Distinguishing Intentional from Unintentional Disfluencies in Psychotherapy

Maayan Babkoff[1], *Simon Betz*[2], *Eshkol Rafaeli*[3], *Gideon E. Anholt*[1]

[1]Ben Gurion University of the Negev, Department of Psychology

[2]Bielefeld University, Linguistics Department, Phonetics Workgroup

[3]Bar-Ilan University, Department of Psychology

babkoffm@post.bgu.ac.il, ganholt@bgu.ac.il

Abstract

In psychotherapy, therapists' disfluencies, such as pauses, repetitions, and elongations, have received little direct study. While these speech patterns are often assumed to reflect cognitive or emotional strain, emerging perspectives suggest they may also serve communicative functions. This pilot study investigates how consistently external raters classify disfluencies in therapy as intentional or unintentional, and how these classifications align with therapists' retrospective reports. Using coded segments from real therapy sessions, we analyzed inter-rater agreement and modeled how contextual and structural features influenced judgments. Results show low agreement overall, with higher consistency during less structured therapeutic segments. Additionally, disfluencies at syntactic boundaries were more likely to be perceived as intentional, suggesting that structural cues might bias interpretation. These findings highlight the need for refined analytic tools to better distinguish intentional from unintended speech phenomena in clinical and psycholinguistic research.

Index Terms: psychotherapy, verbal disfluencies, intentionality, video recall method

1. Introduction

1.1. Verbal communication in psychotherapy

Psychotherapy is an evidence-based approach shown to alleviate a wide range of psychological issues, including anxiety, depression and personality disorders [1, 2]. At its core lies verbal interaction, which enables emotional expression, strengthens the therapeutic bond, and guides the therapeutic process [3]. Although the efficacy of many forms of psychotherapy is supported by research, the specific mechanisms, especially those involving therapist communication, remain under investigation [4].

Speech serves as the primary tool through which therapists connect with clients, support emotional insight, and build alliances [3]. While research has often focused on *what* therapists say, much less attention has been given to *how* they speak, especially in terms of fluency and disfluency [5]. Although disfluencies are often perceived as disruptions, their communicative or cognitive role in therapy remains insufficiently explored [5, 6]

1.2. Disfluencies

Disfluencies refer to moments of speech flow deviating from the ideal, often including fillers, pauses, elongation, and repetitions [7, 8] While traditionally regarded as signs of disruption, more recent research has shown that disfluencies can structure discourse, support listener comprehension, increase task performance, and signal cognitive effort to the listener [8, 9, 10]. These interruptions often occur during processes such as lexical retrieval, syntactic planning, or conceptual formulation, and are no longer viewed solely as signs of communicative failure [8, 11], up to the point that some researchers replaced the term itself by more positively connoted ones like "own communication management" [12] or "fluenceme" [13].

Researchers continue to debate what qualifies as a disfluency, and, moreover, how to refer to phenomena subsumed under this term, reflecting a lack of universal consensus on its definition (see Eklund, 2004, for a thorough overview). Some researchers view them strictly as unintentional breakdowns in speech, while others suggest that certain forms, like fillers or pauses, may serve communicative or pragmatic functions [8, 9, 14]. Regardless of their intentionality, all disfluencies may be perceived and interpreted by the listener, which may infer difficulty [15], (un)certainty [16], (in)competence [17] or (un)truthfulness [18] levels of the utterance due to the presence of disfluency.

This conceptual ambiguity is especially relevant in contexts such as psychotherapy, where the line between spontaneous disruption and deliberate rhetorical use may be particularly blurred (see Section 1.3). These complexities highlight the need for closer investigation of how disfluency manifests in therapeutic contexts, and under which conditions it might take on communicative or strategic significance.

