

Scripted Hesitation in a Financial Customer-Service Chatbot: A Brazilian Portuguese Case Study

Valeria Vieira dos Santos

Graduate Program in Linguistics

Federal University of São Carlos (UFSCar)

São Carlos, Brazil

valeriavdosantos93@gmail.com

Abstract

Financial customer service is a high-stakes conversational domain in which debt negotiation, cancellations, and complaints require pragmatic resources beyond fluent text generation. We present a case study of linguistic hesitation in approximately 3,000 WhatsApp customer-service interaction records (approximately 89k tokens) from one Brazilian financial company, contrasting clients, human agents, and one commercial scripted chatbot. A rule-based pipeline adapted a Conversation Analysis typology to written chat and identified 20,938 pattern occurrences across four formal categories. Client messages account for 59% of detections, but unavailable per-role token totals preclude rate comparisons. Archived within-role distributions place 40% of agent detections in sensitive contexts and 50% of chatbot detections in technical contexts. The chatbot also relies on only 3–5 recurrent surface forms. We interpret these results as descriptive evidence that human and scripted hesitation serve different interactional profiles, not as a general claim about financial chatbots or LLM-based agents. The study illustrates both the value of pragmatic analysis for financial conversational systems and the limits of conclusions drawn from a single-company, single-annotator corpus.

1 Introduction

Conversational agents mediate customer service in the financial sector, where interactions may involve debt negotiation, contract cancellation, and complaints. In Brazil, much of this interaction occurs over WhatsApp, making production dialogues in Brazilian Portuguese (BP) a useful testbed for examining the pragmatic behavior of deployed systems.

This paper examines *linguistic hesitation*. Rather than treating hesitation only as noise, Conversation Analysis relates it to discourse planning,

uncertainty management, and lexical search (Marcuschi, 1999; Clark and Fox Tree, 2002; Levelt, 1983). Before a refusal or other face-threatening act, hesitation may also contribute to politeness and mitigation (Brown and Levinson, 1987). Scripted chatbots sometimes imitate this surface behavior through messages such as “Analisando...” (“Analyzing...”). Whether those scripts reproduce the context-sensitive functions of human hesitation is an empirical question.

We analyze a single-company corpus containing clients, human agents, and one commercial scripted chatbot. We ask: (i) how do the formal and contextual distributions of detected hesitation patterns differ across the three roles? and (ii) what distinguishes the archived human patterns from this chatbot’s fixed scripts? This is a case study, and it does not characterize financial chatbots in general.

The contributions are: (1) a descriptive analysis of 20,938 rule-detected occurrences in approximately 3,000 BP financial customer-service interaction records; (2) an adaptation of Marcuschi’s hesitation typology to written production chat; and (3) a scoped comparison between context-sensitive human patterns and one scripted chatbot. The analysis condenses and revises aggregate results first developed in the author’s master’s thesis (Santos, 2025).

2 Related Work

Disfluency in NLP. Computational work on disfluency builds on Shriberg (1994)’s *reparandum–interregnum–repair* typology for English speech, from noisy-channel models (Johnson and Charniak, 2004) to neural detectors (Zayats et al., 2016; Rocholl et al., 2021). This tradition often treats disfluency as material to remove before downstream processing rather than as a pragmatic signal. For BP, Luna et al. (2025) review methods for identifying and classifying speech disfluencies, and Luna

et al. (2022) evaluate transcription tools with a focus on disfluency detection. Both address speech rather than written customer-service chat. Descriptive work on BP hesitation also draws on spoken corpora such as NURC (Barbosa, 2006). We extend this line of inquiry to a hybrid written-chat genre (Marcuschi, 2000).

Hesitation in conversational systems. Dialogue and speech-synthesis research has explored fillers for naturalness and turn-taking (Skantze, 2021). Work on LLMs separately examines whether models can express calibrated uncertainty in words (Kadavath et al., 2022; Lin et al., 2022; Mielke et al., 2022; Zhou et al., 2023). Our object is narrower: a non-LLM, scripted production chatbot in BP. It therefore provides a case-study baseline, not evidence about contemporary generative agents.

3 Corpus and Method

3.1 Data and units of analysis

The archived corpus comprises approximately 3,000 customer-service interaction records conducted over WhatsApp between April and July 2024 and exported from Zendesk, totaling approximately 89k tokens. Records include three roles: client, human agent, and chatbot. The corpus was sampled to cover technical support, general inquiries, and sales, and each message was indexed by interaction ID, authorship, and turn position.

The corpus-level totals imply roughly 30 tokens per interaction record. The unit labeled an “interaction” is therefore not necessarily a long, sustained three-party conversation; the archive includes brief service exchanges. Likewise, the 20,938 figure below counts rule-detected pattern spans, not tokens. A message may contain multiple detections, and a multi-token repetition or repair is counted as one occurrence. These units should not be combined into a hesitation rate.

