

Intercoder Reliability in Annotating Complex Disfluencies

Peter A. Heeman,¹ Andy McMillin,² J. Scott Yaruss³

¹Center for Spoken Language Understanding, Oregon Health & Science University, Beaverton OR

²Hearing & Speech Institute, Beaverton OR

³Dept. of Communication Sciences and Disorders, University of Pittsburgh, Pittsburgh PA

heeman@cslu.ogi.edu, andym@hearingandspeech.org, jsyaruss@pitt.edu

Abstract

In previous work, we presented an annotation scheme that can describe complex disfluencies. In this paper, we first show the prevalence of complex disfluencies and illustrate the types of distinctions that our scheme allows. Second, we present an annotation tool that allows the scheme to be easily applied. Third, we present the results of a reliability study in annotating complex disfluencies with the annotation tool. We find that subjects, even with a minimal amount of training, achieve high intercoder agreement. This work will help pave the way for speech recognizers to precisely model the structure of disfluencies, both for understanding conversational speech of non-stutterers and for assessing stuttering severity.

Index Terms: disfluencies, stuttering, annotation scheme

1. Introduction

In conversational speech, disfluencies are very common. Hence, it is important to build disfluency modeling into speech recognizers. Modeling disfluencies is also important for dealing with the speech of people who stutter, both for spoken language applications, but also, as part of future automatic stuttering assessment tools. To deal with the disfluencies of both people who do not stutter and those who do, we need an annotation scheme that can capture the full range of disfluencies. The scheme also needs to capture what is happening at the word level in a systematic way, so that sophisticated language models can be built.

In previous work [1], we presented a scheme that covers multi-iteration repetitions, revisions, editing terms, and starters, as well as complex clusterings of these. As the scheme can be difficult to apply, we introduced the *vertical alignment method*, a pen-and-paper method that simplifies the task of determining the annotation codes. The method consists of 4 steps. The first 2 steps involve subjective decisions by the annotator, while the last 2 are purely algorithmic.

In this paper, we review the first 2 steps of the vertical alignment method. We then illustrate the types of disfluency clusters that can be captured with the scheme. We then describe a prototype computer tool for performing the first 2 steps of the vertical alignment method. We then present a study that assesses intercoder reliability in annotating disfluencies with the tool.

2. Related Work

Several schemes have been proposed for annotating disfluencies in stuttered speech, including an extension to the CHAT annotation scheme [2] and SDA [3]. However, these schemes cannot describe the full range of disfluency behavior nor the role that each word plays in a disfluency.

A scheme was developed for disfluencies typical of non-stutterers: single-iteration repetitions and revisions [4]. As illustrated in Fig. 1, disfluencies are decomposed into 4 parts:

(1) a *reparandum*; (2) an *interruption point*, which is where the reparandum ends; (3) optional *editing terms*, such as ‘um’ and ‘let’s see’; and (4) an *alteration*, which is the replacement for the *reparandum*. This scheme identifies all of the words involved in a disfluency and the role that each plays. However, it does not address multi-iteration repetitions nor clustered disfluencies, which are common in stuttered speech [5, 6].

Figure 1: Four parts of repetitions and revisions

Shriberg [7] extended the above scheme so it could describe complex disfluency patterns. She allows the reparandum and alteration of a disfluency to be embedded inside of another disfluency, in which the outer disfluency uses the alteration (but not reparandum) of the inner one. However, Shriberg finds that this nesting assumption does not always hold, and uses an ad-hoc operator, ‘#’, in annotating them. The use of this operator will likely be confusing to annotators, and make it difficult to model the true regularities of overlapping disfluencies.

3. Vertical Alignment Method

In previous work [1], we extended the repair-structure approach so that overlapping disfluencies can be annotated. We differ from Shriberg in that only the reparandum, and not the alteration, of a disfluency needs to be embedded inside of the other disfluency. Fig. 2 shows an utterance with a sound repetition embedded inside of a phrase repetition, with the reparanda marked. Our annotation scheme allows multi-iteration repetitions, revisions, editing terms, and starters to be annotated, as well as complex clusters of them.

Figure 2: Embedded Reparanda

For disfluencies that have overlapping reparanda, it can be difficult for annotators to determine the extent of each reparandum. Hence, we developed the *vertical alignment method* to simplify this task, which consists of 4 steps [1]. The first 2 steps, require subjective decisions on the part of the annotator, and are described below. The last 2 steps, determining the word-level annotation codes from the alignment, are purely algorithmic.