1.3. Psychotherapy as a unique case: Pseudo-disfluency

Psychotherapy presents a unique context in which disfluencies may be intentionally employed as part of the therapeutic method. Therapists sometimes use pauses, hesitations or repetitions to guide reflection, emphasize emotional content, or regulate the rhythm of conversation [5]. In such cases, disfluencies may not indicate difficulty, intentionally or as a linguistic by-product [19] , but rather serve as deliberate rhetorical tools, with the therapist *mimicking* the phonetic surface form of disfluency. This raises the possibility that certain speech patterns commonly classified as disfluencies, should instead be viewed as "pseudo-disfluencies", i.e. intentional hesitations [5], terms, which we thus use interchangeably in this study. This stands in contrast to casual or informational discourse, where such vocal interruptions are more often signs of spontaneous processing difficulty rather than communicative intention. In therapy, by contrast, disfluency can function as a deliberate communicative gesture, signaling

10.21437/DiSS.2025-10

attunement and inviting reflection within the therapeutic exchange. Regardless of whether such forms are classified as disfluencies or something else, the ability to distinguish between intentional and unintentional uses is crucial for understanding the function of speech in psychotherapy.

2. the present study

2. 1. Research aim and rationale

Building on the theoretical distinction between intentional and unintentional disfluencies, the current study aims to examine whether external observers can reliably distinguish between the two in real therapeutic dialogue.

2.2. Research question

The study addresses two primary questions: (1) What is the level of *inter-annotator agreement* among external raters when classifying therapist disfluencies as intentional or unintentional? (2) To what extent do these external classifications align with the *gold standard* of therapist self-assessments regarding the intentionality of their disfluencies? This comparison with a therapist-provided gold standard is a potential contribution to the field, as such internal benchmarks are often absent in related research. Note that gold standard is not exactly the same as ground truth in this case, as speakers might not be able to judge their own performance correctly. Still, we believe that it is a valid gold standard, especially against the background of our understanding of therapists' rhetorical usage potential of disfluencies (cf. Section 1.3).

2.3 Study design overview

To begin addressing these questions, we conducted an initial pilot study using segments from sessions conducted by two trained therapists. In this stage, six external raters independently coded selected disfluencies and provided classifications based on perceived intentionality. The raters were undergraduate psychology student assistants, with basic psycholinguistic knowledge and no prior clinical training or experience as psychotherapy practitioners**.** The sessions were drawn from a larger, still ongoing, clinical intervention study conducted at Bar-Ilan University (Rafaeli, unpublished data). Therapist self-assessments of disfluency intentionality were collected after the original therapy sessions using the video-recall paradigm [20]. Due to logistical constraints related to the broader study timeline, the video recalls were recorded 6-7 months after the sessions. For future research, video recalls will take place shortly after the session. This broader study was designed to evaluate a structured, single-session therapy protocol targeting anxiety and depression, and included repeated measures of therapeutic process and outcome (Rafaeli, unpublished data).

The disfluency segments analyzed in the present study were drawn from three distinct intervention components embedded within that broader protocol: psychoeducation, brainstorming, and imagery-based work (i.e., imagery rescripting). These components were selected for focused analysis due to their distinct conversational and emotional demands, though they represent only part of the full therapeutic structure. In psychoeducation, therapists provided structured explanations about emotional reactions and unmet needs, aiming to normalize and clarify client experiences. Brainstorming segments involved collaborative generation of behavioral or emotional strategies in response to specific challenges raised by the client. Imagery-based work guided clients through rescripting emotionally significant memories via guided visualization, with the goal of transforming maladaptive emotional patterns through experiential reprocessing. For the full instruction script, including rating definitions and examples, see Appendix in Section 5.

In this paper we take a novel, yet exploratory, approach venturing into the effects of intentional and unintentional production of therapists' disfluencies by explicitly asking speakers and listeners about their (perception of) intentionality. Doing so, we hope to gain valuable insights both on the level of foundational linguistics research and on the level of applied psychotherapy, for which it might be crucial to understand the effect of intentionality in disfluency production.

2.4 Tools

To promote consistent classification of disfluencies, both therapists and external raters received a standardized instruction script clarifying how to distinguish between intentional and unintentional disfluencies. Both were asked to judge whether a disfluency served a rhetorical purpose (e.g., guiding the patient or emphasizing a point) or reflected internal processing (e.g., lexical search or emotional hesitation).