The archived analysis grouped occurrences into three communicative contexts: *technical* (product support and error resolution), *administrative* (processes, deadlines, and documentation), and *sensitive* (complaints, debt negotiation, cancellations, and high emotional load). The row-level export is no longer available, so conversation counts by context and within-conversation clustering cannot be reconstructed.

3.2 Hesitation typology

Customer-service chat is a hybrid genre between orality and writing (Marcuschi, 2000). Adapting Marcuschi (1999, 2003), we use four formal categories:

- **Prosodic manifestations:** filled pauses and orthographic lengthening (e.g., “uumm...”);
- **Hesitation markers:** conventional expressions that may hold the floor or attenuate uncertainty (e.g., “então...” / “so...”);
- **Function words:** repeated connectives or pronouns (e.g., “que... que eu preciso” / “that... that I need”);
- **Lexical fragments and micro-repairs:** truncations and immediate adjustments (e.g., “ex... extrato” / “sta... statement”).

The typology originates in Conversation Analysis; our contribution is its operationalization for written production chat. The examples are synthetic illustrations of form and do not reproduce customer content.

3.3 Annotation and validation

Occurrences were identified by a Python pipeline with preprocessing and normalization, rule-based matching, and manual review. Regular expressions covered informal-writing variants such as character repetition and ellipsis- or hyphen-marked truncation. A symbolic approach was used because the markers were formally delimited and no annotated reference corpus existed for BP customer-service chat.

A stratified random sample of approximately 30% of *detected* occurrences, preserving category proportions, was manually re-annotated and showed 87% agreement with the rule labels. Because sampling was conditioned on detection, this result assesses agreement among flagged spans and does not estimate recall or false negatives. The same researcher designed the rules and produced the reference annotation; there was no second annotator and no inter-annotator agreement calculation.

3.4 Analysis scope

The analysis is descriptive. It compares absolute detections, formal categories, turn position, role, archived context proportions, and qualitative pragmatic functions. Per-role token totals are unavailable, so we do not compare hesitation rates or density across roles. The loss of row-level data also

Category	Client	Chatbot	Agent
Prosodic manif.	703	320	419
Hesitation markers	4,911	2,190	1,257
Function words	4,233	1,801	1,023
Lexical fragments	2,500	947	634
Total	12,347	5,258	3,333

Table 1: Rule-detected occurrences by formal category and interlocutor role.

Context	Client	Chatbot	Agent
Technical	44%	50%	35%
Administrative	36%	40%	25%
Sensitive	20%	10%	40%

Table 2: Archived distribution of detections across contexts (% within role).

prevents conversation-level uncertainty estimates or clustering-adjusted inference.

4 Results

4.1 Distribution by interlocutor

Of 20,938 rule-detected occurrences, 12,347 (59%) are in client messages, 5,258 (25%) in chatbot messages, and 3,333 (16%) in human-agent messages. These are shares of detections, not role-normalized hesitation rates. Table 1 reports the category counts.

Client shares within each category are 48.8% for prosodic manifestations, 58.8% for hesitation markers, 60.0% for function words, and 61.3% for lexical fragments. The last three values are relatively close; the table therefore does not support the earlier claim that clients disproportionately produce function words and hesitation markers. Without per-role token denominators, it also cannot establish that clients hesitate more frequently than the other roles.

4.2 Communicative context

Table 2 reproduces the archived within-role percentages reported in the source analysis (Santos, 2025). We omit cell counts because the row-level records are unavailable to determine whether the original counts were direct observations or back-calculated from rounded percentages.

Within these archived distributions, sensitive contexts contain 40% of agent detections, compared with 20% of client and 10% of chatbot detections. Technical contexts contain 50% of chatbot detections. Qualitative inspection associated some

agent patterns in sensitive exchanges with mitigation before refusals or collection-related directives. A synthetic illustration is “*Sim, sim... entendo sua preocupação, mas precisamos resolver esta questão*” (“Yes, yes... I understand your concern, but we need to resolve this issue”). By contrast, the chatbot repeatedly used processing scripts such as “*Analisando... sua solicitação foi recebida*” (“Analyzing... your request has been received”). These contrasts describe this corpus and do not establish a causal pragmatic effect.

4.3 Adaptive and scripted patterns

Additional archived coding reports that 83% of human detections occur in turns longer than 15 words and that 67% precede potentially face-threatening acts. For the chatbot, 78% of detections were coded as independent of request complexity and drew on only 3–5 recurrent surface forms. These figures summarize the original single-annotator analysis; they could not be re-estimated from row-level data and are treated as exploratory descriptions rather than population estimates.

We do not use sentiment as evidence for an interactional cost. The archived material does not provide a common criterion or base rates for comparing reactions after messages with and without hesitation, so the available counts cannot support an association between chatbot hesitation and negative user sentiment.

