

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Faculdade de Letras
Programa de Pós-Graduação em Estudos Linguísticos

Rossana da Cunha Flores

**CONTEXT-ENRICHED NATURAL LANGUAGE GENERATION
MODELS**

Belo Horizonte
2023

Rossana da Cunha Flores

CONTEXT-ENRICHED NATURAL LANGUAGE GENERATION MODELS

Tese apresentada ao Programa de Pós-graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutora em Estudos Linguísticos.

Área de concentração: Linguística Aplicada.
Linha de pesquisa: Estudos da Tradução – 3B.

Orientador: Prof. Dr. Fábio Alves da Silva Júnior

Coorientadora: Prof(a). Dr(a). Adriana Silvina Pagano

Coorientador: Prof. Dr. Thiago Castro Ferreira

Belo Horizonte

2023

F634c

Flores, Rossana da Cunha.

Context-enriched Natural Language Generation Models [manuscrito] /
Rossana da Cunha Flores. – 2023.

1 recurso online (202 f. : il., tabs. (algumas color.)) : pdf.

Orientador: Fábio Alves da Silva Júnior.

Coorientadora: Adriana Silvina Pagano.

Coorientador: Thiago Castro Ferreira.

Área de concentração: Linguística Aplicada.

Linha de Pesquisa: Estudos da Tradução.

Tese (doutorado) – Universidade Federal de Minas Gerais,
Faculdade de Letras.

Bibliografia: f. 157-180.

Apêndices: f. 181-202.

Exigências do sistema: Adobe Acrobat Reader.

1. Linguística aplicada – Teses. 2. Linguística – Processamento de
dados – Teses. 3. Processamento da linguagem natural (Computação) –
Teses. I. Alves, Fábio. II. Pagano, Adriana Silvina. III. Ferreira, Thiago
Castro. IV. Universidade Federal de Minas Gerais. Faculdade de Letras.
IV. Título.

CDD: 401.9



UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

FOLHA DE APROVAÇÃO

Context-enriched natural language generation models

ROSSANA DA CUNHA FLORES

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Doutor em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA APLICADA, linha de pesquisa Estudos da Tradução.

Aprovada em 01 de setembro de 2023, pela banca constituída pelos membros:

Prof(a). Fábio Alves da Silva Júnior - Orientador

UFMG

Prof(a). Adriana Silvina Pagano - Coorientadora

UFMG

Prof(a). Thiago Castro Ferreira - Coorientador

UFMG

Prof(a). Igor Antonio Lourenco da Silva

UFU

Prof(a). Yohan Bonescki Gumiel

PUC-PR

Prof(a). Evandro Landulfo Teixeira Paradelo Cunha

UFMG

Prof(a). Maucha Andrade Gamonal

UFMG



Documento assinado eletronicamente por **Fabio Alves da Silva Junior, Professor do Magistério Superior**, em 09/09/2023, às 22:09, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Evandro Landulfo Teixeira Paradela Cunha, Professor do Magistério Superior**, em 11/09/2023, às 11:11, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Maucha Andrade Gamonal, Usuária Externa**, em 15/09/2023, às 10:22, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Igor Antônio Lourenço da Silva, Usuário Externo**, em 21/09/2023, às 15:47, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Thiago Castro Ferreira, Usuário Externo**, em 23/09/2023, às 10:43, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Yohan Bonescki Gumiel, Usuário Externo**, em 27/09/2023, às 13:43, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Adriana Silvina Pagano, Professora do Magistério Superior**, em 02/10/2023, às 12:31, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2492747** e o código CRC **AA89CD60**.

To Theo, my sunflower.

Acknowledgements

A Ph.D. journey is full of expectations and possibilities. A path full of challenges and unique experiences that I am thrilled to conclude by feeling more confident and pleased with myself. Since 2018, I have started a beautiful family, traveled to the most incredible places, lived the fears during the pandemic, and encountered passion, friendship, love, and kindness in Belo Horizonte and here in Floripa, within different environments that I visited, such as UFMG, UFSC, LETRA, WiNLP, University of Macau, ACL 2019 in Florence, and Kiphu 2019 in Montevideo.

I want to thank my family for all their support during this long endeavor, especially my loving son Theo, who taught me the most profound lessons since 2019. He was the one who gave me the strength to complete this chapter of my life. Thanks, my little boy, for all the love you gave me when I had to be absent due to this research.

To my beloved husband and Theo's father, Arion, thank you for believing in me, especially during the initial years of this journey when I had to leave my job. It would have been harder to walk this path without you.

In memory, I want to thank my grandmother Rosa and my grandfather Teódulo, who raised me, gave me direction, and always encouraged me to study.

To my dear friend Marta, thanks for being by my side whenever possible. Thanks for your dedicated time while reviewing this thesis and leaving encouraging words that motivated me to finish it.

To my friends Grazi, Cris, Patrícia, Clau, Cybelle, and Camila, thanks for the emotional and/or academic support during these last years, and for understanding my frequent absence.

I am grateful to my advisor, Professor Ph.D. Fabio Alves for encouraging me at the AILA 2017 conference to apply for the Ph.D. position at UFMG. He believed in the potential of my research and presented me with different approaches

here in Brazil and in Macau, opening my mind to my very best.

To my co-advisor, Professor Ph.D. Adriana Pagano for walking with me through different paths until we found the best solution to fulfill my ideas. Her insightful lessons and driving force provided me with the tools necessary for the most significant challenges of my academic career.

To my co-advisor, Professor Ph.D. Thiago Castro Ferreira for supporting my work within these two complementary areas. His experience and kindness were essential for me to enter this new world.

To professors Ph.D. Igor A. Lourenço da Silva and Ph.D. Ivandre Paraboni for sharing their knowledge during the qualifying exam that contributed to the improvement of this thesis.

To DaMata and Betinha coworkers, thanks for sharing your knowledge and embracing the challenges in our meetings during the pandemic. To UFMG, POSLIN, and LETRA for making this possible.

I am profoundly grateful for this incredible opportunity.

Abstract

With the advent of conversational agents or chatbots and the widespread digital dissemination of information on critically sensitive domains, such as governance and healthcare, the need to model natural language generation to target human interlocutors has become all the more important. While Natural Language Processing (NLP) models have gained substantial attention, only in recent years has research opened new paths for tackling Human-Computer Interaction (HCI) from the perspective of natural language. Among those, just a few have investigated how to develop language production models using a Human-Centered Design (HCD) perspective to build context-enriched Natural Language Generation (NLG) models from their conceptions. Drawing on the tenets of the HCD methodology, this thesis compiles three studies aimed at enriching NLG models with context in order to generate texts for successful human communication, offering a productive improvement in user comprehension and experience. The first study focuses on a particular task of NLG models known as Referring Expression Generation (REG). We address the gap in state-of-the-art REG models when dealing with entities not seen during their training, compromising the comprehension of the final text. Drawing on the results of the automatic and human evaluation and an ablation study using the WebNLG corpus, this first approach contributes to generating more meaningful referring expressions to unseen entities. The second study collaborates with the development of DaMata, a robot journalist that automatically produces daily and monthly reports about the deforestation of the Legal Amazon area. With the context enrichment of the REG component, DaMata can provide more diverse information and reach wider audiences. Moreover, DaMata’s pipeline framework ensures high-fidelity data publishing and enhances user experience. Finally, in the third study, we focus on developing a human-centered NLG model within a conversational question-answering system in the healthcare domain. We created personas to model interpersonal interaction (expert and non-expert users) in that specific domain. The third approach presented competitive results with a human-centered context enrichment, giving direction to more adequate relations between humans and NLP – Human-Centered NLP. In addition, we showed that an HCD approach benefits language generation from the perspective of “machines”

but mainly humans, who are the primary actors within NLP. Overall, this thesis contributes with new directions to low-resource contexts (languages other than English and sensitive domains) where the need to promote effective communication is essential.

Keywords: Natural Language Generation, Referring Expression Generation, Conversational Systems, Persona, Human-Centered NLP.

Resumo

Com o advento de agentes conversacionais ou chatbots e a disseminação digital generalizada de informações em domínios criticamente sensíveis, como a governança e a saúde, a necessidade de modelar a geração de linguagem natural a fim de alcançar interlocutores humanos tornou-se ainda mais importante. Embora os modelos de Processamento de Linguagem Natural (PLN) tenham ganhado atenção substancial, apenas nos últimos anos a pesquisa abriu novos caminhos para abordar a Interação Humano-computador (IHC) a partir da perspectiva da linguagem natural. Entre esses, poucos estudos investigaram como desenvolver modelos de produção de linguagem através de uma perspectiva do Design Centrado em Humanos (HCD) na construção de modelos Geração de Linguagem Natural (GLN) enriquecidos de contexto desde a concepção. Com base nos princípios da metodologia HCD, esta tese compila três estudos que visam enriquecer modelos GLN com contexto a fim de gerar textos para uma comunicação eficaz, além de oferecer uma melhora produtiva na compreensão e experiência do usuário. O primeiro estudo se concentra em uma tarefa particular de modelos de GLN, conhecida como Geração de Expressão de Referência (REG), abordando a lacuna dos modelos REG de última geração em lidar com entidades que não foram vistas durante seu treinamento e assim prejudicando a compreensão do texto final. A partir dos resultados das avaliações automática e humana e de um estudo de ablação que utiliza o corpus WebNLG, esta primeira abordagem contribui para gerar expressões de referência mais significativas a entidades não vistas. O segundo estudo colabora com o desenvolvimento do DaMata, um robô jornalista que produz relatórios diários e mensais automaticamente sobre o desmatamento na Amazônia Legal. Com o enriquecimento contextual do componente REG, o DaMata pode fornecer informações mais diversificadas e alcançar uma maior audiência. Além disso, a estrutura de *pipeline* do DaMata assegura a publicação de dados de alta fidelidade e promove uma melhor experiência do usuário. Finalmente, no terceiro estudo, o principal objetivo é o desenvolvimento de um modelo GLN centrado em humanos através de uma abordagem de agente conversacional de pergunta e resposta para o domínio da saúde. Foram criadas personas de forma a modelar as relações interpessoais (usuários especialistas e não especialistas) nesse domínio em

particular. A terceira abordagem apresentou resultados competitivos através do enriquecimento de contexto centrado em humanos, e com direcionamentos para implementações mais adequadas às relações entre humanos e PLN – PLN Centrado em Humanos. Além disso, mostramos que uma abordagem HCD pode beneficiar a geração de linguagem da perspectiva das “máquinas”, mas principalmente dos humanos, que são os principais atores do PLN. No geral, esta tese contribui com novas direções para contextos com poucos recursos (idiomas diferentes do inglês e domínios sensíveis) onde a necessidade de promover uma comunicação eficaz é essencial.

Palavras-chave: Geração de Linguagem Natural, Geração de Expressão de Referência, Agentes Conversacionais, Persona, PLN Centrado em Humanos.

List of Figures

Figure 1 – Overview of the thesis and the three studies.	28
Figure 2 – NLG Applications – adapted from Santhanam and Shaikh (2019).	38
Figure 3 – Pipeline architecture proposed by Ferreira et al. (2018a).	38
Figure 4 – Conversational AI – Task-based flow – adapted from Gao, Galley and Li (2019, p. 132).	45
Figure 5 – Conversational AI architecture – adapted from Horn et al. (2019, p. 3).	49
Figure 6 – Seven HCI challenges according to Stephanidis et al. (2019, p. 1231).	51
Figure 7 – User-Centered Design – adapted from Soegaard and Dam (2015).	54
Figure 8 – Referring Expression Generation of the entity <code>Ada_Lovelace</code> : a noun-phrase.	60
Figure 9 – Referring expression generation of an unseen entity <code>Adenan_Satem</code>	78
Figure 10 – Robot-Journalist Pipeline Architecture and a Human-Readable Output Sample.	84
Figure 11 – Human-centered AI framework highlighting groups of users (Ri- bera; Lapedriza, 2019).	99
Figure 12 – User-centered design and how to integrate personas and corpora – adapted from Soegaard and Dam (2015).	101
Figure 13 – Final/Lay-user persona: Eris (created with Canva (2023)).	103
Figure 14 – Final/Lay-user persona: Otto (created with Canva (2023)).	105
Figure 15 – Domain Expert Persona: Alexis (created with Canva (2023)).	107
Figure 16 – Persona-based corpus compilation of lay and expert texts (created with Bizagi (2023)).	116
Figure 17 – Intent and slot-filling annotation of social media post in the Diabetes Mellitus domain (Brazilian Portuguese).	120
Figure 18 – Bete Entities: Number of occurrences, percentage, average, and standard deviation annotation per document sorted in decreasing count order – adapted from Pavanelli (2022).	122

Figure 19 – Bete Relations: Number of occurrences, percentage, average, and standard deviation annotation per document sorted in decreasing count order – adapted from Pavanelli (2022).	122
Figure 20 – NLP models analysis – semantic resources, end-users, domain adaptation requirements, and HCD methodology (created with Bizagi (2023)).	130
Figure 21 – Third approach on enriching NLG models – Human-Centered NLG model pipeline (created with Bizagi (2023)).	131
Figure 22 – Applying HCD methodology to the context enrichment of an NLG model (created with Bizagi (2023)).	133
Figure 23 – Training paradigms within PLMs – adapted from Cai et al. (2023, p. 4).	135
Figure 24 – Main Components of the Dia(Bete) Conversational AI System.	142
Figure 25 – Betinha stories developed with Rasa framework.	143
Figure 26 – Rasa shell during the HCD evaluation.	148

List of Tables

Table 1	– Sample from our corpus with a post-edited question answered to expert and lay users in the diabetes mellitus domain. Some of the simplified sentences are bolded (The English translation and its original form in Brazilian Portuguese).	43
Table 2	– Sample from our corpus of how to answer a question to expert and lay users in the diabetes mellitus domain (The English translation and its original form in Brazilian Portuguese).	48
Table 3	– A set of RDF (Resource Description Framework) triples and their corresponding text. Followed by entities’ mapping and a delexicalized template.	63
Table 4	– Referring Expressions’ Accuracy and String-edit distance (SED), BLEU and Text Accuracy scores of the models, Pronoun – Precision, Recall, and F1-Score of the models in the automatic evaluation. The best results are boldfaced , whereas the second best are <u>underlined</u> .	69
Table 5	– Sample outputs of an unseen domain (Politician) – the original and the generated text of each of the four models. Referring expressions are boldfaced , and constants are “double-quoted” and <i>italicized</i> .	70
Table 6	– Mean and $\pm 95\%$ Confidence intervals on Fluency, Grammaticality, and Semantic Adequacy results for All, Seen, and Unseen entities of the Human Evaluation. The best results are bolded , whereas the second best are <u>underlined</u> .	74
Table 7	– Results of the Ablation Analysis. The best features (with the highest drop) are boldfaced , whereas the second best are <u>underlined</u> .	76
Table 8	– Sample of DaMata’s context enrichment in Brazilian Portuguese and English according to information recovered from Semantic Web.	87

Table 9 – Data Statistics.	88
Table 10 – The strategies and how they were used to represent the numeric referring expressions. They are <i>highlighted</i>	89
Table 11 – English error rates and BLEU score for the 4 numerical strategies and 3 language models - Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.	90
Table 12 – Portuguese error rates and BLEU score for the 4 numerical strategies and 3 language models. Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.	91
Table 13 – English and Portuguese “Incorrect Number” error rates results, grouped by number description strategies – higher values denote more errors. For comparison, we show only results for Monthly and Daily Deforestation, since these domains are present in both languages. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.	92
Table 14 – Samples of expert and non-expert texts in the Diabetes Mellitus domain (Original text in Brazilian Portuguese and the corresponding English translation) – Simplifications are boldfaced	114
Table 15 – Persona-based corpus compilation criteria.	115
Table 16 – Linguistic literacy criteria – based on NILC research (Specia; Aluísio; Pardo, 2008; Aluísio et al., 2008).	119
Table 17 – Annotation sample from our corpus in Brazilian Portuguese and its translation to English.	121
Table 18 – Entities description and examples according to Figure 17 – adapted from Pavanelli (2022, p. 26).	121
Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from Gasperin, Maziero and Aluisio (2010).	123

Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from Gasperin, Maziero and Aluisio (2010).	124
Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from Gasperin, Maziero and Aluisio (2010).	125
Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from Gasperin, Maziero and Aluisio (2010).	126
Table 20 – Sample of the simplification spreadsheet in Brazilian Portuguese and the corresponding translation (some columns were omitted due to ethical considerations).	127
Table 20 – Sample of the simplification spreadsheet in Brazilian Portuguese and the corresponding translation (some columns were omitted due to ethical considerations).	128
Table 21 – NLP models analyzed according to the use of personas, external resources, corpora (persona or user profiles), and HCD approach.	131
Table 22 – Fine-tuned models approach followed by their respective configuration and tasks (NLI – Natural Language Inference and CQA – Conversational AI Question Answering).	136
Table 23 – Persona model’s settings based on the literacy scale presented in Section 16.	138
Table 24 – Persona-based corpus NLI: F1-Score and Accuracy considering two forms of adding persona context – ID and Description. . . .	138
Table 25 – Persona-based corpus statistics considering the original corpus (annotated by Empoder@ research group) and the nutrition corpus (annotated by nutrition students based on the medicine annotation).	139
Table 26 – Conversational AI Frameworks/Platforms.	140
Table 27 – Pre-trained NLI models – F1-Score and Accuracy metrics in NLI task. The best results are boldfaced	144
Table 28 – Pre-trained CQA models – BLEU and SARI metrics in CQA task. The best results are boldfaced	145

