Disfluencies and Human Speech Transcription Errors

Vicky Zayats[1], *Trang Tran*[1], *Richard Wright*[2], *Courtney Mansfield*[2], *Mari Ostendorf*[1],

[1]Electrical & Computer Engineering Department, University of Washington
[2]Linguistics Department, University of Washington
{vzayats,ttmt001,rawright,coman8,ostendor}@uw.edu

Abstract

This paper explores contexts associated with errors in transcription of spontaneous speech, shedding light on human perception of disfluencies and other conversational speech phenomena. A new version of the Switchboard corpus is provided with disfluency annotations for careful speech transcripts, together with results showing the impact of transcription errors on evaluation of automatic disfluency detection.

Index Terms: spontaneous speech, perception, disfluencies

1. Introduction

Human errors in transcribing spontaneous speech provide insights into perception of speech and how humans process spoken language. In this study, we investigate mistranscriptions of spontaneous speech, particularly the co-occurrence of errors (or misperceptions) with disfluencies and coarse-grained word classes. The motivation is twofold. First, the analysis of errors can shed light on human perception of disfluencies and other spontaneous speech phenomena in conversation. Second, a better understanding of where errors occur improves our ability to interpret assessments of automatic algorithms against human transcriptions and associated findings.

Much of the work on misperception has been on read speech in a laboratory environment, where lexical items and grammatical structure can be controlled for. Here, we study misperception of spontaneous speech, taking advantage of two versions of transcriptions associated with the Switchboard corpus. Our work provides new insights by averaging statistics over many instances of different word categories. This allows us to look at phenomena such as disfluencies, where we find higher rates of misperception (assuming mistranscription indicates misperception).

The Switchboard Corpus [1, 2] is a large collection of conversational telephone speech that has been annotated for a number of different types of linguistic structure. It is associated with two sets of human transcriptions. The second represents a careful correction of the first, but much of the linguistic annotation is based on an earlier release.

We pose two hypotheses. First, we expect that words which carry more information are less likely to be misperceived, since they tend to be more clearly articulated. Second, we anticipate that disfluent regions will be disproportionately associated with misperceptions, particularly repetition disfluencies. In order to carry out the second analysis, it is necessary to revise the existing disfluency annotation for the more careful transcriptions. We describe an automatic procedure that gives high quality results and make the resulting annotations publicly available.

The paper makes three main contributions. First, we present distributional analyses of the location of transcription errors that confirm our hypotheses, but also point to informal words used in conversation as a high error category. Second, to support the analysis, a new version of disfluency annotated Switchboard is released. Finally, experiments on disfluency detection show that transcription errors impact performance estimates mainly for repair disfluencies.

2. Human Speech Transcription

With the fast improvements in ASR system performance, there have been multiple studies that evaluate human transcription accuracy on a variety of datasets. One of the early studies on human transcription [3] reports error rates across multiple corpora with different level of difficulty and vocabulary sizes ranging from 10 to 65k words. The human transcription rates in this study varies from 0.1% (on transcribing 10 digits) to 7.4% (keyword spotting) depending on the task and corpora. The study also estimates the error rate for the Switchboard corpus to be 4%, although the reference attributed this number to "personal communication." Later, independent studies [4, 5, 6] re-evaluated this number using professional transcribers, reporting error rates of 5.1-5.9% and 6.8-11.3% on subsets of the Switchboard (21k words) and CallHome (22k words) corpora, respectively, which are part of NIST 2000 CTS evaluation set. The differences in error rate correspond to how careful the transcribers are, with a quality checking pass improving the average error rate by 5-20% [6]. Another study by LDC on quality of the transcription [7] reveals huge differences in the transcription error rate between very careful (4.1-4.5%) and quick transcription (9.6%) evaluated on the RT-03 evaluation set, which contains subsets of Fisher and Switchboard datasets (76k words in total). The authors report that with very careful transcriptions 95% of annotation discrepancies between multiple transcribers are "judgment calls" due to contractions, rapid or difficult speech, or disfluencies. The authors also notice that with the quick transcriptions, the regions of disfluency are by far the most prevalent contributors to transcriber disagreement across the different languages used in the study. Our study confirms these observations with quantitative analysis, but also shows high errors for other spontaneous speech phenomena. Transcription errors in genres other than conversational telephone speech (broadcast news, broadcast conversation, interviews, and meetings) are approximated at rates 1.3-6.3% in English, 6.1-9.5% in Chinese and 3.1-8.3% in Arabic.