5 Discussion

The clearest contrast in this case study is between contextual concentration and surface regularity. Human-agent detections are concentrated in sensitive contexts, where qualitative examples are compatible with mitigation, whereas the scripted chatbot relies on a small repertoire and concentrates detections in technical contexts. This pattern motivates testing whether future systems should condition hesitation-like cues on interactional context rather than inserting fixed processing scripts.

The result has two practical implications. First, designers should distinguish transparent status messages from markers that imply human-like uncertainty or reflection. Grounding such expressions in calibrated model uncertainty may be more defensible than decorative simulation (Mielke et al., 2022; Lin et al., 2022). Second, evaluation should separate formal detection from pragmatic adequacy: a system can reproduce ellipses or fillers while

failing to place them appropriately in sensitive exchanges. Both implications remain hypotheses for future controlled and multi-company studies, not validated prescriptions from the present corpus.

6 Conclusion

This case study compares detected hesitation patterns in client, human-agent, and scripted-chatbot messages from one Brazilian financial company. The archived aggregate results show different contextual profiles and a markedly limited chatbot repertoire. They do not establish role-normalized frequency differences, recall of the rule pipeline, sentiment effects, or general behavior of financial chatbots. The main contribution is a scoped description of how a Conversation Analysis typology can expose the difference between surface imitation and context-sensitive pragmatic use in production chat.

Limitations

The corpus comes from one company, one sector, one language, and one messaging channel; the chatbot is scripted rather than LLM-based. Per-role token totals are unavailable, so absolute counts cannot be interpreted as comparative rates. The 87% validation result is based only on detected occurrences and therefore does not measure recall. Rule development, reference annotation, contextual coding, and pragmatic coding were performed by one researcher without independent annotation. The context table survives only as an archived aggregate cross-tabulation; conversation counts by context, the provenance of its cell counts, and within-conversation clustering cannot be reconstructed. Row-level data are unavailable, preventing new base-rate, uncertainty, or significance analyses. For these reasons, all quantitative comparisons are descriptive. We also exclude sentiment from the evidentiary claims because the proprietary labels lack a common comparison criterion and base rates.

Ethics Statement

The corpus was provided by a Brazilian financial company under confidentiality and anonymization protocols. No personally identifiable or institutional information is reported. The data-sharing agreement prohibits publication or redistribution of the raw conversations, and the original row-level export is no longer accessible. Examples in this paper are synthetic illustrations that preserve relevant

formal properties without reproducing customer content. Only archived aggregate statistics are reported. Because neither the row-level data nor the original executable pipeline can be released, we do not claim full computational reproducibility and rely on the workshop’s exception for privacy and commercial constraints.

References

- Plínio A. Barbosa. 2006. *Incursões em torno do ritmo da fala*. Pontes Editores.
- Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.
- Herbert H. Clark and Jean E. Fox Tree. 2002. Using uh and um in spontaneous speaking. *Cognition*, 84:73–111.
- Mark Johnson and Eugene Charniak. 2004. A TAG-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 33–39.
- Saurav Kadavath, Tom Conerly, Amanda Askell, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Willem J. M. Levelt. 1983. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.
- Alana S. Luna, Ariane Machado-Lima, and Fátima L. S. Nunes. 2022. Analysis of transcription tools for Brazilian Portuguese with focus on disfluency detection. In *Proceedings of the XXI Brazilian Symposium on Human Factors in Computing Systems (IHC)*. SBC.
- Alana S. Luna, Ariane Machado-Lima, and Fátima L. S. Nunes. 2025. [Identification and classification of speech disfluencies: A systematic review on methods, databases, tools, evaluation and challenges](#). *Journal of the Brazilian Computer Society*, 31(1).
- Luiz Antônio Marcuschi. 1999. *Análise da Conversação*. Ática, São Paulo.
- Luiz Antônio Marcuschi. 2000. *Da fala para a escrita: Atividades de retextualização*. Cortez.
- Luiz Antônio Marcuschi. 2003. Gêneros textuais: definição e funcionalidade. In Angela Paiva Dionísio, Anna Rachel Machado, and Maria Auxiliadora Bezerra, editors, *Gêneros textuais e ensino*, pages 19–36. Lucerna.

- Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents’ overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.
- Johann C. Rocholl, Vicky Zayats, Daniel D. Walker, Noah B. Murad, Aaron Schneider, and Daniel J. Liebling. 2021. Disfluency detection with unlabeled data and small BERT models. In *Proceedings of Interspeech*, pages 766–770.
- Valeria Vieira dos Santos. 2025. O papel da hesitação na interação humano-máquina: Uma análise no atendimento via chat. Master’s thesis, Universidade Federal de São Carlos, São Carlos, Brazil.
- Elizabeth E. Shriberg. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. thesis, University of California, Berkeley.
- Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.
- Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi. 2016. Disfluency detection using a bidirectional LSTM. In *Proceedings of Interspeech*, pages 2523–2527.
- Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of EMNLP*.