Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech

Ludivine Crible[a]*, Ivana Didirková[b], Christelle Dodane[c] and Loulou Kosmala[d]

[a]Department of Linguistics, Ghent University, Belgium; [b]UR 1569 TransCrit, Université Paris 8 Vincennes – Saint-Denis, France; [c]UMR 5267 Praxiling, CNRS, Université Paul Valéry Montpellier 3, France; [d]EA 4398, GReMLIN, Université Paris-Ouest Nanterre, France;

ludivine.crible@ugent.be

Towards an inclusive system for the annotation of (dis)fluency in typical and atypical speech

This paper presents an operational annotation system for (dis)fluencies in typical and atypical speech, based on existing standard annotation schemes previously established in the literature. Grounded in a functional approach to (dis)fluency, we address some of the conceptual and technical limitations found in previous annotation models, and offer an integrated and inclusive system which is compatible with different multi-layered annotation software such as Praat or ELAN. Our aim is twofold: to create comparable annotated corpora both in typical and atypical speech, and to provide natural language processing and the health sector with applications for diagnostic and therapy in speech disorders.

Keywords: disfluency; annotation; typical speech; atypical speech

Introduction

All speakers experience hesitations and speech production difficulties in their everyday language use, at an average of around 6% of typical speech (e.g. Bortfeld et al., 2001).[1] Such episodes are often termed 'disfluencies' and can be defined as disruptions of the speech flow, as opposed to the "fluent" state of language, idealised as "smooth, rapid, effortless" (Crystal, 1987, p. 421). In typical speech, however, the presence of so-called disfluencies is not necessarily "disruptive" per se, and is in fact a highly common feature of conversational, spontaneous speech (e.g. Allwood et al., 1990). In atypical speech, on the other hand, great alterations in the speech flow can also be considered as one of the symptoms of a speech disorder such as stuttering.[2] Stuttering is a speech

---

[1] The authors of this paper are ordered alphabetically. All authors contributed equally to all stages of the research project.

[2] Throughout this paper, we use the terms 'typical' and 'atypical' to refer to disfluencies that are common in all speakers vs. those that occur primarily in speakers presenting some speech or

disorder which stems from both neurological (Etchell et al., 2018) and genetic (Frigerio-Domingues et al., 2019) factors, and unlike disfluencies found in typical speech, which are mainly related to speech planning issues, the origins of stuttering-like disfluencies are still under debate and are supposed to be related to issues of the motor or sensory system (especially somatosensory and auditory feedback) (Ambrose, 2004; Monfrais-Pfauwadel, 2014).

Disfluency phenomena have thus been observed for both typical and atypical speech, but there is considerable variation in the conceptual approach, the extent and format of different typologies, which reflects the diversity of (sub)disciplines, languages and target populations (adults vs. children, native vs. non-native, typical vs. atypical speakers). Most frameworks in typical speech consider filled ("uh, uhm") and unfilled pauses (silences), repeats ("I… I was there"), restarts ("I were… was there"), lengthening[3] ("I was theeeere"), and word fragments ("I was at ho… work") as typical disfluencies (e.g. Shriberg, 1994), while studies on stuttering tend to focus on a restricted set of phenomena such as repetitions (mostly identical repetitions of sounds, syllable fragments, syllables, monosyllabic words) and blocks (or silent lengthenings, corresponding to silent pauses accompanied by visible tension) (Lickley, 2017). Some studies have shown that stutterers produce disfluencies which share similar features with typical disfluencies, along with stuttering-like disfluencies (SLD, e.g. blocks), but

---

speech disorder, respectively. By extension, these terms are also used to characterize speech and speakers themselves, although this is by no means meant in a discriminatory way nor does it denies the right to be Neurodiverse.

[3] Note that the terms 'prolongation' and 'lenghtening' are often interchangeable in the literature. For the sake of consistency, we are only using the latter in this paper.

overall, the field lacks an integrated view of disfluencies in typical and atypical speech. Disfluency thus exists in all types of language data, yet it has traditionally been approached through different models for different corpora. Annotation models differ with respect to their terminology (e.g. disfluencies vs. repairs) the categories covered (e.g. whether discourse markers are included or not), structure (hierarchical or not), preferred tool and method (e.g. semi-automatic or manual), which makes it difficult to conduct reliable comparisons across datasets (Crible et al., 2019).

The present proposal addresses this gap and provides an operational annotation system for (dis)fluencies in typical and atypical speech, as part of the ANR-fund *BENEPHIDIRE* project[4], further grounded in a functional approach to (dis)fluency, following Crible et al. (2019). This approach assumes the ambivalence of elements that can be used either "fluently" or "disfluently" depending on the context of production, without restricting disfluencies to mere disruptions or removable errors, hence including a whole range of phenomena. While this type of approach has so far mainly been adopted in typical speech (e.g. Crible, 2018; Kosmala, 2021), the present model[5] also extends it to atypical speech. The system introduced in this paper aims to bring together existing categories and conventions found in some of the previous (dis)fluency models in order to homogenize them in a single scheme which can be applied to various corpora.

---

[4] ANR-18-CE36-0008 Le Bégaiement : la Neurologie, la Phonétique, l'Informatique pour son Diagnostic et sa Rééducation (BENEPHIDIRE, PI : Fabrice Hirsch)

[5] A first version of this model was presented during the 2021 *DiSS* workshop (Didirková et al., 2021).

We will first review a selection of annotation models for disfluency in typical and atypical speech, and address their main limitations. We then introduce our proposal, starting with general principles followed by operational definitions of disfluency categories. We then report inter-annotator agreement and discuss annotated examples to illustrate the benefits and remaining difficulties of the model. Finally, we conclude this paper with a presentation of the next steps, perspectives, and possible applications of this project.

Previous Annotation Models of (Dis)fluency

Models for Typical Speech Production

Seminal work on disfluency annotation was carried out by Shriberg (1994) in English human–human and human–computer interactions from a computational perspective. Her model includes a notation system for repetitions, substitutions, insertions, deletions, filled pauses, explicit editing terms, discourse markers, conjunctions, word fragments, misarticulations, and contractions. It excludes anything that is "arguably part of the speaker's intended utterance" (1994, p.1), such as unfilled pauses or some discourse markers such as "well" or "like". Later works by Eklund (2004) in Swedish, Moniz (2013) in Portuguese, or Christodoulides et al. (2014) in French were largely based upon Shriberg's original proposal. Other major frameworks which adopt a similar approach include the Switchboard corpus (Meteer et al., 1995) or the SimpleMDE guidelines (Strassel, 2003).

Pallaud et al. (2013) developed a more syntactic approach to "self-interruptions" in French, where disfluencies are first decomposed into three structural parts (i.e. *reparandum*, *interregnum*, and *reparans*) borrowing terms from Levelt (1983) and Shriberg (1994). Similarly, Ginzburg et al.'s (2014) model stems from a segmentation

perspective, influenced by theories such as *dialogic syntax* or *Question Under Discussion*. They further implemented a multilingual and multimodal dialogue corpus annotation system for disfluency, exclamations and laughter, as part of their DUEL project (Hough et al., 2016).