Step 1: Determine where the interruption points are, and start a new line after each one.

Step 2: Align words that are replacements for one another into

Figure 3: Complex Pattern of Repetitions

all your pictures are — small weak an-
are of small — animals

Figure 4: Omitted & inserted words

the same columns. Fig. 3 shows an alignment: the 2 instances of ‘in’ are in the same column, so are the 4 instances of ‘a’, the 3 variations of ‘home’s’, and the 2 instances of ‘attic’.

Revisions: A revision is where a speaker backtracks, but does not strictly repeat what was just said, but modifies it. For insertions and omissions, a word in the reparandum or revision, respectively, will not have a corresponding word in its column. This is indicated by putting dashes in the empty cell (Fig. 4).

Editing Terms: Editing terms, as shown in Fig. 1, commonly occur after the interruption point of revisions and single-iteration repetitions. In this case, they are formatted on the same line as its associated reparandum, and displayed in bold (Fig. 5). Editing terms can also occur on their own in what are called covert repairs [4]. We still view them as causing a backtracking: after the editing term, a new line is started with the subsequent words lined up with the beginning of the editing term.

a raindrop **I mean**
a
a
a rainbow happens when light divides into ...

Figure 5: Editing term with a reparandum

Starters: Starters are similar to editing terms in that they are not part of the speaker’s message. They differ in that stutterers use them to help (re)start phonation. Hence, starters typically lead, without pause, into the following word. We format them similar to editing terms, but with them preceding the alteration, and formatted in italics (Fig. 6).

after sunset they c-
and they come out to look for food

Figure 6: Starter with an associated reparandum

4. Corpus

We collected a corpus of read-speech samples from 8 children, 7 of whom stuttered, and one who did not. Each child, ranging in age from 9 to 12 years old, read 9 different stories aloud. A research assistant transcribed the words and word fragments using SpeechView, indicating their start and stop times. The assistant also annotated the individual words with the disfluency codes, derived by applying the vertical alignment method.

5. Complex Disfluency Patterns

In this section, we show the wide variety in which backtracking can overlap with each other to form complex disfluencies, and we show that the vertical alignment, unlike other annotation schemes, can precisely describe their structure. As the corpus of speech is mainly from children who stutter, the results are directly applicable to assessing stuttering severity. It is also applicable to non-stutterers, as they also produce complex disfluencies [5, 6].

For this analysis, we focus on *overlapping backtracks*: backtrackings that share at least one column. Thus the 3 back-trackings of Fig. 5 are one set of overlapping backtrackings. We excluded any backtracking set that includes a prolongation or a block. This gives 544 sets of backtrackings. Of these, 430 fall into the traditional categories used by stuttering researchers: 131 sound repetitions, 105 word repetitions, 33 phrase repetitions, 124 revisions, 5 editing terms, and 32 starters.

it c-
it c-
it can have mountains

Figure 7: Phrase-Sound Repetition

Of the remaining 114 (21%), 31 do not fall into traditional stuttering categories, but can be annotated with the repair structure scheme (as well as our scheme and PLS). This includes 19 that are similar to a phrase repetition, but with the last word of the reparandum being a word fragment (Fig. 7). We refer to these as *phrase-sound* repetitions (cf. [6]). The other 12 consisted of a revision or repetition with an editing term or a starter at the interruption point: 5 had an editing term, and 7 had a starter (Fig. 6).

For the remaining 83 cases, we checked whether they can be written in Shriberg’s scheme. We found that 10 of them, including the one in Fig. 3, violate Shriberg’s embedding assumption, and so require her ad-hoc ‘#’ operator.

To further show the strength of our annotation scheme, we analyzed the backtracking patterns. For ease of analysis, we focused on the 35 of the 83 overlaps that did not include a revision. We analyzed them as to whether the interruption point (a) stays in the same column or moves to a subsequent column, or (b) moves to an earlier column. Fig. 3 illustrates the latter case, as the speaker starts at an earlier column for the last backtracking. We found that 13 of the 35 have this property.

The above distinctions were automatically determined from the annotations. Such distinctions are important, as they might be important factors in assessing the severity of stuttering. For example, disfluencies in which the speaker retreats in subsequent backtrackings (Fig. 3) might be more indicative of stuttering severity. The distinctions are also necessary in order to build a speech recognition language model, which needs to capture how likely different patterns are. None of the other annotation schemes allow all of the above disfluencies to be captured.