When trying to analyze and compare human and machine errors [6, 8] using the NIST 2000 CTS dataset, both papers report function and backchannel words being dominant word categories labeled as errors, though due to a limited data size the statistics on frequent word associated with errors can be unreliable. For example, the most common insertion (by human transcribers) is token "i" with only 10 occurrences [9]. In comparison, our work explores transcription errors using a large scale dataset (1.3M words), which allows identification of more reliable patterns and more fine-grained analysis of error contexts.

3. Data

Switchboard I – Release 2 [1, 2] is a collection of about 2,400 telephone conversations between strangers, of which 1126 conversations were hand-annotated with disfluencies as part of the Penn Treebank Release 3 dataset [10]. Because human transcribers are imperfect, the original transcripts contained errors, some of which were corrected in the Treebank release. Mississippi State University researchers ran a clean-up project which hand-corrected conversations, which we will refer to as MsState transcriptions, and produced alignments between the transcripts indicating the type of errors (missed, extra, or substituted) [11]. They did not re-annotate disfluency or parse structure.

The MsState transcription guidelines were designed for higher consistency, so some of the transcription "errors" reflect a difference in transcription guidelines. In particular, the MsState transcription guidelines differ from the original in asking the transcriber to more faithfully represent the spoken version, e.g. by allowing more variants of words (e.g. "naw" for "no," "gonna" for "going to," and "um-hum" as well as "uh-huh"), encouraging (rather than discouraging) use of contractions, having explicit conventions for pronunciation variants and mispronunciations, and in the form for transcribing word fragments ("w[ent]-" vs. "w-"). The conventions for handling word fragments also differ in terms of conventions for using "I-" in a repetition, and in asking the transcriber to include the fragment even if they do not know what the speaker intended, which leads to a higher rate of word fragments in the MsState transcripts. These differences inflate the substitution error rate, and some are therefore ignored in our analyses.

The original work documenting the segmentation and transcription correction effort [11] states that the human transcription word error rate is reduced from approximately 10% to 2%. Our analysis on the disfluency-annotated subset shows a word difference rate of 5%, using the standard error rate calculation (insertions + deletions + substitutions/total number of words in the MsState transcript), with 2.4% associated with substitutions. (If contractions are split, as in many language processing studies, the word difference rate is 5.2%, with 2.6% substitutions.) The smaller error rate may be due in part to the fact that we did not count differences due to transcription conventions as errors, e.g. ("i-" vs "I") and differences associated with the CONT transcription, used e.g. for acronyms. It is also consistent with other studies described in the previous section. If the 10% error rate is based on the original release of the Switchboard transcripts, then there would also be a difference related to the fact that there were some corrections in the Treebank release.

This error rate includes many word fragments: the MsState transcripts contain roughly 2.5 times more fragments than the Treebank transcripts. Ignoring the single-phone word fragments, which represent roughly 75% of the fragments added in the retranscription, the error rate is 4.7% (5.0% with split contractions). As expected, word fragments are more often missed than inserted: 12.4% vs. 7.2%, respectively (11.8% vs. 7.0% for split contractions).