(Dis)fluency models are also found in the field of Second Language Acquisition, where (dis)fluencies are compared across first language (L1) and second language (L2) speakers. For instance, Götz (2013) offered a three-fold typology, based on different dimensions of (dis)fluency, such as production (e.g. speech rate, or use of filled and unfilled pauses, discourse markers), perception (e.g. idiomaticity, lexical diversity) and gesture (manual gestures, gaze direction). Götz (2013) therefore favored observable features from all linguistic levels. However, her model remains strictly conceptual, and can hardly be used for annotation purposes.

Most recently, Crible et al. (2019) developed an annotation model that encompasses elements from many of the above frameworks. It does not require subjective judgements of relative fluency but is based on formal definitions, in support of the functionally ambivalent view of (dis)fluency. Their (dis)fluency categories include filled and unfilled pauses, discourse markers, explicit editing terms, false-starts, truncations, identical and modified repetitions, and propositional and morphological substitutions. Crible and colleagues offered a fine-grained notation system, which allows the annotators to investigate specific combinations of (dis)fluencies forming recurrent combinatory patterns.

Models for Atypical Speech Production

Annotations of (dis)fluencies in atypical speech are carried out by researchers and speech therapists in order to evaluate the fluency of a speaker presenting a speech disorder. The severity of the disorder can then be characterised by several parameters

(e.g. (dis)fluency type, duration, frequency, among others). For instance, perceptual studies suggest that a naïve listener would tend to consider that longer lengthenings are more likely to be stuttering-related (Didirková et al., 2016).

To the best of our knowledge, among the various speech and language disorders, only stuttering requires specific annotation models because it presents unique types and/or features of disfluencies, such as blocks and mid-syllable pauses (see below). By contrast, disorders such as aphasia or cluttering can be covered with existing models for disfluency annotation developed for typical speech. Aphasia and cluttering do involve specific phonetic and semantic errors, but errors and disfluencies are typically considered as different categories and are annotated using different typologies, similarly to the treatment of error annotation in second-language corpora, for instance (e.g. Gilquin & De Cock, 2011). Therefore, the remainder of this literature review will focus on disfluencies in stuttering.

One of the annotation models used in speech therapy is Systematic Disfluency Analysis (henceforth SDA), used to identify and quantify a full range of disfluent behaviors from a panel of disfluencies ranging from typical patterns to behavior reflecting stuttering problems (Campbell et al., 1991; Campbell & Hill, 1987). SDA procedures are applied to orthographic transcriptions of speech samples obtained in various speaking situations. Besides qualitative judgements of instances of disfluency, other evaluations are made including both audible and visual aspects such as number of repetitions, duration of lengthenings, increases in tension, rate, loudness and pitch, and accompanying behavior such as "tics", syncinesia, or physical tensions, among others (Campbell et al., 1991; Campbell & Hill, 1987). Tension is mostly identified using video data, allowing for observation of the speaker's face. Scoring procedures are applied to the speech samples, which lead to reliable severity ratings and help in

differential evaluation and in evaluation of treatment effectiveness. Interestingly, SDA differentiates between "more stuttering-like" and "less stuttering-like" disfluencies. Within the first category, we find blocks, lengthenings, phoneme and syllable repetitions, and "stuttered word repetition". The second category encompasses "non stuttered" word repetitions, repetitions of word sequences, modifications, interjections, long (one second) hesitations, and incomplete words. This categorization can be problematic, since it considers that some disfluency categories such as word sequence repetitions, interjections or modifications are never due to stuttering, a claim which, we argue, should be taken with a grain of salt. Indeed, a speaker can introduce an interjection or a modification because they anticipate a problematic word, and inserting an interjection can help them to prepare the upcoming sound/syllable. As for word sequence repetitions, those can be stuttered, especially when the words are monosyllabic ("I w.. I w… I was").

Most of the time, however, researchers and clinicians use a simplified annotation system including three main types of "stuttering events" or "stuttering-like disfluencies": phoneme and syllable repetitions, blocks, and lengthenings. Note that some authors differentiate between (voiced) lengthenings and silent lengthenings, while others prefer to use the term "block" to describe a silent lengthening (Guitar, 2019; Lickley, 2017; Riley, 1994, p. 4; Seth & Maruthy, 2019).

The most widely used annotation system for assessing SLD is the Stuttering Severity Instrument (SSI) by Riley. More than a simple annotation system, the SSI (Riley, 2009) is a clinical assessment tool which takes into consideration several parameters. First of all, the tool considers the frequency, which is measured as a percentage of stuttered syllables, as well as disfluency duration. Physical concomitants during speech production, such as the use of distracting sounds, facial grimaces, head

movements, movements of the extremities, are also observed. By summing the subscores, the SSI allows for severity assessment of the person's stuttering (very mild, mild, moderate, severe, very severe). Riley (2009) considers that any silent or audible lengthening, as well as any sound or syllable repetition, is a SLD, while repetitions and reformulations of whole words and phrases are not considered as SLD.

Note that it is commonly acknowledged that people who stutter produce both SLD and other disfluencies. Thus, one of the most important problems for annotating speech in people who stutter is to define precise criteria for differentiating between a lengthening due to stuttering, which is in general supposed to be due to a motor-related problem, e.g. in a transition between the disfluent and the subsequent sound (e.g., Didirková & Hirsch, 2020), and a lengthening due to other linguistic processes, such as formulation (Lickley, 2017).

Limitations of the Previous Models

Based on this brief review of existing models in the (dis)fluency literature, three main shortcomings can be identified. Firstly, existing typologies often exclude a number of phenomena on the grounds that they are "intentional" (see Shriberg, 1994) or "fluent" (e.g. discourse markers are often excluded, as in Meteer et al., 1995). Intentionality as a criterion for inclusion or exclusion is questionable since it can apply to several types of disfluencies (filled and unfilled pauses, some repetitions might be intentional) and can never be asserted with any certainty. These elements are sometimes clearly related to disfluency (e.g. reformulative discourse markers like "well" or "I mean"), so that excluding them is bound to make the coverage of the annotation model only partial and non-exhaustive.

Secondly, most annotation schemes tend to focus exclusively on either typical or

atypical speech, with the exception of the FLUCALC project (Bernstein-Ratner & Brundage, 2019) which studies child data, and is bound to specific conventions (CHAT) as well as the technical format of the CLAN package (MacWhinney, 2000). It is therefore difficult to compare the distribution of (dis)fluencies in typical and atypical speech when the corpora have been annotated following different guidelines. Thirdly, not all models are easily replicable and applicable to other technical formats, given the lack of explicit notation systems in some typologies (e.g. Götz, 2013). Our model aims at addressing these issues and is presented in the next sections.