6. Annotation Tool for Study

To measure the intercoder reliability of the subjective decisions that annotators will need to make, we focus on the first 2 steps of the vertical alignment method. Furthermore, we built a computer tool that lets subjects perform these steps using a graphical interface. We expected that the tool would improve reliability as it ensures that subjects do not make simple mistakes, such as skipping a word from the transcription; and it allows subjects to easily change their annotation. The tool also aided our data analysis as it keeps extensive log files.

The tool, shown in Fig. 8, displays the waveform of the excerpt. The subject can select a region of the waveform, and listen to it. Below the waveform, the words and word fragments that were said are displayed, time-aligned to the waveform. At the bottom is the script the speaker was supposed to read. The subjects did not have to align the transcribed words to the script,

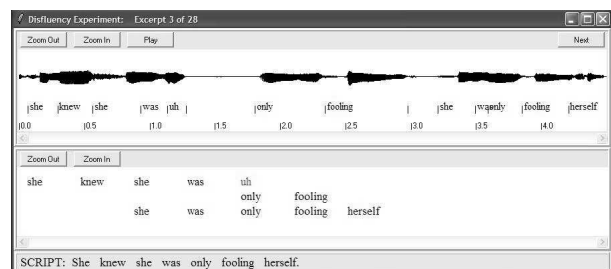


Figure 8: Annotation Tool with First Practice Excerpt

Level	1	2	3	4
Number of Excerpts	4	12	7	9
Components/Excerpt	1.8	2.3	3.0	3.6
Overlapping Components	0.0	46.4	52.4	78.1
Complex Overlaps	0.0	0.0	0.0	15.6

Table 1: Number of Excerpts and Components

and were not told in the instructions how the script should be used. Just above the script is the annotation region. Subjects start each excerpt with the transcription across the first row and align them into rows and columns to form the annotation.

Users align the words by first clicking on a word. Subjects can then press the ‘enter’, ‘space’, ‘dash’, ‘r’, or ‘backspace’ key. ‘Enter’ moves the word, and all following words on that line, onto a new line starting at the first column. ‘Space’ moves the word, and all following words on that line, one column to the right. If the word has preceding words on that line, the vacated cell is filled with dashes, otherwise, it is left blank. ‘Dash’ works similarly, excerpt that the vacated cell is always filled with dashes. ‘R’ toggles the color of a word between black and red, where red denotes editing terms and starters. Finally, ‘backspace’ moves the word, and all following words on that line, one column to the left, as long as that cell is empty or has dashes. If the word is in the first column, the line is joined with the preceding line. Note that the functionality of the ‘enter’, ‘space’ and ‘backspace’ keys is similar to a word processor.

7. Methods

To determine the reliability of the annotation scheme, we conducted an experiment in which subjects annotated excerpts of speech with the vertical alignment method. To focus the study on just this aspect, we did two things. First, we used the annotation tool. Second, we only included sentences in which it was clear what the speaker was saying in order to focus the experiment on disfluency annotation, rather than in interpreting what words the speaker is saying. Hence, subjects were also given a transcription of what words were said, as well as the sentence of the story that was being read.

Excerpts: We selected 32 excerpts from the corpus that contained a complete sentence with at least one repetition or revision. Excerpts were only considered if they did not contain a block, an ambiguous sound, or a word fragment that had no backtracking. Three expert annotators determined the correct annotations. Excerpts were not chosen to reflect the natural distribution of different types of disfluencies, but to emphasize ones with complex backtracking structures. We categorized the excerpts into 4 levels of difficulty. The first level only had excerpts with repetitions in it, either alone, or separated by at least one word. The next 3 levels added revisions, editing terms, starters, and clusters of these. Fig. 2 is a level 2 excerpt, Figs. 5 and 6 are level 3, and Fig 3 is level 4. The first row of Table 1 reports the number of excerpts in each level.

We devised 3 metrics to measure how the excerpts increase in difficulty as the level increases. The second row of Table 1 reports the average number of components per excerpt, where a component is a starter, editing term, or reparandum. The third row reports the percentage of components involved in overlapping backtrackings. The fourth row reports the percentage of components that are reparanda and that cannot be analyzed using an embedded repair structure (e.g., Fig. 3). As expected, the values for each metric increase as the level increases.

