12th edition of the Disfluency in Spontaneous Speech Workshop (DiSS 2025)
4-5 September 2025, Lisbon, Portugal

On variability in the identification and labelling of disfluencies — preliminary results from 23 annotations of the same data

Jürgen Trouvain[1], *Ludivine Crible*[2], *Malte Belz*[3], *Simon Betz*[4], *Štefan Beňuš*[5], *Lorraine Baqué*[6], *Marina Cantarutti*[7], *Jessica Di Napoli*[8], *Ivana Didirková*[9], *Maria Machuca*[10], *Lucia Mareková*[11], *Oana Niculescu*[12], *Pauliina Peltonen*[13], *Aurelie Pistono*[14], *Loredana Schettino*[15], *Vered Silber-Varod*[16], *Simon Williams*[17]

[1]Saarland University, Germany

Corresponding author: trouvain@lst.uni-saarland.de

Abstract

This study provides a preliminary report on a large inter-annotator agreement experiment where 23 expert annotators from various research backgrounds identified and labelled disfluencies in the same speech sample. Each annotator was instructed to analyze the sample according to the framework (definitions, segmentation, labels, etc.) they typically use. The annotations were then processed and compared across three different dimensions: 1) the scope of the chosen typology and the definitions within, 2) the implementation of the typology in terms of annotation tiers and labels, and 3) the temporal alignment of the annotations. Preliminary findings reveal that there are substantial variations between annotators on various levels of annotation. The lack of a common standard becomes particularly evident in more complex segments, such as repairs.

Index Terms: disfluencies, annotation, inter-annotator agreement

1. Introduction

Disfluencies are a central feature of spontaneous speech. They have been at the core of many research agendas and subdisciplines, such as phonetics, computational modelling, learner corpus research, psycholinguistics, conversation analysis and corpus-based pragmatics. This diversity of approaches, while constituting a token of vitality and relevance for the field, is also a challenge for the study of disfluencies since each set of research questions requires (at least in part) its own analytical method. As a result, there is little consensus on the way disfluencies are annotated and analyzed: (partly) different phenomena may be included and/or they can be defined and labelled differently. The proliferation of domain-specific frameworks makes it very difficult to have an overview of the field or to compare results across studies. It can also lead to imprecisions on different levels, ranging from confusion to possibly erroneous annotations of large corpora, for instance [1]. Although this is also true of other fields of study that require some level of interpretation, the situation for disfluencies is particularly striking, as there are currently no international standards available.

An annotation standard for disfluencies would require agreement among many scholars from various research traditions. Such an ambitious goal, if it can ever be realized or if it is even desirable, first requires a precise mapping of the current practices in terms of the breadth and depth of existing annotation models, from theoretical considerations to technical implementation. It is in particular necessary to understand where disagreements occur (whether at the level of scope, definitions, labels, segmentation, or temporal extent, etc.) so that a possible common core can be found and a "pivot" language can be developed.

For this purpose, a large disfluency annotation experiment was initiated, reaching out to disfluency experts from various fields and backgrounds who typically meet at the workshops on Disfluency in Spontaneous Speech (DiSS). They were asked to analyze disfluencies in two samples of audio- and video-recorded conversations in English, following the annotation model they typically use. The annotations were then processed and compared across three different dimensions: 1) the scope of the chosen typology and the definitions within, 2) the implementation of the typology in terms of annotation tiers and labels and 3) the temporal alignment of the annotations. The present paper reports preliminary findings of the first two dimensions of this meta-analysis.

2. Methods

2.1. Data

For this annotation experiment, we selected two excerpts from conversations between native speakers of English who give their opinion about a given topic (here, social media). These excerpts were taken from the SITAF corpus [2], with permission from the authors of the corpus. The excerpts amount to 4:37 minutes of recordings and were provided with both the audio and the video files. However, no orthographic transcription was provided in order to avoid biases and predetermination of, for instance, pause identification or certain articulatory features.

2.2. Annotators

These files were analyzed by 23 annotators (two of them native speakers of English) from 10 countries and 15 different institutions. Co-authors of this paper were all part of the annotator pool. They all had extensive experience with disfluency research, although not everyone uses annotation tools (like Praat [3], used in this study for annotation) to the same extent.

2.3. Tasks

Annotators were given the following set of instructions:

- Annotate what you identify as phonetic exponents of (dis)fluencies in these two excerpts according to the specifications of your framework.
- Include one or two research questions you typically investigate with such annotations.
- Provide your coding scheme (list of abbreviations, full labels and operational definitions, including citations) together with a motivation and research area.
- Label your tiers (if any) with explicit names so that they can be understood by others, and duplicate annotation tiers per speaker.