3.1. Automatic Mapping of Disfluency Annotations

A goal of this research was to align the MsState speech transcripts (for which there are more careful transcripts and good time alignments) with disfluencies that had been hand-annotated on an earlier (less faithful) version of the transcripts. In order to transfer the disfluency annotations to the MsState transcripts, as in the example:

Transcript	Annotation
Treebank	also the [whole + whole] thing
MsState	also the [whole {DEL the } + whole] thing
Mapped	also [the whole + the whole] thing

we used a multi-step process that leverages our previous work on transcript differences [12] and avoids a costly hand-annotation process. Each word in the original Treebank annotation was associated with a disfluency label based on a begin-inside-outside (BIO) tagging scheme that accounted for both reparandum and correction spans following [13]. Using the MsState alignments, we automatically inserted, deleted and substituted words in the disfluency transcripts. For each inserted or substituted words and the window of ± 2 neighbors, two additional (temporary) labels were used: 'D' for representing any state that correspond to being part of disfluency (either reparandum or repair), and 'A' which would allow any state. In addition, we used the 'D' state for words surrounding deletions that were originally annotated as disfluencies, assigned non-disfluency state 'O' for words surrounding insertions that were originally annotated as non-disfluencies, and assigned 'A' for all the rest. Then, we ran automatic disfluency detection with integer linear programming constraints [14, 13] for assigning BIO labels to the words associated with 'A' and 'D' labels. This approach allowed us to identify and add missed disfluencies and remove hallucinated disfluencies. We refer to the resulting annotations as "silver" annotations, since they are strongly constrained by the original hand annotations. While some errors are introduced in this process, most transcription errors are short and isolated, so the constraints of labels on neighboring words are reasonably strong. As discussed in the next section, analysis of a subset of the test set shows that the mapped labels are quite good. We also show, in Section 5, that the automatically corrected data used in training a disfluency detection model (vs. the Treebank annotations) leads to better performance.

3.2. Quality of Automatic Mappings

To assess the quality of the automatically mapped data and the improvements in the mapping associated with the new disfluency model, we initially selected 100 test sentences[1] to hand correct for disfluencies. In annotating these, we observed that there were sentence segmentation alignment errors in some of the cases for which missed words occurred at sentence boundaries. We therefore selected additional sentences to hand annotate, separately computing statistics for those that included missed words at boundaries and those that did not.

Roughly 15% of the 100 test sentences appeared to have missed words associated with a segmentation boundary. Because the available alignments for between the MsState and Treebank transcripts did not preserve sentence boundaries, a mechanism is needed to align deleted (missed) words to sentences. Initially, we arbitrarily assigned these to the sentence following the boundary. Examining 63 sentence pairs for possible boundary alignment problems, we found that 27% involved assignment errors. Of the 17 sentences that had errors, two simple reassignment rules related to unintelligible regions and backchannels addressed 13 cases. Using these rules, the full data set was reprocessed, and the problems seemed to be minimal in the subsequent hand annotation effort.

After automatic refinement of sentence segmentation, we annotated additional sentences, resulting in 453 sentences in to-

[1]For simplicity, we use the term "sentence" rather than sentence-like unit or slash unit (SU), which are sometimes used since they do not always have the complete grammatical structure of written sentences.

tal. For this set, we compared the mapped silver annotations to the gold annotations. The reparandum labels are very good with F-score of 90.1 (90.1 precision, 90.1 recall). The silver interruption points have F1 90.6 (89.4 precision, 91.8 recall). It is difficult to use standard disfluency detection scoring to characterize the quality of the original Treebank transcriptions because of the word sequence differences. However, we can easily assess the detected interruption points, and we find that the original Treebank annotations have much lower quality, with F1 79.7 (precision 88.9, recall 72.2). Most of the difference is in recall, consistent with the hypothesis that many transcription errors are in disfluent (reparandum) regions.

4. Analysis of Transcription Errors

In the analyses below, we distinguish between two types of misperceptions: i) a word that appears in the careful MsState transcript but does not appear in the Treebank transcript, referred to as a 'miss,' and ii) a word that appears in the Treebank transcript but not in the MsState transcript, referred to as a 'hallucination.' For differences in the two transcripts that correspond to a substitution error, the word that is in the Treebank transcript is counted as a hallucination, and the word that is in the MsState transcript is counted as a miss.