List of Abbreviations

AI	Artificial Intelligence
AIML	Artificial intelligence Mark-up Language
ANN	Artificial Neural Network
API	Application Programming Interface
ASR	Recognize/transcribe speech
ATS	Automatic Text Simplification
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CNN	Convolution Neural Network
COLING	International Conference on Computational Linguistics
COVID-19	Coronavirus Disease of 2019
CQA	Conversational AI Question-Answering
CSV	Comma-Separated Values
DARPA	Defense Advanced Research Projects Agency
DCU	Dublin City University
DETER	Brazil's Real-time Deforestation Detection System
DIET	Dual Intent and Entity Transformer
DL	Deep Learning
DNN	Deep Neural Networks
EHR	Electronic Health Records
EMNLP	Empirical Methods in Natural Language Processing Conference
FAQ	Frequently Asked Questions
GDPR	Union's General Data Protection Regulation
GLN	<i>Geração de Linguagem Natural</i> – NLG
GPT	Generative Pre-training
HAI	Human-Centered Artificial intelligence
HCD	Human-Centered Design
HCI	Human-Computer Interaction
HiSAN	Hierarchical Self-Attention Network

HXAI	Human-centered eXplainable Artificial Intelligence
IAH	Inteligência Artificial Centrada em Humanos – HAI
IBM	International Business Machines Corporation
ID	Identification
IHC	<i>Interação Humano-Computador</i> – HCI
INAF	<i>Indicador Nacional de Alfabetismo Funcional</i> (National Indicator of Functional Literacy)
INLG	International Natural Language Generation Conference
INPE	<i>Instituto Nacional de Pesquisas Espaciais</i> (Brazilian National Institute for Space Research)
IR	Information Retrieval
ISO	International Standardization Organization
JSON	JavaScript Object Notation
LAPPIS	<i>Laboratório Avançado de Produção, Pesquisa e Inovação em Software</i> (Laboratory for Advanced Software Production, Research, and Innovation)
LARA	Latin America Research Awards
LETRA	Laboratory for Experimentation in Translation
LGPD	<i>Lei Geral de Proteção de Dados</i> (Brazil’s General Data Protection Law)
LLM	Large Language Model
LS	Lexical Simplification
LSTM	Long-Short Term Memory layers
LT4SG	Language Technologies for Social Good
LUIS	Language Understanding Intelligent Service
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MT	Machine Translation
NAACL	North American Chapter of the Association for Computational Linguistics
NER	Named Entity Recognition
NILC	<i>Núcleo Interinstitucional de Linguística Computacional</i> (The Interinstitutional Center for Computational Linguistics)

NLG	Natural Language Generation
NLI	Natural Language Inference
NLP	Natural Language Processing
NLU	Natural Language Understanding
NMT	Neural Machine Translation
NN	Neural Network
NoSQL	Not Only SQL (Structured Query Language)
PLM	Pre-trained Language Model
PLN	<i>Processamento de Linguagem Natural</i> – NLP
POSLIN	Graduate Program in Linguistics and Applied Linguistics
QA	Question-Answering
RDF	Resource Description Framework
RE	Relation Extraction
REG	Referring Expression Generation
RQE	Recognizing Question Entailment
SAMSA	Simplification Automatic evaluation Measure through Semantic Annotation
SARI	System output Against References and Input sentence
SED	String-Edit distance
SMT	Statistical Machine Translation
SQL	Structured Query Language
SS	Sentence Simplification
SUS	<i>Serviço Nacional de Saúde</i> (Brazilian National Health Service)
TAIS	<i>Assistente Virtual da Lei de Incentivo à Cultura</i> (Virtual Assistant for the Culture Incentive Law)
TER	Translation Edit Rate
TS	Text Simplification
TSAR	Workshop on Text Simplification, Accessibility, and Readability
TTS	Synthesize Speech
UC	University of California
UCD	User-Centered Design
UFMG	<i>Universidade Federal de Minas Gerais</i> (Federal University of Minas Gerais)

UFRGS	<i>Universidade Federal do Rio Grande do Sul</i> (Federal University of Rio Grande do Sul)
UnB	<i>Universidade de Brasília</i> (University of Brasília)
URI	Uniform Resource Identifier
USP	<i>Universidade de São Paulo</i> (University of São Paulo)
XAI	eXplainable Artificial Intelligence
XML	Extensible Markup Language

Contents

1	INTRODUCTION	26
1.1	Thesis Overview	31
2	TOWARDS CONTEXT-ENRICHED NLG MODELS	34
2.1	Thesis Questions	35
2.2	Objectives	35
3	BACKGROUND	37
3.1	Natural Language Generation	37
3.1.1	Discourse Ordering	39
3.1.2	Text Structuring	39
3.1.3	Lexicalization	40
3.1.4	Referring Expression Generation	40
3.1.5	Textual Realization	41
3.1.6	Application: Robot Journalist	41
3.2	Text Simplification Systems	41
3.2.1	Lexical simplification	42
3.2.2	Machine Translation (MT)	43
3.2.3	Hybrid approach	44
3.3	Conversational AI / Dialogue Systems	44
3.3.1	Chatbots	46
3.3.2	Question Answering Systems	47
3.4	Intersections of NLG and Human-Centered Approaches	50
3.4.1	Human-Computer Interaction	50
3.4.2	Human-Centered Artificial Intelligence	52
3.4.3	Human/User-Centered Design	53
3.4.3.1	Persona versus User Profiles	55
3.4.3.2	Persona versus Implied Reader	56
3.4.3.3	Persona versus Personality	56

4	ENRICHING AN NLG MODEL FROM A REFERRING EXPRESSION GENERATION APPROACH	57
4.1	Introduction	59
4.2	Related work	61
4.2.1	Data	62
4.2.2	Model	64
4.2.2.1	Encoder	65
4.2.2.2	Decoder	66
4.2.2.2.1	Attention Mechanism	66
4.2.2.2.2	Decoding	67
4.2.2.2.3	Copy Mechanism	67
4.2.2.2.4	Loss	68
4.2.3	Automatic Evaluation	68
4.2.3.1	Data	68
4.2.3.2	Model Settings	69
4.2.3.3	Baselines	70
4.2.3.3.1	OnlyNames	70
4.2.3.3.2	NeuralREG+Catt	71
4.2.3.3.3	ProfileREG	71
4.2.3.4	Metrics	71
4.2.3.5	Results	71
4.2.4	Human Evaluation	72
4.2.4.1	Method	72
4.2.4.2	Results	73
4.2.5	Ablation Study	74
4.2.5.1	Method	75
4.2.5.2	Results	75
4.2.6	Discussion	76
4.2.6.1	Surrounding Context	77
4.2.6.2	Copy Mechanism	77
4.2.6.3	Gender and Type Entity Representations	77
4.2.6.4	Future Work	78
4.2.6.5	Conclusion	79

5	ENRICHING AN NLG MODEL FROM A ROBOT-JOURNALISM APPROACH	81
5.1	Introduction	83
5.1.1	System Overview	84
5.1.1.1	Content Selection	84
5.1.1.2	Discourse Ordering	85
5.1.1.3	Text Structuring	85
5.1.1.4	Lexicalization	85
5.1.1.5	Referring Expression Generation	85
5.1.1.6	Textual Realization	86
5.2	Enriching DaMata's Context	86
5.3	Conclusion	93
6	ENRICHING AN NLG MODEL FROM A HUMAN-CENTERED APPROACH	94
6.1	Addressing the Gap in HCD Approaches to NLG	96
6.1.1	Applying HAI and HCD Methodologies to NLG	98
6.1.2	Final/Lay User Persona: Eris	102
6.1.2.1	Eris' Biography	103
6.1.2.2	Eris' Goals	104
6.1.2.3	Eris' Frustrations	104
6.1.3	Final/Lay User Persona: Otto	104
6.1.3.1	Otto's Biography	105
6.1.3.2	Otto's Goals	106
6.1.3.3	Otto's Frustrations	106
6.1.4	Domain Expert Persona: Alexis	106
6.1.4.1	Alexis' Biography	107
6.1.4.2	Alexis' Goals	108
6.1.4.3	Alexis' Frustrations	108
6.2	Addressing the Lack of In-Domain Data	108
6.3	Addressing the Lay-Expert Communication Gap	110
6.4	Persona-Based Corpus	112
6.4.1	Corpus and Annotation Process	114

6.4.2	Corpus	115
6.4.3	Compilation and Automatic Correction	115
6.4.4	De-Identification	117
6.4.5	Post-edition	117
6.4.6	Annotation	119
6.4.7	Simplification	122
6.4.8	Expert Evaluation	128
6.5	Revisiting the NLG Model from an HCD Approach	129
6.5.1	Hybrid Approach	132
6.5.2	Context	133
6.5.3	Fine-tuning	134
6.5.3.1	Baseline Models	135
6.5.3.2	Fine-tuned Models Settings	137
6.5.4	Classification – Natural Language Inference	137
6.5.5	Question-Answering & Conversational AI	139
6.5.6	Conversational AI	140
6.5.7	Evaluation	143
6.5.7.1	NLI Evaluation	144
6.5.7.2	CQA Evaluation	145
6.5.7.3	HCD & Human Evaluations	145
6.5.8	Results & Discussion	146
6.5.8.1	Metrics	146
6.5.8.2	Strategies	146
6.5.8.3	Pre-trained Models	146
6.5.9	Conclusion	147
6.5.9.1	Limitations and Future Work	148
7	CONCLUSION	150
7.1	Contributions	152
7.1.1	External Knowledge	152
7.1.2	Interpersonal Relations	153
7.1.3	Enriching NLG Models with HCD Approaches	153
7.2	Future Work	154

7.2.1	External Knowledge	154
7.2.2	Interpersonal Relations	155
7.2.3	Enriching NLG Models with HCD Approaches	155
7.3	Ethical Considerations	155
	REFERENCES	157

APPENDIX 181

APPENDIX A – DAMATA’S UNDER-REVIEW PAPER . .	182
---	------------

APPENDIX B – PROTOCOLO DE SIMPLIFICAÇÃO (SIM- PLIFICATION PROTOCOL)	195
--	------------

1 Introduction

“Intelligent computing requires learning to operate with the semantic practices of natural language” predicted linguists M.A.K. Halliday and C. Matthiessen far back in the 1990s, in an early attempt to model natural language generation based on parameters of language user and use. *What* is said, as well as *how* it is said, depending on *who* is being addressed, are all conflated in [Halliday and Matthiessen \(1999\)](#) theory as intricately woven components of language production. In this regard, exploring the intricacy of natural language was put forward by the authors as a model to get machines to output written text for humans, at that time through automatically generated weather reports and cooking recipes ([Halliday; Matthiessen, 1999](#); [Halliday; Matthiessen, 2014](#)). Twenty-four years later, with the advent of conversational agents or chatbots, the most recent prompting Artificial Intelligence (AI) as ChatGPT¹ ([OpenAI, 2022](#); [OpenAI, 2023](#)), and the widespread digital dissemination of information on critically sensitive domains, such as governance and healthcare, the need to model natural language generation to target human interlocutors has become all the more important.

While natural language AI models have gained substantial attention in recent years, most Human-Computer Interaction (HCI)² and Human-Centered Artificial Intelligence (HAI)³ research remains in the context of the usability⁴ of interfaces, i.e., additionally to websites and applications. Regardless, only in recent years has research opened up new paths for tackling human-computer interaction from the perspective of natural language. Among those are [Li et al. \(2016\)](#); [Luan et al. \(2017\)](#); [Chu and Wang \(2018\)](#); [Fikri, Takamura and Okumura \(2018\)](#); [Ram](#)

¹ [<https://openai.com/blog/chatgpt>](https://openai.com/blog/chatgpt)

² Human-Computer Interaction (HCI) is “an area of research and practice that emerged in the early 1980s, initially as a specialty area in computer science embracing cognitive science and human factors engineering” ([Carroll, 2009](#), p. 1).

³ For more information about Human-centered Artificial Intelligence (HAI), please refer to HAI Section 3.4.2.

⁴ According to [Sharp, Preece and Rogers \(2019, p. 20\)](#), usability “refers to ensuring that interactive products are easy to learn, effective to use, and enjoyable from the person’s perspective”.

et al. (2018); Salminen et al. (2020); Wang et al. (2021a); Majumder et al. (2021); Gonzalez (2021); Paranjape (2022); and Majumder (2023), who investigated how to develop language production models, also known as Natural Language Generation (NLG) models, using a Human-Centered Design (HCD) perspective.

According to Norman (2023, p. 2), HCD is “a vital approach for accommodating real users – real people”. The author (2023) emphasizes the four principles of HCD: (i) being people-centered⁵ (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 form⁶. Within HCD, another frequently used term is User-Centered Design (UCD)⁷, which is an “approach aiming to develop more usable interactive systems by focusing on the use of the system as well as by applying human factors/ergonomics and usability knowledge and techniques” (ISO, 2019, p. 2). Sharp, Preece and Rogers (2019) argue that this is a methodology where end-users could influence how a design takes shape, allowing the enrichment of NLG models to produce texts more tuned to domain expertise and interpersonal relations.

Drawing on the tenets of the HCD methodology, this thesis compiles three studies aimed to enrich NLG models with context in order to generate texts for successful expert-lay communication, offering a productive improvement in user comprehension and experience. The purpose of enriching context through different approaches is to positively affect users’ understanding by providing texts more suitable to users’ expectations in order to present relevant information and at the same time produce a text accepted by these users to engage them with different NLG applications. Please refer to Figure 1 for an overview of this thesis’s main areas.

⁵ <https://www.interaction-design.org/literature/topics/people-centered-design>

⁶ <https://www.interaction-design.org/literature/topics/human-centered-design>

⁷ According to Rubin and Chisnell (2008, p. 12), user-centered design “represents the techniques, processes, methods, and procedures for designing usable products and systems, but just as important, it is the philosophy that places the user at the center of the process”.

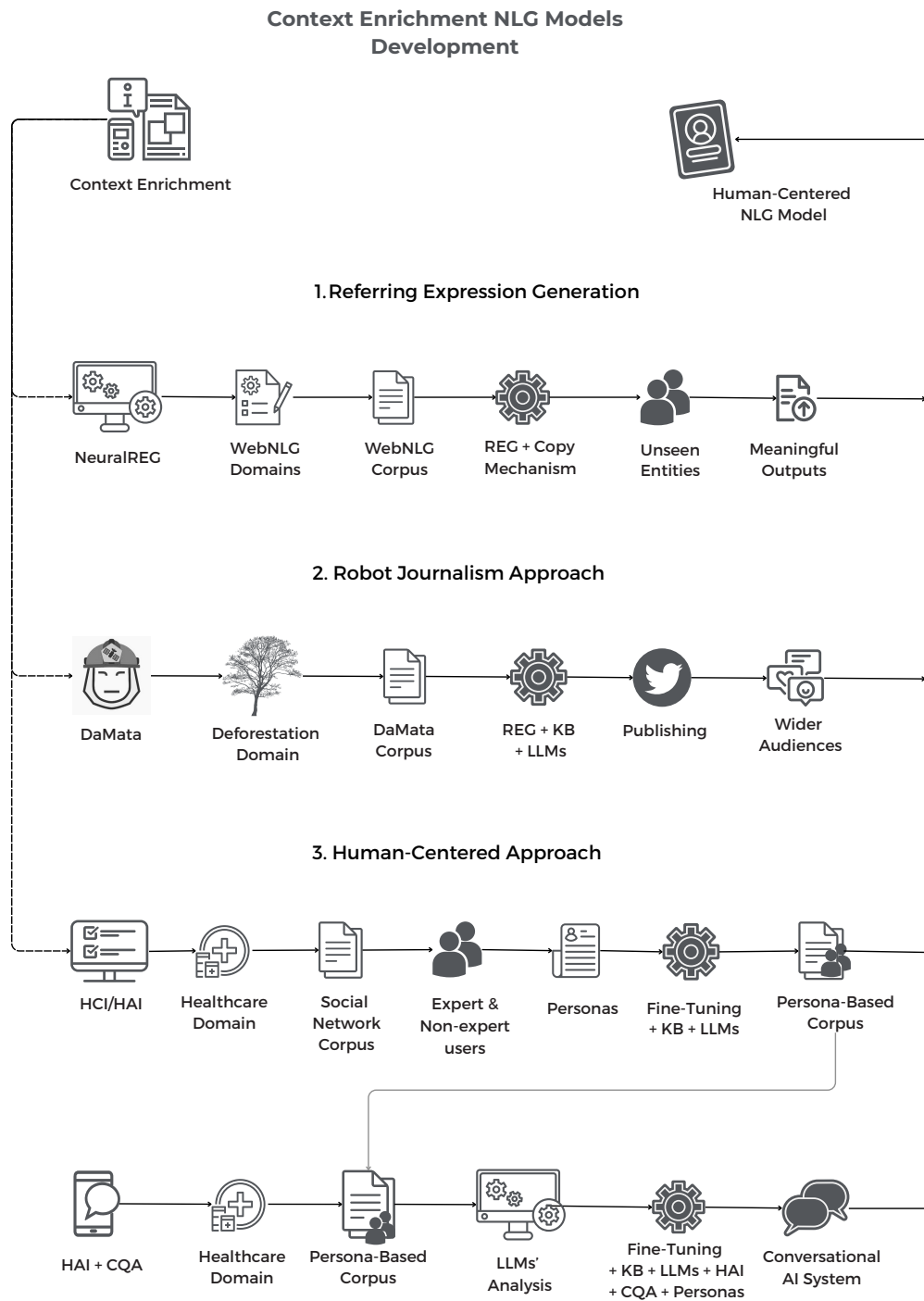


Figure 1 – Overview of the thesis and the three studies.