10.21437/DiSS.2025-12

- Provide the annotations in the TextGrid format (as used in Praat).

We emphasized that the term "(dis)fluency" might be understood or used differently by different annotators and that it was presently used as a broad cover term for all phenomena historically categorized as such, including, for example, repairs, false starts, self-corrections, repetitions, pauses, filled pauses, filler particles, clicks, prolongations, etc.

This experiment thus resulted in a dataset of 23 TextGrid files per sample, as well as 23 "meta-files" containing the detailed annotation model used by each annotator, as well as details about each annotator's background and approach.

2.4. Data analysis

We proceeded with the analysis in three stages. The first focused exclusively on the meta-files and aimed at comparing 1) the scope of the annotation models (how many phenomena, which were annotated) and 2) the exact definitions used to describe each phenomenon. The goal of the first stage was to look for commonalities and differences in the annotation models and to link these back to the annotators' fields of research in order to identify clusters that could possibly explain further differences found in the next two stages of the analysis.

In the second stage, the goal was to compare units of segmentation, numbers and types of tiers, and the format of labels, and to see whether additional information was annotated such as position and duration. Both the meta-files and the TextGrids were taken into account for this stage of the analysis.

In the third stage, the focus is on the degree of agreement regarding the temporal alignment of annotated phenomena. This stage is currently ongoing and will not be reported here.

3. Dimensions of variability

Various forms of variability occurred between annotations at different levels of analysis. One common form of disagreement we observed was a mismatch between labels used to annotate a given phenomenon and the definitions or interpretations provided for them.

3.1. Same label but different meanings

One mismatch that occurred was where annotators used the same label to describe a given phenomenon but different meanings were provided to describe the phenomenon. For example, the label SILENCE, or some variation on this or the equally frequent term PAUSE, such as [sil] or [sp] for SILENT PAUSE or [p] for PAUSE, was used by all but one annotator in their annotation. However, the use of the label varied somewhat across annotators.

First of all, there was a clear distinction observed between the application of the label SILENCE to refer to silence more generally, and to disfluent or hesitant silence. A minority of annotators explicitly labelled only *disfluent* instances of silence (8 out of 23). Two additional annotators made a distinction between disfluent and fluent occurrences of silence, while the remaining annotators marked silences more generally where they occurred, typically labelling them as separating units of speech such as IPUs (inter-pausal units) and/or speaker turns, without providing a specific indication of disfluency. This results in a great degree of variation in terms of how to interpret the label SILENCE in the annotations.

Additionally, the acoustic definition of what was labelled as SILENCE varied across annotations. Some annotators relied on the perception of *perceived* silences, while others looked specifically at intervals devoid of speech or vocalization. There were also differences between annotators regarding whether or not the label SILENCE could include other sounds such as breath noises, clicks, and physiological noises such as swallowing. Some annotators explicitly distinguished between entirely silent intervals and labelled the other phenomena separately, while for others, breath noises and clicks were annotated as part of SILENCE in the sense of a non-speech section.

3.2. Different label but same meaning

There were also cases where annotators used different labels to describe the same phenomenon. One example includes the annotation of FILLER PARTICLES. Here, annotators typically used one or more different terms: FILLER PARTICLE, FILLER, NON-LEXICAL FILLER and/or FILLED PAUSE. All annotators included fillers in their annotations, but they chose different ways to represent them. Typically, annotators either relied on one of the four terms listed above, or they used the lexical forms of fillers in English ('uh' and 'um' or 'er' and 'erm') as labels in their annotations.

3.3. Same label but only partially same meaning

Conversely, there were also instances where annotators used the same label to describe various phenomena. For instance, some annotators used the label [REP] for the *reparandum*, i.e. the portion of speech that is corrected in a repair, whereas others used it for the *reparans*, i.e. for the correction itself. Similarly, the term FALSE START was used to label retraced but also non-retraced false starts.

3.4. Same label but different interpretation

In some cases the same label occurred with different interpretations or categorisations. For example, CLICKS were categorised as both NON-VERBAL VOCALISATION and DISFLUENCY.

3.5. Different interpretations of items

One other source of variability comes from the broad category of LEXICALIZED FILLERS or DISCOURSE MARKERS. This includes words and phrases such as 'like', 'you know', and 'sort of', and was labelled on a separate, dedicated tier by only six of the 23 annotators. Even though all six of these annotators included tokens of 'like' or 'you know', their interpretations varied in two respects. First, the label names used by the annotators suggest that they treat these words differently: some label them as DISCOURSE MARKERS, some as FILLERS, and some use a hybrid form of LEXICALIZED FILLED PAUSE.