With these definitions, 4.7% of the words in the MsState transcripts are missed in the Treebank version, and 3.2% of the words in the Treebank transcript are hallucinations. Over 80% of the hallucinations are substitutions; i.e., words are misheard rather than invented. Of the missed words, 55% are substitutions. Ignoring substitutions, it is four times more likely for a word to be missed than invented. Words that are invented are often grammatical corrections, e.g. (hallucinations in brackets)

she had to be put in [a] nursing home
[it] was good talking to you.
you talked about the telephone calls [and] people coming and soliciting [and] selling things

4.1. Word Category

We first looked at misperceptions depending on word category, classifying words as lexical, function, fragment, and other. The "other" category comprises words that are characteristic of conversational speech (vs. written text), including words that function as backchannels (uh-huh, um-hum, huh, ...), filled pauses (um, uh), interjections (oh, ooh), and single word responses that can play the role of a backchannel (yeah, nope, huh, nah). Our hypothesis was that these categories would differ substantially in the tendency for transcribers to miss or hallucinate the words.

Figures 1 and 2 show the log relative frequency of missed and hallucinated instances of each word, with different colors/symbols indicating the word category. The blue line shows the linear regression fit to the function word statistics for words with log frequency greater than -9, showing that on a log scale, the relative frequency of a particular word being missed or hallucinated is generally proportional to the frequency of that word, for both content and function words. Word fragments follow a similar trend with an offset associated with an overall higher rate, which is primarily due to transcription substitutions, e.g. 'th-' vs. 'thi-'. Defining an outlier as a word where the difference from the function word prediction is more than 5 times the function word RMS error, only 2 function words and 4 content words are outliers. The exceptions involve transcription conventions associated with reduction ('wanna' vs. 'want to', 'gonna' vs. 'going to', 'till' vs. 'until'), substitutions with phonetically similar words ('time' vs. 'type'), or are frequently in a high error context (adjacent to a disfluency).

The "other" words seem to behave differently, with atypically high frequency of being misperceived. Of the 22 "other" words with log frequency greater than -9, 10 are outliers. Some cases involve transcription conventions ('yep' vs. 'yes'), but other errors reflect different meanings ('uh-huh' vs. 'hm'). The results also show differences between the two filled pauses: 'um' is infrequently missed and almost never hallucinated, unlike 'uh'. This is consistent with previous observations that the two filled pauses tend to be used differently.

In general, it appears that spontaneous speech phenomena – fragments, backchannels and interjections – are associated with higher error rates. This may be because people are not consciously aware of these phenomena (though they may unconsciously use them in interpreting the intent of the interlocutor), and thus they need more training to transcribe them. It has been observed that transcript errors on 'um' and 'uh' are substantially higher on average for those people who transcribed only a few conversations compared to those who had transcribed a large number [15]. We hypothesize that the same will be true for disfluencies in speech that involve repetition or correction.

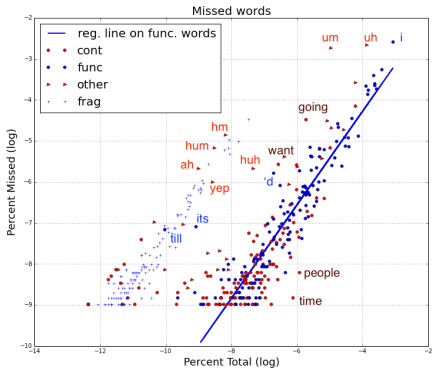

Figure 1: *Log relative frequency of missed instances of different words compared to their overall frequency in the corpus, distinguishing between function, content, fragment, and other.*

