Bridging the Language Gap: Integrating Language Variations into Conversational AI Agents for Enhanced User Engagement

Marcellus Amadeus
Stanford University / São Paulo, Brazil
7marcellus@gmail.com

José Roberto Homeli da Silva
University of São Paulo / São Paulo, Brazil
jose.homeli.silva@alumni.usp.br

João Victor Pessoa Rocha
Federal University of Minas Gerais / Belo Horizonte, Brazil
joaovcpr@ufmg.br

Abstract

This paper presents the initial steps taken to integrate language variations into conversational AI agents to enhance user engagement. The study is built upon sociolinguistic and pragmatic traditions and involves the creation of an annotation taxonomy. The taxonomy includes eleven classes, ranging from concrete to abstract, and the covered aspects are the instance itself, time, sentiment, register, state, region, type, grammar, part of speech, meaning, and language. The paper discusses the challenges of incorporating vernacular language into AI agents, the procedures for data collection, and the taxonomy organization. It also outlines the next steps, including the database expansion and the computational implementation. The authors believe that integrating language variation into conversational AI will build near-real language inventories and boost user engagement. The paper concludes by discussing the limitations and the importance of building rapport with users through their own vernacular.

1 Introduction

Conversational agents have become more widespread among end users in recent years. Skjuve et al. (2022) explained that conversational agents are expected to have "intelligent" behavior and create relationships with users. Apple's Siri and Amazon's Alexa are great examples of how we interact with conversational AI.

To foster effective communication and rapport, speaking in a similar manner to AI users is imperative. Vernacular is the spoken language style through which people communicate when they are relaxed, and their level of monitoring is low (Wardhaugh, 2005). Thus, conversational AI professionals should strive to incorporate the target user's vernacular into their agent inventory. Our efforts go in the direction of how AI agents can respond in the target dialect. For instance, if a chatbot is set up to talk and write as a *Paulista*,[1] which lexical items and phrases would be relevant and representative of the *Paulista* dialect?

For that reason, this paper aims to outline the Language Variation Project at Alana AI (not operating anymore), which encompasses creating a database of expressions from Brazilian Portuguese (PT-BR) vernacular, especially those that vary according to the region and situation. As Labov informally stated, we understand variation as "different ways of saying the same thing" (Guy et al., 2007). Ultimately, it will be possible to use this database of expressions to build some sort of synonymy dictionary, enabling the AI agent to adapt its language according to the target end user's dialect. Further applicability of this labeled data involves using it as an instruction dataset that allows for fine-tuning a Large Language Model (LLM).

This study is organized as follows: in section 2 a brief discussion about language variation and AI is made; next, section 3 lists the procedures for data collection; in the section 4, we present the annotation taxonomy; and finally, we discuss expectations for the next steps of the project.

2 Related Work

Considered the father of modern Sociolinguistics, Labov asserted that, to understand language structure, linguists should study language variation in its social context (Agnihotri, 2013). Language variation is influenced by several elements, and one of

[1] Demonym to someone who was born in the state of São Paulo, Brazil.

Proceedings of the 1st Worskhop on Towards Ethical and Inclusive Conversational AI: Language Attitudes, Linguistic Diversity, and Language Rights (TEICAI 2024) , pages 16–20

them is social change. For instance, there are significant differences in speech between citizens from the state of Minas Gerais (Brazil) and those from the state of Santa Catarina (Brazil). This contrast happens due to the influence of cultural, geographical, and historical elements in their language.

Moreover, variation is a complex linguistic process. It can be multi-layered in the sense that it affects all language subsets: idiolects ("individual"), registers ("situational"), sociolects (group), dialects (region), and languages. Finegan and Biber (1994) went further and explained that the same patterns that motivate register variation also prompt sociolectal variation.

Therefore, our focus was on diatopic and diaphasic variation. Diatopic variation refers to the language variation related to geographical region differences, which is highly related to dialects; for example, the previously mentioned contrast between Minas Gerais and Santa Catarina. From another perspective, diaphasic variation[2] concerns the variation that is established depending on the communicative context (Raso and Mello, 2012); for instance, the distinction between formal and informal situations.