In this study, we follow HCD principles (Norman, 2023, p. 2):

- Focus on the generation of texts to people within their contexts – e.g., in the third study, people are interacting through social network platforms in order to seek information about a health condition.
- Address the lack of information and generate data based on literacy guidelines in order to people’s better understanding – e.g., in the third study, texts are presented in a simplified version to address healthcare lay users.
- Connect people’s needs with research knowledge production to better human-AI cognitive interactions – e.g., the application of a human-centered approach.
- Small and simple interventions – e.g., building this study in three complementary parts, each addressing features that aim for the same purpose, that is focusing on people’s needs.

The first study focuses on a particular task of NLG models known as Referring Expression Generation (REG), which aims to automatically generate correct noun phrases (e.g., *The mathematician Ada Lovelace*) to refer to entities (e.g., `Ada_Lovelace`) along the automatically generated discourse (e.g., “__ was the first to recognize that the machine had applications beyond pure calculation”). In particular, we noticed that state-of-the-art REG models are unable to generate referring expressions to entities not seen during their training, compromising the fluency of the final text. To solve the issue, we extended a state-of-the-art REG model by providing a more descriptive input representation of the entity to be referred to. Furthermore, expanding NLG models (Krahmer; Theune, 2002; Gatt; Krahmer, 2018a) is a current practice within NLG to increment an algorithm. In this study, we extend a REG model with a novel approach in which the target entity representation is enriched with information retrieved from the Semantic Web, such as the type of the entity (e.g., person, location, organization, etc.), as well as, its gender (e.g., female or male) when the entity is a person. Based on an automatic and human evaluation as well as in an ablation study, we concluded that the enriched entity input representation allowed our REG approach to create more meaningful referring expressions, increasing the comprehension of the final text.

In the second study, we worked with DaMata, a robot journalist who automatically produces daily and monthly reports about the deforestation of the Legal Amazon area. In order to enhance text fluency for users, we experimented with extending referring expressions to geographical entities like Brazilian states, cities, and conservation areas within the Legal Amazon region. This project has been conducted in collaboration with the Graduate Program in Electrical Engineering at the Polytechnic School of the University of São Paulo (USP). After the context-enrichment initial development, we conducted a second investigation with different Large Language Models (LLMs)⁸ in order to address the issue of Numerical Referring Expressions. We present the first outcome of an ongoing project with USP and Dublin City University (DCU) and discuss some improvements to our proposal.

Finally, in the third study, we developed a human-centered Natural Language Generation model within a conversational AI question-answering application for the Diabetes Mellitus domain. To develop the approach, first, we created personas (Nielsen, 2011; Nielsen, 2019), which are target user groups constructed to guide the production of texts that can be understood by real users according to the context of the investigated situation (expert-expert, expert-lay user) and text type (complex and simplified). These created personas were central to the development of a context-enriched NLG model as well as guiding the compilation of a monolingual persona-based corpus (more details in Section 6.1.2) within the conversational AI question-answering approach. This corpus was annotated by a group of professionals (AI specialists and domain experts) within the Empoder@ project⁹, following selected Ribera and Lapedriza (2019) framework guidelines. The Empoder@ project was conducted by a group of researchers at LETRA (Laboratory for Experimentation in Translation)¹⁰ together with the School of Nursing and the Department of Statistics

⁸ Please refer to each study in order to verify the language model in use. For more information about Language Models, address Chapter 15 (Paes A.; Vianna; Rodrigues, 2023) for Brazilian Portuguese and RNNs and LSTMs' Chapter (Jurafsky; Martin, 2023) for English.

⁹ *Projeto Empoder@ – Protótipo conceitual e metodológico para avaliação de intervenções orientadas ao autocuidado em diabetes* (Empower project – Methodological innovation in educational practices aimed at autonomy in healthcare). According to Chaves et al. (2017, p. 2), the group “has been carrying out studies on the translation, adaptation, and validation of healthcare instruments (...) for application in the Brazilian cultural context”. More information about the project: <<http://www.letras.ufmg.br/empodera/>>.

¹⁰ <<http://www.letras.ufmg.br/site/en-GB/laboratories>>

at UFMG for application in the Brazilian cultural context from 2015 to 2022.

After the creation of personas and compilation of a corpus, we compared and classified NLG models that consider multiple domains¹¹, external knowledge features, and HCD, i.e., a human-centric approach. With the lessons learned from the first study, we added semantic and syntactic resources (e.g., personas from the compiled corpus; Semantic Web; Open Data, as ontologies¹², and information from DBPedia¹³) to the selected model, in order to provide more consistent data considering limited or non-existent domain corpora for interpersonal relations (expert and non-expert). Finally, we applied automatic and human metrics side-by-side with human-centered practices to evaluate text production and verify our proposal. The third and last approach presented competitive results with a human-centered context enrichment, giving directions to more adequate relations between humans and NLP – Human-Centered NLP (see more information in Sections 3 and 6). In addition, we showed that an HCD approach benefits language generation from the perspective of “machines” but mainly humans, who are the primary actors within NLP. Overall, the three approaches contribute with new directions to low-resource contexts (languages other than English and sensitive domains) where the need to promote effective communication is essential.

1.1 Thesis Overview

With the permission of the Graduate Program in Linguistics and Applied Linguistics (POSLIN) at the Federal University of Minas Gerais (UFMG), this thesis is submitted in a journal format. Accordingly, Chapters 4 and 5 are in research papers format¹⁴, and Chapter 6 follows a similar structure. Despite the

¹¹ According to [Fernandes \(2006, p. 92\)](#), the term “domain refers to the area of .. language in which the corpus focuses”. As far as domain is concerned, there are two main types of corpora: general and specialized ([Baker, 1995, p. 229](#)). This definition is used interchangeably with the one used by [Koehn \(2020, p. 92\)](#): domain is “defined by a corpus from a specific source, and may differ from other domains in topic, genre, style, level of formality, etc.”.

¹² For the Diabetes domain: <http://biportal.bioontology.org/ontologies/DMTO/?p=classes&conceptid=root>

¹³ <https://wiki.dbpedia.org/>

¹⁴ The layout, references, and appendices of the research paper’s version are adapted accordingly to each chapter.

fact that all three studies' chapters are designed in a standalone format, they are key and complementary elements for the unfolding of the history of this thesis. This document is organized as follows:

1. **Chapter 2** describes the main goals and research questions in this research;
2. **Chapter 3** provides a Background of the main concepts this thesis builds on, such as Natural Language Generation, Referring Expression Generation, Text Simplification, Human Artificial Intelligence, Human-centered Design, and Human-centered NLP. Then it presents how the parts of this study interact with each other;
3. **Chapter 4** describes the first context enrichment of an NLG model – a Referring Expression Generation (REG) model. The chapter presents a systematic review of REG models, the advances in the area, and the implementation of a REG model that is able to generate referring expressions to unseen entities. The main content of this chapter is adapted from [Cunha et al. \(2020\)](#) paper:

Cunha, R.; Ferreira, T. C.; Pagano, A.; Alves, F. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 2261–2272. Available on: <https://aclanthology.org/2020.coling-main.205>.

In order to fulfill the requirements of this thesis, information about how this study relates to the scope of this thesis is provided along with the adapted research paper.

4. **Chapter 5** explains DaMata's project and our second approach to generating context-enriched NLG models. The basis of this chapter is the INLG paper ([Teixeira et al., 2020](#)) that was submitted during DaMata's project and where this thesis' author has developed the context extension to the Referring Expression Generation part of DaMata Reporter. With the permission of the

authors, this chapter is adapted from parts of the [Teixeira et al. \(2020\)](#) paper in order to relate this second study with the objectives of this thesis.

Teixeira, A. L. R.; Campos, J.; **Cunha, R.**; Castro Ferreira, T.; Pagano, A.; Cozman, F. DaMata: A robot-journalist covering the Brazilian Amazon deforestation. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, 2020. p. 103–106. Available on: <https://aclanthology.org/2020.inlg-1.15>.

In addition, this chapter presents additional information about a follow-up study of DaMata regarding numerical referring expressions to the deforestation domain in conjunction with texts from a healthcare domain (COVID-19).

5. **Chapter 6** describes the third study and the last approach to context enrichment of NLG models. The chapter has two complementary sections: a Persona-based corpus and a Human-centered enriched NLG model. This study applies the lessons learned from the previous studies presented in Chapters [4](#) and [5](#) in relation to context enrichment and a more human-centered natural language generation;
6. **Chapter 7** recaps the three studies presented in this thesis, starting with the concluding remarks, the main contributions, future work, and concluding with ethical considerations;
7. **References** and **Appendices** are included in the final sessions of the document.

2 Towards Context-enriched NLG Models

In this thesis, we aim to enrich Natural Language Generation (NLG) models with context in order to improve how texts are presented to end-users. The context enrichment draws into the importance of developing NLG systems aligned to Human-computer Interaction (HCI) principles that provide clear and understandable texts in light of human cognitive expectations, actual human needs, and background knowledge. The study addresses the target user needs as support for text generation with application to a Conversational Artificial Intelligence Question-answering (CQA) system: NLG and conversational AI outputs that are more tuned to interpersonal relations (lay users, experts, and their overlaps¹) in decision-making and informative domains (i.e., governance: deforestation, and healthcare: diabetes mellitus).

To the best of our knowledge, limited research has investigated the effectiveness of enriching NLG models with context, that considers human-centered approaches from the perspective of natural language, and offers a productive improvement in user comprehension and experience through more suitable and relevant texts that are more tuned to interpersonal relations. In summary, this thesis compiles three studies aimed at enhancing NLG models in order to generate texts for successful expert-lay communication. Considering the aforesaid, the following questions are posed.

¹ These overlaps happen when different types of users combine certain characteristics, for example, when lay users are also domain experts.

2.1 Thesis Questions

1. What strategies add external knowledge to NLG models within a specific domain with a central concern of generating understandable texts?
2. How are NLG models designed to provide an adequate production of meanings and a better understanding within critically sensitive domains, such as healthcare and governance?
3. How a human-centered approach can be applied to NLG models to improve users' comprehension and experience?

In the next section, the general and specific objectives outlined for this research are presented.

2.2 Objectives

This research has the general objective of enriching NLG models with context to produce texts for thriving expert-lay communication by considering end-users and their needs from the perspective of natural language in critically sensitive domains, such as governance and healthcare. The following specific objectives are outlined:

1. Enrich hybrid computational approaches to specific domains with application to NLG and conversational AI, by revisiting solutions from pipeline and end-to-end approaches, and adding external resources;
2. Produce texts in informative and decision-making domains, i.e., healthcare (specifically, in the context of diabetes mellitus) and governance (deforestation), by investigating human-centered design and how these could be applied to current NLG solutions;
3. Generate texts that contribute to a relevant comprehension within particular interpersonal relations within the domain: expert to lay users, complex to

simplified texts, by revisiting Referring Expression Generation, Robot Journalism, Text Simplification, and Conversational AI applications guided by a Human-Centered Design approach.

The next chapter presents the theoretical framework behind this thesis, starting with the umbrella term of Natural Language Processing.

3 Background

Natural language processing (NLP) is the field that investigates how human languages are handled and understood by computers (Yee, 2015). Furthermore, NLP is intrinsically related to the area of Artificial Intelligence (AI), maintaining close connections to computational linguistics, machine learning, and speech and language processing (Yee, 2015; Goldberg, 2017; Manning, 2015; Eisenstein, 2018; Jurafsky; Martin, 2023; Caseli; Nunes, 2023). Additionally, NLP is an umbrella term that comprises several modules, such as information extraction, machine translation, and natural language generation. This last one will be addressed in the following section.

3.1 Natural Language Generation

Data-to-text Natural Language Generation (NLG)¹ is the computational process of generating natural language in the format of text, voice, or both from non-linguistic data (Reiter; Dale, 2000; Gatt; Krahmer, 2018b). This process involves different resolutions and selections from matching non-linguistic sources (e.g., raw data, numerical data, ontologies, etc.) to a linguistic form (how to present the information in written, spoken, and/or visual outputs). From Bateman and Zock (2016, p. 1–2) perspective, NLG has a broad application that goes from planning an action to verbalizing it. In addition, NLG is a sub-component of several tasks as can be seen in Figure 2, including, e.g., machine translation (MT), conversational AI (that will be addressed in this study), image captioning, automated writing assistance, creative writing, summarization, and simplification.

According to Eisenstein (2018, p. 457), most data-to-text and end-to-end NLG systems share some of the challenges presented by pipeline architectures (Reiter; Dale, 2000; Ferreira et al., 2018a). These systems need to identify parts of

¹ As noted by Manishina (2016, p. 20), Natural Language Generation is “the area of natural language processing which handles the production of natural language utterances (either written or spoken) from a machine representation form: conceptual, logical, symbolic, etc”.

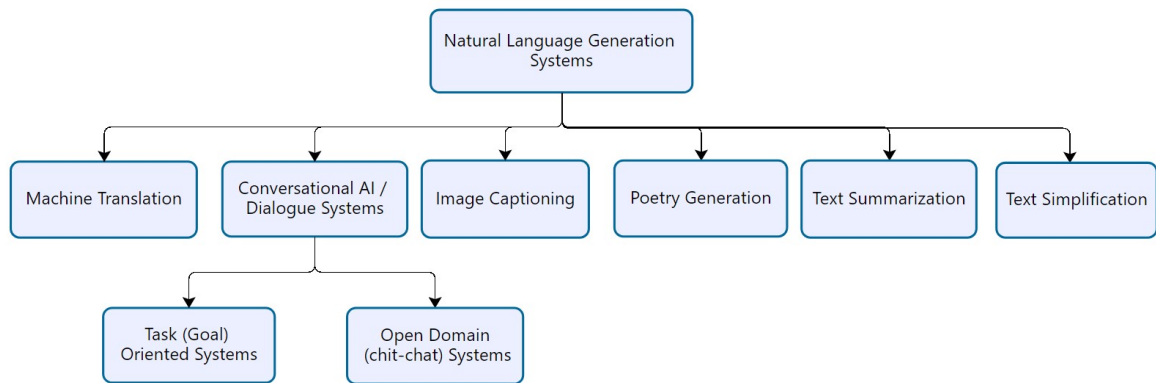


Figure 2 – NLG Applications – adapted from [Santhanam and Shaikh \(2019\)](#).

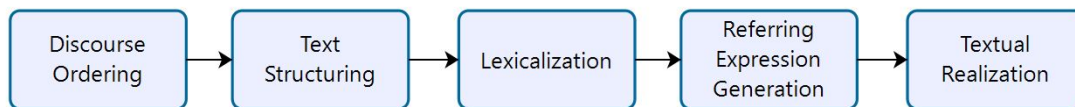


Figure 3 – Pipeline architecture proposed by [Ferreira et al. \(2018a\)](#).

the data they want to describe, how they plan to present the information, and then lexicalize this information into sentences or paragraphs. One example of pipeline architecture is the one depicted in Figure 3 and proposed by [Ferreira et al. \(2018a\)](#). In this architecture, after the content selection, the non-linguistic representation of the content is converted into text in five steps: Discourse Ordering, Text Structuring, Lexicalization, Referring Expression Generation, and Textual Realization. To give a broader perspective of these tasks within our study, we will briefly explain each one of them.

3.1.1 Discourse Ordering

Discourse Ordering concerns the ordering of events in discourse to construe a logical timeline that enhances clarity and reading comprehension (Jurafsky; Martin, 2023). This stage is responsible for deciding which order the source text should be verbalized in the target text. According to Gatt and Krahmer (2018a, p. 12), the ordering preference is implied by the application domain constraints. As an example, within the deforestation domain, the source text is not ordered and constituted by the following information:

```
[The main cause of the deforestation, the total deforestation within the
month, and the deforestation area]
```

After the “discourse ordering” stage, the system generates the ordered target text presented as the following monthly report:

```
MONTHLY_ALERT -> TOTAL_DEFORESTATION -> CAUSE
```

3.1.2 Text Structuring

Text Structuring is responsible for structuring the information in sentences and paragraphs (Gatt; Krahmer, 2018b; Ferreira et al., 2019), bearing in mind how intents should be ordered. Following our deforestation example, the text will be structured:

```
MONTHLY_ALERT
<paragraph>
  <sentence>
    TOTAL_DEFORESTATION
  </sentence>
  <sentence>
    CAUSE
  </sentence>
</paragraph>
```

3.1.3 Lexicalization

Lexicalization occurs after the intents are ordered and structured. The lexicalization module chooses a template that verbalizes the text where phrases and words are selected. For example:

```
<paragraph>
  <sentence>
    In MONTH_1 YEAR_1, AREA_1 of the LOCATION_1 deforest,
    according to data from INSTITUTE_1.
  </sentence>
  <sentence>
    The main cause of deforestation was CAUSE_1, responsible
    for AREA_1 of deforested area.
  </sentence>
</paragraph>
```

3.1.4 Referring Expression Generation

Referring Expression Generation is a traditional micro-planning task ([Krahmer; Deemter, 2019](#)) which aims to automatically generate correct references to entities mentioned as discourse unfolds. Given our deforestation example, the system will generate the following referring expressions:

```
<paragraph>
  <sentence>
    <reference tag='MONTH_1'>June</reference>
    <reference tag='YEAR_1'>2019</reference>
    <reference tag='AREA_1'>914 sq km</reference>
    <reference tag='LOCATION_1'>the Legal Amazon</reference>
    <reference tag='INSTITUTE_1'>the National Institute for
    Space Research (INPE)</reference>
  </sentence>
  <sentence>
    <reference tag='CAUSE_1'>clear-cut deforestation, which leaves
    the soil clear of any vegetation</reference>
    <reference tag='AREA_1'>835 sq km</reference>
  </sentence>
</paragraph>
```

3.1.5 Textual Realization

Textual Realization is the final step in the pipeline architecture, converting machine intermediate representations into text. De-tokenization, contractions, as well as verbal and nominal inflections are solved according to a rule-based strategy. The generated output from our example would be:

```
<paragraph>
  <sentence>
    In June 2019, 914 sq km of the Legal Amazon were deforested,
    according to data from the National Institute for Space Research (INPE).
  </sentence>
  <sentence>
    The main cause of deforestation was clear-cut deforestation,
    which leaves the soil clear of any vegetation, responsible
    for 835 sq km of deforested area.
  </sentence>
</paragraph>
```

3.1.6 Application: Robot Journalist

An application of the data-to-text NLG pipeline is a robot journalist. This kind of system uses non-linguistic data, which are high volumes of structured data streams available nowadays, in order to generate news reports. We can find these systems in both academia and industry as the ones developed by [Theune et al. \(2001\)](#), [Lee, Krahmer and Wubben \(2017\)](#), and [Leppänen et al. \(2017\)](#). In this study, our robot journalist DaMata converts non-linguistic data into text in six steps: Content Selection, Discourse Ordering, Text Structuring, Lexicalization, Referring Expression Generation, and Textual Realization. More information about the system and each stage within the NLG pipeline is available in [Chapter 5](#).

