A Persona-Based Corpus in the Diabetes Self-Care Domain – Applying a Human-Centered Approach to a Low-Resource Context

Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano and Fabio Alves
Arts Faculty, Federal University of Minas Gerais, Brazil
{rossanacunha, thiagocf05, apagano, fabio-alves}@ufmg.br

Abstract

While Natural Language Processing (NLP) models have gained substantial attention, only in recent years has research opened new paths for tackling Human-Computer Design (HCD) from the perspective of natural language. We focus on developing a human-centered corpus, more specifically, a persona-based corpus in a particular healthcare domain (diabetes mellitus self-care). In order to follow an HCD approach, we created personas to model interpersonal interaction (expert and non-expert users) in that specific domain. We show that an HCD approach benefits language generation from different perspectives, from machines to humans – contributing with new directions for low-resource contexts (languages other than English and sensitive domains) where the need to promote effective communication is essential.

Keywords: Natural Language Generation, Healthcare, Human-Centered Design, Persona, Human-Centered NLP

1. Introduction

The use of Human-Centered Design (HCD) in the context of Natural Language Processing (NLP), more specifically Natural Language Generation (NLG), has grown since 2015 (Vinyals and Le, 2015; Serban et al., 2018; Ram et al., 2018) with the development of NLP and NLG models that incorporate in their engines: personas (Li et al., 2016; Wolf et al., 2019), user profiles (Cao and Cheung, 2019), and/or corpus built with the use of personas (Zhang et al., 2018; Mazaré et al., 2018), during the training of NLG applications.

According to Norman (2023, 2), HCD is "a vital approach for accommodating real users – real people". The author (2023) emphasizes the four principles of HCD: (i) being people-centered[1] (i.e., focusing on people within their context); (ii) understanding and solving the source of problems; (iii) assuming that everything is an interconnected system; and (iv) doing small and straightforward interventions to refine solutions in an iterative form[2].

To the best of our knowledge, few studies have explored HCD and NLG in the context of low resource conditions – a sensitive domain and a language other than English. With the intent of providing a resource that could help the development of more human-centered NLP systems and also overcome the lack of in-domain data, we developed a corpus called Dia(Bete) in the Brazilian cultural context for applications in the diabetes mellitus domain applying an HCD approach as presented in Figure 1. Code and data are publicly available[3].

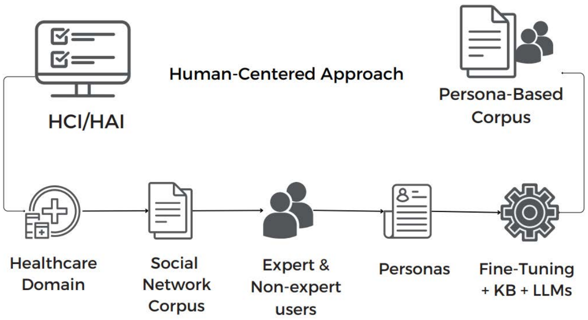


Figure 1: Human-Centered approach to a low-resource context.

To develop our corpus, we first created personas (Nielsen, 2011, 2019), which are target user groups constructed to guide the production of texts that can be understood by real users according to contextual features of the investigated situations (expert to expert, expert to lay user communication) and text type (complex and simplified). The corpus is annotated with their corresponding personas based on characteristics of literacy as well as simplified/complex versions, and additional notes from linguists and physicians, in the diabetes mellitus domain.

2. Human-Centered Approach

To design a corpus with an HCD perspective in mind, we need to understand our users' expectations and take into account their needs and concerns. In order to achieve this, we turn to an HCD method for user models: personas. According to Cooper et al. (2014), personas are "user models that are represented as specific, individual human

[1] https://www.interaction-design.org/literature/topics/people-centered-design

[2] https://www.interaction-design.org/literature/topics/human-centered-design

[3] https://github.com/Dia-Bete/PersonaBasedCorpus.

LREC-COLING 2024, pages 1353–1369
20-25 May, 2024.

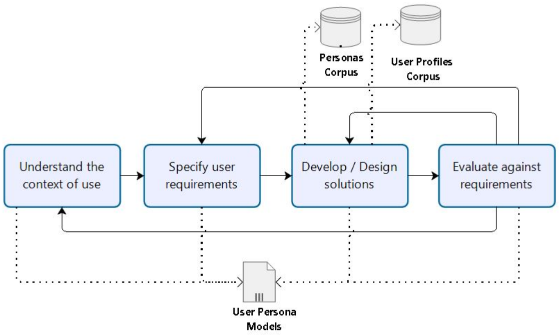

Personas Corpus
User Profiles Corpus
Understand the context of use
Specify user requirements
Develop / Design solutions
Evaluate against requirements
User Persona Models

Figure 2: User-Centered Design and how to integrate personas and corpora – adapted from Soegaard and Dam (2015).

beings". We addressed this need for a more user-driven generation by applying a group of personas that are more frequently encountered within the healthcare domain ranging from more expert to non-expert users. Among the benefits of using personas is the creation of a common reference (e.g., the user can be referred to as a persona instead of drawing on a vague assumption of who the end-user could be) to guide the development and promote the engagement of the real end-users within the healthcare domain. Recent studies regarding personas (Nielsen, 2019; Salminen et al., 2020, 2021; Yang, 2023) reiterate the benefits of using them within an NLP context.

According to Salminen et al. (2021, 1685), data-driven personas are created "from the behaviors and demographics of user segments". Additionally, Suojanen et al. (2015) emphasize that personas can be a very constructive tool when discussing solutions, "which can sometimes be reduced to generalities or platitudes if discussed without a specific target user in mind". In line with this research, we developed a persona corpus based on the demographic data and groups identified in the results of studies conducted within the project Empoder@[4]. For more information, please refer to Pagano (2015); Abd-Alrazaq et al. (2019); Vieira et al. (2017); Chaves et al. (2019); Neves et al. (2021); Cortez et al. (2022), and Nunes et al. (2023).

In the next subsections, we present the personas created for a particular healthcare domain (diabetes mellitus self-care) and their application within the creation of the persona-based corpus.

Non-Expert Persona One of the designed personas in our approach is Eris, a non-expert user regarding diabetes knowledge for self-care. We can see more characteristics of this persona in Figure 3. Key features of this persona are middle-aged woman; not yet diagnosed with diabetes; having problems finding information in digital apps due to eyesight impairment and limited technology literacy (for full persona please refer to Figure 3).

Expert Persona Another persona in our approach is Alexis, a *domain expert* user in the sense that she herself is a healthcare provider. Please refer to the expert persona in Figure 8 (Appendix Section 10.3). This persona has a higher education level and expects to be provided with more technical information regarding diabetes.

In our project, we created three personas based on the demographic data collected from the Empoder@ Project (Chaves et al., 2017). However, in order to develop our corpus, we selected two of them, which are representative of expert and non-expert users in terms of domain, i.e., diabetes self-care. Our corpus was meant to target users like Eris and Alexis by providing answers on two levels of complexity. That is why each answer drafted in response to a user post or question has two versions, one of them being more simplified than the other. For detailed personas created in this project see the Appendix Section 10.3.

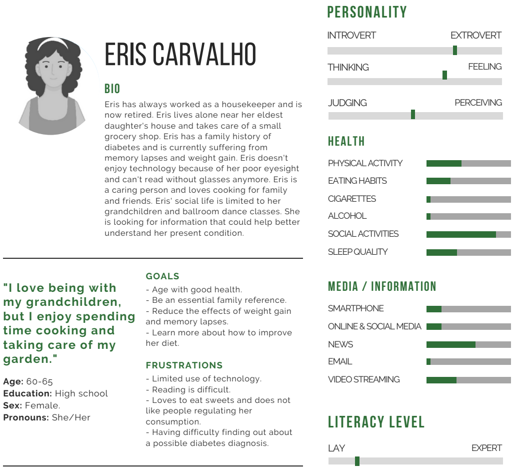
2.1. Addressing the Lay-Expert Communication Gap

Text Simplification is a commonly used task within NLG (Erdem et al., 2022). The main idea behind this task is to perform transformations on an input text to provide a more accessible text for readers with different literacy levels, cognitive impairment, and reading and writing learning skills (Aluísio et al., 2008; Specia et al., 2008; Siddharthan, 2014; Saggion, 2017; Alva-Manchego et al., 2020). In the same vein, this study addresses text simplification as a human-centered task where end-users of a particular domain can access a more readable text in accordance with their literacy level.