Towards an Inclusive Model: Introducing ANODIS

General Principles of the Present Model

We present an annotation model which overcomes methodological, technical, and conceptual limitations found in the previous frameworks. Firstly, we cover any element that is *potentially* disfluent, including at the discourse level. This means that we do not discriminate between fluent or disfluent (uses of) pauses or repetitions prior to the annotation. What we do exclude from the annotation are disfluencies that are caused by another speaker's intervention (other-interruptions, repetitions or lengthenings due to overlapping speech). Secondly, the scope of our model is broad and inclusive: while we only report data from native normal adult and stuttering adult speech in this paper, the proposal is designed to be applicable to native and non-native, typical and atypical, adult and child data alike.

Thirdly, in terms of format, we opted for a multi-layered hierarchical annotation structure (instead of enriched transcriptions, as in the CHAT conventions) and a user-friendly notation system that is compatible with natural language processing applications. While the model defines several categories and attributes, it is flexible

enough so that it allows the analyst to choose the level of granularity that is required by their specific research interests. It is hierarchical in two main ways: i) the annotation covers information of different levels, namely the main disfluency category on the one hand and specific attributes on the other (position, extent, subtype); ii) in complex sequences (e.g. multiple embedded repetitions), the annotator identifies the "main" encompassing structure while still accounting for its internal components (see the following section for details and illustrations).

Our system also includes eight verbal (dis)fluency categories, some of which can be subdivided for finer distinctions if required. It is described in detail in the following section.

Annotation categories

Simple disfluencies

We make a first distinction between simple and compound disfluencies. The former are local, one-part disfluencies. They are annotated on the [disf] tier (see next section), except for lengthenings which are annotated on a separate [leng] tier.

Lengthening - LG. This label applies to a non-phonological lengthening of the duration of a phoneme (a syllable) that is perceived as abnormal. Phonological lengthenings (e.g. final lengthenings in French, vowel quality in other languages) are not annotated. Lengthenings can apply to consonants (LG:c) or vowels (LG:v). They can also apply to other disfluencies such as discourse markers or filled pauses (see below). This is why lengthenings are annotated on a separate tier, so that the annotation is aligned to the lengthened segment (for automatic measurements). We propose to use the term "lengthening" both for typical and atypical speech (see below), and we do not

use the term “prolongation” in our annotation system (e.g. Eklund, 2001), for strictly terminological reasons.

Block - BL. A block is a disruptive silent pause characterised by visible or audible tension before or within a word. Blocks can affect consonants (BL:c) or vowels (BL:v). They can occur between words (BL:wo), between syllables (BL:sy) or within syllables, between phonemes (BL:pho). The smallest unit of reference is identified: a mid-syllable block is by definition mid-word, so we only annotate the lower level (syllable). Note that the term “block” is here used for disfluencies which are sometimes called “silent lengthenings” in the literature. A block can mainly be identified using a video recording associated with audio data. However, other studies also describe blocks from an articulatory viewpoint and allow for their identification based on articulatory data (e.g., a direct observation of supraglottic articulatory movements of the tongue; see Didirková et al., 2021, for more details).

Discourse marker - DM. A discourse marker is a syntactically optional expression that performs a pragmatic function, such as French *donc* ‘so’, *ben* ‘well’ or *tu vois* ‘you know’. To avoid the issue of discourse marker identification, we recommend using a closed list of expressions. Discourse markers can occur between utterances (DM:ut) or between words, within utterances (DM:wo). If an utterance starts (or ends) with multiple discourse markers and/or filled pauses, all clustered DMs are considered to be in the “between utterances” position, as long as they are outside the syntactic boundaries of the utterance. We still annotate each DM in separate aligned intervals.

Silent pause – PS. A silent pause is any absence of vocalization, without a predefined threshold. Only within-turn pauses are annotated. Silent pauses can occur between utterances (PS:ut), between words (PS:wo), between syllables (PS:sy) or

between phonemes (PS:pho). They can be distinguished from blocks by a total absence of perceptible tension.

Filled pause – PF. A filled pause is any non-lexical vocalization such as *euh* (/ø/) in French. Filled pauses can occur between utterances (PF:ut), between words (PF:wo), between syllables (PF:sy) or between phonemes (PF:pho).

Self-interruptions – SI. A self-interruption occurs when a lexical or discourse unit is left incomplete by the speaker without external cause (such as overlapping speech or other speakers' interventions). Self-interruptions are annotated at the final interval before the interruption. Interruptions caused by other speakers are not annotated. Self-interruptions can affect words (i.e. word fragments, truncations ; SI:wo) or utterances (SI:ut); in the latter case, no word is interrupted but the utterance is syntactically incomplete and never completed later. Self-interruptions thus differ from modifications (see below) mainly at the syntactic level: the new utterance cannot be integrated in the self-interrupted one, the two segments (the interrupted and the new one) do not have the same structure or the same function. With modifications, on the contrary, one (or several) item clearly replaces another one.

If a word fragment is then completed (e.g. *la mai- maison*), it is treated as a type of repetition (see below). If a word fragment is then substituted by a different word and the utterance continues (e.g. *j'étais dans la mai- le pavillon de mes parents*), it is treated as a modification (see below).

Compound disfluencies

These are multi-part disfluencies that involve at least two words. They are annotated on the [rm] tier.

Repetition – RI, RX. A repetition is the reiteration of one or several elements of various sizes without modification. The repetition can be identical (RI), in which case

the repeated elements are immediately adjacent and without any alteration or addition (e.g. no pause, no discourse marker between the repeated parts). The repetition can also be mixed (RX), in which case some non-propositional element can be found between the repeated parts (silent and filled pauses or discourse markers). Any sequence containing repeated elements and the addition, deletion or modification of at least one of the elements is treated as a modification (see below). Repetitions can apply to whole utterances (RI/RX:ut), multiple words (RI/RX:mult), single words (RI/RX:wo), syllables (RI/RX:sy) or phonemes (RI/RX:pho). Multiple repetitions of different levels (phonemes, word, multiple words) are often embedded, in which case it is technically impossible to align the intervals with the repeated material. To address this, we suggest to only annotate the first part of the repetition if the final repeated item is part of another repetition:

1. c/ c/ cet élé/ cet élément (*th- th- this ele- this element*)

   RI:pho RI:mult

In this constructed example, there is a phoneme repetition ("c/"); in the third instance of the repetition, the word is completed ("cet"): we align the annotation to the first two phones and label it "RI:pho". We also see a second repetition that covers two words "cet élément"; in the first instance of the repetition, one of the words is truncated ("élé/"): we use the "mult" tag here to account for the scope of the repetition over multiple words and/or multiple types of segments (here a word and a syllable).