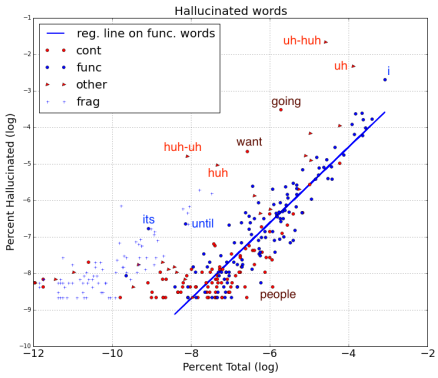

Figure 2: *Log relative frequency of hallucinated instances of different words compared to their overall frequency in the corpus, distinguishing between function, content, fragment, and other.*

Table 1: *Relative frequency of different disfluency types and the PMI associated with different reparandum word error categories ($m*$ = complete miss, m = miss, h = hallucinate).*

		PMI(x, y)		
x	$P_M(x)$	$(x, m*)$	(x, m)	(x, h)
restart	0.003	0.49	0.64	0.60
repetition	0.032	0.61	0.41	0.22
repair	0.025	0.70	0.48	0.22
complex	0.003	0.61	0.40	0.54
fluent	0.879	-0.19	-0.08	-0.01

4.2. Disfluent Regions

We hypothesized that transcribers would be more likely to both miss and mistranscribe words in the reparandum of a disfluency. In Table 1, we provide the overall rate (relative frequency) of words associated with the different disfluency types x, together with the pointwise mutual information (PMI) of types and different error categories: PMI$(x, c) = \log[P(x|c)/P(x)]$. PMI greater than zero implies that the transcription error is more likely to occur in region x than would be predicted by its overall relative frequency. Relative frequencies of words occurring in the reparandum of different disfluency types (P_M) are computed based on the MsState transcripts using the mapped disfluency labels. For most disfluency types, words that are completely missed (excluding substitutions) are even more likely to be in a disfluency reparandum.

The results show that transcription errors of all categories are more often associated with disfluent regions than with fluent speech, consistent with the overall rate of disfluencies being slightly higher in the MsState transcripts than in the Treebank transcripts (12.1% vs. 10.8%, respectively). As expected, the effect is stronger for missed words than hallucinations. The high rate of errors in restarts may be indicative of these being less attended to by human listeners or a higher incidence of fragments here, but this is also the category where inter-annotator agreement are least reliable.

While words seem to be more frequently missed in the reparandum of a disfluency, it may be that some disruption is still perceived. For example, if the repetition *'I I I'* is transcribed as *'I I,'* it is clear that the transcriber perceived a disfluency. Hallucinations of disfluent regions are much less frequent, based on the PMI analysis above and the analysis of interruption point errors in Sec. 3.2. The PMI results and analysis of gold annotations suggest that hallucinations associated with disfluencies most often involve restarts.

5. Disfluency Detection Experiments

The experiments in this section leverage a text-based neural disfluency detection system described in [16]. The model uses multiple levels to process a sentence: the first level calculates similarities for each word with words in the surrounding window; the second level uses the similarities as input to a convolutional neural network and max-pooling layer that learns local pattern matches; then we flatten the resulting outputs and concatenate with the word embeddings; and finally the resulting vector is input to a bidirectional LSTM-CRF.

We train two versions of the model (one using the original Switchboard transcripts and disfluency annotations, and the other using the corrected transcripts with silver mapped annotations) and assess performance on the original and silver versions of the full test set plus a subset that has been fully hand-corrected (gold). The results are presented in Table 2.

Table 2: *Disfluency detection results training and testing on different versions of the annotations*

		Training Data	
Test set	Transcript	Original	Silver
Full	Original	87.80	87.23
	Silver	88.67	86.96
Gold	Original	88.69	88.54
Subset	Silver	89.17	87.00
	Gold	88.69	89.73

The result with the original training and test transcripts is comparable to other reported results.[2] In the gold subset, comparing the standard configuration to testing on gold transcripts with silver training gives an indication of the noise in performance estimates associated with less careful transcripts, and it shows that prior work under-estimates performance a little. Using the silver test transcripts also under-estimates performance. The differences are small, but the more careful annotation may be useful in experiments that leverage acoustic cues.