3.2 Text Simplification Systems

Text Simplification (TS) is a commonly used task within NLG ([Erdem et al., 2022](#)). The main idea behind this task is to perform transformations on the input

text without altering the real meaning while providing a more accessible text for readers with different literacy levels, cognitive impairments, and reading and writing learning skills (Aluísio et al., 2008; Specia; Aluísio; Pardo, 2008; Siddharthan, 2014; Saggion, 2017; Alva-Manchego; Scarton; Specia, 2020). In the same vein, this study addresses TS as a human-centered task (see more information in Section 3.4) where end users of a particular domain can access a more readable text according to their literacy.

According to TS research (Shardlow, 2014; Siddharthan, 2014; Saggion, 2017; Alva-Manchego; Scarton; Specia, 2020; Štajner, 2021; Al-Thanyyan; Azmi, 2021), text simplification has distinct strategies. Nonetheless, we introduce the categories related to the level of the simplification task, such as sentence (Alva-Manchego; Scarton; Specia, 2020), paragraph (Devaraj et al., 2021), and document-level (Cripwell; Legrand; Gardent, 2023) simplification. Moreover, TS advocates various approaches to these levels, including lexical, syntactic, monolingual machine translation, and hybrid techniques. These approaches are listed as follows.

3.2.1 Lexical simplification

Lexical simplification (LS) aims to make the text easier to read by replacing complex words with synonyms without changing the original meaning of the sentence (Siddharthan, 2014; Paetzold; Specia, 2017; Saggion, 2017; Al-Thanyyan; Azmi, 2021). According to Paetzold and Specia (2017, p. 550), the LS task usually follows a four stages pipeline, which comprises (i) the complex word identification; (ii) the generation of substitution candidates; (iii) the choice of which candidates fit the context; and (iv) the ranking of the remaining substitutes according to their simplicity. Much research focuses on LS, including the investigated domains in the present study. Moreover, a simplification example from our persona-based corpus is presented in Table 1 where the difference between expert and lay readers is highlighted.

Regarding the developments in LS, in 2022 a shared task was conducted

Q: Good morning people, I would like to know if anyone in the group has already drank lemon with turmeric in the morning on an empty stomach to reduce diabetes. Tell me if it was a good experience.

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.*

A (Expert): Lemon with turmeric itself may not even cause any harm. However, **they do not have benefits** to reduce blood glucose* (amount of **carbohydrates** in the blood).

R (Expert): 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 (Lay): Consumption of lemon with turmeric does not cause any harm, but **neither does it bring any benefit in terms of** reducing blood glucose (amount of **sugar** 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).

Table 1 – Sample from our corpus with a post-edited question answered to expert and lay users in the diabetes mellitus domain. Some of the simplified sentences are **bolded** (The English translation and its original form in Brazilian Portuguese).

during the TSAR-2022 workshop² at EMNLP³, the results presented new resources as multilingual benchmarks and discussions about the development of LS (Štajner et al., 2022; Saggion et al., 2022), especially concerning Portuguese with the publication of North, Zampieri and Ranasinghe (2022). For more information about LS research, address the surveys from Shardlow (2014), Paetzold and Specia (2017), Al-Thanyyan and Azmi (2021), North et al. (2023).

3.2.2 Machine Translation (MT)

The Machine Translation (MT) area lends advances within MT regarding the use of LLMs and other technologies to the simplification task. In addition, the complex text is considered the source text of a monolingual translation task, where the target text is the simplified version (Shardlow, 2014; Siddharthan, 2014; Alva-Manchego; Scarton; Specia, 2020; Al-Thanyyan; Azmi, 2021). In this study, we

² Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022) – <<https://taln.upf.edu/pages/tsar2022-ws/>>

³ <<https://2022.emnlp.org/>>

investigate some of the translation models' advancements in order to find a better solution to the simplification task within different domains as will be presented in Section 6.2.

3.2.3 Hybrid approach

The hybrid approach aims to address the gap in the rule-based lexical simplification approach in relation to its performance (Al-Thanyyan; Azmi, 2021). Accordingly, Saggion (2017, p. 2) states that lexical and syntactical simplification can be explained individually, however, they are naturally related.

In addition, the period from 2012 to 2023 has been a breakthrough with the advancements in Large Language Models (LLMs), more specifically with the Deep Learning approaches. Moreover, North et al. (2023) investigate these advances in relation to LS, but TS, in general, has been changing rapidly with the developments in the area, such as Transformers, GPT (Brown et al., 2020; OpenAI, 2022; OpenAI, 2023), and most recent LLMs. For more information about the developments in the text simplification area, please refer to the advances presented at the 2022 TS' shared task (Saggion et al., 2022) and the discussion about LS with Deep Learning approaches (North et al., 2023).

3.3 Conversational AI / Dialogue Systems

Conversational AI or Dialogue Systems are NLP systems that communicate with users using textual, speech, or both, utterances. We can classify conversational AI within two categories: open-domain (chit-chat) and task-oriented (goal-oriented) (Eisenstein, 2018). Open-domain systems, like chatbots, are more related to extended conversational goals or chit-chat. According to Jurafsky and Martin (2023, p. 297), these kinds of systems “mimic the unstructured conversations or ‘chats’ characteristic of human-human interaction”. The main objective of open-domain systems is entertainment, engaging users or just chatting with them, such as Cleverbot⁴ (Carpenter, 1997) or Microsoft XiaoIce system (Zhang et al., 2018a), an advance from first rule-based systems like ELIZA (Weizenbaum, 1966) and

⁴ <<http://www.cleverbot.com>>

PARRY (Colby; Weber; Hilf, 1971). Despite the technological advances in current conversational agents, Eisenstein (2018, p. 467) points out the “significant gap between research and practice”, the author emphasizes that plenty of practical systems are still not flexible, whereas research ones are more driven to “on-the-fly” changes and more tuned to the user’s intents.

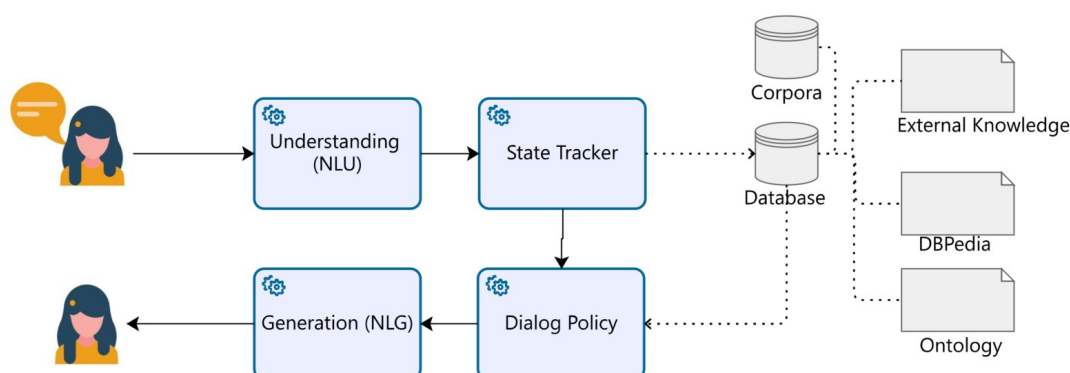


Figure 4 – Conversational AI – Task-based flow – adapted from Gao, Galley and Li (2019, p. 132).

The task-oriented conversational AI’s main goal is to help a user accomplish a specific task (see Figure 4 for more details). For instance, these systems are commonly used to book flights, add calendar tasks, make calls, and recently are used for social good, answering questions about a specific domain, as one of the goals we are trying to accomplish with our proposal. Other examples of these task-based systems are the ones included in digital assistants (Amazon Alexa, Apple Siri, Google Home, and Microsoft Cortana, etc.), working as conversational agents in different domains (Eisenstein, 2018; Jurafsky; Martin, 2023). In 2022, during the

semifinals of the Alexa Prize TaskBot Challenge 2⁵, eight teams presented their solutions to the new version of the task-based conversational AI system, i.e., a multimodal conversational AI to help users with multiple steps tasks.

A common task-based flow is exemplified in Figure 4 ((Gao; Galley; Li, 2019, p. 132)), where a human starts interacting with a conversational agent in a written or speech format. The user's input is received by the conversational agent's first module, i.e., Natural Language Understanding (NLU), which is responsible for interpreting users' input, identifying their intent, and adjusting the information to the next step. The second module is the state tracker, which is responsible for storing the conversation content and its stages. In addition, this component can search within the knowledge base (i.e., corpora, database, and external knowledge – DBPedia and/or ontologies) for any extra information that can facilitate the next step of the dialogue. Then, the third module verifies which action suits the current stage of the conversation (i.e., validating the information so far within the knowledge base) and serves as input to the fourth and last module – the generation one (NLG) – where a response is generated, giving the information to the human user and ending the dialogue flow.

As highlighted by Jurafsky and Martin (2023, p. 326), conversational AI is the area that has more connections with human-computer interaction within NLP. This characteristic is related to the essential role that the user plays in the development of these kinds of systems. The authors (2023) suggest when building conversational AI systems (i) study the user and the task by understanding real and possible users; (ii) build simulations and prototypes, and finally (iii) perform iterative tests in order to check the design on users⁶.

3.3.1 Chatbots

As discussed before, chatbots are usually open-domain systems. They respond according to the user's input, generating a relevant answer in order to engage the user and have a conversation (Eisenstein, 2018; Jurafsky; Martin, 2023). Most

⁵ <<https://www.amazon.science/alexa-prize/taskbot-challenge/2022>>

⁶ For more information about Human-centered Design (HCD) within conversational AI systems, please refer to Pearl (2017), Nielsen (2019), Salminen et al. (2020), and Salminen et al. (2021).

recent studies have been combining characteristics of task-driven in open-domain applications as an attempt to provide a goal-oriented chatbot with “*human likeness*” skills (Vinyals; Le, 2015; Serban et al., 2018; Ram et al., 2018; Zhang et al., 2018a). In general, chatbots are applications that are more concerned about customer experience. This singular characteristic has gained visibility since the Alexa Prize challenge (Ram et al., 2017; Ram et al., 2018), where systems had to be able to promote engaging and coherent conversations. Within the challenge, several participants presented approaches related to better customer experience as the studies of Fang et al. (2017) and Curry et al. (2018). However, they frequently opted for applying a more NLP-driven perspective (i.e., addressing first the NLP task), such as sub-dialogues⁷ (Pichl et al., 2018), intent classification⁸ (Finch et al., 2020) –, instead of combining HCD methodology with NLP (Ram et al., 2017; Ram et al., 2018; Fang et al., 2017; Curry et al., 2018; Gonzalez, 2021; Paranjape, 2022; Majumder, 2023).

3.3.2 Question Answering Systems

One of the components of conversational AI systems is question-answering (QA) tasks/sub-modules. These systems appeared in the 1960s with the intent to answer questions about baseball statistics or scientific facts (Jurafsky; Martin, 2023). QA systems are used in a variety of applications, from tourism to bio-medicine domains (Dimitrakis; Sgontzos; Tzitzikas, 2019), and they vary in their implementation as information retrieval-based⁹ to knowledge-based systems (Eisenstein, 2018). The main objective of QA systems is to answer questions about a given matter. For those systems implemented within an information retrieval architecture, they parse the query and select or generate an answer correspondingly. In addition, knowledge-based QA systems often extract information from large

⁷ Pichl et al. (2018) applied the concept of sub-dialogues, where the conversation is split into smaller dialogues within a Dialog Manager component to decide which sub-dialogue is the most appropriate to answer.

⁸ Jurafsky and Martin (2023, p. 15) refer to “intent classification” as the dialogue task of determining the user’s intent by analyzing the information and inferring “what general task or goal is the user trying to accomplish”.

⁹ According to Jurafsky and Martin (2023, p. 102), information retrieval (IR) “is the task of finding the document d from the D documents in some collection that best matches a query q ”.

Q: What causes nocturnal hunger syndrome?

A (Expert): The nocturnal hunger syndrome is characterized by a lack of appetite in the morning, insomnia, and excessive hunger at night. The person feels hungry even though he has already eaten at night. There is no exact cause of the appearance of this syndrome. Nevertheless, some researchers point to the occurrence of an imbalance of the hormones that regulate the body's night and day cycles, with possible genetic influence. Many studies suggest that the syndrome is associated with periods of stress.

A (Lay): The exact cause of this problem is not yet known. Some hypotheses are an imbalance of hormones that regulate the body's night and day cycles, stress, and genetic influence.

P: *O que causa a síndrome da fome noturna?*

R (Experto): *A síndrome da fome noturna, é caracterizada pela falta de apetite pela manhã, insônia e excesso de fome à noite. A pessoa sente fome mesmo que já tenha se alimentado à noite. Não se sabe exatamente o que causa o aparecimento dessa síndrome, mas hipóteses apontam para a ocorrência de um desequilíbrio dos hormônios que regulam os ciclos noturno e diurno do organismo, com possível influência genética. Também se sabe que ela está associada à períodos de estresse.*

R (Leigo): *Ainda não se sabe a causa exata desse problema. Algumas hipóteses são: desequilíbrio dos hormônios que regulam os ciclos noturno e diurno do organismo, estresse e influência genética.*

Table 2 – Sample from our corpus of how to answer a question to expert and lay users in the diabetes mellitus domain (The English translation and its original form in Brazilian Portuguese).

corpora (Eisenstein, 2018; Jurafsky; Martin, 2023; Dimitrakis; Sgontzos; Tzitzikas, 2019). Finally, these QA systems are usually part of larger systems and try to help with common frequently asked questions to more domain-driven issues as the example presented in Table 2.

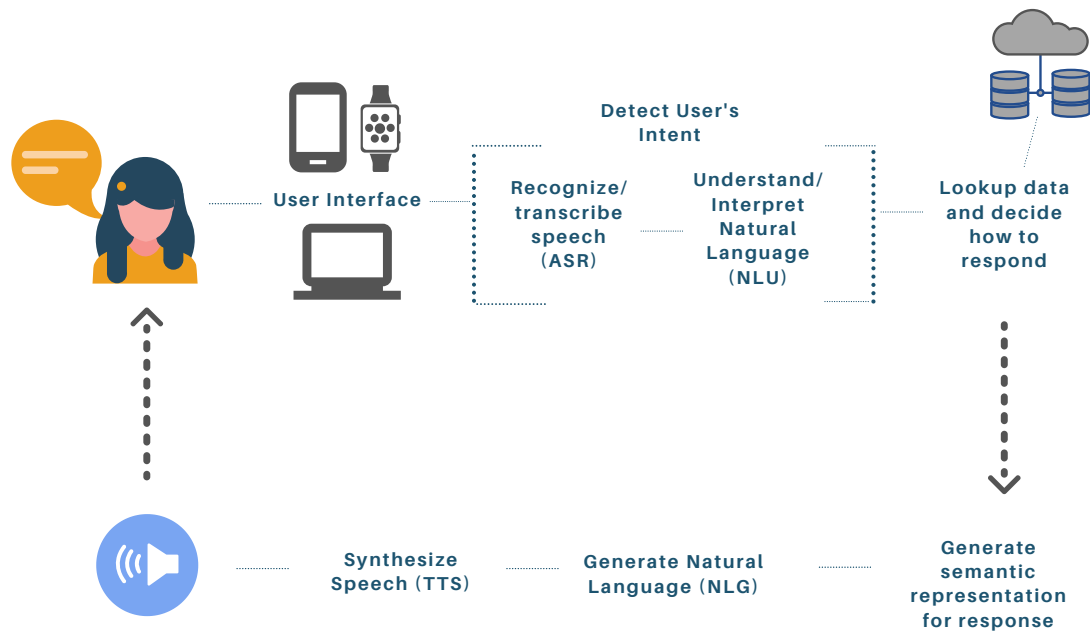


Figure 5 – Conversational AI architecture – adapted from [Horn et al. \(2019, p. 3\)](#).