Modification – MO, MR. A modification is a substitution, addition or deletion of linguistic material in the context of a repair sequence. The modification can either involve repeated elements (MR) or not (MO). The modification can be caused by a grammatical (MO:g ; MR:g), a lexical (MO:l ; MR:l) or a phonological issue (MO:p ; MR:p). In addition, the modification can apply to units of different natures: whole

utterances (e.g. MR:l:ut), multiple words (MR:l:mult), single words (MR:l:wo), syllables (MR:l:sy) or phonemes (MR:p:pho). We use the extension "mult" if at least one of the two parts of the modification includes two words or more or if the modified/added material is at least two words. The cause and extent of the modification are optional tags, depending on the research question. In the case of a long modification including repetition(s), we give priority to the alignment of the repetitions and hence only create an interval for the MR aligned to the "modifying" segment:

2. je n'ai pas euh _ je n'ai pas entrepris de de le faire enfin je n'ai pas choisi de le faire et euh _ et voilà (*I didn't uh I didn't start to to do it I mean I didn't choose to do it and euh and that's it*)

In this excerpt (Figure 1), there is first a mixed repetition of several words (RX:mult) covering "je n'ai pas" and the silent and filled pause in between. Then, we see a word repetition of "de" (RI:wo). Lastly, we can see a modification with repetition, caused by a lexical element and covering the whole passage ("je n'ai pas entrepris de le faire" repaired by "je n'ai pas choisi de le faire"). Because of the numerous embedded structures, we only annotate the modification (MR:l:mult) on the repairing segment, here "choisi". This illustrates once more the hierarchical decision-making process of the annotation. We use MR:l:mult here because the modification applies to multiple words, even though only one word is modified. The repeated material within the modification ("je n'ai pas … de le faire") is thus not aligned.

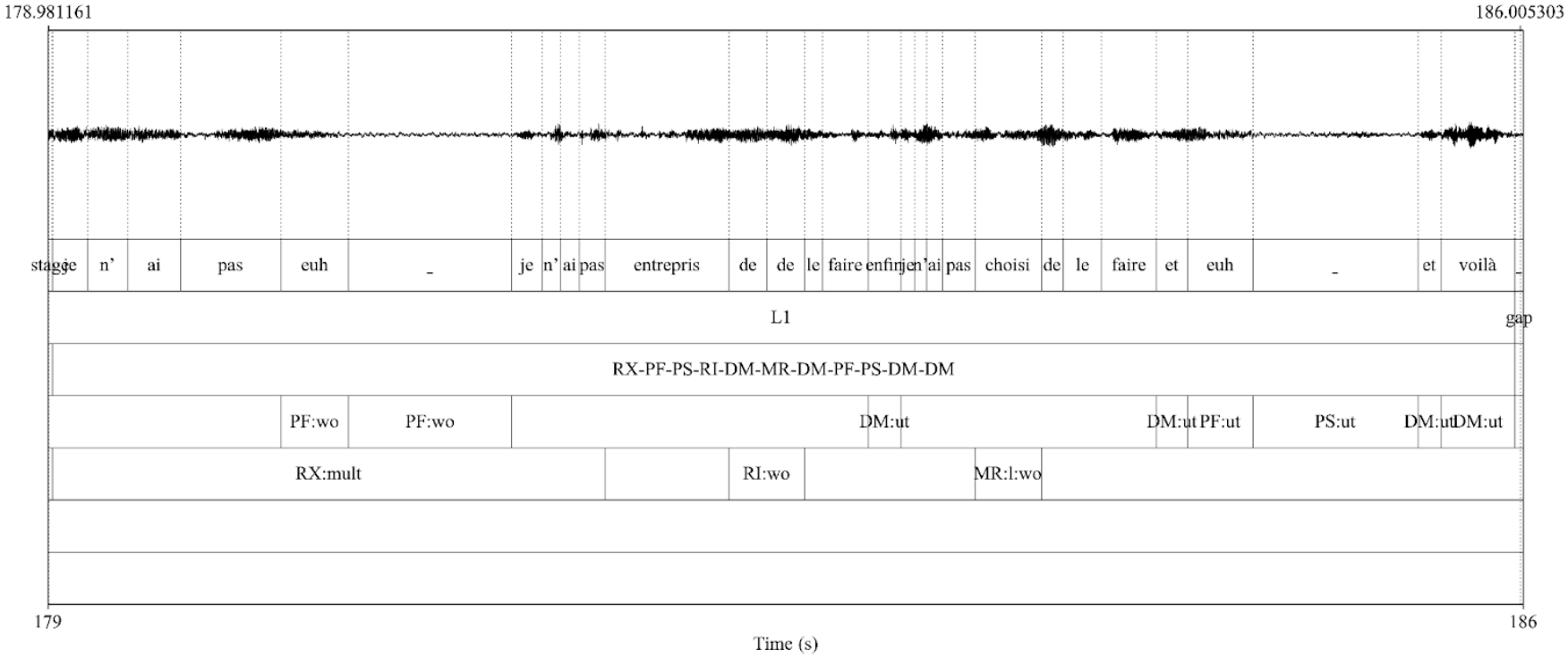

Figure 1: Annotation of example 2. Top to bottom tiers: word, speaker, seq, disf, rm, leng, para.

There can be a long distance between repeated words (i.e. a lot of inserted lexical material between the two parts of the repetition):

3. moi je suis euh enfin j'ai arrêté mes études ici en janvier donc euh je me suis inscrite en tant que demandeuse d'emploi et je suis à la recherche
   (*me I was uh I mean I quit my studies here in January so uh I enrolled as a job seeker and I am looking*)

The first "je suis" '*I am*' can be interpreted either as a self-interruption (SI:ut) or a modification with repetition (MR: the speaker decided to add some information before starting the sentence with "je suis à la recherche"). In those cases, the annotator has to decide how likely it is that the second repeated item was intended as a modified repetition of the first interrupted segment. In the example above, given the long distance between the two "je suis", and the high frequency of "je suis" in French, we would suggest a conservative bias to not over-interpret the utterance and only annotate it as a self-interruption.

Segmentation and annotation procedure

The specifications of scope (e.g. RI:wo, RI:mult), position (e.g. DM:ut) or repair source (e.g. MO:g) are optional and may not be necessary for all research questions, in which case the annotation only reports the main category label such as "RI" or "DM". The disfluency categories defined above are annotated and aligned to the transcript at any level that is relevant for the research question of the analyst (word or smaller). In addition to the three tiers mentioned above ([disf] for simple disfluencies, [leng] for lengthenings, [rm] for compound disfluencies), two further tiers may need to be added: one for the whole disfluent sequence [seq] and one to flag atypical disfluencies [path]. Optional tiers can also be added for paraverbal activity and speaker information.

Disfluent sequences – [seq]

On a separate tier, an interval is created to cover all adjacent disfluencies from the categories above. The tags for each interval take up the main disfluency categories (without indicators of position, extent or sound category) in their linear order in the sequence. Each tag is separated from the next by a hyphen (-).

We also repeat in this tier isolated disfluencies, in which case the tag appears on its own (e.g. PF). If a disfluency is affected by another disfluency (typically, lengthening), we specify this and write the "main" element from the [disf] tier first and the diacritic element second (e.g. PF-LG).

A sequence stops at any word interval that is not affected by any disfluency. If the transcript is segmented below word-level (phones, syllables) and some phones/syllables are unaffected by a disfluency, the disfluent sequence still continues over these fluent segments; in other words, our reference unit is the orthographic word.