Implementing these processes into a conversational AI is highly challenging. Chaves et al. (2019) discussed a case study in which they implemented register analysis in order to help a chatbot understand how to speak to the user, and they concluded that the user reaction was better after the implementation. On the other hand, LLMs can fail with low frequency or new regional expressions. We asked[3] ChatGPT 3.5 to define the word *bruguelo* (meaning: baby). Not only it did not provide a definition but also it said there is no such a word in Portuguese. Google's Bard was tested[4] with the word *bruguelo* as well. Although Bard retrieved a reasonable answer, there was some kind of bug that mixed Portuguese and Persian.

Customer services agent and client interactions were also tested. The initial prompt described that the user and ChatGPT will simulate a virtual attendant-client interaction and it should respond as if it were a *Mineiro*.[5] The client's problem was "my computer broke and no one from the company responded to me."[6] ChatGPT's response[7] sounded unnatural considering the *Mineiro*'s dialect, it perpetuated racial slurs (*caboclo*) and the general tone of the message was not professional and polite.

In our case, the challenge is the high dependence on the context that regional expressions have. In the case of Brazilian Portuguese (PT-BR), this can be seen with the word "trem" in Minas Gerais, which can be associated with "train" as a means of transportation or an anaphoric referent to non-human concrete entities (Amaral, 2014). Therefore, the primary difficulty is the annotation of such words: what elements of interaction should be accounted for; which extra- and intra-linguistic factors should be included; how polysemous words should be classified?

If the annotation problem were solved, there would still be the issue of computational processing. Socially informed elements are quite complex to be handled computationally because

> *chatbots would need to be enriched with computational models that can evaluate the conversational situation and adapt the chatbot's linguistic choices to conform with the expected register, which is similar to the subconscious humans' language production process* (Chaves et al., 2021, p. 13-14).

Having computational handling in mind, the discussion of which computational procedure is suitable for this project is still in discussion but a viable option is described in section 5.

3 Data Collection

Guided by a corpus-driven approach,[8] some regional expressions were collected in order to build a coherent taxonomy. In the initial attempt, we analyzed websites and academic papers focusing on regionalisms, compiling the expressions they featured. Further details about the taxonomy will be explored in section 4.

[2]This type of variation does not cover only register variation, but for our purposes, we simplified it to register variation.

[3]You can see at the following link that it could answer well about *paraíba*—at the beginning—but not about *bruguelo*. Link: https://chat.openai.com/share/31148b96-b852-49f7-acc7-52b8f4ae7ac7

[4]Check out the conversation at https://g.co/bard/share/0c49a91600ea

[5]Demonym to someone who was born in the state of Minas Gerais, Brazil.

[6]Original text: *meu computador quebrou e ninguém da empresa me responde*.

[7]https://chat.openai.com/share/9aea38f3-3e92-417d-8bf1-a187ddc977d4

[8]McEnery and Hardie (2012) claims that a corpus-driven approach lets the corpus/data itself be the source of a "theory of language."

In the second trial, we listed some criteria to collect sources of expressions to have more reliable classifications. The established criteria for collecting sources of expressions include:

1. having scientific evidence: sociolinguistic studies tend to concentrate on lexical variation, which is our focus so far;
2. being posted in a regional means of communication (e.g., city newspaper): regional media are prone to use their region dialectal expressions;
3. or, as the last resource, being in accordance with the annotator's native speaker experience: the annotator has seen an expression in a website, in the media, or in a book that they think pertains to a certain region or situation. However, they must be sure that this is statistically relevant.[9]

Alongside the expressions, such as "caô" (similar to "a lie" or "a bluff"), the annotator would also get an example of the expression in a sentence from the expression source or a social media post; for example, *vamo ver se ele tá de caô ou não* ("let's see if he's lying or not," literally). The example sentence was also collected so that the annotator could analyze the meaning in context and do adequate annotation. The final course of action in the data collection phase is (i) selecting sources to extract expressions, (ii) listing the expressions found, and (iii) adding examples of sentences. Thus, the annotation is done based on examples taken from sociolinguistics academic articles, regional newspapers, or blogs. Hence, the tendency is to collect empirical data whether in its written or transcribed forms, in the case of speech data.