In this study, we developed a context-enriched NLG model and extended it to a QA context, specifically, an automatic conversational AI question generation module according to the architecture presented in Figure 5, which takes into account the user's input and detects user's intent in order to answer accordingly. The decision-making process and the generation of the semantic representation consider the domain expertise and user literacy levels. The conversational QA part of the model is used in the diabetes mellitus domain and is responsible for answering questions about managing the diabetes mellitus condition, e.g., questions related to treatments, risk factors, and symptoms as presented in Section 6.

In the next section, we present overlaps between NLG and human-centered approaches, crucial perspectives to producing more satisfactory chatbots as well as providing fluent texts to the users.

3.4 Intersections of NLG and Human-Centered Approaches

Recent studies (Läubli et al., 2020; Paranjape, 2022; Majumder, 2023) have provided evidence that suggests the growing interest in combining methods from Human-Computer Interaction and User-Centered techniques in relation to NLG. In this section, we describe some of these methodologies and how they can be used in the development process of context-enriched NLG models.

3.4.1 Human-Computer Interaction

Human-computer Interaction (HCI) is “an area of research and practice that emerged in the early 1980s, initially as a specialty area in computer science embracing cognitive science and human factors engineering” (Carroll, 2009). With technological development, a lot has changed in the HCI research focus, which evolved, according to Lazar, Feng and Hochheiser (2017, p. 4), from “task performance in statistical software to increase the collaboration, connections, emotion, and communication”. In 2023, HCI’s primary attention is about engaging people to use an interface in different environments, that is, being aware of the groups of users in terms of age, domain, educational, and technical experience (Lazar; Feng; Hochheiser, 2017). Additionally, Stephanidis et al. (2019) highlight that HCI’s goal converges with the trend of a more human-centered AI, which focuses on humans and their needs as well as ensuring that the technology serves its purpose.

HCI and its tendency to be user-centered since the dissemination of the term by Norman and Draper (1986), has led to new challenges in the HCI field with a mixture of providing a more trustful relationship between humans and technology, and a shift in users’ behavior, that is, users becoming more critical and demanding regarding the use of technology (Stephanidis et al., 2019). Besides, Stephanidis et al. (2019, p. 1232) point out seven challenges within the HCI area that can be seen in Figure 6, which are:

- human-technology symbiosis (how humans will live harmoniously together with technology);



Figure 6 – Seven HCI challenges according to [Stephanidis et al. \(2019, p. 1231\)](#).

- human-environment interactions (interactions with entire technological ecosystems);
- ethics, privacy, and security (moral principles that govern behavior);
- well-being, health, and eudaimonia (physical, mental, and social well-being);
- accessibility and universal access (design for people with disabilities);
- learning and creativity (gaining knowledge and producing original ideas);
- social organization and democracy (stable structure of relations inside a group).

The presented perspective is designed in order to achieve a human-centered approach, a term that is interchangeable with user-centered and people-centered

approach where the human, more particularly, the user, is at the center of the interaction development and design process, in an attempt to provide support to human’s well-being (Stephanidis, 2015; Stephanidis et al., 2019; Garibay et al., 2023). HCI theoretical, as well as, empirical studies have developed a methodology aiming at profiling user models to promote better user experience.

In the same vein, the practice of finer-tuning resources to user needs and their preferences is changing conventional modalities of human-centered AI, as will be discussed in the next section.

3.4.2 Human-Centered Artificial Intelligence

Human-centered artificial intelligence (HAI) is a perspective on Artificial Intelligence (AI) and Machine Learning (ML)¹⁰ that “algorithms must be designed with an awareness that they are part of a larger system consisting of humans” (Riedl, 2019, p. 33). Due to the increasing concern about AI’s impact on human lives and how this can unpredictably affect humans, HAI research institutes started to be created by renowned universities like Stanford University, UC Berkeley, and MIT, to highlight humanistic and ethical frontiers with AI (Xu, 2019, p. 43). Another objective of research within HAI is to decrease the black-box phenomenon¹¹ of ML, promoting more fairness, transparency, and interpretability to users by providing means of explaining the decisions behind these intelligent systems, as well as, designing these systems with social responsibility in mind (Riedl, 2019; Xu, 2019). Holley and Becker (2021) emphasize that the principle of HAI research “is that AI should enhance human lives and thoughts, not replace them”.

Within HAI, an explainable-driven area has gained visibility in order to support high-consequence human decisions: the eXplainable AI (XAI), particularly in decision-making fields like healthcare, law, and governance. The primary focus of XAI is “ensuring that all stakeholders who must use AI directly or indirectly understand how the AI works” (Holley; Becker, 2021). Furthermore, research like the

¹⁰ According to Riedl (2019, p. 33–34), Machine Learning (ML) is “a particular approach to the design of intelligent system in which the system adapts its behavior based on data”.

¹¹ Riedl (2019, p. 34) highlights that this black-box phenomenon “refers to the situations wherein the user cannot even know what algorithms it uses, or that the system is so complicated as to defy easy inspection”.

ones conducted by the Explainable AI DARPA program (Gunning, 2017; Gunning; Aha, 2019)¹² are contributing to the development of new strategies to facilitate and support the correct use of technology among non-expert users. To elucidate the *explanation* term in the context of Human-Centered AI, Arrieta et al. (2020, p. 85) clarifies that “[g]iven a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand”.

According to Xu (2019, p. 44), AI’s new trend is emerging as a human-centered approach with technological innovation. Additionally, the focus on human-centered design and user-centered design approaches is crucial to provide technology that will be able to support users’ needs by being more aligned to user requirements (Stephanidis et al., 2019). Finally, the necessary technology and user needs to build efficient intelligent systems are well-aligned to the development of natural language and speech processing systems, where the need for understanding humans is a fundamental feature to perform accurately (Jurafsky; Martin, 2023; Riedl, 2019; Garibay et al., 2023).

In the context of our research, the view of HAI will be applied in different stages of our study in order to align the development of context-enriched NLG models to humans. In addition, our aim is to provide our end-users with relevant information within the environment they are living in and consider their needs. We will discuss more about this human-centered perspective within each approach we are proposing.

3.4.3 Human/User-Centered Design

Human-centered Design (HCD), specifically, User-centered Design (UCD), is the approach, to developing and designing interactive systems, “that aims to make systems usable and useful by focusing on the users, their needs and requirements, and by applying human factors/ergonomics, and usability knowledge and techniques” (ISO, 2019, p. 2). The UCD approach, presented in Figure 7, emerged from HCI and

¹² According to Gunning and Aha (2019, p. 45), the goal of XAI was “to create a suite of new or modified ML techniques that produce explainable models that, when combined with effective explanation techniques, enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems”.

involves designing software that will improve human well-being, user experience, accessibility, and sustainability (Cooper et al., 2014; Stephanidis, 2015; Sharp; Preece; Rogers, 2019; ISO, 2019).

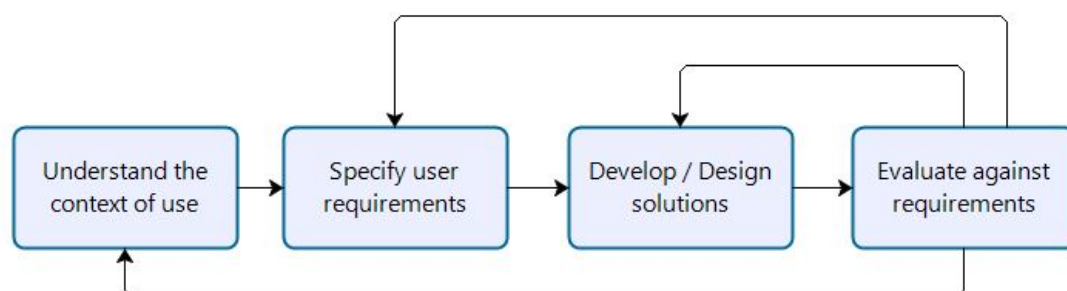


Figure 7 – User-Centered Design – adapted from Soegaard and Dam (2015).

In addition, UCD is an iterative process as can be seen in Figure 7 that focuses on users and their needs by understanding the context of use (e.g., diabetes mellitus domain); specifying user requirements (e.g., different proficiency levels); developing and/or designing solutions (e.g., building an NLG model with a question-answering application); and finally, evaluating against requirements (e.g., applying satisfaction questionnaires). This process is about making knowledge of the target users through different research methods and techniques as will be shown in the following sections.

To design or develop an NLG system with a UCD 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 a UCD method, specifically, to user models: personas. According to Cooper et al. (2014), personas are “user models that are represented as specific, individual human beings”. Additionally, personas provide us with starting point information about how groups of users behave, how they think, what they want to accomplish, and why (Cooper et al., 2014; Nielsen, 2019). Furthermore, personas are fictitious users, based on data and real-world observation, that are representations of real users according to the investigated situation and context (Nielsen, 2011; Nielsen, 2019).

One of the primary benefits of the use of personas during the design or

development of an NLG system is that it helps us communicate about different types of users and their needs (Cooper et al., 2014). Each persona represents a group of users and is designed with specific characteristics, encapsulating a distinct set of behavior and usage patterns regarding the use of a particular software (Cooper et al., 2014; Nielsen, 2019). Moreover, Nielsen (2019, p. 5) adds that the persona method “tries to break with the automated perception and instead create empathetic descriptions of the user”. According to Harley (2015, p. 1), “if users are described [...] as broad profiles, that information will simply not lodge itself as deeply in team members’ brains as a distinct persona will”. The persona model helps (i) determine how the NLG system should behave, (ii) communicate with team members, (iii) help the development team to identify with target users, (iv) provide an agreement of opinions and commitment, (v) measure how this will work with real people, and (vi) contribute to other aspects involving the design or development of an NLG system (Cooper et al., 2014).

3.4.3.1 Persona versus User Profiles

On the one hand, a user profile is regularly based on a stereotype and primarily can include a role, a job function, or a group membership (ISO, 2008; LeRouge et al., 2013; Cooper et al., 2014). On the other hand, personas generally have information on behaviors, motivations, background, expectations, and (experience, life, end) goals. Additionally, personas are “narrative tools to help better communicate the real underlying data” and also “represent a broad cross-section of users” (Cooper et al., 2014). We hypothesize using data-based personas within a domain, we “highlight specific details and important features of the group” (Harley, 2015) and avoid common stereotypes as the ones created for female conversational AI systems (Curry; Rieser, 2020), such as Microsoft’s XiaoIce persona¹³ (Zhou et al., 2020).

¹³ According to Zhou et al. (2020), they “design the XiaoIce persona as an 18-year-old girl who is always reliable, sympathetic, affectionate, and has a wonderful sense of humor”.

3.4.3.2 Persona versus Implied Reader

According to Suojanen, Koskinen and Tuominen (2015b, p. 63), implied readers are “significant as a tool for analyzing the source text and planning the target text”. The authors (2015b) complement that “[t]he addressee (or target audience) of any text or translation is not a real person but a concept” (Suojanen; Koskinen; Tuominen, 2015b, p. 63). Accordingly, an “implied reader is a textual tool, a way to analyze how readers and their characteristics are visible in texts” (Suojanen; Koskinen; Tuominen, 2015a, p. 62–65), whereas personas are “fictive archetypes of users based on empirical information on real users” (Suojanen; Koskinen; Tuominen, 2015b, p. 70).

3.4.3.3 Persona versus Personality

In order to promote a better understanding of the target users, some studies (Schwartz et al., 2013; Oraby, 2019; Fadhil; Schiavo, 2019) add a “personality dimension”, that is, a psychological model of personalities to characterize their personas and help bring them to life (Nielsen, 2019). Even though there are some intersections between HCI and the psychology field, we can create personas with or without the use of personality characteristics. Besides, several NLP/NLG models (Woods et al., 2017; Oraby et al., 2018; Oraby, 2019; Harrison et al., 2019; Ramos; Monteiro; Paraboni, 2020) implement the use of personality in their architectures with an attempt to meet user needs and expectations.

4 Enriching an NLG Model from a Referring Expression Generation Approach

The content of this chapter is adapted from Cunha et al. (2020) following paper:

Cunha, R.; Ferreira, T. C.; Pagano, A.; Alves, F. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 2261–2272. Available on: <<https://aclanthology.org/2020.coling-main.205>>.

Summary of the First Approach

In this chapter¹, the first approach of the context-enrichment of NLG models is presented with an application to the Referring Expression Generation. This chapter was adapted from the accepted paper at COLING 2020 (Cunha et al., 2020) in order to meet the requirements of this thesis. The context enrichment was conducted by proposing two extensions to NeuralREG, a state-of-the-art encoder-decoder REG model. The first is a copy mechanism² (i.e., a component of the REG model responsible for deciding if the next referring expression should be generated from the output vocabulary or copied from the input representation of the target entity), whereas the second consists of representing the gender and type of the referent as inputs to the model. Drawing on the results of the automatic and human evaluation as well as an ablation study using the WebNLG corpus, we contend that our proposal contributes to the generation of more meaningful referring expressions to unseen entities than the original system and related work. The context-enrichment of NeuralREG with the two proposed extensions is an essential component of the development of this thesis where the outcome was replicated in the two following context-enrichment approaches. Besides, this thesis' author was responsible for almost the full development, except for the annotation and human evaluation phases. The additional contributions followed conducting an ablation study and automatic evaluations, and implementing the necessary code to this first context-enrichment approach. This thesis's author was responsible for writing, submitting, and reviewing the accepted COLING 2020 paper. Due to its paper format, this chapter can be read independently. However, the REG model's context-enrichment is considered the first achievement of this thesis regarding the production of texts more driven to interpersonal relations. Code and all produced data are publicly available^{3 4}.

¹ This chapter was adapted from the accepted paper at COLING 2020 (Cunha et al., 2020). The author of this thesis is the first author of the paper.

² Please refer to the description of the copy mechanism in Section 4.1.

³ <<https://github.com/rossanacunha/NeuralREG>>

⁴ <<https://github.com/ThiagoCF05/NeuralREG/tree/improvements>>

4.1 Introduction

Data-to-text Natural Language Generation (NLG) is the computational process of generating natural language in the form of text or voice from non-linguistic data. A traditional micro-planning task within the pipeline data-to-text architecture is referring expression generation (REG) (Krahmer; Deemter, 2019), which aims to automatically generate relevant noun phrases (e.g., *The mathematician Ada Lovelace*) to refer to entities (e.g., `Ada_Lovelace`) mentioned as discourse unfolds (e.g., “__ was the first to recognize that the machine had applications beyond pure calculation.”).

Traditionally, REG systems produce references to discourse entities in two explicit steps. First, they decide on the referential form, i.e., choosing whether a referring expression should be a pronoun (*She*), a proper name (*Ada Lovelace*), a description (*The mathematician*), etc. Once the choice is made, such systems textually realize the referring expression based on the chosen referential form and discourse context. If the first step selects a proper name as the form to refer to `Ada_Lovelace` for instance, the ensuing step is responsible for deciding, among *Ada*, *Ada Lovelace*, or another text realization, i.e., the one that is the most relevant referring expression to that entity in a given discourse context (please refer to this example in Figure 8).

With the advent of large amounts of data, REG systems have undergone a significant change in their architecture. From being rule-based modular, they have become data-driven end-to-end systems that aim to perform the choice of referential form and surface realization jointly. An example of these more integrated approaches is NeuralREG (Ferreira et al., 2018a), an end-to-end neural REG model that produces referring expressions deciding on form and content jointly based on representations of the referent and its surrounding context.

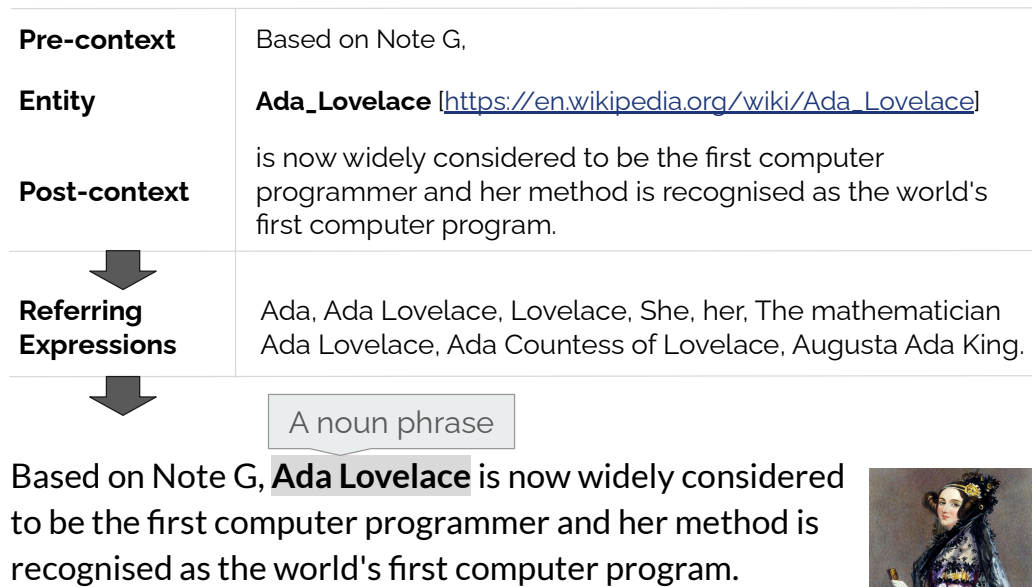


Figure 8 – Referring Expression Generation of the entity `Ada_Lovelace`: a noun-phrase.