Atypical disfluency – [path]

A "path" tag (for "pathological") will be assigned to any disfluency annotated on the other tiers if it is perceived as caused by a language, communication or voice disorder. "Pathological" disfluencies show specific characteristics that are disorder-dependent.

Paraverbal events – [para]

In this optional tier, paraverbal activity can be annotated, such as laughter, coughs or tongue clicks. This category further includes non-lexical vocalizations (*mm,* creaky voice) and respiratory conduct (tongue clicks, sigh, inbreaths, and other types of mouth noise). It can also cover secondary fluency-related behaviors such as averted gaze or so-called "explicit editing terms" such as "I can't speak" or "sorry" (Shriberg, 1994).

Speaker information – [spk]

In case of dialogues, the speakers can be identified in a [spk] tier, where the intervals will be aligned to speaker turns. Tags such as "spk 1" and "spk 2" can be used to refer to the different speakers.

Testing the model: agreement analysis and examples

Inter-annotator agreement measure

Four untrained annotators (undergraduate students with little to no experience in Praat and annotation) were asked to test our model on two different sound files containing semi-spontaneous speech:

- a young (female) adult with no known disorder who talks about her studies and job hunting to a (female) counsellor (5 min 07 sec);
- a young (male) adult who stutters with severe stuttering (4 min 26 sec) describing a typical day. His stuttering is mainly characterised by blocks.

It should be noted that the different time intervals for the disfluency tiers were not identified prior to the test, which means that the annotators had to delimit the tiers' boundaries themselves, which could have made the annotation process more difficult. This is further discussed below.

We calculated the Inter-Annotator Agreement (IAA) using the Fleiss' Kappa, a statistical measure allowing us to assess the reliability of agreement between the four coders when assigning the different categories of disfluencies available in the ANODIS model and particularly relevant for a group of multiple annotators.[6] First, we will

[6] Krippendorff (2004) points out several limitations of Fleiss' Kappa, such as its inability to account for individual preferences of annotators for particular categories. Despite its

present the results of IAA for the file corresponding to the adult speaker with no disorder, and then for the file corresponding to the adult speaker who stutters, both in their L1. This calculation was made on the [disf], [rm], and [leng] tiers using the {irr} package (Gamer et al., 2019) in RStudio (RStudio Team, 2021). If the raters are in complete agreement, $\kappa = 1$ and if there is no agreement among the raters, then $\kappa \leq 0$ (Landis & Koch, 1977)[7].

Typical speech

On the [disf] tier, annotators show a moderate agreement when it comes to identifying the presence vs. absence of a disfluency ($\kappa = 0.409$, $z = 11.6$, $p = 0.000$), with 134 items to annotate. On this tier, the annotators reach a moderate agreement as to the specific category of disfluency ($\kappa = 0.469$, $z = 18.7$, $p = 0.000$). Zooming in on individual labels, PF (30 in the corpus) are the most reliable to identify ($\kappa = 0.692$, Figure 2 on the left), while the other categories cause more problems to the raters, such as the 10 DM ($\kappa = 0.218$, fair, Figure 2, on the right), PS – also 10 – ($\kappa = 0.297$, fair) and the 2 BL ($\kappa = 0.16$, slight). This suggests that DM, PS and BL are quite challenging to annotate. This comes as no surprise: discourse markers are notoriously difficult to identify with a high degree of consensus, since they do not form a fixed grammatical category but are mainly defined on functional criteria (see Fischer, 2006). Silent pauses, in turn, appear to be confused with blocks by some of the annotators. In addition, the annotators were

---

drawbacks, Kappa is widely used in corpus linguistics and is therefore useful to compare with previous proposals.

[7] If $\kappa < 0$: less than chance agreement; $\kappa = 0.01$-0.20: slight agreement; $\kappa = 0.21$-0.40: fair agreement; $\kappa = 0.41$-0.60: moderate agreement; $\kappa = 0.61$-0.80: substantial agreement; $\kappa = 0.81$-0.99: almost perfect agreement; $\kappa = 1$: perfect agreement.

hesitant to identify pauses which they did not perceive as "disfluent", even though the instructions of the model specify that no judgment of fluency should be made prior to the annotation. The well-known multifunctionality of pauses (physiological, articulatory, structuring, rhetorical) and their varying duration might thus explain the low agreement score for PS, as already shown in previous studies (Oehmen, Kirsner & Fay, 2010; Zellner, 1995). As for blocks, our annotators are training to be speech therapists and, as such, are familiar with stuttering-specific disfluencies; they might have been thrown off by having to work on a non-stuttering speaker for this category, not knowing whether to focus on it or not.

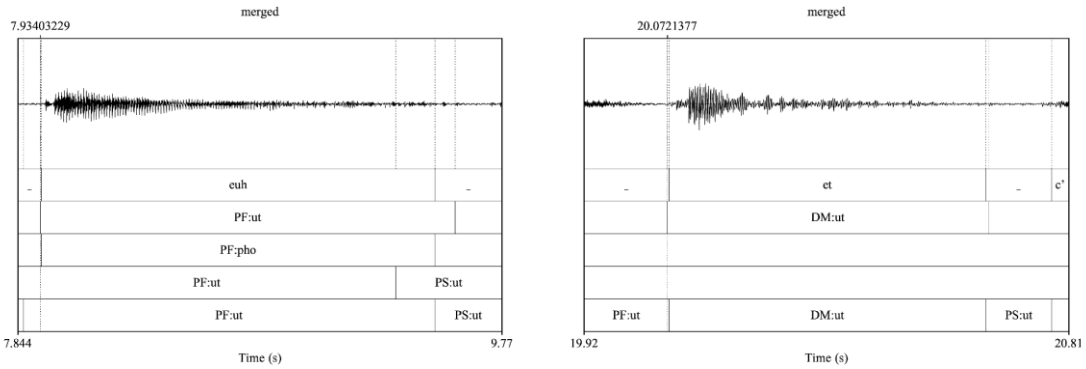

Figure 2: On the left, 4/4 coders have identified a PF: "*euh*" and in the right, 2/4 coders have correctly identified a DM in the utterance "*et- c'est vrai que (...)*" ("and – it's true that (…)").

In addition, we found poor agreement on the [leng] tier with 17 lengthening occurrences in the corpus ($\kappa = -0.117$, $z = -1.43$, $p = 0.152$), which reflects the well-known difficulty of identifying lengthenings (Eklund, 2001). As Eklund (2001) and Rohr (2016) discussed, it is very difficult to tell exactly when a segment is "prolonged", based on perception alone. In addition, lengthenings are very speaker-specific, which means that the average duration of a syllable may slightly differ according to the speaker, which makes the identification of a so-called "disfluent" lengthening more difficult.