According to text simplification research (Shardlow, 2014; Siddharthan, 2014; Saggion, 2017; Alva-Manchego et al., 2020; Štajner, 2021; Al-Thanyyan and Azmi, 2021), the task can be addressed by distinct strategies. In general, text simplification aims to redraft a text to promote a better understanding of knowledge and be received by the interlocutor so that the information is clear and easy to use. Within Applied Linguistics, there are several solutions for simplifying texts. They vary from innovative directions to more conservative ones. All of them aim at achieving better communication between lay and expert users (Aluísio and Gasperin, 2010; Pagano, 2015; Araujo, 2021; Oliveira, 2022).

Our study focuses on sentence simplification (Alva-Manchego et al., 2020) within a hybrid ap-

[4] http://www.letras.ufmg.br/empodera/



ERIS CARVALHO

BIO

Eris has always worked as a housekeeper and is now retired. Eris lives alone near her eldest daughter's house and takes care of a small grocery shop. Eris has a family history of diabetes and is currently suffering from memory lapses and weight gain. Eris doesn't enjoy technology because of her poor eyesight and can't read without glasses anymore. Eris is a caring person and loves cooking for family and friends. Eris' social life is limited to her grandchildren and ballroom dance classes. She is looking for information that could help better understand her present condition.

"I love being with my grandchildren, but I enjoy spending time cooking and taking care of my garden."

Age: 60-65
Education: High school
Sex: Female.
Pronouns: She/Her

GOALS

- Age with good health.
- Be an essential family reference.
- Reduce the effects of weight gain and memory lapses.
- Learn more about how to improve her diet.

FRUSTRATIONS

- Limited use of technology.
- Reading is difficult.
- Loves to eat sweets and does not like people regulating her consumption.
- Having difficulty finding out about a possible diabetes diagnosis.

Figure 3: Non-expert persona: Eris.

proach and a specific domain – healthcare (Pagano, 2015). Similarly, Grabar and Cardon (2018) created a French medical corpus to help design and test methods for text simplification tasks – CLEAR[5]. They compiled a corpus from French comparable datasets containing texts from encyclopedia articles, drug leaflets, and scientific summaries. The final corpus contains 16,190 document-level pairs with a subset of 663 manually annotated sentence-level pairs. The authors concluded their work by emphasizing the endeavor made to perform semantic annotation. The CLEAR corpus served as the basis for the development of a French biomedical text simplification corpus (Cardon and Grabar, 2020).

A Portuguese medical corpus in the domain of Parkinson's disease was developed by Zilio et al. (2020) in conjunction with a simplification web-based tool: MedSimples[6]. The corpus was created in order to evaluate MedSimples and contains a total of 570 annotated sentences and a total amount of 2080 instances.

Devaraj et al. (2021) explored automated simplification in the medical domain with a fine-tuning approach. The authors introduced an English dataset with pairs of technical and simplified texts derived from the Cochrane database (McIlwain et al., 2013)[7]. The results were satisfactory when training with BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019), followed by a better performance with BART (Lewis et al., 2019). A similar approach was conducted by Patel et al. (2022) with different pre-trained language models.

Some of the advances in Medical text simplification are summarized as follows: (i) Simple TICO-19, an automated simplification approach using machine translation and simplification of COVID-19 texts (Shardlow and Alva-Manchego, 2022); (ii) EASIER, a corpus comprising 260 annotated documents in Spanish (Alarcon et al., 2023); (iii) TESLEA, an automated corpus of medical text simplification fine-tuned with pre-trained models (Phatak et al., 2022); as well as works focused specifically on fine-tuning models in order to improve patient-doctor interactions (Yunxiang et al., 2023; Han et al., 2023). In addition, a shared task was conducted during the 2023's MEDIQA[8] at the 5th Clinical Natural Language Processing Workshop[9], to perform summarization and simplification of medical clinical notes. The results presented new resources as multilingual benchmarks and discussions about the development of medical NLP (Ben Abacha et al., 2023).

3. Persona-Based Corpus Development

Due to the difficulty in finding an automatic solution that could provide domain-adapted versions of texts to end-user literacy, we first manually compiled a persona-based corpus following similar approaches (Grabar and Cardon, 2018; Oraby et al., 2019; Ramos et al., 2020). Table 1 shows sample texts from our corpus.

Within the Brazilian healthcare context, several studies tried to explore and solve the communication difficulty between patients and medical staff (Pagano, 2015; Chaves et al., 2017, 2019; Oliveira, 2022; Nunes et al., 2023). In the same vein, our study follows an interdisciplinary approach, recruiting expertise from different areas, such as applied linguistics, pharmacy, medicine, nutritional science, and computer science, to address the communication concerns related to language encountered by experts and non-experts in the diabetes mellitus domain. Our corpus was meant to be used in Question-Answering and Conversational systems. Therefore, it was designed to comprise questions or posts retrieved from public online forums on diabetes self-care, to which answers were drafted by medical and nutritional science students, and supervised by their tutors. Each answer was drafted in both a more simplified and a less simplified version.

Regarding simplification, we draw on the project called PorSimples[10] (Aluísio et al., 2008) developed

[5] http://natalia.grabar.free.fr/resources.php

[6] https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/

[7] https://consumers.cochrane.org/PLEACS

[8] MEDIQA is a series of shared tasks on Medical NLP – available at https://sites.google.com/view/mediqa2023/home

[9] ACL-ClinicalNLP 2023 – https://clinical-nlp.github.io/2023/

[10] PorSimples – Simplification of Portuguese Texts

Persona Experta	Persona Leiga
*Na cetoacidose os sintomas são: poliúria (**aumento da urina**), polidipsia (aumento da sede), aumento da frequência respiratória, náuseas e vômitos, **hálito cetônico**, cefaleia (dor de cabeça), confusão mental, **dor abdominal**.*	*Na cetoacidose os sintomas são: poliúria (**maior vontade de urinar**), polidipsia (aumento da sede), respiração acelerada, náuseas e vômitos, **hálito de "maçã podre"**, cefaleia (dor de cabeça), confusão mental, **dor na região do abdômen**.*
Expert Persona	**Non-expert Persona**
In ketoacidosis, the symptoms are polyuria (**increased urine**), polydipsia (increased thirst), **increased respiratory rate**, nausea and vomiting, **ketonic breath**, cephalalgia (headache), mental confusion, and **abdominal pain**.	In ketoacidosis, the symptoms are polyuria (**increased desire to urinate**), polydipsia (increased thirst), **rapid breathing**, nausea, vomiting, and **"rotten apple" breath**, cephalalgia (headache), mental confusion, **and pain in the region of the abdomen**.

Table 1: Sample texts targeting expert and non-expert personas (Original text in Brazilian Portuguese on top with English gloss below) – Simplifications are **boldfaced**.

at the University of São Paulo (USP). This project was able to produce a series of NLP technologies targeting text adaptation for Brazilian Portuguese text simplification to lay readers, including people with cognitive impairments, in order to promote accessibility and digital inclusion for people with different levels of literacy (Aluísio et al., 2008).

The PorSimples project converges with our main goal concerning the generation of easier-to-read texts with a view to bridging the gap between specialists and patients. Through more accessible language, the patient can better understand what is deemed necessary for self-care. Language becomes a fundamental tool for accessing self-care and increases learning about their diabetes condition (Pagano, 2015; Oliveira, 2022).

3.1. Corpus and Annotation Process

This section focuses on the development of the corpus and how an HCD method, i.e., the personas, was able to guide the development, compilation, annotation, simplification, validation, and evaluation of the corpus. The persona-based corpus aims to provide support for informative and decision-making domains like healthcare (i.e., diabetes self-care) based on demographic data collected in the project Empoder@.

We will first introduce how we collected data to build the corpus. Next, we will describe compilation and automatic correction before corpus annotation. Finally, we describe the annotation process in more detail and which criteria have been followed by the annotators.