Our collection also covers toxic and inappropriate terms, such as the derogatory "boiola" (similar to "faggot"). By including these terms, our conversational agent will have a tailored stop-word list, enabling it to block messages and comments of toxic content efficiently. This customization guarantees the agent to identify and filter out specific toxic terms that might go unnoticed by more general toxicity tools.

4 Taxonomy

As previously stated, the taxonomy is data-oriented. We created a first draft of the taxonomy based on the collected data. There are eleven classes, ranging from more concrete to more abstract ones. `INSTANCE` refers to the expression itself; `TIME` points out to when the expression can be used (be it morning, afternoon, or night); `SENTIMENT` associates with polarity. On the other hand, `STATE` and `REGION` relate to where the expression is more used. Moreover, `TYPE` is the specific meaning the instance portrays, while `GRAMMAR` is the grammatical "status" of who is speaking (male or female; singular or plural). Finally, `POS TAG` is the instance's morphological category; `MEANING` refers to the broader pragmatic meaning, and `LANGUAGE` is the language the expression is used, in this case, PT-BR. Table 1 displays how the taxonomy is organized with *vou chegar* as an example. This expression can be used in the following context:

Speaker 1 (S1): *Muito bom te ver, S2.* ***Vou chegar*** *agora porque minha mãe está esperando.*
Great to see you, Lucas. **I'm going to leave** now because my mother is waiting.

Speaker 2 (S2): *Beleza, S1. Conversamos mais depois*
Cool, Alice. We talk more later.

Class	Attribute
INSTANCE	"vou chegar"
TIME	all day
SENTIMENT	neutral
REGISTER	informal
STATE	MG
REGION	Southeast
TYPE	I'm leaving
GRAMMAR	singular-noGender
POS TAG	verb
MEANING	farewell
LANGUAGE	PTBR

Table 1: Taxonomy organization with the `INSTANCE` *vou chegar* as an example.

One of the most demanding and probably important classes is `MEANING`. Some of its attributes are greeting if the expression is used to start an interaction and farewell if the expression is used to end an interaction. This class deals with the instance's pragmatic value; thus, as one can predict, as long as new expressions are collected, new attributes will be added to `MEANING`. Although it may generate an extensive list, our belief is that it can account for

[9] It could be done by searching the expression on social media like X.

differentiating the various meanings in polysemous expressions and successfully conveying an expression's pragmatic meaning. The current annotation process involves a considerable amount of manual labor, especially concerning the TYPE and MEANING classes. This manual annotation holds significance as it reveals the challenges that humans have while classifying and, very likely, that a machine would encounter too. To address this, we are contemplating the implementation of LLMs for annotation to accelerate the process but have humans in the role of annotation reviewers.

With the classes at hand, we decided to do a bottom-up annotation from the most concrete (INSTANCE) to the most abstract classes (LANGUAGE). This direction is useful because: (i) it helps the annotator grasp the context in which the instance can be used; (ii) it is not so cognitively loaded since it starts from something specific and material.[10]

The annotators are trained linguists in our team. To mitigate problems with biases, the linguists were instructed to focus on the meaning of the expressions as well as to get the region and state from the data source. Especially in academic papers on lexical variation, the meaning and the region are explicitly mentioned; thus, the annotator will simply indicate them in the classification.

5 Final Words

This paper has presented a straightforward way of integrating language variation into conversational AI. As a pilot study, the first steps towards this integration were described, following the sociolinguistic tradition and common practices in Computational Linguistics.