By contrast, in the [rm] tier, annotators showed a substantial agreement in identifying whether or not a disfluency was produced on 47 intervals ($\kappa = 0.578$, $z = 11.9$, $p = 0.000$). As for the specific disfluencies identified, there is a moderate agreement between the four annotators ($\kappa = 0.474$, $z = 12.8$, $p = 0.000$). More specifically, the 14 RI are considerably easier to identify ($\kappa = 0.539$, $z = 11.124$, $p = 0.000$, moderate) than the 2 RX ($\kappa = 0.048$, $z = 0.997$, $p = 0.319$, slight), and the 2 MR ($\kappa = 0.130$, $z = 2.681$, $p = 0.007$, slight). Only 1 MO was identified in the corpus ($\kappa = -0.011$, $z = -0.220$, $p = 0.826$, poor). However, some of these scores are not significant because of the very low number of occurrences in our sample (in particular for RX and MO). Still, given the complexity of some of the repetition sequences in our sample (see section below) and the lack of training of our annotators, we take these results as encouraging.

Atypical speech (person who stutters)

Overall, the four annotators reach a lower agreement on atypical speech, with only fair agreement on the [disf] tier ($\kappa = 0.326$, $z = 13.7$, $p = 0.000$ on all items). BL is the most reliable category to identify, with a moderate agreement on the 50 occurrences in the corpus ($\kappa = 0.423$, $z = 12.852$, $p = 0.000$), while there is a fair agreement for PF (11 occurrences, $\kappa = 0.311$, $z = 9.452$, $p = 0.000$) and for PS (10 occurrences, $\kappa = 0.311$, $z = 9.452$, $p = 0.000$) and a slight agreement for the 23 DM ($\kappa = 0.143$, $z = 4.346$, $p = 0.000$). This, somehow low agreement, is however due to a confusion related to combined SLD (e.g., h…hhh…hello), where a phoneme repetition can be interrupted by what can be perceived as a block, and/or prolonged. This issue can be easily addressed by using all the available tiers.

For the [leng] tier, the agreement is slight ($\kappa = 0.174$, $z = 3.97$, $p = 0.000$), which was already been noted for typical speech. For the [path] tier, the agreement is moderate

(κ = 0.577, z = 16.7, p = 0.000), and fair for the [rm] tier overall (κ = 0.272, z = 7.7, p = 0.000). On this last line, there was almost perfect agreement when raters had to identify a disfluency or no disfluency (κ = 0.8, z = 13.6, p = 0.000). More specifically, there is a fair agreement for RI (κ = 0.333, z = 7.516, p = 0.000), a slight agreement for RX (κ = 0.176, z = 3.978, p = 0.000) and a poor agreement for MR (κ = -0.009, z = -0.201, p = 0.841).

In sum, ANODIS shows encouraging agreement scores considering the lack of training of the annotators and the complexity of the task, with a many fine-grained categories. Still, areas of improvement remain for this model and are expected: notoriously difficult items to identify (PF, DM) and/or complex embedded structures (repetitions). We will focus on the latter in the following section.

Focus on repetitions

One of the most innovative features of ANODIS, besides its coverage of both typical and atypical data, is the treatment of repetitions and modifications, and in particular, how to deal with the alignment and classification of complex embedded sequences. In addition to the examples provided in the previous section, we would like to illustrate the benefits of ANODIS and compare it with other annotation models in order to stress how it deals with such structures and hopefully improves their annotation compared to previous models.

Our first example, taken from a non-stuttering speaker, can be seen in the following screenshot and is transcribed in (4):

4. et q/ quel _ quel euh _ [tongue click] _ comment est-ce que vous en êtes arrivée à ce ce choix d'études-là
   and wh- what what uh how did you get to this this choice of studies

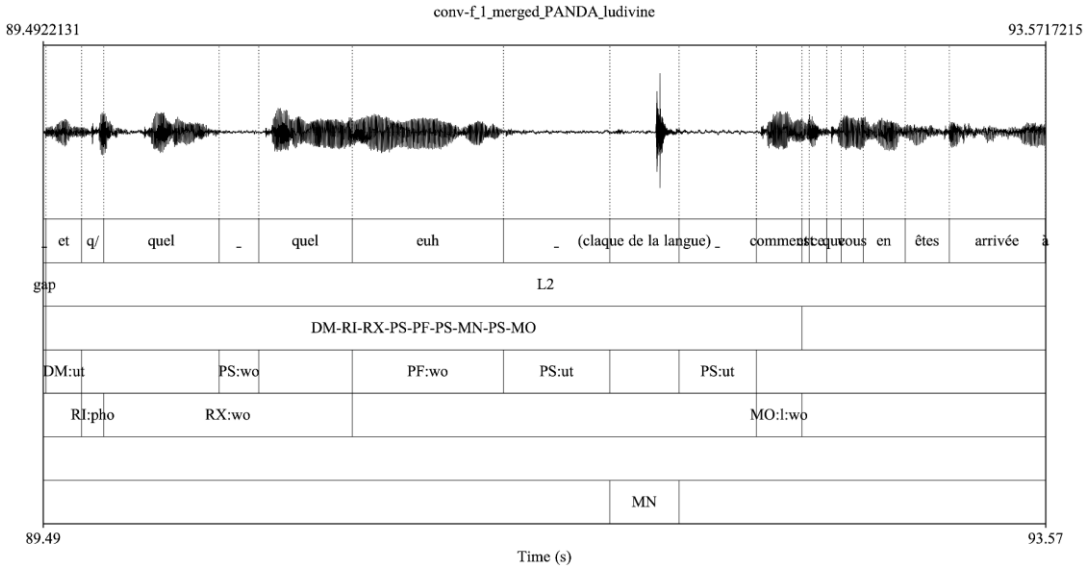

Figure 3: Example 4 annotated using ANODIS. Top to bottom tiers: word, speaker, seq, disf, rm, leng, para.

As can be seen on the [rm] tier (third from bottom, Figure 3), three events are aligned in this excerpt: an identical repetition of a phoneme (“q/”), a mixed repetition of a word (“quel”) also containing a silent pause, and a lexical modification of a word (“quel” substituted by “comment”). In terms of text-to-sound alignment, as explained above, only the first element of the identical repetition (the first “q/”) is aligned, since the second element is part of the next repetition (“quel quel”). Similarly, for the modification, only the substituting word “comment” is aligned, since the substituted element “quel” is part of a repetition, as is very often the case. Thus, we give priority to the exact alignment of repetitions (as this information might be interesting for researchers looking at repetition durations) over that of modifications, which can take many different forms. This precedence of repetitions for alignment purposes simplifies the annotation procedure and maintains high precision for the analysis of repetition duration, all of which is made possible by the hierarchical nature of the system. The only drawback to this alignment format is that it is not explicit in the annotation what is substituted by “*comment*”, although the [seq] tier does indicate that everything from “*et*

q/" to "*comment*" is part of a single disfluency sequence. The benefit, in turn, is that we keep track of the three separate phenomena, with precise information as to the subtype (RI/RX), extent (phoneme/word) and source (here, lexical), without over-complicating the annotation. As mentioned above regarding inter-annotator agreement, the [rm] tier received relatively high scores, which shows the ease and robustness of the process.