Second, while many tokens were identified as belonging to their respective categories by all six annotators, some tokens were not uniformly included by all. For example, our dataset includes 31 tokens of 'like', which were not all regarded as disfluencies: only eight of them (26%) were identified by all six annotators as disfluencies, twelve of them (39%) by five annotators, and the remaining 11 tokens (35%) by fewer than five annotators. This suggests that annotators differ in their interpretation of the effect of context on the function of these lexical tokens.

3.6. Same label but on different tiers

There were also cases where annotators differed in how they structured their annotation tiers and labelled disfluencies within

them. For instance, some annotators included REPAIRS as part of a broader tier which included other types of disfluency, while others annotated REPAIRS on a separate, dedicated tier.

4. Complex disfluencies

Complex combinations of disfluencies are a notorious challenge regardless of context. What actually constitutes a COMPLEX DISFLUENCY? We identified at least two types: one consists of a concatenation of disfluencies in a row, whereas the other one is nested. The former is of a syntagmatic nature (see Figure 1) and the latter has a paradigmatic character (see Figure 3).

In the following examples, Fig. 4 and Fig. 2 are based on the speech sections shown in Fig. 3 and Fig. 1, respectively. The table-like overview enables a better readability than just showing the TextGrid files. Terminological differences in the labels illustrate the contrasting concerns of annotators.

4.1. Syntagmatic complexity

Most annotations allowed for syntagmatic complexity, with annotation of disfluency in the same (or different) tiers. Sequences of repeating or alternating disfluencies were observable throughout the annotations.

In Fig. 1 and Fig. 2 an extract of 22 seconds labelled by four randomly selected annotators is depicted for the following transcribed segment, with '-' indicating a boundary marker used by the annotators, and 'P' marking a pause, followed by its duration in seconds (given in parentheses):

> "the latter makes it - P(.24) - throat clearing - P(.54) - click - makes you - a more lonely person - P(.53) - filler - P(.84) - so I don't know - P(.11) - I guess - P(3.58) - I feel I - like - P(1.48) - agree with it - P(.62) - but - like - not - P(1.46) - I don't know"

All four annotators agree on one FILLER PARTICLE (between 'person' and 'so') – but that is the only agreement between all four annotators. There is much agreement on SILENT PAUSES between annotators 1, 2 and 4, but not for 3. There is a complex pause sequence of approximately 1 second at the beginning that includes a throat clearing and a tongue click. This is seen as one SILENT PAUSE by annotator 1, and as part of a larger REPAIR section by annotator 2. The click alone is seen as a FILLER PARTICLE by annotator 3, and annotator 4 looks only at the first silent part.

Two annotators (2 and 4) agree on one instance of 'like' as a LEXICAL FILLER, but not on the second instance of 'like'. With respect to LENGTHENING, only annotator 4 refers to this phenomenon in an explicit way.

4.2. Repairs and paradigmatic complexity

Sections of REPAIR are notoriously difficult to describe. Annotators use different labels, and are interested in different phenomena. In addition, the temporal extension of selected phenomena differs, and, during wider repair segments, there can be instances of repair that happen within another repair.

The forms of agreement and disagreement amongst annotators who label REPAIRS are illustrated in Fig. 3 and in Fig. 4 in an extract of 7.5 seconds for the following transcribed segment. As in the example before, '-' indicates a boundary marker used by the annotators, and 'P' marks a pause, followed by its duration in seconds (given in parentheses). In addition, capital letters indicate clearly accented words:

> "communicate with MORE people - it's just - P(.45) - I guess - there's like - P(.52) - you have MORE - you can have - P(.25) - there's potential for MORE - relationships - but - P(.60)"

The following analysis illustrates how annotators' labelling of repairs varies in segmentation, terminology, and attention to pauses, but broadly aligns on the key speech events.

The content in Fig. 4 refers to the labels of eight randomly selected annotators showing substantial variability. For example, annotator 1 (first tier in Fig. 4) gives a single label of [substitution repair] to a larger extract, i.e. starting with the four words, "communicate with more people" (not visible in Fig. 4) that completes the preceding utterance "but also it gives you the opportunity to". All other annotators begin no earlier than "it's just".

Annotator 2 divides the utterance into two repair segments: "I guess there's like" and "there's potential for more", and labels both as reformatting; the intervening "you have more - you can have" is unlabelled. Annotator 3 takes a similar approach to Annotator 1 and labels the stretch, "there's like - you have more - you can have - there's potential for more relationships" as a new start repair with different syntactic structure. Annotator 4 adopts Levelt's [4] term REPARANDUM to label "it's just", "I guess - there's like", and "you can have more - you can have", withouth labelling the reparans. The next annotator (no. 5), sees REPAIRS as traced or non-retraced FALSE STARTS, and labels "it's just" and "there's like" as non-retraced false starts, and "you can have more - there is potential for more" as a retraced FALSE START.