Following insights from the pilot, we revised the instructions provided to both therapists and raters, as it became clear that the term "intention" was not consistently understood. In the updated instructions, we explicitly defined in which contexts a disfluency should be considered intentional, such as when used to prompt patient reflection or manage conversational rhythm, and in which contexts it should not, such as when arising from lexical retrieval difficulties or cognitive overload. The revised version aimed to reduce ambiguity and promote more consistent judgments across raters. (see Appendix in Section 5 for both versions)

To analyze the disfluencies in the recorded sessions, we used the *Praat* software [21] to segment and annotate speech data. Each disfluency was coded according to its syntactic position, occurring either within or between syntactic units, as well as its formal type, including silent pauses, filled pauses, elongations, self-editing and repetition [8, 11]. These linguistic features were used in subsequent statistical models to examine their relation to rater judgments and therapist self-assessments. We reviewed each excerpt with the therapist or rater, prompting them to reflect on whether the interruption appeared intentional or unintentional. Separately, disfluencies were also coded for their type (e.g., silent pause, repetition, editing) and syntactic position (e.g., within or between clauses), allowing for further analysis of their contextual placement.

2.5 Preliminary pilot results

Initial findings involving two therapists and two therapy sessions revealed limited agreement among external raters when classifying disfluencies as intentional or unintentional. Inter-rater agreement was generally low: Pairwise Cohen's Kappa values varied widely, and the overall Fleiss' Kappa across all raters was 0.295 indicating only fair reliability.

Notably, the syntactic position of a disfluency appeared to influence raters' judgments: disfluencies occurring between syntactic units were more often classified as intentional compared to those within clauses. Out of 78 disfluencies coded in session 1, 49 occurred within syntactic units and 29 between units. Of the between-unit disfluencies, 76% were classified as intentional by the majority of raters, compared to only 48% of within-unit disfluencies. Additionally, agreement levels varied across intervention types, with the highest reliability observed during imagery-based interventions (Fleiss' Kappa = 0.45, n = 14) and lower consistency during psychoeducation ($\kappa = 0.21$, n = 10) and brainstorming phases ($\kappa = 0.18$, n = 9).

Additionally, the logistic regression models demonstrated high predictive accuracy (AUC ≈ 0.91), underscoring the influence of linguistic cues on rater perceptions. We modeled whether a disfluency would be classified as intentional by the majority of external raters. Predictor variables included syntactic position (within vs. between syntactic units), disfluency type (e.g., filled pause, repetition, elongation, silent pause), and the type of therapeutic intervention. We also examined interaction effects between disfluency type and intervention context. The model revealed that disfluencies occurring at syntactic boundaries, particularly filled pauses during imagery-based interventions, were significantly more likely to be perceived as intentional. These findings suggest that certain structural and contextual cues may bias raters toward attributing intentionality

To assess whether aggregating judgments could enhance classification accuracy, we analyzed majority-vote outcomes. The majority classification aligned with the therapist's own labeling in 92.5% of disfluencies for one therapist and 68.5% for the other, yielding an average agreement rate of 81% across both.

However, a closer analysis revealed a systematic pattern in misclassifications: most occurred in disfluencies positioned between syntactic units. In these instances, raters tended to interpret the disfluency as intentional, even when therapists had not. This suggests that while majority voting improves overall accuracy, it may also amplify specific biases. In particular, syntactic position appears to skew perception, leading raters to over-attribute intentionality in structurally salient contexts.

Examples:
Unintentional disfluencies:
1. "We need to have a uhm uhm sense. that we can do something." (coded as unintentional by all raters and therapist)
2. "Can you just describe to me uhm what could go wrong?" (, coded as unintentional by five of six raters and therapist)
Intentional disfluencies:
1. "That means that after, let's say - [pause] - two months, you can switch to a different technique." (coded as intentional by all raters and therapist)
2. "It's now eight in the evening - [pause] - you see on the clock - [pause] - that it's time..." coded as intentional by all raters and therapist)

2.6 Next Steps & Future Directions

To strengthen and extend these preliminary findings, we plan to replicate the study with additional therapists and a larger sample of therapy sessions. This expansion will allow us to assess the generalizability of rater agreement patterns and to examine variability across different therapists and therapeutic styles. We also aim to include raters with clinical experience to explore whether familiarity with therapeutic discourse improves classification accuracy.

