and reference sequences are shown in Figure 5. A positive surprisal difference reflects hypothesized words being more expected than the reference, on average. There are notable distribution differences in deletions and substitutions. In deletions, humans have more mass distributed in the negative range, particularly around -2 to -7. This mass is explained in part by isolated hesitations and backchannels. Humans were 3.5 times more likely to miss these tokens than ASR. The distribution of substitutions also varies between humans and ASR. The surprisal difference for ASR substitutions is higher ($\bar{x}=0.38$) than that of human substitutions ($\bar{x}=0.23$). This may be due to the fact that the surprisal language model is similar to that used by the ASR system and humans are using their world knowledge.

4. Conclusion

In summary, our analysis shows fine-grained differences between recognition errors in human and machine transcriptions on CTS data, although average WERs are similar. Human transcribers are more likely to delete while the ASR system more often inserts and misrecognizes words. In addition, there are clear differences in terms of recognition performance between Switchboard and CallHome, with its particularly informal style. Not only did both human transcribers and ASR systems produce less faithful transcriptions of CallHome, as has been noted in previous studies, but our work shows that systematic differences between human and ASR transcriptions are more extreme.[6] This has implications for building more robust ASR systems which are capable of adapting to different conversational contexts. Humans are able to dynamically adapt to both global and local changes in context. In order for ASR systems to achieve parity on human performance, they must be able to adapt to a variety of contexts.

Conversational phenomena have the highest rates of error and there are particularly large discrepancies in processing these words. Human transcribers are likely to miss conversational words, which may reflect that conversational words are processed differently by listeners. While listeners may not be consciously aware of these words, they have been shown to help aid in speech processing, for instance, allowing a listener to identify novel referents [26] or process disfluencies [27] more efficiently. Listeners use conversational words as cues, although they may not accurately transcribe such words. By contrast, ASR must accurately recognize these cues in order for a dialogue system to make use of them. For spoken language technology to reach human-level conversational skills, WER may need to surpass that of human transcribers.

5. Acknowledgements

---

[5]The reference distributions show a secondary peak around -7.5, corresponding to some backchannels and hesitations, transcribed and distributed differently between the Hub5 data and Fisher corpus LM.

[6]It has been noted that Switchboard results are optimistic because of the overlap of speakers in the training and test sets. Since our findings show that overall rates of human errors are comparable to ASR for both corpora, we expect that the different findings reported here are mainly related to the different speaking style.

6. References

[1] M. L. Glenn, S. M. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription methods for consistency, volume and efficiency," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC10)*. Valletta, Malta: European Language Resources Association (ELRA), May 2010. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2010/pdf/849_Paper.pdf

[2] M. Przybocki and A. Martin, "2000 NIST Speaker Recognition Evaluation: LDC2001S97," 2001.

[3] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription," in *Proc. Interspeech*. ISCA - International Speech Communication Association, 8 2017, pp. 137–141.

[4] W. Xiong, J. Droppo, X. Huang, F. Seide, M. L. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[5] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim *et al.*, "English conversational telephone speech recognition by humans and machines," *Proc. Interspeech 2017*, 2017.

[6] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of Switchboard," in *Fifth International Conference on Spoken Language Processing*, 1998.

[7] V. Zayats, T. Tran, R. Wright, C. Mansfield, and M. Ostendorf, "Disfluencies and human speech transcription errors," in *Proc. Interspeech 2019*, 2019, pp. 3088–3092. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3134

[8] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[9] R. Munot and A. Nenkova, "Emotion impacts speech recognition performance," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 2019, pp. 16–21.

[10] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[11] J. Fiscus, W. M. Fisher, A. F. Martin, M. A. Przybocki, and D. S. Pallett, "2000 NIST evaluation of conversational speech recognition over the telephone: English and Mandarin performance results," in *Proc. NIST Speech Transcription Workshop*, 2000, pp. 1–5.

[12] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1. IEEE, 1992, pp. 517–520.

[13] A. Canavan, D. Graff, and G. Zipperlen, "Callhome American English speech," *Linguistic Data Consortium*, 1997.

[14] P. Kingsbury, S. Strassel, C. McLemore, and R. McIntyre, "CALLHOME American English Transcripts," *University of Pennsylvania: Linguistic Data Consortium*, 1997.

[15] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016.

[16] J. Yang and Y. Zhang, "NCRF++: An open-source neural sequence labeling toolkit," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. [Online]. Available: http://aclweb.org/anthology/P18-4013

[17] A. Taylor, M. Marcus, and B. Santorini, "The Penn treebank: an overview," in *Treebanks*. Springer, 2003, pp. 5–22.

[18] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: a resource for the next generations of speech-to-text," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA), May 2004. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/767.pdf

[19] J. Hale, "A probabilistic Earley parser as a psycholinguistic model," in *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, 2001, pp. 1–8.

[20] R. Levy, "Expectation-based syntactic comprehension," *Cognition*, vol. 106, no. 3, pp. 1126–1177, 2008.

[21] B. Roark, A. Bachrach, C. Cardenas, and C. Pallier, "Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 2009, pp. 324–333.

[22] I. F. Monsalve, S. L. Frank, and G. Vigliocco, "Lexical surprisal as a general predictor of reading time," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 398–408.

[23] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[24] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[25] V. Pham, T. Bluche, C. Kermorvant, and J. Louradour, "Dropout improves recurrent neural networks for handwriting recognition," in *14th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2014, pp. 285–290.

[26] J. E. Arnold, M. Fagnano, and M. K. Tanenhaus, "Disfluencies signal theee, um, new information," *Journal of psycholinguistic research*, vol. 32, no. 1, pp. 25–36, 2003.

[27] S. E. Brennan and M. F. Schober, "How listeners compensate for disfluencies in spontaneous speech," *Journal of Memory and Language*, vol. 44, no. 2, pp. 274–296, 2001.