With this type of work, we aim to advance the area of semantic and pragmatic modeling, as well as foster innovation in AI agent development. When incorporated into conversational AI, we believe language variation will not only build up near-real language inventories but also boost user engagement.

By the time of production of this paper, our database has:

- 11 classes;
- 80 pre-set attributes;
- 170 expressions fully annotated;
- 639 expressions to be annotated;
- 9 toxic expressions to be annotated.

Moving forward, we intend to expand our database with the source materials in our backlog. Moreover, the computational implementation has to be chosen alongside the engineering team. One of the possible alternatives is creating a key for each group of synonyms, but further investigation is needed in order to confirm its feasibility.

Our next steps also cover automatizing the annotation process by using LLMs to see if they can somehow accelerate the annotation process in any of the classes. This technique would involve the compilation of multiple sentences containing the expressions collected. These gathered sentences can be employed as input for an LLM. Finally, the LLM can be fine-tuned using our annotated database, consequently enhancing its performance to the specific subtleties present in the regional expressions.

We hope to raise awareness of the importance of building rapport with users through their own vernacular. Speaking like the users may not only create a good relationship between users and AI agents—consequently, the brand, the person, the company, or else that uses it as its voice—but also can make the message clearer since it is in a language variety the user understands the most.

Limitations

Our taxonomy was construed based on the research tradition in Sociolinguistics and Pragmatics. However, language is highly diverse and variable, and expressions may not fit well in the taxonomy. Of course, some level of revision and validation is expected, but it can lead to extensive and specialized manual work. Moreover, the taxonomy is able to cover a great range of expressions. Nevertheless, a challenge emerged: multi-word expressions (MWE). Since this project is in its early stages, we decided to annotate solely single-word expressions, even though we also collect MWEs. MWEs need a different computational treatment (Ramisch, 2023). Hence, further analysis is necessary to incorporate them into our annotated database.

On the other hand, the automatization of these processes can also generate issues. While an algorithm or an LLM can be a good sentiment annotator for general words, they may not work well with a deeply informal regional expression that is not statistically present in their training texts.

[10] Language can be considered a material.

Acknowledgments

We would like to thank the rest of the Computational Linguistics team at Alana AI for their enriching feedback and additions to this project.

References

Rama Kant Agnihotri. 2013. Labov's concept of the vernacular speech: The site of language structure, acquisition and change. *Contemporary Education Dialogue*, 10(1):99–122.

Eduardo Tadeu Roque Amaral. 2014. Análise de um nome geral na fala dos mineiros: Para que serve esse trem? *Trama*, 10(20):27–44.

Ana Paula Chaves, Eck Doerry, Jesse Egbert, and Marco Gerosa. 2019. It's how you say it: Identifying appropriate register for chatbot language design. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 102–109.

Ana Paula Chaves, Jesse Egbert, Toby Hocking, Eck Doerry, and Marco Aurelio Gerosa. 2021. Chatbots language design: the influence of language variation on user experience. *arXiv preprint arXiv:2101.11089*.

Edward Finegan and Douglas Biber. 1994. Register and social dialect variation: An integrated approach. volume 315, page 347. Oxford University Press New York.

Gregory R Guy, R Bayley, and C Lucas. 2007. Variation and phonological theory. *Sociolinguistic variation*, page 1.

Tony McEnery and Andrew Hardie. 2012. Corpus-based versus corpus-driven linguistics.

Carlos Ramisch. 2023. *Multiword expressions in computational linguistics*. Habilitation à diriger des recherches, Aix Marseille Université (AMU).

Tommaso Raso and Heliana Mello. 2012. C-ORAL-BRASIL I: corpus de referência do Português Brasileiro falado informal. A general presentation. *Speech and Corpora*, page 16.

Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2022. A longitudinal study of human–chatbot relationships. *International Journal of Human-Computer Studies*, 168:102903.

R. Wardhaugh. 2005. *An Introduction to Sociolinguistics*. Blackwell Textbooks in Linguistics. Wiley.