Following insights from the pilot, we revised the instructions provided to both therapists and coders, as it became clear that the term "intention" was not consistently understood. In the updated protocol (see Section 2.4 (Tools)), we explicitly defined in which contexts a disfluency should be considered intentional, such as when used to prompt patient reflection or manage conversational rhythm, and in which contexts it should not, such as when arising from lexical retrieval difficulties or cognitive overload. This clarification was intended to reduce ambiguity and promote more consistent judgments across raters.

Unlike the pilot phase, in which therapists' self-assessments of disfluency intentionality were collected 6–7 months after the therapy sessions, future therapist self-assessments will be collected immediately after the recorded session, in close proximity to the speech event itself. We recognize that this extended delay may have limited the accuracy of therapists' retrospective judgments, particularly regarding subtle speech features such as disfluency. By reducing the temporal gap, we expect to obtain more reliable and ecologically valid data, and to observe stronger alignment between therapists' reports and external ratings. Finally, increasing the dataset will enable more robust modeling of the linguistic and contextual factors that shape how disfluencies are intended and perceived.

3. Discussion

The current study explored whether external raters can reliably distinguish between intentional and unintentional therapist disfluencies, and how these classifications align with therapists' own judgments. Preliminary findings revealed low to moderate agreement between raters, suggesting that identifying the speaker's intent behind disfluencies is not always straightforward and requires nuanced interpretation. At the same time, certain patterns emerged: disfluencies located between syntactic units were more likely to be perceived as intentional, and agreement was higher in emotionally immersive intervention segments, such as imagery work. These findings point to the potential role of both linguistic structure and therapeutic context in shaping how disfluencies are interpreted.

These findings (see Section 2.5) intersect with a broader theoretical debate in the linguistics literature regarding how disfluencies are defined and whether they must, by definition, be unintentional. The concept of pseudo-disfluency may be interpreted either as an extension of the disfluency category or as a distinct communicative form. Regardless of classification, our findings underscore the need to distinguish between intentional and unintentional disruptions in speech.

In the therapeutic context, this distinction becomes particularly relevant: therapists may deliberately adopt speech patterns that simulate hesitation, often associated with cognitive or emotional processes like uncertainty, in order to guide the patient, mark empathy, or create space for reflection. This communicative use of disfluency reflects a core characteristic of therapeutic speech,

its deliberate layering of form and meaning. In this setting, what may superficially appear as hesitation can actually serve as a relational signal or empathic gesture. Recognizing this helps clarify why disfluencies in therapy cannot be treated as mere processing artifacts, as they often carry communicative intent that is embedded within the therapeutic alliance itself.

Inconsistencies in rater agreement further suggest that pseudo-disfluencies blur the boundary between planned and unplanned speech, revealing the interpretative ambiguity surrounding speaker intent. This underscores the need for clearly defined coding criteria and well-calibrated task instructions; as our revised protocol aims to demonstrate, more precise definitions of "intentionality" may lead to greater inter-rater reliability in future phases of the study.

Beyond their theoretical implications, these findings may have practical relevance for clinical training and supervision. If disfluencies can serve intentional functions within therapeutic dialogue, such as regulating emotional tone, encouraging patient reflection, or signaling relational attunement, then helping therapists become aware of their own speech patterns may enhance their communicative precision. Integrating sensitivity to disfluency into training could also improve therapists' ability to attune to clients' nuanced responses, particularly in emotionally charged or complex sessions. Moreover, cultivating awareness of pseudo-disfluencies might offer an additional layer of self-reflection for therapists, supporting more deliberate and empathic interventions.