To better visualize the contributions of ANODIS against existing annotation models, Table 1 reproduces the same example following Crible et al.'s (2019) approach:

et	q/	quel	_	quel	euh	tongue click	_	comment
DM+RI+UP+FP+RM								
<DM>	<TR<RI0	TR>RI1><RI0	<UP>	RI1><SP0	<FP>		<UP>	SP1>
			<WI>		<WI>		<WI>	

Table 1. Example annotation from Crible et al.'s (2019) model

A number of differences can be observed between the two annotation formats:

- the intricate system of brackets and numbers in Crible et al. (2019) makes the transcript harder to read and possibly to annotate reliably;
- no distinction is made between the phoneme repetition and the word repetition;
- the same unit can be labeled with a cluster of three (or potentially more) labels, such as the first "*quel*", whereas ANODIS prioritizes and applies hierarchical rules that alleviate the annotation;
- although the numbers and brackets make it possible in principle to align and identify all parts of a compound disfluency in Crible et al.'s (2019) system, in practice it seems quite hard to do so automatically;
- no label is provided to account for the tongue click in their model ([paraverbal] tier in ours).

In sum, while both models have their drawbacks and advantages, ANODIS appears as both more economical (fewer labels) yet more precise (position of simple disfluencies, extent of repetition, cause of modification).

Another interesting comparison relates to the status of repeated elements in modifications. ANODIS distinguishes between, on the one hand, repetitions without modification of the content (RI, RX) and, on the other, repetitions which are part of substitutions (MR), themselves distinct from modifications without any repeated elements (MO, as in "*le _ la maison*"). The following excerpt (from the same conversation) presents a case of MR, where the preposition "*au*" is substituted by its feminine counterpart "*à la*" (thus, a grammatical source for the modification). As can be seen in the screenshot, the extent of the modification is a word (MR:g:wo), but a longer segment is aligned in order to account for the repetition of "*vraiment*".

5. euh vraiment confrontée au _ vraiment à la réalité

   uh really confronted to really to the reality

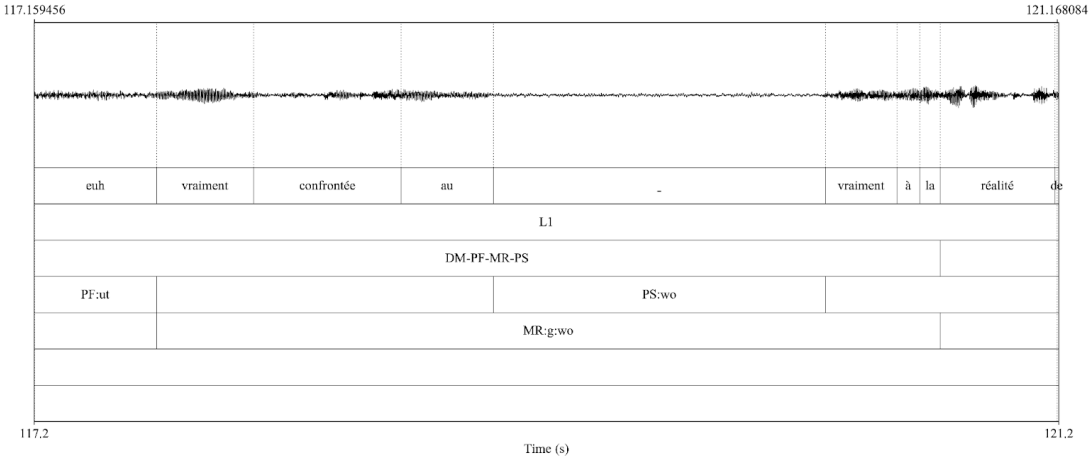

Figure 4: Example 5 annotated using ANODIS. Top to bottom tiers: word, speaker, seq, disf, rm, leng, para.

In doing so, we indicate i) that the whole MR interval is caused by a grammatical change, ii) that the main phenomenon is the modification and iii) that the repetition is only a side-effect of that modification (Figure 4). As previously, Table 2

reproduces the same example annotated according to the guidelines in Crible et al. (2019):

euh	vraiment	confrontée	au	_	vraiment	à	la
FP+RM+DE+SM+UP							
<FP>	<RM0	<DE>	<SM0	<UP>	RM1>	SM1	SM1>
				<WI>			

Table 2. Example annotation of a modification in Crible et al.'s (2019) model

The resulting annotation format is much more complex (eight intervals instead of three) without being necessarily more informative: the repetition of "*vraiment*" is labelled separately from the substitution, with the information that it is a "modification repetition" (RM). We believe that this is slightly confusing, since it is not the repeated element itself that is modified (in fact, "*vraiment*" is repeated exactly), but it is included in a modification. The main added value to the system by Crible et al. (2019) is the annotation of the deleted word "*confrontée*" (<DE> for "deletion"), which is not accounted for in ANODIS, although we argue that this brings only limited information.

In sum, our proposal leads to a change of perspective towards repetitions that operates a major distinction between repetitions as the main disfluency event (RI, RX) and repetitions as the result or side-effect of a modification, in which case the annotated label reflects a different disfluency main event (MR).

Conclusion

In this paper, we present ANODIS, a new and inclusive annotation model for disfluencies in typical and atypical speech, which addresses a number of limitations from previous proposals. In particular, ANODIS offers a hierarchical, flexible system for multi-layered annotation that i) integrates disfluencies related to stuttering, ii)

provides a solution for complex embedded sequences, iii) captures fine-grained information about position, subtype and extent of disfluencies and iv) pays particular attention to the alignment and status of repetitions (within or outside modifications). Its technical format was designed with the perspective of automatic post-treatment; annotated data will be used to train models for automatic disfluency annotation. We also showed that ANODIS can be reliably used by untrained annotators, although, like any annotation model, it requires some practice. The number of categories and diacritics is necessary to account for the complexity of spontaneous language use (in particular frequent embedded sequences) without losing too much precision.

We believe that the present model fills a gap in the comparison of typical and atypical speech and will benefit researchers from many fields with an interest for disfluencies. Large-scale annotation campaigns using ANODIS on various types of corpora (other speech and language disorders besides stuttering, first and second-language speakers, children and adults, multiple languages) will not only further attest the reliability of the model, but it will also broaden our understanding of how various factors impact on speech fluency, which can in turn lead to applications in the medical and pedagogical sectors.

While the inclusive approach of ANODIS has several benefits in terms of annotation reliability and analytical possibilities (e.g. easy comparison of categories across speakers), one might argue that the functional motive behind disfluencies such as modifications can be very different in a person who stutters and one who does not. In clinical settings, it would still be useful to be able to identify whether a word was modified because it was not exactly the intended one, or because it felt as “difficult” to produce. However, we believe that such a functional level of analysis can be added at a

later stage, if needed for research or clinical purposes, but should remain independent from the more neutral, formal identification stage that we propose in this paper.

Another avenue of this project is to combine it with gesture annotation. Whether it is in typical or atypical speech, (dis)fluency should not be seen as a strictly verbal process, but as a multimodal one as well, marked by visible resources such as eye gaze, facial expressions, and manual gestures (see Seyfeddinipur, 2006; Graziano & Gullberg, 2018; Kosmala, 2021). Preliminary work on (dis)fluency and gesture has in fact already been conducted as part of this project (see Dodane & Didirková, 2021) and showed differences in gesture use between two speakers (one stutterer and one non-stutterer), calling for further research in this area.