Although NeuralREG is able to generate adequate referring expressions to discourse entities seen during the training phase, the model does not generalize to unseen ones, i.e., it can not generate referring expressions to entities that were not seen during its training. This study aims to fill this gap by proposing two extensions to the model's original architecture. The first is a copy mechanism, which may decide at each decoding timestep whether the next token of the referring expression should be generated from the output vocabulary or copied from the input representation of the target entity. We thereby hypothesize that the model will be able to generate a token from the vocabulary for seen entities and to copy tokens from the input representation in the case of unseen ones. The second extension consists of representing the gender and type of the entity as input to the model. Such information can be easily extracted from the Semantic Web and may help the model to generate pronominal (e.g., *She*) and descriptive (e.g., *The country*)

referring expressions to unseen entities.

To evaluate our approach, we conducted experiments relying on a delexicalized version (Ferreira et al., 2018b) of the WebNLG corpus (Gardent et al., 2017b). We first compare our proposal with the original NeuralREG and other related approaches as ProfileREG (Cao; Cheung, 2019). Second, to assess the quality of the texts generated by our model, we conducted a supplementary evaluation with human judges. Next, we follow the rationale of ablation studies to analyze the importance of each feature in our model within the process of referring expression generation. Finally, we discuss some advantages of the introduced features and how they interact to improve accuracy, variety, and generalization.

4.2 Related work

Given an entity to be referred to in a particular context, traditional REG methods have addressed this task in two steps. The first one concerns the choice of referential form, i.e., deciding whether the target reference is more likely to be a proper name (*Belo Horizonte*), a description (*The city*), a pronoun (*It*), or another referential form. Regarding this step, Reiter and Dale (2000) suggested always choosing a full proper name as the first reference to a particular entity in a given context, whereas pronouns may be used for its subsequent references if there is no other entity with the same person, gender and number in-between the target reference and its antecedents. More recently, Ferreira, Krahmer and Wubben (2016) proposed a naive Bayes method, which is able to non-deterministically choose a referential form to a particular reference. The model’s choice is conditioned upon discourse features that the psycholinguistic literature has shown to impact this choice, such as grammatical position, givenness, and recency of the target reference.

Once the referential form is chosen, the second step of traditional REG models focuses on the surface realization of the reference. Most part of the literature on this step focuses on the generation of descriptions (Dale; Reiter, 1995) although some studies have approached the generation of proper names (Siddharthan; Nenkova; McKeown, 2011; Deemter, 2016; Ferreira; Krahmer; Wubben, 2017).

In contrast to previous proposals that have focused on selecting referential

form or referential content, [Ferreira et al. \(2018a\)](#) proposed an end-to-end approach: NeuralREG, a referring expression generator able to perform the choice of referential form and the surface realization in an end-to-end style using a neural encoder-decoder architecture. Given an entity to be referred to in a particular textual context, the approach first encodes the entity identifier and the text prior (pre-context) and subsequent to the reference (post-context) to later decode this representation into a correct referring expression using attention ([Bahdanau; Cho; Bengio, 2015](#)).

Although NeuralREG ([Ferreira et al., 2018a](#)) can generate suitable referring expressions to entities seen during training, it presents certain problems when referring to unseen ones. To overcome this limitation, [Cao and Cheung \(2019\)](#) presented a profile-based model. Their solution uses information from both profile⁵ (i.e., information retrieved from the entity’s Wikipedia page) and context (pre- and post-contexts jointly) to generate suitable references to unseen entities. The authors conclude that their approach is more successful in determining the most suitable referring expression to a particular entity.

In contrast to [Cao and Cheung \(2019\)](#) solution, in order to address the limitations of dealing with unseen relations and entities, our proposal uses a combination of a copy mechanism together with representations of gender and type of referent as input to the model. We expect both extensions to make the model able to produce suitable referring expressions in particular to unseen entities. We describe our model in more detail in the next sections, based on the data used to investigate and evaluate our approach.

4.2.1 Data

We evaluated our proposal based on an enriched version ([Ferreira et al., 2018b](#)) of the WebNLG corpus ([Gardent et al., 2017a](#)). The original resource is a parallel corpus with sets of RDF⁶ (Resource Description Framework) triples⁷ and their corresponding verbalizations. Each RDF triple set consists of subject-

⁵ https://en.wikipedia.org/wiki/Ada_Lovelace

⁶ RDF (Resource Description Framework) is a W3C “standard designed as a data model for metadata” ([W3C, 2014](#)). Available at <https://www.w3.org/RDF/>.

⁷ According to [W3C \(2014\)](#), triple is “the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link”.

predicate-object (e.g., `Adenan_Satem | birthPlace | Japanese_occupation_of_British_Borneo`), which is illustrated in Table 3 and can be verbalized in different forms. Each subject and object is a *Uniform Resource Identifier* (URI), which can be represented by a Wikipedia ID (e.g., `Adenan_Satem`, `Abdul_Taib_Mahmud`) or a literal value like a date, number, or constant (e.g., “*Barisan Raayat Jati Sarawak*”), followed by a predicate (e.g., `birthPlace`) which is a relation between these entities (Gardent et al., 2017a; Gardent et al., 2017b).

RDF Triples		
Subject	Predicate	Object
<code>Adenan_Satem</code>	<code>birthPlace</code>	<code>Japanese_occupation_of_British_Borneo</code>
<code>Abdul_Taib_Mahmud</code>	<code>successor</code>	<code>Adenan_Satem</code>
<code>Abdul_Taib_Mahmud</code>	<code>residence</code>	<code>Sarawak</code>
<code>Abdul_Taib_Mahmud</code>	<code>party</code>	<code>“Barisan Raayat Jati Sarawak”</code>

↓

Corresponding Text

Adenan Satem was born in **Japanese Occupied British Borneo**. His successor was Abdul Taib Mahmud, who, resides in Sarawak and is a member of the “*Barisan Raáyat Jati Sarawak*” party.

Tag	Entity/Constant	Referring Expression
BRIDGE-1	<code>Adenan_Satem</code>	Adenan Satem
PATIENT-1	<code>Japanese_occupation_of_British_Borneo</code>	Japanese Occupied British Borneo
BRIDGE-1	<code>Adenan_Satem</code>	His
AGENT-1	<code>Abdul_Taib_Mahmud</code>	Abdul Taib Mahmud
PATIENT-2	<code>Sarawak</code>	Sarawak
PATIENT-3	<code>“Barisan Raayat Jati Sarawak”</code>	Barisan Ra’ayat Jati Sarawak

Template

BRIDGE-1 was born in PATIENT-1. BRIDGE-1 successor was AGENT-1, who, resides in PATIENT-2 and is a member of the PATIENT-3 party.

Table 3 – A set of RDF (Resource Description Framework) triples and their corresponding text. Followed by entities’ mapping and a delexicalized template.

The WebNLG dataset is an NLG benchmark that differs from other datasets (Novikova; Dušek; Rieser, 2017; Mille et al., 2018) due to its data diversity in terms of attributes, patterns, and shapes (i.e., RDF tree shapes from DBPedia). The corpus contains 25,298 English texts verbalizing sets of 1 to 7 RDF triples in 15 different domains. The dataset has five domains exclusive to the test set, providing

adequate means to evaluate our model’s performance regarding the generation of referring expressions to unseen entities.

We used an enriched version of the WebNLG corpus obtained by a delexicalization process (i.e., mapping each entity to a generic tag and later replacing their corresponding referring expressions in discourse with these tags) which was created by [Ferreira et al. \(2018b\)](#). Table 3 shows an example of a set of 4 triples and corresponding text, together with the intermediate representations obtained in the delexicalization process, such as general tags, Wikipedia IDs (entity/constant), referring expressions, and the delexicalized template.

To train and evaluate our approach, we have a pre-processing stage where we extract a collection of referring expression entries from the enriched version of WebNLG ([Ferreira et al., 2018b](#)). This stage is performed once, where we map the WebNLG corpus information as the basis to obtain external information without adding new features or changing the WebNLG structure. We avoid a possible impact on the evaluation results since all entities are available on DBPedia. Each entry consists of a Wikipedia ID, i.e., a target entity (`Adenan_Satem`), a truecased tokenized referring expression (`Adenan Satem` or `His`), and lowercased tokenized pre- (`Adenan_Satem was born in`) and post-contexts (`Adenan_Satem successor was Abdul_Taib_Mahmud, who, resides in Sarawak and is a member of the Barisan raayat jati sarawak party.`), indicating the surrounding context of the target reference.

4.2.2 Model

Our approach was based on NeuralREG ([Ferreira et al., 2018a](#)) and aims to generate a referring expression $y = \{y_1, y_2, \dots, y_N\}$ with N tokens to refer to a target entity, given the textual context prior to the reference

$$X^{(pre)} = \{x_1^{(pre)}, x_2^{(pre)}, \dots, x_M^{(pre)}\} \quad (4.1)$$

with M tokens (e.g., pre-context) and subsequent to the reference

$$X^{(post)} = \{x_1^{(post)}, x_2^{(post)}, \dots, x_L^{(post)}\} \quad (4.2)$$

with L tokens (e.g., post-context). Unlike [Ferreira et al. \(2018a\)](#), where the target entity is represented by a single token, our approach describes the referent by an identifier

$$X^{(wiki)} = \{x_1^{(wiki)}, x_2^{(wiki)}, \dots, x_T^{(wiki)}\} \quad (4.3)$$

with T tokens and its entity type E and gender G . To generate the referring expression given the description of the target entity and its surrounding context, we implemented an *encoder-attention-decoder* architecture with a copy mechanism, sharing the same input word-embedding matrix V , as explained in the following sections.

4.2.2.1 Encoder

In order to generate feature representations for the inputs, the model starts by encoding the identifier of the target entity as well as the pre- and post-contexts, using three different bidirectional Long-Short Term Memory⁸ layers (LSTM) ([Hochreiter; Schmidhuber, 1997](#)). The identifier of the target entity

$$X^{(wiki)} = \{x_1^{(wiki)}, x_2^{(wiki)}, \dots, x_T^{(wiki)}\} \quad (4.4)$$

is represented by the forward and backward hidden-state vectors

$$(\vec{h}_1^{(wiki)}, \dots, \vec{h}_m^{(wiki)}) \quad (4.5)$$

and

$$(\overleftarrow{h}_1^{(wiki)}, \dots, \overleftarrow{h}_m^{(wiki)}) \quad (4.6)$$

To form its final feature representation, forward and backward hidden-state representations at each timestep t are concatenated as

$$h_t^{(wiki)} = [\vec{h}_t^{(wiki)}, \overleftarrow{h}_t^{(wiki)}] \quad (4.7)$$

Using the two remaining bidirectional LSTMs, the same process is repeated for the textual context surrounding the reference, resulting in the final pre- and post-context representations

$$([\vec{h}_1^{(pre)}, \overleftarrow{h}_1^{(pre)}], \dots, [\vec{h}_m^{(pre)}, \overleftarrow{h}_m^{(pre)}]) \quad (4.8)$$

⁸ According to [Jurafsky and Martin \(2023, p. 198\)](#), LSTMs (Long-Short Term Memory) “divide the context management problem into two subproblems: removing information no longer needed from the context, and adding information likely to be needed for later decisionmaking”.

and

$$([\vec{h}_1^{(post)}, \overleftarrow{h}_1^{(post)}], \dots, [\vec{h}_m^{(post)}, \overleftarrow{h}_m^{(post)}]) \quad (4.9)$$

respectively. Finally, the type and gender of the target entity are encoded into their respective vector representations, V_{type} and V_{gender} , by looking up their entry in the sharing word-embedding matrix V .

4.2.2.2 Decoder

Once the information about the target entity and its surrounding contexts are encoded, their vector representations are fed into an LSTM decoder, augmented with attention and copy mechanisms⁹, in order to produce an adequate referring expression to the target entity according to the context. The process is explained in detail in the following sections.

4.2.2.2.1 Attention Mechanism

The decoder process starts with the attention mechanism (Bahdanau; Cho; Bengio, 2015), which aims to compute a vector c_t at each timestep t . The mechanism first computes the energies $e_{tj}^{(wiki)}$, $e_{tj}^{(pre)}$ and $e_{tj}^{(post)}$ based on the encoder states $h_t^{(wiki)}$, $h_t^{(pre)}$ and $h_t^{(post)}$, together with the decoder state s_{t-1} . The softmax function is then applied over these energies, resulting in the final attention probabilities $\alpha_t^{(wiki)}$, $\alpha_t^{(pre)}$ and $\alpha_t^{(post)}$. Equations 4.10 and 4.11 show the computation of the energies and final attention probabilities, where $k \in \{wiki, pre, post\}$ and the matrices $W_a^{(k)}$ and $U_a^{(k)}$ as well as the attention vectors $v_a^{(k)}$ are training parameters.

$$e_{tj}^{(k)} = v_a^{(k)T} \tanh(W_a^{(k)} s_{t-1} + U_a^{(k)} h_j^{(k)}) \quad (4.10)$$

$$\alpha_{tj}^{(k)} = \frac{\exp(e_{tj}^{(k)})}{\sum_{n=1}^N \exp(e_{tn}^{(k)})} \quad (4.11)$$

⁹ Please refer to the next sections about each augmentation. For more information, please access the Language Models Chapter (Paes A.; Vianna; Rodrigues, 2023) for Brazilian Portuguese and RNNs and LSTMs Chapter (Jurafsky; Martin, 2023) for English.

At each decoding step t , a final context vector $c_t^{(k)}$ is computed based on the sum of the encoder states $h_t^{(k)}$ weighed by the attention probabilities $\alpha_{tj}^{(k)}$, as the following equation expresses:

$$c_t^{(k)} = \sum_{j=1}^N \alpha_{tj}^{(k)} h_j^{(k)} \quad (4.12)$$

Finally, in order to obtain the final context vector c_t , we follow the concatenative approach of NeuralREG, where the attention vectors $c_t^{(wiki)}$, $c_t^{(pre)}$ and $c_t^{(post)}$ are simply concatenated, such as $c_t = [c_t^{(wiki)}, c_t^{(pre)}, c_t^{(post)}]$.

4.2.2.2.2 Decoding

After attending the representations of the target entity and its surrounding contexts, the resulting attention vector c_t is concatenated with the previous decoding state s_{t-1} , the word-embedding of the previously generated token $V_{y_{t-1}}$ and the vector representations of the type and gender of the target entity, V_{type} and V_{gender} . This concatenation is then fed into the decoding layer, which produces its next state s_t . Finally, a softmax layer is applied over the decoding state s_t to generate a probability distribution over the output vocabulary. Equations 4.13, 4.14 and 4.15 summarize this process:

$$s_t = \Phi_{\text{dec}}(s_{t-1}, [c_t, V_{y_{t-1}}, V_{type}, V_{gender}]) \quad (4.13)$$

$$z_t = W_b s_t + b \quad (4.14)$$

$$P_{\text{vocab}}(w) = \frac{\exp(z_{ti})}{\sum_{j=1}^J \exp(z_{tj})} \quad (4.15)$$

4.2.2.2.3 Copy Mechanism

To enable the approach to generate referring expressions to unseen entities, we also implemented a copy mechanism during the decoding process, similar to the one presented by See, Liu and Manning (2017). This mechanism first computes a probability p_{gen} based on the attention vector of the target entity $c_t^{(wiki)}$, the decoding state s_{t-1} and the word-embedding of the previously generated token $V_{y_{t-1}}$, as the following equation expresses:

$$p_{\text{gen}} = \text{sigmoid}(W_c c_t^{(wiki)} + W_d s_{t-1} + W_e V_{y_{t-1}} + b) \quad (4.16)$$

p_{gen} is used to decide between (i) choosing the token with the highest probability in the softmax probability distribution $P_{vocab}(w)$ in Equation 4.15 or (ii) copying the token from the description of the entity $X^{(wiki)}$ with the highest probability according to the attention weights $\alpha_t^{(wiki)}$. The final probability distribution to choose the next token at each timestep t is given by the following Equation:

$$P(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} \alpha_{ti}^{(wiki)} \quad (4.17)$$

In this context, we expect the model to learn that p_{gen} should have a higher value when the target entity was seen during training, and a lower one when a referring expression should be generated for an unseen entity.

4.2.2.2.4 Loss

During training time, the approach has its training parameters updated in order to minimize the following loss function:

$$J(\theta) = - \sum_t P(y_t) \quad (4.18)$$

4.2.3 Automatic Evaluation

4.2.3.1 Data

We used the delexicalized version of the WebNLG corpus described in Section 4.2.1. In particular, we used version 1.5 of the corpus, which is publicly available¹⁰. This version of the corpus contains 67,027, 8,278, and 19,210 referring expression instances in training, development, and test sets, respectively. Training and development domains have instances of 10 semantic domains, whereas the test set has instances of those 10 domains, plus 5 unseen ones in the former sets.