Annotator 6 adopts a variety of terms: FALSE START ("I guess - there's like"); REPARANDUM ("you have more", "you can have"); and REPAIR ("there's potential for more"). Annotator 7 is broadly similar but also labels INTERREGNA: LEXICALISED FILLED PAUSE and SILENT PAUSE (after "it's just"); unspecified interregnum (after "there's like"); and SILENT PAUSE (after "you can have"). "You can have" is labelled as a REPARANDUM and followed by a substitution type REPARANS ("there's potential for more"). Annotator 8 labels two SELF-CORRECTION DELETIONS: "there's like" and "you have - more you can have".

Broadly, annotators' syntagmatic labelling of repairs is sometimes similar, i.e. they comment on the substantive content differing only in the level of detail; but their paradigmatic labelling depends on their recognition of interregna. In other words, despite the variety of terms and foci that annotators adopt, they tend to label the same speech events, many of which could be described either as abandoned utterances (FALSE STARTS) or as FILLERS, depending on the semantic content, while differences relate to whether they label or ignore the spaces in between, i.e. the SILENT PAUSES, with or without FILLERS.

What we can sometimes observe as well is that the differences between annotators go beyond specifying the details and that annotators show a different understanding as to what constitutes repair itself.

5. Discussion and conclusion

The examples given above very clearly show that we are far from using a uniform standard of labelling for disfluency phenomena. Although such a uniform labelling scheme would be very helpful for comparing findings across studies, its adoption would very much depend on labelling traditions on the one hand

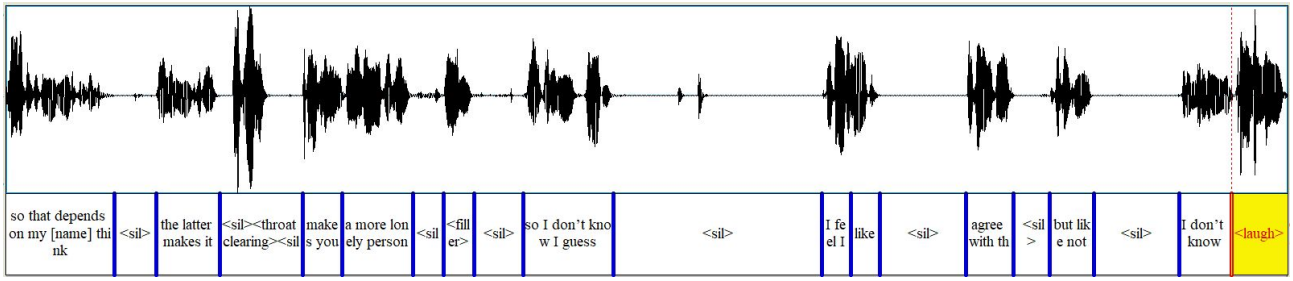

Figure 1: *Waveform and text annotation of an extract of 22 seconds illustrating a concatenation of many disfluent and fluent sections.*

	the latter makes it	P	throat clearing	P	click	makes	you	a more lonely person	P	FP	P	so, I don't know	P	I guess	P	I feel I	like	P	agree with it	P	but	like	not	P	I don't know
1			p						sp	fp	sp		sp		sp			sp		sp				sp	
2	repair								sp	fp	sp		sp		sp		lfp	sp		sp		lfp		sp	
3					fp					fp																
4			sp				lng		sp	fp	sp		sp				lfp	sp		sp				sp	

Figure 2: *Section from Fig. 1 with 4 out of 23 annotations in tabular and schematic form; first column denotes annotators; labelling scheme: sp = silent pause, fp = filler particle, lng = lengthening, lfp = lexical filler particle.*