Criteria	Attribute
Period	2015-2020
Type of Relation between Texts	Comparable
Subject Area	Language register
Domain	Specialized
Mode and Medium	Written
Temporal Restriction	Synchronic
Number of Languages	Monolingual
Extra-linguistic Information of Texts	Personas

Table 2: Persona-based corpus compilation criteria.

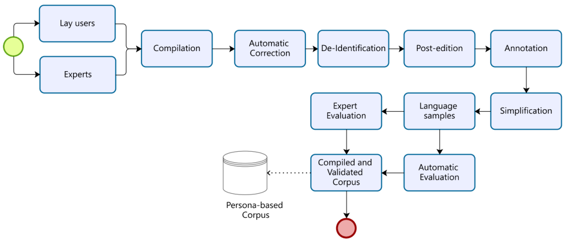


Figure 4: Persona-based corpus compilation.

3.2. Corpus Compilation Criteria

As presented in the HCD Approach's Section 2, our corpus is oriented by the personas constructed with the data collected from studies within the Brazilian healthcare context. After building the personas, we decided to compile a Brazilian Portuguese monolingual corpus, i.e., a corpus that contains texts in a particular language (Zanettin, 2011), and in the same vein is able to represent "varieties and genres of a particular language" (McEnery and Hardie, 2011, 19), specifically, texts targeting expert and non-expert users in the diabetes self-care domain. Our corpus also qualifies as an intralingual corpus, which according to Jakobson (2000, 30–31) has the characteristic of "rephrasing of a text in one language into another in the same language", that is, the simplification of a technical text into language understandable by lay readers. A summary of the criteria adopted (drawing on Fernandes (2004, 2006)) for the decisions made to compile our corpus can be found in Table 2.

3.3. Compilation and Automatic Correction

From an HCD perspective, we followed the corpus compilation pipeline shown in Figure 4. First, texts (comments/questions) were collected from from posts written in Brazilian Portuguese retrieved from public forums in Brazil focused on diabetes self-care. This task was done by members of our research group, who were responsible for manually collecting posts. Then, we carried out semi-automatic preprocessing and cleaning to remove numerous noisy words from the collected corpus such as repeated punctuation, emojis, wrong cap-

for Digital Inclusion and Accessibility. For more details, please refer to `http://www.nilc.icmc.usp.br/nilc/index.php/porsimples`.

italization, and misspelled words. Finally, the corpus was de-identified, post-edited, annotated, and simplified as will be discussed in the following sections.

3.4. De-Identification

To preserve individuals' privacy, we followed Brazil's General Data Protection Law (*Lei Geral de Proteção de Dados* – LGPD)[11] established in 2018, regarding privacy and data protection guidelines – an analogous regulation is the European Union's General Data Protection Regulation (GDPR)[12]. LGPD provides comprehensive guidelines on how to process personal data, individual rights, and obligations. In our context, the de-identification process follows steps to prevent re-identification, reflecting the balance between privacy protection and data utilization within a sensitive information domain (Berman, 2016; Liu et al., 2017; Pagano et al., 2023). De-identification comprised the following steps:

1. Removing any individual's identifier, such as name, phone, organization, or location;

2. Changing quasi-identifiers: age was randomly changed in an interval of $[\text{age}-5$ or $\text{age}+5]$, and relative mentions followed similar age references: parent → uncle, parent in law, sibling → cousin, partner, daughter, son → niece, nephew;

3. Replacing any consecutive characters or uppercase characters, or punctuation (e.g., "Aaaaaaaa" and "Aa", "*BOM DIAAAAA*" to "*Bom dia*", and "???" to "?', respectively)[13];

4. Paraphrasing (see Section 2.1 for more information).

3.5. Post-Edition

During the post-editing phase, normalization of texts and punctuation was performed. The original text was classified according to a proficiency scale developed following literacy and simplification directions drawing on:

- Systemic-Functional Linguistics (Halliday and Matthiessen, 2014) regarding lexical density and grammatical metaphoricity;

- Syntax studies (Szmrecsanyi, 2004) discussing criteria to assess text complexity – text length, phrase complexity, coordinated, subordinated, and embedded clauses;

- Guidelines in the Manual for Portuguese Simplification (Specia et al., 2008; Aluísio et al., 2008) developed by NILC[14] within the PorSimples project;

- The National Indicator of Functional Literacy (INAF – *Indicador Nacional de Alfabetismo Funcional*)[15].

The INAF index considers five main literacy categories applying to the Brazilian population, as follows:

1. **Illiteracy**: lack of ability to fulfill simple tasks involving reading, with possible ability to read everyday numerical expressions (e.g., telephone numbers, house numbers, and prices);

2. **Literacy – elementary level**: ability to locate explicit information, stated explicitly in non-metaphorical language, in short texts comprising sentences or words that construe meanings related to everyday situations, such as advertisements, and to compare, read and write everyday numerical expressions (e.g., schedules, prices, bills/coins, telephones);

3. **Literacy – basic level**: ability to locate one or more chunks of information in medium-length texts, solve basic operations with numbers, and also understand meaning exchanges in everyday situations;

4. **Literacy – intermediate level**: the ability to find information stated explicitly in a non-metaphorical language in medium-length texts (e.g., journalistic, scientific), solve mathematical problems involving percentage and proportion, interpret and prepare a summary of different texts, and perform very basic inferences;

5. **Literacy – proficient level:** ability to read long texts of higher complexity (e.g., description, exposition, and argumentation), find multiple

[11] *Lei Geral de Proteção de Dados Pessoais* (LGPD) – Lei nº $13.709/2018$. Available at https://www.gov.br/mds/pt-br/acesso-a-informacao/lgpd

[12] European Union's General Data Protection Regulation (GDPR) – available at https://gdpr-info.eu/recitals/no-159/

[13] These patterns can be used to identify the author of a post (e.g., "*GOOD MOOOOORNING!!!!*" has fewer occurrences than its conventional form "*Good morning!*" (Berman, 2016; Liu et al., 2017; Pagano et al., 2023).

[14] NILC – *Núcleo Interinstitucional de Linguística Computacional* (The Interinstitutional Center for Computational Linguistics) at https://sites.google.com/view/nilc-usp/.

[15] *Indicador Nacional de Alfabetismo Funcional* (National Indicator of Functional Literacy) – available at https://alfabetismofuncional.org.br/

Proficiency Scale	
1	very poorly structured comment/question
2	comment/question poorly structured and/or with many language errors
3	comment/question reasonably well structured and with some language errors
4	well-structured comment/question with errors especially related to domain
5	well-structured and error-free comment/question

Table 3: Text-based proficiency scale.

Question
Bom dia. A minha filha é diabética do tipo 2, ela neste momento está com muita dor de cabeça, não consegue dormir, tonturas, visão turva e fortes dores no estômago. O que posso fazer para reduzir as dores?
Good Morning. My daughter has type 2 diabetes, she is currently having episodes of intense headaches, she can't manage to sleep, dizziness, blurred vision, and severe stomach pain. What can I do to help her reduce her pain?

Table 4: Sample question from our corpus (Original text in Brazilian Portuguese on top with English gloss below).

types of information related to tasks in different contexts, compare different texts, recognize effects of meaning, and perform inferences.

A proficiency scale was designed following a Likert scale format ranging from 1 to 5, as presented in Table 3.

Besides proficiency level, posts were classified according to the target designed persona into expert and non-expert regarding domain knowledge. The researchers involved during the post-edition phase participated in a workshop in order to learn the particulars of the personas and receive training on how to validate each post in the corpus.

Category	Description	Examples of annotated entities
Complication	diseases and health conditions	type 2
Symptoms	physical or mental condition experienced by the patient	headache, can't sleep, dizziness, blurred vision, severe stomach pain

Table 5: Categories of annotated entities, description and examples – adapted from Pavanelli (2022, 26).

3.6. Annotation

We first annotated the corpus for domain categories capturing the main topic of each post, e.g., symptoms, *food-and-diet*, *general-information*, *blood-sugar*, etc. We can see an example of a post in Table 4. In a second round of annotation, we selected the original version of the corpus, considering all emojis and extra punctuation, in order to annotate the posts with a sentiment analysis scale – very positive, positive, neutral, negative, and very negative (Gumiel et al., 2021).