Each instance of the sets is formed by the target entity, a referring expression, and pre- and post-contexts. Pre- and post-contexts are represented in their lowercase and tokenized forms, whereas the referring expression is in its truecased and tokenized one. Moreover, references to different discourse entities are represented by their Wikipedia IDs. In contrast, numbers, dates, and other constants are

¹⁰ <<https://github.com/ThiagoCF05/webnlg>>

Entities	Model	RE.Acc.	SED	BLEU	T.Acc.	Prec.	Recall	F1
All	OnlyNames	<u>0.51</u>	<u>4.21</u>	<u>65.48</u>	<u>0.14</u>	-	-	-
	NeuralREG	0.38	9.86	48.10	0.11	0.72	<u>0.63</u>	<u>0.67</u>
	NeuralREG+Copy	0.59	3.53	66.14	0.16	<u>0.79</u>	0.53	0.63
	ProfileREG	0.42	7.05	54.08	0.09	0.82	0.86	0.84
Seen	OnlyNames	0.53	4.32	66.39	0.16	-	-	-
	NeuralREG	<u>0.70</u>	<u>3.07</u>	<u>70.21</u>	<u>0.20</u>	<u>0.78</u>	<u>0.76</u>	<u>0.77</u>
	NeuralREG+Copy	0.73	2.50	71.74	0.24	0.79	0.71	0.75
	ProfileREG	0.69	3.11	69.11	0.17	0.79	0.90	0.84
Unseen	OnlyNames	0.50	4.10	63.97	0.11	-	-	-
	NeuralREG	0.07	16.71	25.42	0.00	0.67	<u>0.55</u>	<u>0.61</u>
	NeuralREG+Copy	<u>0.46</u>	<u>4.57</u>	<u>59.15</u>	<u>0.08</u>	<u>0.79</u>	0.42	0.54
	ProfileREG	0.14	11.03	36.45	0.00	0.84	0.83	0.84

Table 4 – Referring Expressions’ Accuracy and String-edit distance (SED), BLEU and Text Accuracy scores of the models, Pronoun – Precision, Recall, and F1-Score of the models in the automatic evaluation. The best results are **boldfaced**, whereas the second best are underlined.

represented by one-word ID replacing white spaces with underscores and eliminating double quotes (Ferreira et al., 2018a; Ferreira et al., 2018b). To represent the target entity $X^{(wiki)}$ as described in Section 4.2.2, we lowercase the Wikipedia ID of the target entity, remove all special characters, and then split it into a list based on underscores (e.g., Abdul_Taib_Mahmud $\rightarrow$ [abdul, taib, mahmud]). Accordingly, all target entities’ gender (female, male, neutral) and type (person, organization, etc.) used by our approach were automatically retrieved from DBpedia¹¹.

4.2.3.2 Model Settings

Regarding the model parameters, we followed most of the settings from NeuralREG’s set-up of Ferreira et al. (2018a). We trained the model with 60 epochs with a dropout of 0.2. Furthermore, we set the early stopping of the neural networks to 10 and the beam size to 1. We applied a maximum output limit generation of 30. Moreover, we set the batch, state, and attention sizes to 80, 256, and 256, respectively. Additionally, we set pre-context, post-context, and entity word embeddings to be 128D each.

¹¹ <http://dbpedia.org/page/Ada_Lovelace>

Original: **Adenan Satem** was born in **Japanese Occupied British Borneo**. His successor was **Abdul Taib Mahmud**, who, resides in **Sarawak** and is a member of the “*Barisan Raáyat Jati Sarawak*” party.

OnlyNames: **Adenan Satem** was born in **Japanese Occupied British Borneo**. **Adenan Satem** successor was **Abdul Taib Mahmud**, who, resides in **Sarawak** and is a member of the “*Barisan Raáyat Jati Sarawak*” party.

NeuralREG: **The Boeing light combat** was born in **Abilene, in Texas**. They successor was **the hal of the astronaut**, who, resides in **the state of Grenada** and is a member of the “*Barisan Raáyat Jati Sarawak*” party.

NeuralREG+Copy: **Adenan Satem** was born in the **Japanese Occupied British Borneo**. His successor was **Abdul Taib Mahmud**, who, resides in **Sarawak** and is a member of “*Barisan Raáyat Jati Sarawak*” party.

ProfileREG: **258.2 Satem** was born in the **Japanese**. Its successor was the **Taib the Moro**, who, resides in the **Sarawak** and is a member of the “*Barisan Raáyat Jati Sarawak*” party.

Table 5 – Sample outputs of an unseen domain (Politician) – the original and the generated text of each of the four models. Referring expressions are **boldfaced**, and constants are “double-quoted” and *italicized*.

4.2.3.3 Baselines

We compared our proposal (*NeuralREG+Copy*) against three baselines: the concatenative attention version of the original model *NeuralREG*, OnlyNames (Ferreira; Krahmer; Wubben, 2016), and ProfileREG (Cao; Cheung, 2019).

4.2.3.3.1 OnlyNames

OnlyNames correlates an entity that will be referred to by its Wikipedia ID. This baseline exclusively works with proper names by replacing entities’ underscores with white spaces (e.g., Ada_Lovelace to “Ada Lovelace”). Instead of working exclusively with proper names, our approach implements other referential forms, such as pronouns and descriptions. In addition, uses a context-based disambiguation method (e.g., when encountering entities with the same proper name: “John_Smith_(explorer)” and “John_Smith_(lexicographer)”), consequently yielding a more natural discourse flow in the texts produced.

4.2.3.3.2 NeuralREG+Catt

NeuralREG+Catt is an end-to-end deep neural network model that uses both form and content to generate texts. The model works with a delexicalized version of the WebNLG corpus by first encoding pre- and post-contexts as a reference. In contrast to our proposal, NeuralREG+Catt does not implement a copy mechanism and does not consider any external knowledge when selecting the best-referring expression.

4.2.3.3.3 ProfileREG

ProfileREG encodes information from a local context and an external profile to generate references to a given entity. This model is able to determine the best reference to an entity by selecting from existing vocabulary, pronouns, or the entity’s profile. Contrary to ProfileREG, our model uses selected entity features and different architectures in order to evaluate the best scenario for generating referring expressions.

4.2.3.4 Metrics

We calculated Accuracy and String Edit Distance ([Levenshtein, 1966](#)) in order to measure the quality of the generated referring expressions in comparison with the gold-standard ones. To evaluate the models’ performance in realizing pronouns, we also computed the accuracy, precision, recall, and F1-Score, based on a concise difference between the gold-standard referring expressions to the ones produced by the model. Finally, we compared the original texts against the references lexicalized through the models by computing text accuracy (i.e., the ratio of correctly generated referring expressions) and BLEU (Bilingual Evaluation Understudy) score ([Papineni et al., 2002](#)).

4.2.3.5 Results

Table 4 presents the results of our model in comparison with the baselines for *all* entities as well as for *seen* and *unseen* ones. In terms of referring expression accuracy, string edit distance, text accuracy, and BLEU score, our proposed

approach outperforms the three baselines considering *all* entities and *seen* ones only. Regarding *unseen* entities, our model presents higher results for the same metrics in comparison with all models, except for OnlyNames one. Regarding pronouns, ProfileREG introduces the best results, while the OnlyNames model is not considered, since this model is not able to generate this form of reference.

Table 5 shows an example of a text lexicalized with referring expressions generated by our proposal and the three baselines. The text was extracted from the test set of the data in the Politician domain, not present in the training and development sets. By comparing our approach (*NeuralREG+Copy*) to the baseline OnlyNames, we can see that our model is able to generate more variation in referring mechanisms since it makes use of a pronoun as a referential form, while OnlyNames uses repetition of proper names. The outputs for the **Adenan_Satem** for NeuralREG and ProfileREG models show generation problems, namely entities completely unrelated to the references (e.g., **The Boeing light combat** and **258.2 Satem**, respectively).

4.2.4 Human Evaluation

To assess the quality of the texts generated by our proposal and the three baselines, we conducted a supplementary evaluation with human judges.

4.2.4.1 Method

Two applied linguists were recruited to rate the texts from the Laboratory for Experimentation in Translation (LETRA) at the Faculty of Letters at the Federal University of Minas Gerais (UFMG). They are proficient in English and have over 20 years of expertise as translators and language advisers.

We selected 75 instances of the delexicalized version of the WebNLG corpus, considering a unique occurrence for each combination between the number of triples (ranging from 1 to 7) and domain (10 seen and 5 unseen ones). After selecting the set of triples, we collected the corresponding produced versions of each investigated model introduced in this study (our proposal and three baselines). Finally, we randomly ordered the final trial set of ($4 \times 75 =$) 300 sentences to decrease the bias

of having the 4 generated texts together during the evaluation.

The performed evaluation followed the best practices suggested by [Lee et al. \(2019\)](#) and the guidelines in [Novikova, Dušek and Rieser \(2018\)](#) regarding human evaluations of NLG systems. For instance, we chose well-defined criteria to assess text quality and a well-established scale for assessment. The participants were asked to rate the automatically generated sentences with respect to three criteria: fluency, i.e., whether the text flow was acceptable; grammaticality, i.e., whether grammaticality and lexical patterns were close to human language patterns; and semantic adequacy, i.e., whether the information in the output text matched that of the input representation. In addition, a 5-point Likert scale was used (1 – very low, 2 – low, 3 – medium, 4 – high, and 5 – highly/fully adequate).

4.2.4.2 Results

Table 6 summarizes the results of human evaluation regarding fluency, grammaticality, and semantic adequacy for all, seen, and unseen entities. Our proposed model outperformed the previous version of NeuralREG and presented competitive results compared to the current state-of-the-art in the literature. Regarding the grammaticality metric, our model presents the best results for all, seen, and unseen entities considering the three baselines. Regarding fluency and semantic adequacy, the human evaluation showed similar scores for our proposal and OnlyNames.

Despite OnlyNames’ limitations in referring expression generation, OnlyNames baseline performed very well, which can be accounted for by the fact that WebNLG is a corpus made up of texts potentially used to yield encyclopedia entries, which allow for repetition of proper nouns, unlike other types of text. Hallucination¹² and repetition are often present in neural models ([Rohrbach et al., 2018](#); [Moryossef; Goldberg; Dagan, 2019](#); [Holtzman et al., 2018](#)) can also account for OnlyNames’ good performance which does not suffer from this problem. An example of this can be seen in the RDF triple set:

¹² According to [Ji et al. \(2023, p. 3\)](#), the term “hallucination” denotes a specific type of error when non-existing texts are detected or localized incorrectly at their expected position.

Metric	Model	All	Seen	Unseen
Fluency	OnlyNames	4.03 $\pm$ 0.16	4.05 $\pm$ 0.18	3.95 $\pm$ 0.36
	NeuralREG	3.31 $\pm$ 0.30	3.91 $\pm$ 0.21	1.53 $\pm$ 0.40
	NeuralREG+Copy	<u>4.00 $\pm$ 0.16</u>	<u>4.05 $\pm$ 0.17</u>	<u>3.84 $\pm$ 0.39</u>
	ProfileREG	3.53 $\pm$ 0.26	3.87 $\pm$ 0.21	2.53 $\pm$ 0.64
Grammaticality	OnlyNames	3.99 $\pm$ 0.14	4.05 $\pm$ 0.16	3.79 $\pm$ 0.28
	NeuralREG	3.36 $\pm$ 0.29	3.92 $\pm$ 0.22	1.71 $\pm$ 0.34
	NeuralREG+Copy	4.02 $\pm$ 0.16	4.05 $\pm$ 0.18	3.92 $\pm$ 0.32
	ProfileREG	3.51 $\pm$ 0.24	3.83 $\pm$ 0.20	2.55 $\pm$ 0.59
Semantic Adequacy	OnlyNames	4.85 $\pm$ 0.10	4.88 $\pm$ 0.08	4.76 $\pm$ 0.34
	NeuralREG	3.70 $\pm$ 0.37	4.50 $\pm$ 0.24	1.34 $\pm$ 0.25
	NeuralREG+Copy	<u>4.63 $\pm$ 0.15</u>	<u>4.75 $\pm$ 0.14</u>	<u>4.29 $\pm$ 0.44</u>
	ProfileREG	3.89 $\pm$ 0.32	4.46 $\pm$ 0.24	2.18 $\pm$ 0.55

Table 6 – Mean and $\pm 95\%$ Confidence intervals on Fluency, Grammaticality, and Semantic Adequacy results for All, Seen, and Unseen entities of the Human Evaluation. The best results are **bolded**, whereas the second best are underlined.

[Alfa_Romeo_164 | assembly | Milan. Alfa_Romeo_164 |
relatedMeanOfTransportation | Saab_9000]

OnlyNames output was “Alfa Romeo 164, which is assembled in Mi-lan, is a related means of transportation to Saab 9000, in that they are both cars”, whereas *NeuralREG+Copy* produced the following output “Romeo Romeo 164, which is assembled in Milan, is a related means of transportation to Saab, in that they are both cars”. Despite eventual hallucination problems, the human evaluation showed that our model has more consistent performance, improving overall quality.

4.2.5 Ablation Study

We also performed an ablation study¹³ in order to analyze the performance of the different features used in our proposal.

¹³ An ablation study (Meyes et al., 2019) usually refers to removing some component/attribute of the NLG model to optimize the model by calculating how the removal impacted the performance.

4.2.5.1 Method

We evaluated the copy mechanism, pre- and post-contexts, as well as gender and type embeddings in order to determine which feature best influences the model. The performance of every single feature was analyzed by running the model without it and measuring loss, according to the referring expression accuracy metric in the test part of the data, considering all entities as well as only *seen* and only *unseen* ones. When removing the copy mechanism, the model had a similar performance to the original NeuralREG, though with the target entity also represented by the entity’s embeddings for gender and type.

4.2.5.2 Results

Table 7 depicts the results of our ablation analysis. The removal of the copy mechanism feature (*Ablation 1*) causes the highest decrease in the referring expression accuracy for *all* entities as well as only *seen* and *unseen* ones, validating this feature as the most efficient within the model. In addition, the copy mechanism relevance for generalizing to unseen entities is proved by a high accuracy drop in the analysis. Furthermore, removing entity embeddings for the entities’ gender and type (*Ablation 2*) causes a negligible drop in all and unseen scores, particularly when generating referring expressions to unseen entities. Regarding context, pre-context (*Ablation 3*) does not yield the expected performance regarding accuracy for seen entities, since the produced referring expressions to this type of entities proved better without this feature for all and unseen entities. Nevertheless, we can stress the importance of pre-context for unseen entities, since referring expression accuracy for this kind of entity decreases with the removal of this feature. Finally, post-context (*Ablation 4*) causes the second-highest decrease, being validated as the second-best feature.

Model	Setup	Accuracy		
		All	Seen	Unseen
Proposal	pre, copy, post, entity type, and gender embeddings	0.59	0.73	0.46
Ablation 1	without copy	0.40	0.70	0.09
Ablation 2	without entity embeddings	<u>0.50</u>	0.71	<u>0.28</u>
Ablation 3	without pre-context	0.54	<u>0.67</u>	0.42
Ablation 4	without post-context	<u>0.50</u>	0.64	0.36

Table 7 – Results of the Ablation Analysis. The best features (with the highest drop) are **boldfaced**, whereas the second best are underlined.

4.2.6 Discussion

This study is the first approach to content-enrichment of NLG models with an application to a REG model. This approach addresses a limitation of NeuralREG, a state-of-the-art encoder-decoder referring expression generation system, that is a drawback to generating references to entities not previously seen during its training. To solve the problem, we proposed two extensions to the original approach: a copy mechanism and a multi-token representation for the referent as well as its gender and type.

Considering pre- and post-contexts where an entity should be referred to and information about the entity’s gender and type, at each decoding step, our model decides whether the next token of the referring expression should be generated from the output vocabulary or copied from the multi-token input representation of the entity.

Although our approach set out to improve the generation of referring expressions to unseen entities only, an automatic evaluation suggests that it presents competitive results regarding all the models when comparing overall performance and also for seen entities. Regarding the generation of pronouns and references to unseen entities, our model outperforms ProfileREG and OnlyNames. Furthermore, human evaluation, conducted to rate the automatically generated sentences showed that our model achieved the best results regarding grammaticality. Regarding fluency and semantic adequacy, *NeuralREG+Copy* and OnlyNames presented similar results as shown in Table 6. The similarities between both models are

striking, which demonstrates that OnlyNames remains a competitive baseline in REG since this baseline is less prone to produce hallucinations. In order to understand these results and have a deeper insight into the performance of each feature, we also conducted an ablation analysis, which showed different results in referring expression generation for seen and unseen entities.

4.2.6.1 Surrounding Context

Pre- and post-contexts seem to perform different roles when used as input features to generate referring expressions. Based on our ablation analysis, post-context plays a crucial role, being ranked the most important feature when generating referring expressions to seen entities and third to unseen ones. On the other hand, pre-context seems to have a slight contribution only to the generation of references to unseen entities. However, when not used, the approach performs worst for the generation of referring expressions to seen entities.

4.2.6.2 Copy Mechanism

Among the input features, the copy mechanism proved an essential feature of the model. Its importance was supported by the results of the ablation analysis, which pointed to this feature as the most important for the generation of referring expressions to unseen entities. This confirms the copy mechanism to be a productive addition to NeuralREG in order to make it able to work with entities not seen during training.

4.2.6.3 Gender and Type Entity Representations

Besides the copy mechanism, we sought to make NeuralREG generalize to unseen entities by feeding it with embedding representations of the referent’s gender and type. Among the motivations to use these features, we considered how easy it is to access this information in the Semantic Web since entities in WebNLG are represented by their URIs. Second, we hypothesized that such representations would allow the model to generate pronominal and descriptive referring expressions to unseen entities. To some extent, the pronominal reference **his** to the unseen entity **Adenan_Satem** produced by our approach and depicted in Table 5 shows that

the representations may indeed help. Ablation results also showed that they are the second most important feature in referring expression generation to unseen entities, only behind the copy mechanism (another extension proposed by our study). Such a result confirms our hypothesis and the benefits of enriching an NLG model with entities' representations aligned with the corresponding domain. Please refer to Figure 9 for a more user-friendly presentation of the unseen entity **Adenan_Satem**.