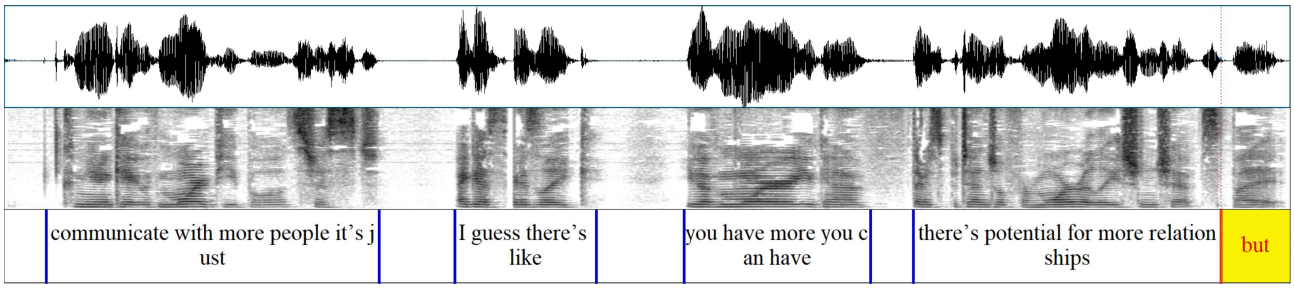

Figure 3: *Waveform and text annotation of a repair section of 7.5 seconds illustrating nested disfluencies.*

	... people	it's just	P	I guess	there's like	P	you have more	you can have	P	there's potential for more	relationships	but	P
1	repair: new start; different syntactic structure												
2			repair segment - reformatting					repair segment - reformatting					
3				repair: new start; different syntactic structure									
4		reparandum			reparandum			reparandum					
5		non-retraced false start			non-retraced false start			retraced false start					
6				false start			reparandum	reparandum		repair			
7		interregnum (i.r.)				i.r.		reparandum	i.r.	reparans			
8					self-correction deletion		self-correction deletion						

Figure 4: *Repair section from Fig. 3 with 8 out of 23 annotations in tabular and schematic form; first column denotes annotators.*

and on different research questions on the other. Standardisations are *in principle* an appreciated instrument for many different types of categorization. For instance, the International Phonetic Alphabet (IPA) helps a lot when reading phonetic script. However, labelling an extract of speech will very often lead to very different transcriptions – despite a standardized set of IPA symbols.

Though standardization is not the primary goal of this initiative, it was very useful that *different* annotators worked with *identical* data. It showed that *there are differences* in annotations, *where* those differences concretely occur, and *which phenomena* are concerned. Differences in *temporal alignment* can also be observed. The examples given above show that the differences on all listed levels can be dramatic. One important lesson learned with this multiple-annotator study is how important it is to describe annotation procedures in publications, given the possible differences in labels used and interpretation of these labels.

Future research will show how annotation schemes are shaped by different perspectives, research questions, and approaches within research areas and how those annotations can be "translated" to other annotation schemes (cf. the "pivot language" metaphor from the Introduction). As a concrete next step in this project we will inspect more closely the inter-rater agreement, which will be calculated on categories and on the temporal extensions of those categories.

6. Acknowledgements

We thank all our annotators in this initiative.

The list of annotators include all co-authors with the following affiliations:

[2]Ghent University, Belgium
[3]Humboldt-Universität zu Berlin, Germany
[4]Bielefeld University, Germany
[5]Constantine the Philosopher University, Slovakia
[6]Autonomous University of Barcelona, Spain
[7]University of York, United Kingdom
[8]RWTH Aachen University, Germany
[9]Université de Montpellier Paul-Valéry, France
[10]Autonomous University of Barcelona, Spain
[11]Constantine the Philosopher University, Slovakia
[12]Romanian Academy Institute of Linguistics, Romania
[13]University of Turku, Finland
[14]University of Toulouse, France
[15]Free University of Bozen-Bolzano, Italy
[16]Tel Aviv University, Israel
[17]University of Sussex, United Kingdom

In addition, the following colleagues provided their annotations: Angelika Braun (University of Trier, Germany), Daniel Duran (then Leibniz-Zentrum Allgemeine Sprachwissenschaft, now Bielefeld University, Germany), Nathalie Elsässer (then University of Trier, Germany, now Austrian Academy of Sciences, Austria), Melissa Hildebrand (University of Trier, Germany), Farhat Jabeen (Bielefeld University, Germany), Heini Kallio (Tampere University, Finland), Valentin Kany (Saarland University, Germany), Beeke Muhlack (then Saarland University, now Landeskriminalamt Bayern, Germany), Ludger Paschen (Leibniz-Zentrum Allgemeine Sprachwissenschaft, Germany).

7. References

[1] V. Zayats, T. Tran, R. Wright, C. Mansfield, and M. Ostendorf, "Disfluencies and human speech transcription errors," in *Interspeech 2019*, 2019, pp. 3088–3092.

[2] C. Horgues and S. Scheuer, "Why some things are better done in tandem?" in *Investigating English Pronunciation: Trends and Directions*, J. A. Mompeán and J. Fouz-González, Eds. Basingstoke and New York: Palgrave Macmillan, 2015, pp. 47–22.

[3] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," 2025. [Online]. Available: http://www.praat.org/

[4] W. J. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, no. 1, pp. 41–104, 1983.