Importantly, the ability to distinguish pseudo-disfluencies from genuinely unintentional ones may also enable future research to explore the role of spontaneous disfluencies in therapy more directly. Once pseudo-disfluencies are identified and excluded, unintentional disfluencies may serve as indicators of therapist cognitive load, emotional strain, or even intervention difficulty, potentially offering an implicit marker of therapist skill and in-session demands.

Several limitations should be noted when interpreting the current findings. First, the sample size was small and limited to two therapists, which restricts the generalizability of the results. In addition, inter-rater variability may have been influenced by differences in evaluators' familiarity with therapeutic discourse, as most raters were not clinicians themselves. The study also relied on retrospective therapist self-assessments, which may have been affected by memory or hindsight bias. Lastly, although our definitions of intentionality were clarified through updated instructions, subjective interpretation of disfluencies remains a challenge and may still vary depending on context, personal bias, or training background.

An open theoretical direction for future studies concerns the underlying motivation or communicative strategy behind therapists' intentional disfluency production. Do therapists produce disfluencies merely to mimic their phonetic surface form, as a way to align with the patient's conversational rhythm, or do they also mimic their function, that is, create the impression of cognitive or emotional difficulty, in order to signal deeper engagement with the patient's internal experience? In most cases, such interpretations remain speculative, as naturalistic speech data does not typically allow access to speakers' intentions. However, in paradigms such as the one employed in the present study, where therapists are interviewed about their perceived intentions, these layers of meaning become accessible, offering a promising opportunity to explore disfluency not only as a behavior but as a form of strategic social signaling, opening new paths for research at the intersection of psycholinguistics and therapeutic interaction.

4. Conclusions

This study offers an initial step toward clarifying the role of intentionality in therapist disfluencies and highlights the need to distinguish between spontaneous and rhetorically motivated disruptions in speech. By exploring how external raters interpret therapist disfluencies, and to what extent their judgments align with therapists' own assessments, we begin to uncover the complex, context-dependent nature of speech in psychotherapy. Although agreement between raters was limited, these findings raise important questions about whether certain disfluencies can be reliably identified as intentional by listeners, and under which conditions. The introduction of the *pseudo-disfluency* category provides a framework for understanding intentional disfluency as a communicative tool unique to the therapeutic setting. Future research will be essential for refining this distinction and for exploring how different types of disfluencies reflect therapist skill, cognitive load, and relational engagement. Improved understanding of these phenomena may contribute to both theoretical models of therapeutic communication and the practical training of clinicians.

5. Appendix

The first version instruction script: "*I will now go over parts of a therapy session with you. When we speak, sometimes our flow of speech is interrupted for various reasons. Now I would like you to tell me every time I ask you whether the flow was intentionally interrupted, meaning for the conversation, for the therapeutic goal, or for any other reason, or not. That is, if it was not intentional. I need you to decide yes/no every time, even if it's difficult. But I would also appreciate it if you could share with me when it's less clear to you.*"

The updated instruction script: "*I'm going to go over a few parts of the session with you now. When we speak, our fluency is sometimes interrupted for various reasons. Now, I'd like you to tell me, each time I ask, whether the interruption in speech was intentional or unintentional. When I say intentional, I mean that it was used as a rhetorical device, for emphasis, managing the conversation, guiding the patient, or inquiry. When I say unintentional, I mean that the fluency was interrupted due to a thought process, a change in strategy, searching for words or refining them, or stemming from uncertainty, self-criticism, an emotional response, or any other reason. I will need you to decide yes or no each time, even if it's difficult. However, I'd appreciate it if you shared with me when it is less clear to you or if you feel the answer is not definitive.*"

6. References

[1] A. Dragioti, A. Karathanos, M. Gerdle and P. Evangelou, "Does psychotherapy work? An umbrella review of meta-analyses of randomized controlled trials," European Psychiatry, vol. 41, pp. 60–71, 2017.

[2] S. D. Miller, B. L. Hubble and D. H. Duncan, The Heart and Soul of Change: Delivering What Works in Therapy, 2nd ed. Washington, DC: American Psychological Association, 2013.