Instruction Guide: We created an instruction guide that explains the vertical alignment method, and how to use the anno-

Level	1	2	3	4
Correct	95.7	80.6	69.0	68.5
Format	0.0	2.8	2.4	3.7
Region	0.0	6.9	21.4	13.0
Wrong	4.3	9.7	7.1	14.8
Time (seconds)	34.3	48.1	63.8	65.1
Altered	4.3	15.3	33.3	37.0
Component Score	97.4	90.5	90.5	91.7

Table 2: Performance by level of difficulty

tation tool. It also guides the user through 3 practice excerpts. The guide is 5 and a half pages long.

Subjects: Six annotator-subjects participated in the experiment. All were familiar with computers, using them daily for at least word processing tasks. All had a bachelor’s degree, but none had a background in communication disorders.

Sessions: Subjects were given the instruction guide to read and the 3 practice excerpts, for which the research assistant was present to answer any questions. The subjects then annotated the 32 test excerpts, presented from level 1 through 4, with the order in each level randomized. The session lasted one hour.

8. Data Analysis

For each excerpt annotation, we scored each reparandum, editing term, and starter, separately using the following 4 scores.

Correct: Identical to the gold standard or has an insignificant difference. For example, for the revision of ‘looked’ by ‘took one look’, the gold standard had ‘looked’ aligned with ‘look’; but we also allow ‘looked’ to be aligned with ‘took’.

Format: The annotation is not legal, but there is an obvious way to fix it, which results in the correct annotation. For example, a few of the annotations included a column that just had dashes in it, which has no meaning in the annotation scheme; the removal of the column resulted in a correct annotation.

Region: The extent of the component is properly identified, but is miscoded; for example, an editing term is coded as a starter or vice versa, or an editing term or starter is coded as a reparandum or vice versa. Also included are reparanda that have a word aligned in the wrong column.

Wrong: There is a mistake in identifying the extent of a component; for example, the extent of a reparandum is incorrect, a backtracking is not coded, an editing term or starter is not coded, or an extra backtracking is included.

We computed the overall score for an excerpt annotation by taking the worst score achieved by any component of the annotation. For example, if there were three reparanda and one editing term, with two scoring **correct**, one scoring **region**, and one scoring **wrong**, the entire excerpt is scored as **wrong**.

9. Results

We report the results by level of difficulty, by subject, and by each component of the disfluencies.

Difficulty of Excerpts: We first examine how well the subjects did on the excerpts across the 4 levels. One subject double-clicked the next button, causing the tool to skip one excerpt (from level 1); this annotation was excluded from the analyses. As expected, subjects did better on the easiest excerpts, with 95.7% of the excerpts annotated correctly (Table 2). For the most difficult excerpts, only 68.5% were annotated correctly, with a further 3.7% having only minor formatting errors.

We also analyzed how subjects produced the annotations. The row labeled ‘Time’ reports the average amount of time that

Subject	4	3	6	2	5	1
Correct	87.5	87.5	83.9	78.1	71.9	65.6
Region	12.5	6.3	9.7	6.3	18.8	12.5
Wrong	0.0	6.3	6.5	15.6	9.4	21.9
Time	46.8	41.7	51.1	48.0	84.1	56.6
Altered	9.4	15.6	22.6	31.3	31.3	34.4
Played	0.0	0.0	0.0	0.0	100.0	3.1

Table 3: Subject Performance

subjects spent on each excerpt. As the excerpts became more complex, subjects spent more time annotating them. This suggests that subjects took the experiment seriously and gave each excerpt the necessary time to annotate it.

The row labeled ‘Altered’ reports the percentage of excerpts in which subjects changed their annotation. This was determined by checking the log files for whether subjects used the backspace key or toggled the same word multiple times. There are two explanations for why a subject changed an annotation. The first is that a subject made a mistake in using the tool, such as pressing the ‘space’ too many times. Given that there was only a single altered annotation in the level 1 excerpts, which were presented at the beginning of the sessions, this explanation is unlikely. The more likely explanation is that subjects changed their mind as to how to annotate the excerpt, and so were experimenting with alternate annotation. Thus it seems that, as the difficulty of the excerpts increased, subjects made more use of the tool to experiment with alternate annotations.

Individual Subject Performance: We next examine the performance of each of the 6 subjects (Table 3). Here, we collapsed the ‘correct’ and ‘format’ categories. The subjects are arranged by their accuracy. We see that subjects 3, 4, 6, and 2 have similar performance, ranging from 87.5% to 78.1% correct, while subjects 5 and 1 did not perform as well.