Furthermore, these posts were classified regarding the five stages of grief: denial, anger, bargaining, depression, and acceptance (Kübler-Ross et al., 1972) and according to their level of knowledge regarding the diabetes mellitus domain. Entities were annotated according to the ontology proposed by Ben Abacha and Zweigenbaum (2015), some examples of which are presented in Table 5 accordingly. A summary of the Dia(Bete) corpus entities as well as their relations in numbers are presented in the Appendix 10.1 (Figures 5 and 6, respectively).

3.7. Simplification

Our study focused on sentence simplification (Alva-Manchego et al., 2020) within a hybrid approach in order to develop simplification guidelines drawing on Szmrecsanyi (2004) concerning syntactic complexity, the Manual for Portuguese Simplification (Specia et al., 2008) developed by NILC, and the INAF index (please refer to Post-edition Section 3.5).

Following the documentation provided by the PorSimples project and Gasperin et al. (2010), we developed guidelines for annotators to perform the simplification task in our persona-based corpus. The guidelines are summarized as follows:

- Identify relative clauses, sentences with non-finite verbs, and discourse markers inside a sentence;
- Identify negative sentences and check if they can be replaced by affirmative ones;
- Perform noun-to-verb conversion in the case of abstract nouns;
- Rewrite sentences following the pattern subject-verb-object;
- Replace technical words with simpler and/or more colloquial ones.

As already stated, a multidisciplinary team of students with different backgrounds (applied linguistics, pharmacy, medicine, nutritional science, and computer science) worked on answers in response to each post in our corpus (comment/question). Answers were written by medical and nutritional science students and curated by their supervisors. A more simplified version of the answers was also developed in order to meet end-users needs. The simplification performed was meant to preserve the meaning of the answers, even if this required a longer answer than the less simplified one. In some cases, simplified versions were expanded with additional information (e.g., explanations or a warning in case of sensitive questions) in line with healthcare simplification research (Grabar and Cardon, 2018; Cardon and Grabar, 2018; Shardlow and Nawaz, 2019; Van den Bercken et al., 2019; Cardon and Grabar, 2020; Koptient and Grabar, 2020).

The idea of having answers formulated by two distinct disciplinary fields – medicine and nutritional science – was meant to increase the occurrences for each persona. Answers targeting expert and non-expert users were aligned. For more information, please refer to a sample of this corpus in Table 8 (Appendix 10.2). Additionally, Table 7 displays some statistics of our corpus. Further samples of our corpus are presented in the Appendix Section 10.2.

4. Evaluation

Validation was done upon each corpus compilation iteration (for more information, please refer to Figure 4), where the verification was performed with the use of personas in order to identify possible persona mismatches (as explained in Subsection 2). The initial validation plan was aligned with the number of collected posts. After achieving the expected number of posts (i.e., around 1,000 in total), experts with different backgrounds (applied linguistics, pharmacy, medicine, nutritional science, and computer science) performed validation.

Additionally, the human metric took into account the personas created for this research, according to users' characteristics, in addition to the information about the investigated domain. After finishing the iterations, a human-centered evaluation and validation of the results were done in order to contrast the text generation against a human-centered perspective, using personas and heuristics (Suojanen et al., 2015; Cunha, 2016; Cunha and Fernandes, 2016). In relation to metrics, as the corpus is manually created and annotated, a human evaluation was performed on the annotation part within a study of the Empoder@ project. On the annotation, the human evaluation was done by Pavanelli (2022) during the development of the Bete NER model as aforementioned in Section 3.6. Moreover, the corpus has been investigated regarding question-answering entailment by Castro Ferreira et al. (2021). Subsequently, a Sentiment Analysis study was conducted by Gumiel et al. (2021).

Original Post	Omitted due to ethical considerations (please refer to the De-Identification Section 3.4).
Corrected Post	*Olá! Como vou saber se o aparelho de medir a glicemia está com defeito!*
Post-edited Post	*Olá! Como vou saber se o aparelho de medir a glicemia está com defeito?*
Gloss	Hello! How will I know if the device for measuring blood glucose is failing?
Answer	*É recomendado que se calibre o glicosímetro (**aparelho de medir a glicemia**) sempre que for abrir um novo pacote de fitas de testagem. Essa calibragem pode ser realizada em farmácias.*
Non-expert Answer	*É recomendado que se calibre o glicosímetro (**aparelho que mede a quantidade de açúcar no sangue**) sempre que for abrir um novo pacote de fitas de testagem. Essa calibragem pode ser realizada em farmácias.*
Gloss	We recommend you calibrate the glucose meter (**device for measuring blood glucose**) whenever you open a new package of testing strips. You can calibrate the glucose meter at a pharmacy.
Gloss for Non-expert Answer	We recommend you calibrate the glucose meter (**a device that measures the amount of sugar in your blood**) whenever you open a new package of testing strips. You can calibrate the glucose meter at a pharmacy.

Table 6: Sample of questions and answers (expert and non-expert) drafted by medical students (some columns were omitted due to ethical considerations).

Persona-based corpus				
Origin	**Train**	**Dev**	**Test**	**All**
Medical	881	109	109	1099
Nutrition	655	81	82	818
Total	1536	190	191	1917

Table 7: Corpus statistics of answers annotated by medical students and answers annotated by nutritional science students.

5. Results and Discussion

Our corpus was compiled and annotated manually and automatically. After achieving a total of 1,099 pairs (question and answer), we performed validation of the collected data by selecting samples from the corpus. In order to assess the comparability and validity of the aligned instances, an expert evaluation was carried out by some of the participants in our research team. The result was a monolingual corpus aligned with expert and non-expert texts in the diabetes self-care domain. The statistics of the Dia(Bete) corpus are presented in Table 7.

As we can see in Table 6, expert and non-expert texts differ in sentence length and use (or not) of terms pertaining to the diabetes domain. That is, some of the non-expert texts are longer than expert ones due to explanations added to non-expert texts, mostly considering the domain and the sensitive context. However, we have observed that some non-expert occurrences are shorter than expert ones (please refer to more samples in the Appendix Section 10.2). Regarding terminology, it can be observed that simplified answers provide explanations and more colloquial words between parentheses.

6. Conclusion

We reported on the development of a corpus by employing human-centered approaches: HCD and Text Simplification. We first created personas in the context of the healthcare domain, more specifically diabetes mellitus self-care. These personas were created based on data collected in a project within the healthcare context. We applied the verified personas during the compilation, de-identification, annotation, simplification, validation, and evaluation of the corpus. This corpus has been compiled semi-automatically and is intended to assist the training of future human-centered NLP models within sensitive domains. The HCD approach adopted in our study proved a valuable tool to develop resources in low-resource conditions and also provides benefits to the people involved in the process – from the development team to the prospective end-users.

For future work, we plan to augment our training data by drawing on persona-based English corpora (Mazaré et al., 2018) translated into Portuguese, with means to explore out-of-domain data with the human-centered NLP approach. We also plan to use datasets of clinical corpora, e.g., MEDIQA (Ben Abacha et al., 2023), by aligning with respective personas: ($non - expert \rightarrow patient$) and ($expert \rightarrow doctor$).

7. Ethics Statement

This study was exempted from any data consent request as the corpus was compiled from posts in public forums not requiring any signup procedures. All personal data in the posts was treated through a de-identification process conducted during this research (please refer to Section 3, more specifically in the De-Identification Section 3.4). All project members were assigned an equal share of labor. Code and data are publicly available.

8. Acknowledgements

Research funded by the National Council for Scientific and Technological Development (CNPQ) under grant No. 313103/2021-6), the Foundation for the Coordination and Improvement of Higher Education Personnel (CAPES) (Postdoctoral grant No. 88887.508597/2020-00) and the State Funding Agency of Minas Gerais (FAPEMIG) under grant APQ-01.461-14.

9. Bibliographical References

Alaa A Abd-Alrazaq, Mohannad Alajlani, Ali Abdallah Alalwan, Bridgette M Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics*, 132:103978.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. Easier corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53.

Sandra Aluísio, Lucia Specia, Thiago AS Pardo, Erick G Maziero, and Renata PM Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the eighth ACM symposium on Document engineering*, pages 240–248.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Rodrigo Araujo. 2021. *Systemic-Functional modeling of text complexity in Brazilian Portuguese*. Ph.D. thesis, Universidade Federal de Minas Gerais.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Asma Ben Abacha and Pierre Zweigenbaum. 2015. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, 51(5):570–594.