[3] D. E. Orlinsky, E. Heinonen and A. Hartmann, "The psychotherapeutic process: The common core and its variants," in Bergin and Garfield's Handbook of Psychotherapy and Behavior Change, 6th ed., M. J. Lambert, Ed. New York: Wiley, 2015, pp. 31–63.

[4] B. E. Wampold and Z. E. Imel, The Great Psychotherapy Debate: The Evidence for What Makes Psychotherapy Work, 2nd ed. New York: Routledge, 2015.

[5] H. Sacks and G. E. Anholt, "Therapeutic Disfluency: Disruption as Method," Discourse Studies, vol. 25, no. 4, pp. 420–439, 2023.

[6] B. Douglass, A. Baptista, R. H. Horwitz and L. L. Foster, "Hesitation and fluency in psychotherapy discourse: Markers of empathy and engagement," Psychotherapy Research, vol. 30, no. 6, pp. 651–664, 2020.

[7] R. Eklund, Disfluency in Swedish Human–Human and Human–Machine Travel Booking Dialogues, Ph.D. dissertation, Linköping University Electronic Press, 2004.

[8] R. J. Lickley, "Fluency and disfluency," in The Handbook of Speech Production, M. A. Redford, Ed. Hoboken, NJ: Wiley, 2015, pp. 445–469.

[9] H. H. Clark and J. E. Fox Tree, "Using uh and um in spontaneous speaking," Cognition, vol. 84, no. 1, pp. 73–111, 2002.

[10] S. Betz, B. Carlmeyer, P. Wagner and B. Wrede, "Interactive hesitation synthesis: Modelling and evaluation," Multimodal Technologies and Interaction, vol. 2, no. 1, p. 9, 2018.

[11] G. Christodoulides, "Disfluency detection using deep neural networks," in Proc. INTERSPEECH 2016, pp. 3608–3612.

[12] J. Allwood, "Reasons for management in spoken dialogue," NATO ASI Series F: Computer and Systems Sciences, vol. 142, pp. 241–254, 1995.

[13] S. Götz, Fluency in Native and Nonnative English Speech, Amsterdam: John Benjamins Publishing, 2013.

[14] P. E. Engelhardt, "Disfluency and the cognitive processes of speech production," in Research in Language and Cognition, L. Carlson and C. Hölscher, Eds. Springer, 2017, pp. 275–292.

[15] J. E. Arnold, C. L. Hudson Kam and M. K. Tanenhaus, "If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension," Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 33, no. 5, pp. 914–930, 2007.

[16] S. Betz, S. Zarrieß, É. Székely and P. Wagner, "The greennn tree-lengthening position influences uncertainty perception," in Proc. INTERSPEECH 2019, Graz, Austria, pp. 3990–3994.

[17] K. Fischer, O. Niebuhr, E. Novák-Tót and L. C. Jensen, "Strahlt die negative Reputation von Häsitationsmarkern auf ihre Sprecher aus?," in Proc. DAGA 2017, Kiel, Germany, pp. 1450–1453.

[18] B. De Keersmaecker, R. J. Hartsuiker and A. Pistono, "(Don't) believe me, I'm telling the truth! Speech disfluency and eye contact as cues to veracity, intention, and truth judgement," Language, Cognition and Neuroscience, vol. 39, no. 10, pp. 1263–1277, 2024.

[19] I. R. Finlayson and M. Corley, "Disfluency in dialogue: An intentional signal from the speaker?," *Psychonomic Bulletin & Review*, vol. 19, no. 5, pp. 921–928, 2012.

[20] O. Türk, S. Lazarov, Y. Wang, H. Buschmeier, A. Grimminger, and P. Wagner, "Predictability of understanding in explanatory interactions based on multimodal cues," in *Proc. 26th Int. Conf. on Multimodal Interaction*, 2024, pp. 449–458.

[21] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer program]," Version 6.3.09, 2023. [Online]. Available: http://www.praat.org/