References

Allwood, J., Nivre, J., & AhIsén, E. (1990). Speech management - on the non-written life of speech. *Nordic Journal of Linguistics*, *13,* 1–48.

Bernstein-Ratner, N., & Brundage, S. B. (2019). *A clinician's complete guide to CLAN and PRAAT*. https://www.talkbank.org/manuals/Clin-CLAN.pdf

Bortfeld, H., Leon, S., Bloom, J., Schober, M., & Brennan, S. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*(2), 123–147.

Campbell, J., & Hill, D. (1987). Systematic disfluency analysis: Accountability for differential evaluation and treatment. *Miniseminar Presented to the Annual Convention of the American Speech-Language-Hearing Association.* November 1987. New Orleans, LA.

Campbell, J., Hill, D., & Driscoll, M. (1991). Systematic Disfluency Analysis: Using SDA to determine stuttering severity. *Annual Convention of the American Speech-Language-Hearing Association.* November 1991. Anaheim, CA.

Christodoulides, G., Avanzi, M., & Goldman, J.-P. (2014). DisMo: A morphosyntactic, disfluency and multi-word unit annotator. An evaluation on a corpus of French spontaneous and read speech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, 26–31.

Crible, L. (2018). *Discourse Markers and (Dis)fluency: Forms and functions across languages and registers*. Amsterdam: John Benjamins.

Crible, L., Dumont, A., Grosman, I., & Notarrigo, I. (2019). (Dis)fluency across spoken and signed languages: Application of an interoperable annotation scheme. In L. Degand, G. Gilquin, & A. C. Simon (Eds.), *Fluency and Disfluency across Languages and Language Varieties* (Corpora and Language in Use-Proceedings 4). Louvain-la-Neuve: Presses universitaires de Louvain.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language* (2nd ed.). Cambridge: Cambridge University Press.

Didirková, I., & Hirsch, F. (2020). A two-case study of coarticulation in stuttered speech. An articulatory approach. *Clinical Linguistics & Phonetics*, 34(6), 517-535.

Dodane C., & Didirková, I. (2021) Gestes manuels pendant les disfluences normales et bègues : Une étude préliminaire. *Séminaire AFCP – Phonétique Clinique.* May 2021. France (online).

Domingues, C. E. F., & Drayna, D. (2015). *The Genetics of Stuttering*. Hoboken, NJ: John Wiley & Sons.

Eklund, R. (2001). Prolongations: A dark horse in the disfluency stable. *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*. August 2001. Edinburgh, Scotland, UK.

Eklund, R. (2004). *Disfluency in Swedish human–human and human–machine travel booking dialogues* [PhD Thesis]. Linköping University Electronic Press.

Etchell, A. C., Civier, O., Ballard, K. J., & Sowman, P. F. (2018). A systematic literature review of neuroimaging research on developmental stuttering between 1995 and 2016. *Journal of Fluency Disorders*, *55*, 6–45.

Frigerio-Domingues, C. E., Gkalitsiou, Z., Zezinka, A., Sainz, E., Guttierez, J., Byrd, C., Webster, R., & Drayna, D. (2019). Genetic factors and therapy outcomes in persistent developmental stuttering. Journal of Communication Disorders, 80, 11-17.

Gilquin, G., & De Cock, S. (2011). Errors and disfluencies in spoken corpora: Setting the scene. *International Journal of Corpus Linguistics*, *16*(2), 141.

Ginzburg, J., Fernández, R., & Schlangen, D. (2014). Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, *7*, 9–1.

Götz, S. (2013). *Fluency in native and nonnative English speech*. Amsterdam: John Benjamins.

Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology*, *9*, 1-17. https://doi.org/10.3389/fpsyg.2018.00879

Guitar, B. (2019). *Stuttering: An integrated approach to its nature and treatment* (5th Ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Hough, J., Tian, Y., De Ruiter, L., Betz, S., Kousidis, S., Schlangen, D., & Ginzburg, J. (2016). Duel: A multi-lingual multimodal dialogue corpus for disfluency, exclamations and laughter. *10th Edition of the Language Resources and Evaluation Conference*.

Kosmala, L. (2021). *A multimodal contrastive study of (dis)fluency across languages and settings: Towards a multidimensional scale of inter-(dis)fluency* [Unpublished PhD thesis]. Sorbonne Nouvelle.

Krippendorff, K. (2004). Reliability in content analysis some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, *14*, 41–104. https://doi.org/10.1016/0010-0277(83)90026-4

Lickley, R. (2017). Disfluency in typical and stuttered speech. *Fattori Sociali e Biologici Nella Variazione Fonetica-Social and Biological Factors in Speech Variation*, 373–387. https://doi.org/10.17469/O2103AISV000019

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Hove, UK: Psychology Press.

Meteer, M. W., Taylor, A. A., MacIntyre, R., & Iyer, R. (1995). *Dysfluency annotation stylebook for the switchboard corpus*.[PhD thesis]. University of Pennsylvania Philadelphia, PA.

Monfrais-Pfauwadel, M. C. (2014). *Bégaiement, bégaiements. Un manuel clinique et thérapeutique*. De Boeck-Solal.

Moniz, H. G. S. (2013). *Processing disfluencies in European Portuguese* [Unpublished PhD thesis]. Universidade de Lisboa.

Oehmen, R., Kirsner, K, & Fay, N. (2010). Reliability of the manual segmentation of pauses in natural speech. In H. Loftsson, E. Rögnvaldsson & S. Helgadóttir (Eds), *Advances in Natural Language Processing*, 263–268. Berlin: Springer.

Pallaud, B., Rauzy, S., & Blache, P. (2013). Auto-interruptions et disfluences en français parlé dans quatre corpus du CID. *TIPA. Travaux interdisciplinaires sur la parole et le langage*, *29*. https://doi.org/10.4000/tipa.995

Riley, G. (1994). *The Stuttering Severity Instrument for Adults and Children (SSI-3)*. PRO-ED.

Rohr, J. L. (2016). *Acoustic and Perceptual Correlates of L2 Fluency: The Role of Prolongations* [PhD Thesis]. University of Texas.

Seth, D., & Maruthy, S. (2019). Effect of phonological and morphological factors on speech disfluencies of Kannada speaking preschool children who stutter. *Journal of Fluency Disorders*, *61*, 105707.

Seyfeddinipur, M. (2006). *Disfluency: Interrupting speech and gesture*. [PhD Thesis], Radboud University.

Shriberg, E. E. (1994). *Preliminaries to a Theory of Speech Disfluencies* [PhD Thesis]. University of California.

Strassel, S. (2003). *Simple Metadata Annotation Specification Version 5.0–May 14, 2003*.

Zellner, B. (1995). Pauses and the temporal structure of speech. In E. Kellner (ed.), *Fundamentals of Speech Synthesis and Speech recognition*, 41–62. Chichester: John Wiley.