Jules J. Berman. 2016. Chapter 5 - identifying and deidentifying data. In Jules J. Berman, editor, *Data Simplification*, pages 189–231. Morgan Kaufmann, Boston.

Meng Cao and Jackie Chi Kit Cheung. 2019. Referring expression generation using entity profiles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3154–3163.

Rémi Cardon and Natalia Grabar. 2018. Identification of parallel sentences in comparable monolingual corpora from different registers. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 83–93.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716.

Thiago Castro Ferreira, João Victor de Pinho Costa, Isabela Rigotto, Vitoria Portella, Gabriel Frota, Ana Luisa A. R. Guimarães, Adalberto Penna, Isabela Lee, Tayane A. Soares, Sophia Rolim, Rossana Cunha, Celso França, Ariel Santos, Rivaney F. Oliveira, Abisague Langbehn, Daniel Hasan Dalip, Marcos André Gonçalves, Rodrigo Bastos Fóscolo, and Adriana Pagano. 2021. Evaluating recognizing question entailment methods for a Portuguese community question-answering system about diabetes mellitus. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 234–243, Held Online. INCOMA Ltd.

Fernanda Azeredo Chaves, Sumaya Giarola Cecilio, Ilka Afonso Reis, Adriana Pagano, and Heloísa de Carvalho Torres. 2019. Translation and cross-cultural adaptation of the behavior change protocol for educational practices in diabetes mellitus. *Revista latino-americana de enfermagem*, 27.

Fernanda Figueredo Chaves, Ilka Afonso Reis, Adriana Pagano, and Heloísa de Carvalho Torres. 2017. Translation, cross-cultural adaptation and validation of the diabetes empowerment scale–short form. *Revista de Saúde Pública*, 51:16.

Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About face: the essentials of interaction design*. John Wiley & Sons.

Daniel Nogueira Cortez, Débora Aparecida Silva Souza, Andreza Oliveira-Cortez, Ilka Afonso Reis, and Heloisa de Carvalho Torres. 2022. Efeito individual das atitudes e empoderamento para o autocuidado em diabetes mellitus. *Saúde Coletiva (Barueri)*, 12(80):11302–11317.

Rossana Cunha. 2016. Avaliação de um sistema com base em corpus para a pesquisa, ensino e prática da tradução através de critérios de usabilidade e ergonomia cognitiva [*Evaluation of a corpus-based system for the research, teaching and practice of translation under the perspective of ergonomics and usability*]. Master's thesis, Federal University of Santa Catarina, Florianópolis/SC - Brazil.

Rossana Cunha and Lincoln P. Fernandes. 2016. Avaliação comparativa de ferramentas web de apoio à tradução através de critérios de usabilidade e ergonomia cognitive. *Tradução em Revista*, 21:67–90.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara

Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, et al. 2022. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, 73:1131–1207.

Lincoln P. Fernandes. 2004. A portal into the unknown: designing, building, and processing a parallel corpus. *CTIS Occasional Papers*, 4:21–43.

Lincoln P. Fernandes. 2006. Corpora in translation studies: revisiting baker's tipology. *Fragmentos: Revista de Língua e Literatura Estrangeiras*, 30.

Caroline Gasperin, Erick Maziero, and Sandra M Aluisio. 2010. Challenging choices for text simplification. In *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9*, pages 40–50. Springer.

Natalia Grabar and Rémi Cardon. 2018. CLEAR – simple corpus for medical French. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 3–9, Tilburg, the Netherlands. Association for Computational Linguistics.

Yohan Bonescki Gumiel, Isabela Lee, Tayane Arantes Soares, Thiago Castro Ferreira, and Adriana Pagano. 2021. Sentiment analysis in portuguese texts from online health community forums: data, model and evaluation. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 64–72. SBC.

Michael Alexander Kirkwood Halliday and Christian MIM Matthiessen. 2014. *Halliday's introduction to functional grammar*. Routledge.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

Roman Jakobson. 2000. On linguistic aspects of translation. *On translation*, 3:30–39.

Anaïs Koptient and Natalia Grabar. 2020. Rated lexicon for the simplification of medical texts. In *The fifth international conference on informatics and assistive technologies for health-care, medical support and wellbeing healthinFO 2020*.

Elisabeth Kübler-Ross, Stanford Wessler, and Louis V Avioli. 1972. On death and dying. *Jama*, 221(2):174–179.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Zengjian Liu, Buzhou Tang, Xiaolong Wang, and Qingcai Chen. 2017. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, 75:S34–S42.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Tony McEnery and Andrew Hardie. 2011. *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Catherine McIlwain, Nancy Santesso, Silvana Simi, Maryann Napoli, Toby Lasserson, Emma Welsh, Rachel Churchill, Tamara Rader, Jackie Chandler, David Tovey, et al. 2013. Standards for the reporting of plain language summaries in new cochrane intervention reviews (pleacs).

Natália Wilcesky Tosini Neves, Jéssica da Silva Cunha Breder, Joaquim Barreto Antunes, Heloisa Carvalho Torres, Andrei Carvalho Sposito, and Maria Helena de Melo Lima. 2021. Knowledge of self-care practices in diabetes: compasso. *Research, Society and Development*, 10(5):e41410515062–e41410515062.

Lene Nielsen. 2011. Personas in co-creation and co-design. In *Proceedings of the 11th Human-Computer Interaction Research Symposium*, pages 38–40.

Lene Nielsen. 2019. *Personas-user focused design. 2nd edition*. Springer.

Don Norman. 2023. *What is Human-Centered Design (HCD)?*, chapter What is Human-Centered Design (HCD)? Interaction Design Foundation - IxDF.

Laura Barbosa Nunes, Jéssica Caroline dos Santos, Ilka Afonso Reis, and Heloísa de Carvalho Torres. 2023. Avaliação do programa comportamental em diabetes mellitus tipo 2: ensaio clínico randomizado. *Ciência & Saúde Coletiva*, 28:851–862.

Francieli Silvéria Oliveira. 2022. *O discurso do autocuidado em saúde: uma descrição de gêneros na covariação experto-leigo*. Ph.D. thesis, Universidade Federal de Minas Gerais.

Shereen Oraby, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. 2019. Curate and generate: A corpus and method for joint control of semantics and style in neural nlg. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5938–5951.

Adriana Pagano. 2015. A linguagem na construção das práticas educativas nas ciências da saúde. *Heloísa de Carvalho Torres, Ilka Afonso Reis, and Adriana Pagano. Empoderamento do pesquisador nas ciências da saúde. Belo Horizonte: FALE/UFMG*, pages 19–36.

Adriana Pagano, Claudia Moro, Elisa Terumi Rubel Schneider, Lilian Mie Mukai Cintho, and Yohan Gumiel. 2023. *PLN na Saúde*, book chapter 21. BPLN.

Shashank Patel, Rucha Nargunde, Shobhit Verma, and Sudhir Dhage. 2022. Summarization and simplification of medical articles using natural language processing. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Lucas Aguiar Pavanelli. 2022. An end-to-end model for joint entity and relation extraction in portuguese. Master's thesis, PUC-Rio.

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, and Vijay Mago. 2022. Medical text simplification using reinforcement learning (teslea): Deep learning–based text simplification approach. *JMIR Medical Informatics*, 10(11):e38095.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. 2018. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

Ricelli MS Ramos, Danielle S Monteiro, and Ivandré Paraboni. 2020. Personality-dependent content selection in natural language generation systems. *Journal of the Brazilian Computer Society*, 26(1):1–21.

Horacio Saggion. 2017. *Syntactic Simplification*, pages 33–45. Springer International Publishing, Cham.

Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J Jansen. 2021. A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, 37(18):1685–1708.

Joni Salminen, Rohan Gurunandan Rao, Soon-gyo Jung, Shammur A Chowdhury, and Bernard J Jansen. 2020. Enriching social media personas with personality traits: A deep learning approach using the big five classes. In *Artificial Intelligence in HCI: First International Conference, AI-HCI 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, pages 101–120. Springer.

Iulian Vlad Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. 2018. A survey of available corpora for building data-driven dialogue systems. *Dialogue & Discourse*, 9(1):1–49.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, 4(1).