From row ‘Time’, we see that annotation correctness does not vary directly with how much time the subjects spent; rather, subjects with higher correction rates spent less time on each excerpt. From row ‘Altered’, we see that these subjects also made fewer changes to their annotations. This suggests that the subjects with higher correction rates grasped the concepts better than the others, and so spent less time and made fewer changes.

The row ‘Played’ reports the percentage of excerpts for which each subject played at least part of the audio. Surprisingly, only one subject consistently used this feature, while a second used it for only one excerpt, but did not use it correctly. Two other subjects used it only during the practice sessions. Hence, the instructions probably did not adequately explain how this feature works or why it should be used. This also shows that the annotations for a majority of the excerpts can be done correctly without even listening to the audio. However, the lack of audio might be why higher accuracy was not achieved.

Disfluency Components: The more difficult excerpts have multiple components in the disfluencies. To better understand the correction rates, we analyzed how accurately each component was annotated. The accuracy was computed as the number of components that are correct or have a minor formatting error less the number of incorrectly hypothesized components (of which there were 2 in the entire study), and divided by the number of components. From the last row of Table 2, we see that the level 1 excerpts have an accuracy of 97.4%, which reflects that they only have non-clustered repetitions. For the more difficult excerpts, subjects achieved a component accuracy of 90.9%.

We also analyzed the accuracy by component type (Table 4). For reparable that cannot be described using an embedded

	Count	Correct	Region	Wrong
Non-nesting Reparanda	48	83.3	6.3	10.4
Repetitions	279	98.6	0.4	1.1
Revisions	138	87.7	9.4	2.9
Editing Terms	36	86.9	5.6	5.6
Starters	36	69.5	19.4	11.1

Table 4: Analysis of Components

repair structure, subjects achieved 76.7% accuracy; thus they did have problems with complex clusters. Repetitions that occurred alone or that could be written with an embedding structure were annotated at 98.6% accuracy. Revisions were annotated at 87.7% accuracy; most of the errors were with matching words with their replacements, rather than in identifying the extent of the reparandum. Editing terms were annotated at 88.9% accuracy, while starters were at 69.4%. Most of the errors with starters were labeling starters as editing terms or interjections.

10. Conclusion

In this paper, we illustrated the range of disfluencies that our annotation scheme can annotate. We also showed that subjects can annotate complex disfluencies with good intercoder reliability. Such disfluencies are important for speech recognition of conversational speech, and for the automatic assessment of stuttering severity. We believe that the positive results are because (a) the scheme has full coverage and is systematic; (b) the first 2 steps of the vertical alignment method are simple and intuitive; and (c) the annotation tool prevents many simple mistakes and allows experimentation with alternative annotations. The above results were achieved with only minimal training. Further training should include the use of audio cues, and how to identify starters and align revisions.

Due to the success of this study, we plan to build a full version of the annotation tool. The tool will allow users to transcribe the words in audio files, and correct the transcription after the disfluency annotations have been started. The tool will also allow the annotation of blocks and prolongations.

The results reported are not directly comparable to results in the stuttering literature. This is because subjects in our study are doing a much more detailed annotation, but are also given the word transcription. In future work, we will determine how well users agree in transcribing words and word fragments.

11. References

- [1] P. Heeman, A. McMillin, and S. Yaruss, “An annotation scheme for complex disfluencies,” in *ICSLP*, 2006.
- [2] N. Bernstein Ratner, B. Rooney, and B. MacWhinney, “Analysis of stuttering using CHILDES and CLAN,” *Clinical Linguistics and Phonetics*, 10:169–187, 1996.
- [3] H. Gregory, J. Campbell, C. Gregory, and D. Hill, *Stuttering Therapy: Rationale and Procedures*. Pearson Allyn & Bacon, 2003.
- [4] W. Levelt, “Monitoring and self-repair in speech,” *Cognition*, 14:41–104, 1983.
- [5] C. Hubbard and E. Yairi, “Clustering of disfluencies in the speech of stuttering and nonstuttering preschool children,” *J. of Speech and Hearing Research*, 31:228–233, 1988.
- [6] L. LaSalle and E. Conture, “Disfluency clusters of children who stutter: Relation of stutters to self-repairs,” *J. of Speech and Hearing Research*, 38:965–977, 1995.
- [7] E. Shriberg, “Preliminaries to a theory of speech disfluencies,” U. C. Berkeley, Doctoral dissertation, 1994.