6. Summary

In summary, this study has shown that human transcribers tend to misperceive words proportionately to the log frequency of those words, confirming the hypothesis that words which carry more information are less likely to be misperceived. Notable exceptions include disfluencies and words that are characteristic of spontaneous speech (filled pauses, interjections, and backchannels), which are all misperceived at higher than expected rates. The higher error rates may be due to low information load of these words and/or lack of conscious awareness of spontaneous speech phenomena. Lack of awareness would explain the need for annotator training. Further study is needed. To support this and future analyses, this work has provided a new version of Switchboard disfluency annotations. These annotations support more exploration of prosodic cues and disfluency detection [17].

This work was motivated in part by a prior study showing that transcription errors impact findings related to the usefulness of prosodic features in parsing [18], i.e., a significant fraction of the cases where prosody seems to hurt parsing are associated with transcription errors. The availability of the new disfluency annotations will make it possible to explore this question for disfluencies.

For speech recognition applications, having high accuracy transcriptions does not seem to be critical. However, in spoken language processing and translation, disfluencies can impact performance. In addition, there are medical and educational applications where detected disfluencies may provide useful information about the speakers cognitive state.

Acknowledgements

This work was funded in part by the US National Science Foundation, grant IIS-1617176. Any opinions, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

[2]Direct comparisons are difficult because tokenization is not standardized among the different studies.

7. References

[1] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 517–520.

[2] J. J. Godfrey and E. Holliman, *Switchboard-1 Release 2 LDC97S62*, Linguistic Data Consortium, 1993.

[3] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.

[4] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[5] 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 (TASLP)*, vol. 25, no. 12, pp. 2410–2423, 2017.

[6] 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," *arXiv preprint arXiv:1703.02136*, 2017.

[7] M. L. Glenn, S. M. Strassel, H. Lee, K. Maeda, R. Zakhary, and X. Li, "Transcription methods for consistency, volume and efficiency." in *LREC*. Citeseer, 2010.

[8] A. Stolcke and J. Droppo, "Comparing human and machine errors in conversational speech transcription," *arXiv preprint arXiv:1708.08615*, 2017.

[9] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. Conf. Int. Speech Communication Assoc. (INTERSPEECH)*, 2002.

[10] M. P. Marcus, B. Santorini, M. A. Marcinkiewicz, and A. Taylor, "Treebank-3," Linguistic Data Consortium, Tech. Rep., 1999.

[11] N. Deshmukh, A. Gleeson, J. Picone, A. Ganapathiraju, and J. Hamaker, "Resegmentation of SWITCHBOARD," in *Proc. ICSLP*, 1998.

[12] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Unediting: Detecting Disfluencies Without Careful Transcripts," in *Proc. NAACL*, 2015.

[13] V. Zayats, H. Hajishirzi, and M. Ostendorf, "Disfluency Detection using a Bidirectional LSTM," in *Proc. Interspeech*, 2016.

[14] K. Georgila, "Using Integer Linear Programming for Detecting Speech Disfluencies," in *Proc. NAACL*, 2009.

[15] E. Le Grezause, "Um and uh, and the expression of stance in conversational speech," Ph.D. dissertation, Université Sorbonne Paris Cité; University of Washington, 2017.

[16] V. Zayats and M. Ostendorf, "Robust cross-domain disfluency detection with pattern match networks," *arXiv preprint arXiv:1811.07236*, 2018.

[17] ——, "Giving attention to the unexpected: using prosody innovations in disfluency detection," in *Proc. Conf. North American Chapter Assoc. for Computational Linguistics (NAACL)*, 2019.

[18] T. Tran, S. Toshniwal, M. Bansal, K. Gimpel, K. Livescu, and M. Ostendorf, "Parsing speech: a neural approach to integrating lexical and acoustic-prosodic information," in *Proc. NAACL*, 2018, pp. 69–81.