Matthew Shardlow and Fernando Alva-Manchego. 2022. Simple tico-19: A dataset for joint translation and simplification of covid-19 texts. In *Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France, June. European Language Resources Association*.

Matthew Shardlow and Raheel Nawaz. 2019. Neural text simplification of clinical letters with a domain specific phrase table. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 380–389, Florence, Italy. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Mads Soegaard and Rikke Friis Dam. 2015. The encyclopedia of human-computer interaction, 2nd ed.

Lucia Specia, Sandra Maria Aluísio, and Thiago AS Pardo. 2008. Manual de simplificação sintática para o português. *Txosten teknikoa NILC-TR-08-06, Sao Carlos-SP*.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Tytti Suojanen, Kaisa Koskinen, and Tiina Tuominen. 2015. *User-centered translation*. Routledge.

Benedikt Szmrecsanyi. 2004. On operationalizing syntactic complexity. In *Le poids des mots. Proceedings of the 7th international conference on textual data statistical analysis. Louvain-la-Neuve*, volume 2, pages 1032–1039.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Gisele de Lacerda Chaves Vieira, Adriana Pagano, Ilka Afonso Reis, Júlia Santos Nunes Rodrigues, and Heloísa de Carvalho Torres. 2017. Translation, cultural adaptation and validation of the diabetes attitudes scale-third version into brazilian portuguese. *Revista Latino-Americana de Enfermagem*, 25.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Diyi Yang. 2023. Cs329x: Human-centered nlp – human-centered nlp. Accessed: 13 July 2023.

Li Yunxiang, Li Zihan, Zhang Kai, Dan Ruilong, and Zhang You. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*.

Federico Zanettin. 2011. Translation and corpus design. *SYNAPS - A Journal of Professional Communication*.

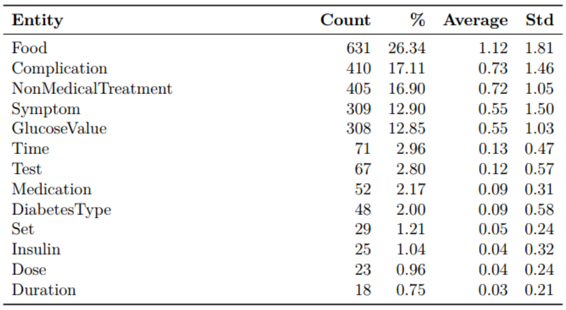
Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Leonardo Zilio, Liana Braga Paraguassu, Luis Antonio Leiva Hercules, Gabriel Ponomarenko, Laura Berwanger, and Maria José Bocorny Finatto. 2020. A lexical simplification tool for promoting health literacy. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 70–76, Marseille, France. European Language Resources Association.

10. Appendix

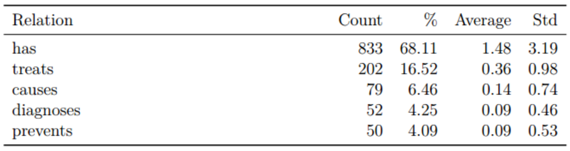
10.1. Annotation – Number of Entities and Relations

A summary of entities in (Dia)Bete corpus [16] is presented in Figure 5 and the number of relations in Figure 6.



Entity	**Count**	**%**	**Average**	**Std**
Food	631	26.34	1.12	1.81
Complication	410	17.11	0.73	1.46
NonMedicalTreatment	405	16.90	0.72	1.05
Symptom	309	12.90	0.55	1.50
GlucoseValue	308	12.85	0.55	1.03
Time	71	2.96	0.13	0.47
Test	67	2.80	0.12	0.57
Medication	52	2.17	0.09	0.31
DiabetesType	48	2.00	0.09	0.58
Set	29	1.21	0.05	0.24
Insulin	25	1.04	0.04	0.32
Dose	23	0.96	0.04	0.24
Duration	18	0.75	0.03	0.21

Figure 5: Dia(Bete) Entities: Number of occurrences, percentage, average, and standard deviation annotation per document sorted in decreasing count order – adapted from Pavanelli (2022).



Relation	Count	%	Average	Std
has	833	68.11	1.48	3.19
treats	202	16.52	0.36	0.98
causes	79	6.46	0.14	0.74
diagnoses	52	4.25	0.09	0.46
prevents	50	4.09	0.09	0.53

Figure 6: Dia(Bete) Relations: Number of occurrences, percentage, average, and standard deviation annotation per document sorted in decreasing count order – adapted from Pavanelli (2022).

10.2. Corpus Samples

From the Medical corpus, we show the sample of answers to expert and non-expert users in Table 9. In addition, we present samples from the Nutrition corpus in Tables 11 and 10.

10.3. Personas

Eris and Alexis were introduced in our paper (please refer to Figures 7 and 8). An additional non-expert persona is presented in Figure 9. For more information about this persona, please see the following Subsection 10.4.

10.4. Non-expert Persona: Otto

Our last persona is a software developer and non-expert in the context of the diabetes mellitus domain. We highlight that even though this persona

[16] Please refer to NER Bete corpus at https://github.com/pavalucas/Bete

Post-edited Post	*Boa noite, amigos e amigas! Depois de uma janta, neste frio, vou de chimarrão quentinho.*
Translated Post-edited Post	Good evening my friends! After having dinner, in this cold weather, I go for a warm *chimarrão*.
Expert Answer	*É importante lembrar que tudo em excesso faz mal e que mais importante que comer ou deixar de comer certos alimentos, é* ***ter uma dieta balanceada com os nutrientes e calorias adequadas para cada pessoa.*** *Por isso, é importante se consultar com um nutricionista (que estão disponíveis gratuitamente no SUS) que poderá montar uma dieta ideal para as suas necessidades.*
Non-expert Answer	*Boa ideia! Chás são uma ótima alternativa aos sucos e refrigerantes, mas lembre-se:* ***ter uma dieta equilibrada em nutrientes e calorias é essencial para o controle do diabetes.***
Gloss for Expert Answer	It is important to remember that everything in excess is harmful and that rather than choosing or avoiding certain types of food the important thing is **having a balanced diet with adequate nutrients and calories for every person**. Therefore, make an appointment with a nutritionist (available free of charge at SUS) who can prepare a dietary plan suited to your needs.
Gloss for Non-expert Answer	Good idea! Tea is a great alternative to juice and soft drinks. Remember, **having a balanced diet is essential for controlling your diabetes.**

Table 8: Sample of questions and answers (expert and non-expert) drafted by nutritional science students.

P: *Bom dia, povo. Gostaria de saber se alguém do grupo já tomou limão com açafrão de manhã, em jejum, para abaixar a diabetes. Me conta se a experiência foi boa.* Q: Good morning people. I would like to know if anyone in the group has already had lemon with turmeric in the morning on an empty stomach to reduce diabetes. Let me know if it was a good experience.
R (Experto): O limão com o açafrão em si pode até não causar nenhum mal. No entanto, **não possuem benefícios** para reduzir a glicemia* (quantidade de **carboidratos** no sangue). A (Expert): Lemon with turmeric may not cause any harm. However, **they are not beneficial** to reduce blood glucose* (amount of **carbohydrates** in the blood).
R (Leigo): O consumo de limão com açafrão não causa nenhum mal, **mas também não traz nenhum benefício no sentido de** reduzir a glicemia (quantidade de **açúcar** no sangue). A (Non-Expert): Consuming lemon with turmeric does not cause any harm, but **it is not beneficial either in terms of** reducing blood glucose (amount of **sugar** in the blood).

Table 9: Sample from the Medical corpus with a post-edited question answered to expert and non-expert users in the diabetes mellitus domain. Some of the simplified sentences are **bolded** (Original text in Brazilian Portuguese with English gloss).

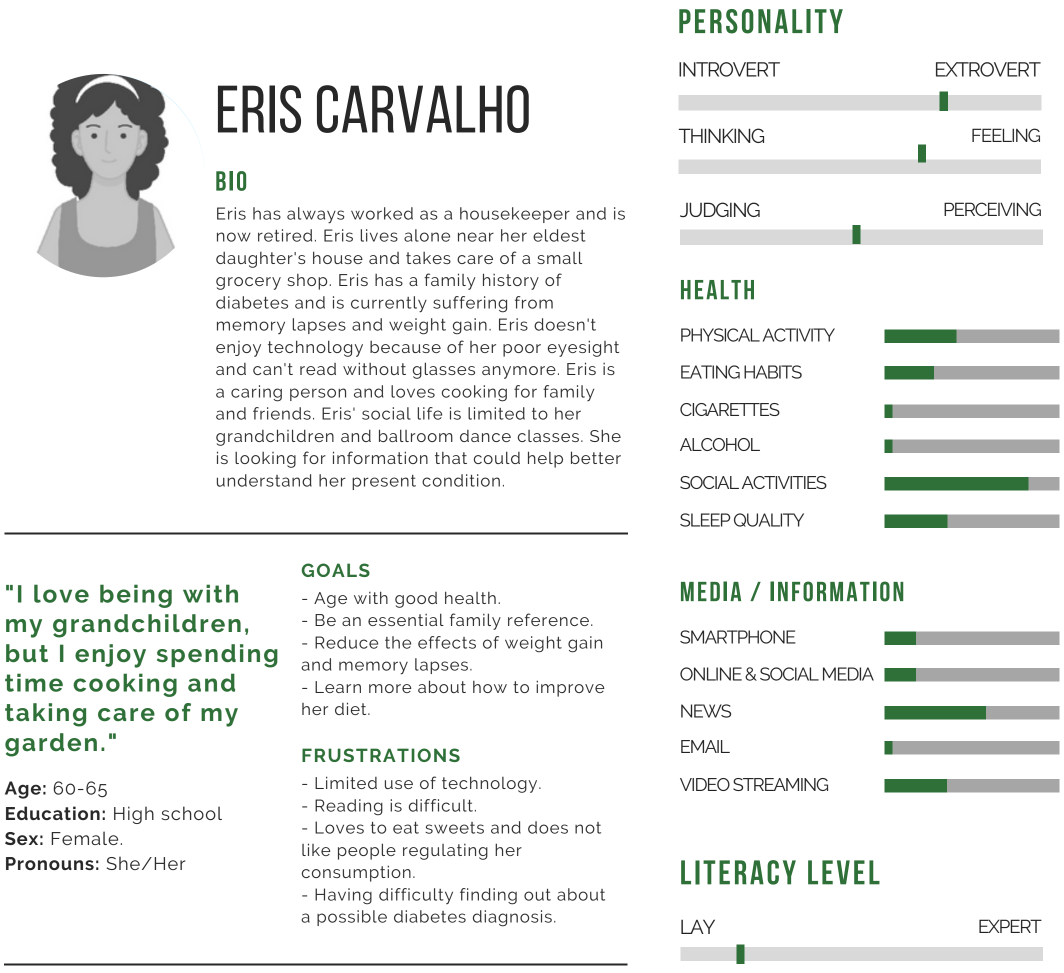
is an expert, he is not an expert within the diabetes mellitus domain. Please refer to this persona in Figure 9. Moreover, the proficiency level addresses mid-texts, an overlap between simplified and more technical texts in the domain.

Post-edited post	*Boa noite. Qual trigo vocês usam? E farinha para empanar? Por favor.*
Gloss for post-edited post	Good evening. What flour do you use? And flour for breading? Please.
Medical Corpus - Expert Answer	*O ideal é que se utiliza trigo integral, já que estes possuem menor índice glicêmico (demora mais tempo para a glicose ser absorvida). De qualquer forma, é importante lembrar que tudo em excesso faz mal e, mais importante que comer ou não determinado alimento, é ter uma dieta balanceada com os nutrientes e calorias adequadas para cada pessoa. Por isso, é importante se consultar com um nutricionista (que estão disponíveis gratuitamente no SUS) que poderá montar uma dieta que seja ideal para você.*
Gloss for Expert Answer	The best is to use whole wheat due to its lower glycemic index (i.e., it takes longer for glucose to be absorbed). However, remember that everything in excess is harmful and that rather than choosing or avoiding certain types of food, the important thing is having a balanced diet with adequate nutrients and calories for every person. Therefore, make an appointment with a nutritionist (available free of charge at SUS) who can prepare a dietary plan suited to your needs.
Nutrition Corpus - Non-Expert Answer	*A farinha de trigo é fonte de amido, ou seja, carboidrato. Ela deve ser consumida com moderação, bem como as frituras, que são ricas em gorduras. Sendo assim, ela pode ser utilizada para empanar, mas se atente à quantidade de farinha e da fritura a ser consumida. Para uma orientação mais específica o ideal é que você se consulte com um nutricionista (que estão disponíveis gratuitamente no SUS) pois ele poderá montar uma dieta ideal para as suas necessidades.*
Gloss for Non-Expert Answer	Wheat flour serves as a significant source of starch, that is, carbohydrates. It should be consumed in moderation, much like fried food, which has a high-fat content. While wheat flour can indeed be used for breading purposes, it's crucial to monitor both the quantity of flour used and the frequency of fried food consumed. For individual guidance, it is best to make an appointment with a nutritionist (available free of charge at SUS) who can prepare a dietary plan suited to your needs.

Table 10: Sample from the Nutrition corpus – Non-Expert Persona (Original text in Brazilian Portuguese with gloss in English).

Post-edited Post	*Alguém que faça contagem de carboidratos?*
Gloss	Anyone who counts carbs?
Expert Answer	*Existe uma calculadora de carboidratos aqui mesmo no site!*
Non-expert Answer	*Existe uma calculadora de carboidratos aqui mesmo no site. Para uma orientação mais específica o ideal é que você se consulte com um nutricionista (que estão disponíveis gratuitamente no SUS) pois ele poderá montar uma dieta ideal para as suas necessidades.*
Gloss for Expert Answer	There is a carbohydrate calculator right here on the website!
Gloss for Non-expert Answer	There is a carbs calculator right here on the website. For individual guidance, it is best to make an appointment with a nutritionist (available free of charge at SUS), who can prepare a dietary plan suited to your needs.

Table 11: Corpus sample from the Nutrition corpus (Original text in Brazilian Portuguese on top with English gloss below).



ERIS CARVALHO

BIO

Eris has always worked as a housekeeper and is now retired. Eris lives alone near her eldest daughter's house and takes care of a small grocery shop. Eris has a family history of diabetes and is currently suffering from memory lapses and weight gain. Eris doesn't enjoy technology because of her poor eyesight and can't read without glasses anymore. Eris is a caring person and loves cooking for family and friends. Eris' social life is limited to her grandchildren and ballroom dance classes. She is looking for information that could help better understand her present condition.

"I love being with my grandchildren, but I enjoy spending time cooking and taking care of my garden."

Age: 60-65
Education: High school
Sex: Female.
Pronouns: She/Her

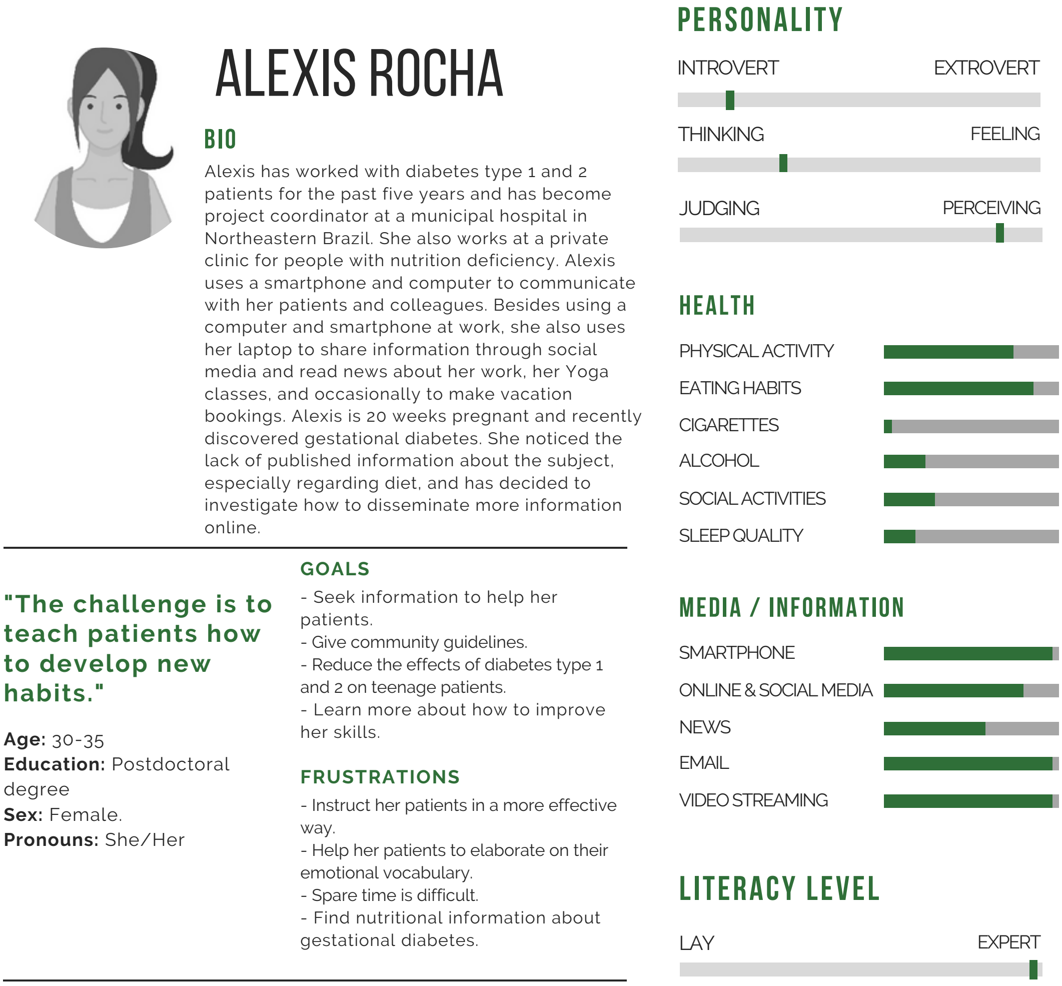
GOALS

- Age with good health.
- Be an essential family reference.
- Reduce the effects of weight gain and memory lapses.
- Learn more about how to improve her diet.

FRUSTRATIONS

- Limited use of technology.
- Reading is difficult.
- Loves to eat sweets and does not like people regulating her consumption.
- Having difficulty finding out about a possible diabetes diagnosis.

Figure 7: Non-expert Persona: Eris.



ALEXIS ROCHA

BIO

Alexis has worked with diabetes type 1 and 2 patients for the past five years and has become project coordinator at a municipal hospital in Northeastern Brazil. She also works at a private clinic for people with nutrition deficiency. Alexis uses a smartphone and computer to communicate with her patients and colleagues. Besides using a computer and smartphone at work, she also uses her laptop to share information through social media and read news about her work, her Yoga classes, and occasionally to make vacation bookings. Alexis is 20 weeks pregnant and recently discovered gestational diabetes. She noticed the lack of published information about the subject, especially regarding diet, and has decided to investigate how to disseminate more information online.

"The challenge is to teach patients how to develop new habits."

Age: 30-35
Education: Postdoctoral degree
Sex: Female.
Pronouns: She/Her

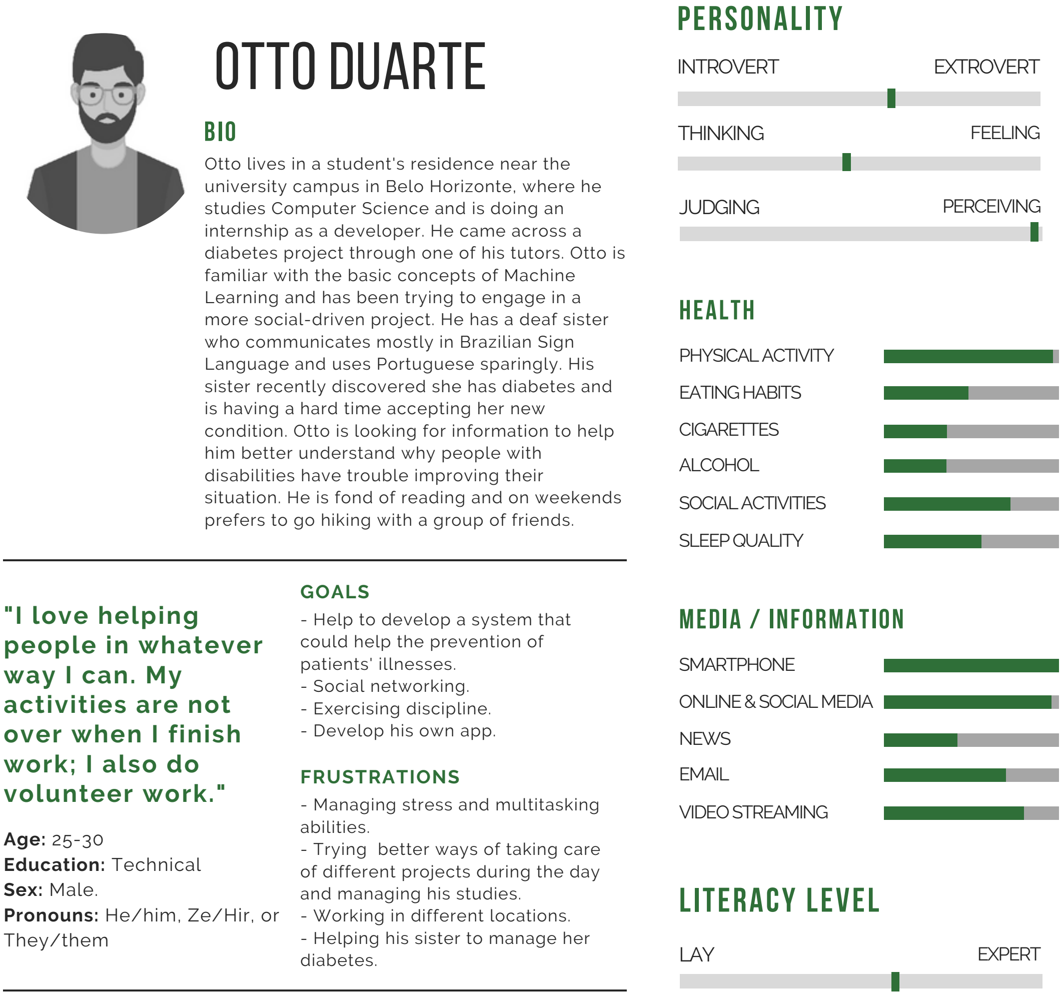
GOALS

- Seek information to help her patients.
- Give community guidelines.
- Reduce the effects of diabetes type 1 and 2 on teenage patients.
- Learn more about how to improve her skills.

FRUSTRATIONS

- Instruct her patients in a more effective way.
- Help her patients to elaborate on their emotional vocabulary.
- Spare time is difficult.
- Find nutritional information about gestational diabetes.

Figure 8: Domain Expert Persona: Alexis.



OTTO DUARTE

BIO

Otto lives in a student's residence near the university campus in Belo Horizonte, where he studies Computer Science and is doing an internship as a developer. He came across a diabetes project through one of his tutors. Otto is familiar with the basic concepts of Machine Learning and has been trying to engage in a more social-driven project. He has a deaf sister who communicates mostly in Brazilian Sign Language and uses Portuguese sparingly. His sister recently discovered she has diabetes and is having a hard time accepting her new condition. Otto is looking for information to help him better understand why people with disabilities have trouble improving their situation. He is fond of reading and on weekends prefers to go hiking with a group of friends.

"I love helping people in whatever way I can. My activities are not over when I finish work; I also do volunteer work."

Age: 25-30
Education: Technical
Sex: Male.
Pronouns: He/him, Ze/Hir, or They/them

GOALS

- Help to develop a system that could help the prevention of patients' illnesses.
- Social networking.
- Exercising discipline.
- Develop his own app.

FRUSTRATIONS

- Managing stress and multitasking abilities.
- Trying better ways of taking care of different projects during the day and managing his studies.
- Working in different locations.
- Helping his sister to manage her diabetes.

PERSONALITY

HEALTH

MEDIA / INFORMATION

LITERACY LEVEL

Figure 9: Non-expert Persona: Otto.