Adenan_Satem was born in **Japanese_Occupied_British_Borneo**. **Adenan_Satem** successor was **AbdulTaib_Mahmud**, who, resides in **Sarawak** and is a member of the "Barisan Raayat Jati Sarawak" party.

Pre-context	Adenan_Satem was born in Japanese_Occupied_British_Borneo.
Entity	Adenan_Satem (https://en.wikipedia.org/wiki/Adenan_Satem)
Post-context	successor was AbdulTaib_Mahmud, who, resides in Sarawak and is a member of the "Barisan Raayat Jati Sarawak" party.
Entity information	Adenan_Satem → [adenan, saten] Sex: male Type: person
Copy mechanism	The entity <i>Adenan_Satem</i> was seen at the beginning of the phrase, the mechanism chooses to use a pronoun instead of repeating the text.

Adenan Satem was born in Japanese Occupied British Borneo. **His** successor was Abdul Taib Mahmud, who, resides Sarawak is a member of the "Barisan Raayat Jati Sarawak" party.

Figure 9 – Referring expression generation of an unseen entity **Adenan_Satem**.

4.2.6.4 Future Work

Although our two proposed extensions allowed NeuralREG to perform better when compared to its previous version and also generate superior referring expressions to unseen entities, OnlyNames performed slightly better than our approach in the generation of referring expressions to this kind of entities as well as produced similar results to our model during human evaluation. We assume that part of this result is related to known issues in semantic neural models, such

as hallucination, which could influence both automatic and human evaluation results, in particular for unseen entities. To fix this issue, we aim to investigate the generation of synthetic referring expression data to augment the training data and better tune our approaches¹⁴.

Results also show that ProfileREG outperformed our model in the generation of pronouns. We hypothesize these results in an impact of incorrect gender and type information for some entities extracted from DBpedia. For instance, the entity BBC in DBpedia¹⁵ is also considered of the type **Person**, leading to the generation of inaccurate descriptions (*The person*) and pronominal (*She* or *He*) outputs. In future work, we aim to manually inspect all type and gender information extracted from DBpedia in order to avoid errors. Additionally, to generate better pronominal referring expressions, we will enhance our approach by using the “profile” computed by the ProfileREG model.

4.2.6.5 Conclusion

We have proposed the context enrichment of the NeuralREG model with two extensions to its implementation to overcome shortcomings in not being able to generalize to entities not seen during the training process when generating referring expressions. We can conclude that our proposal contributes to generating more significant referring expressions to unseen entities, besides seen ones. Furthermore, our study provides a new version of a strong baseline within the NLG area. A future direction in our work is to implement the improvements discussed in this study in order to match OnlyNames’ performance for unseen entities.

Addressing the unseen entities’ limitation of REG models has driven this thesis to understand how to adapt contextual information in order to enhance the accuracy of language comprehension and interpretation. In addition, the ablation study pointed to the direction of effectively surpassing the challenging implementation of the knowledge enhancement approach. Furthermore, external knowledge bases, such as DBpedia and Wikipedia, contribute to a more accurate representation

¹⁴ According to Jurafsky and Martin (2023, p. 258), data augmentation is “a statistical technique for dealing with insufficient training data, by adding new synthetic data that is generated from the current natural data”.

¹⁵ <<http://dbpedia.org/page/BBC>>

of real-world knowledge due to their vast repository of structured and semantically interconnected information. In essence, refining the REG models' understanding of entities, relationships, and events, enhances their ability to generate more contextually relevant and coherent text. Finally, the lessons learned from this first context enrichment approach are applied to the following studies and presented in Chapters 5 and 6.

5 Enriching an NLG Model from a Robot-Journalism Approach

The content of this chapter is adapted from [Teixeira et al. \(2020\)](#) following paper:

Teixeira, A. L. R.; Campos, J.; **Cunha, R.**; Castro Ferreira, T.; Pagano, A.; Cozman, F. DaMata: A robot-journalist covering the Brazilian Amazon deforestation. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, 2020. p. 103–106. Available on: <https://aclanthology.org/2020.inlg-1.15>.

Summary of the Second Approach

In this chapter, we introduce the second approach to context-enrichment of an NLG Model – DaMata Reporter¹, a robot journalist covering deforestation in the Brazilian Amazon. The robot journalist is based on a pipeline architecture of Natural Language Generation, which yields daily and monthly multilingual reports based on the public data provided by DETER, a real-time deforestation satellite monitor developed and maintained by the Brazilian National Institute for Space Research (INPE). Brazilian Portuguese and English automatically generated reports by DaMata are published on the Twitter platform. The context enrichment of DaMata with Wikidata and DBPedia data was conducted by this thesis' author, who was responsible for developing, testing, and implementing the necessary code for the enrichment of the Referring Expression Generation component of DaMata. This thesis' author contributions followed writing and reviewing the accepted INLG 2020 paper and participating in annotating and validating DaMata's generated output. This chapter is adapted from the accepted paper with the permission of the other authors and members of DaMata's development team. The key areas developed of the robot journalist related to the content of this thesis are the context enrichment and the results of this study which have been a crucial part of developing this thesis's second approach. In addition, during an experiment of Damata with Large Language Models, we identified a limitation of these models in relation to Numerical Referring Expressions. We add information about the ongoing research and present key parts of an under-reviewed paper regarding the issue. This chapter has the format of a paper and can be read independently from the rest of the thesis. However, as an essential part of the full study, this chapter gives a general perspective of one of the applications of the Natural Language Generation within a social and political context, the advantages and the burden of automated journalism.

¹ This chapter was adapted from the INLG 2020 paper (Teixeira et al., 2020), available at <https://aclanthology.org/2020.inlg-1.15/>. The author of this thesis is the third author of the paper. The paper was adapted to fulfill the requirements of this thesis. This thesis' author contributions were related to context enrichment, being responsible for developing the code, testing, annotation, validation, and producing the necessary data, especially for the Referring Expression Generation component of DaMata.

5.1 Introduction

Robot-Journalism is one of the most promising Natural Language Generation (NLG) applications thanks to the high volume of structured data streams available nowadays, which enables automated systems to report recurrent information with high fidelity and low latency (Carlson, 2015). The *robot-journalism* or *automated journalism* can be explained as “algorithmic processes that convert data into narrative news texts with limited to no human intervention beyond the initial programming” (Carlson, 2015, p. 2). As happens with any technology, robot journalism brought to our attention the positive and negative impact of this kind of text production on journalism itself, journalists, and news readers.

On the one hand, automation can unburden journalists from tedious data reporting tasks, thus enabling human agents to devote efforts to more investigative coverage (Graefe, 2016). On the other hand, ethical concerns permeate the use of automated journalism, such as the implications of who is the real author behind the news, the journalistic labor, the question of authority, and negative effects on readers as pointed out by Montal and Reich (2017). Despite the unethical uses of robot journalism and its impact on journalists’ careers, the application to social-related domains can be a great ally in producing data from non-readable human sources and being a great tool to the social side of journalism (Carlson, 2015; Montal; Reich, 2017).

In this social context, the health of the Amazon rain forest has drawn global news coverage attention lately when soaring deforestation raised public awareness worldwide (Escobar, 2020). Accurate information about deforested land in the Brazilian part of this territory is yielded by Brazil’s National Institute for Space Research (INPE)² through several satellites and other monitoring systems. One of the most helpful is DETER³, a real-time deforestation warning surveillance system (Diniz et al., 2015).

Despite being publicly available⁴, DETER’s deforestation data output is in graphical and numerical style, which renders it less accessible to general audiences

² <<http://www.inpe.br/>>

³ DETER: Brazil’s Real-time Deforestation Detection System (Diniz et al., 2015).

⁴ <<http://terrabrasilis.dpi.inpe.br>>

and demands coverage by human journalists specialized in the domain. To address this issue, we introduce a robot journalist based on an NLG pipeline architecture, which generates daily and monthly reports about deforestation in the Brazilian Amazon, both in Brazilian Portuguese and English, and publishes them on Twitter.

5.1.1 System Overview

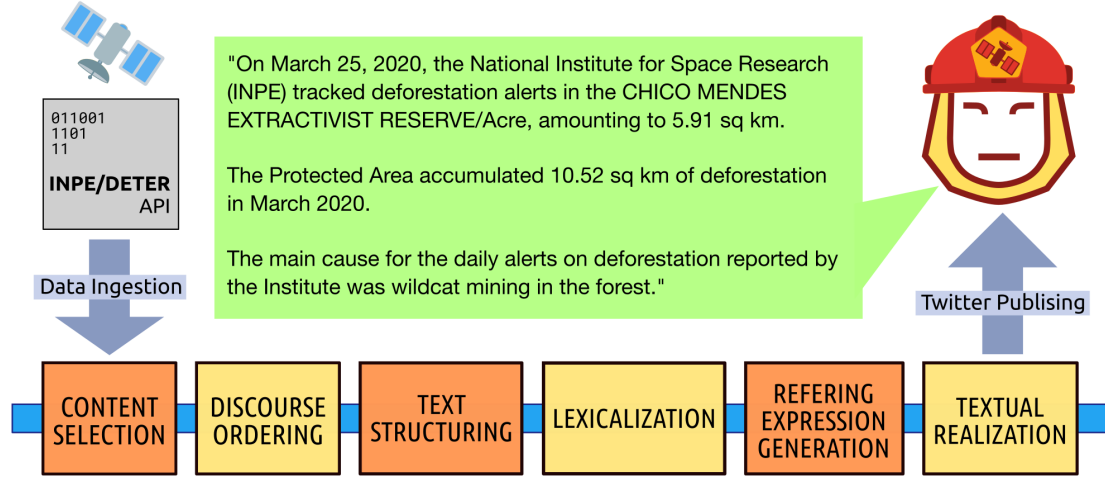


Figure 10 – Robot-Journalist Pipeline Architecture and a Human-Readable Output Sample.

Unlike novel end-to-end systems, which may hallucinate content and are more suitable to fictional domains, our system follows a pipeline architecture as proposed by [Ferreira et al. \(2018b\)](#), which converts non-linguistic data into text in six steps: Content Selection, Discourse Ordering, Text Structuring, Lexicalization, Referring Expression Generation, and Textual Realization. We have added a seventh step – publication – responsible for sharing the generated news on Twitter. The following paragraphs describe each module.

5.1.1.1 Content Selection

This module is responsible for selecting relevant messages to be verbalized in the reports. It draws on raw deforestation data provided by DETER through

its API⁵. For a daily report, our robot-journalist selects data on the amount of deforested area and the main cause of deforestation in a given city or nature-protected area as well as the total area hitherto deforested in the respective month for the given region. For a monthly report, the content selector retrieves the amount of total deforestation in the Brazilian Amazon; month and annual variation; the main cause of deforestation; and the state, city, and protected area with the highest deforested numbers in the respective month.

5.1.1.2 Discourse Ordering

We arrange our discourse messages by a rule-based approach to convey the relative importance of each bit of information, as it is usually done in journalistic texts. For example, for the messages selected for a daily report - daily deforested area, the main cause of deforestation, and total deforestation in the month, a likely outcome order would be: `DAILY_ALERT` $\rightarrow$ `TOTAL_DEFORESTATION` $\rightarrow$ `CAUSE`.

5.1.1.3 Text Structuring

In our robot-journalist, *DaMata*, this step is also solved by a rule-based approach, which, for our previous daily report example, could return: `<P> <S> DAILY_ALERT CAUSE </S> </P> <P> <S> TOTAL_DEFORESTATION </S> </P>`.

5.1.1.4 Lexicalization

Once the order and structure of the intents are defined, our system chooses one lexicalization template that verbalizes the structure of the text. For the sake of variety and to ensure audience engagement, this module randomly picks one template among the several manually annotated ones. For example: `TOTAL_DEFORESTATION` $\Rightarrow$ Deforestation amounted to AREA, in UC.

5.1.1.5 Referring Expression Generation

At this step in the pipeline, the system tracks entity tags and maps the relevant referring expressions, replacing entities in context accordingly. As there are

⁵ <http://terrabrasilis.dpi.inpe.br/homologation/file-delivery/download/deter-amz/daily>

multiple choices to refer to the same entity throughout a piece of discourse, upon first introduced, an entity may be picked up by means of synonymous expressions (e.g., Legal Amazon, the forest, the rainforest, it, etc.), see Figure 10. In order to provide referring expressions that are more suitable to DaMata’s end-users, i.e., a Twitter audience with different proficiency levels, we enrich DaMata with context in order to address this need. This approach was developed by this author’s thesis and is addressed in more detail in Section 5.2.

5.1.1.6 Textual Realization

The final step in our pipeline sequence performs the remaining adjustments to transform the abstract machine intermediate representations into human-readable text. Nominal and verbal inflections, contractions, and de-tokenization are performed according to each particular grammar. The final product is shown in the robot dialog box in Figure 10.

5.2 Enriching DaMata’s Context

To address the gap in generating more diverse outputs in DaMata’s context, first, we gathered information from DBpedia⁶ as external features. We added attributes collected from Web Semantic content, such as DBpedia⁷ and Wikidata⁸ of each entity from the deforestation domain. This context was added to the original pipeline in order to enhance the variety of the texts for DaMata’s end-users. The experiment collected referring expressions to geographical entities, such as Brazilian states, cities, and conservation areas within the Legal Amazon region. The original number of entities in DaMata’s REG module was 480. After the context enrichment with DBpedia/WikiData, the total of entities increased to 1044. Please refer to Table 8 for a more user-friendly presentation of the context-enrichment approach.

⁶ An example of the *information* we used in our approach: for the “Chico Mendes Extractive Reserve” entity, we gathered data from DBpedia – http://dbpedia.org/page/Chico_Mendes_Extractive_Reserve and Wikipedia – https://en.wikipedia.org/wiki/Chico_Mendes_Extractive_Reserve.

⁷ <https://wiki.dbpedia.org/>

⁸ <https://wikidata.org>

Entity	Chico_Mendes_Extractive_Reserve
Type	Extractive_reserve__(Brazil)
WikiData ID	Q10359666
Portuguese Label	<i>Reserva Extrativista Chico Mendes</i>
Portuguese Comment	<i>A Reserva Extrativista Chico Mendes é uma unidade de conservação federal do Brasil categorizada como reserva extrativista e criada por Decreto Presidencial em 12 de março de 1990 numa área de 970.570 hectares no estado do Acre. É administrada pelo Instituto Chico Mendes de Conservação da Biodiversidade (ICMBio).</i>
English Label	Chico Mendes Extractive Reserve
English Comment	Chico Mendes Extractive Reserve (<i>Portuguese: Reserva Extrativista Chico Mendes</i>) is an extractive reserve in the state of Acre, Brazil.
Referring Expressions (Portuguese)	UC_1: [a unidade de conservação, esta unidade de conservação, a região, esta região, RESERVA EXTRATIVISTA CHICO MENDES, ela, A reserva, Esta reserva, Reserva Extrativista Chico Mendes, o Reserva Extrativista Chico Mendes, a Reserva Extrativista Chico Mendes]

Table 8 – Sample of DaMata’s context enrichment in Brazilian Portuguese and English according to information recovered from Semantic Web.

To enrich DaMata’s contextual capabilities across more varied responses and consequently foster a more human-centered experience, we analyzed how to integrate state-of-the-art architectures, known as Large Language Models (LLMs) (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020) into DaMata’s existing pipeline modules. While the integration of LLMs and rule-based systems holds great promise, LLMs often hallucinate (Ji et al., 2023) in domains such as deforestation. In order to address how a numerical-sensitive domain can leverage the benefits of LLMs with minimal numerical hallucinations, we conducted a study in collaboration with the University of São Paulo (USP) and Dublin City University (DCU). We present the main parts of the study due to the fact that this is the first outcome of the research group. The detailed investigation is presented in Appendix A Section.

	Portuguese			English		
	Train	Dev	Test	Train	Dev	Test
Daily Deforest.	2,031	252	242	10	10	10
Month Deforest.	162	10	11	159	8	10
Daily Fire	471	54	54	-	-	-
COVID-19	532	61	54	-	-	-
Total	3,196	377	361	169	18	20

Table 9 – Data Statistics.

This study addresses some LLMs, such as BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and GPT-2 (Radford et al., 2019) in two domains: governance (deforestation) and healthcare (COVID-19). Table 9 introduces the number of instances per language and domain, split into training, development, and test sets. Each instance in the corpus consists of a meaning representation and a corresponding gold-standard verbalization in Brazilian Portuguese or English representing the sentence of a report. For both languages, the verbalizations were automatically generated by the pipeline system described in Section 5, and in the studies of Teixeira et al. (2020) and Campos et al. (2020).

To evaluate the best strategy to represent numeric expressions, each of the three LLM surface realizers was trained in four versions with:

1. numbers represented by digits in the input meaning representation and the reference texts (i.e., no descriptions – *no desc*);
2. numbers described in their spell-out forms (i.e. *fifty-five*) in the input meaning representation but as digits (i.e. *55*) in the target references (i.e., description in the source – *desc src*);
3. numbers represented by digits (i.e. *55*) in the meaning representations but in their spell-out forms (i.e. *fifty-five*) in the target references (i.e., description in the target – *desc trg*);
4. numbers in their spell-out forms (i.e. *fifty-five*) in both meaning representations and target references (i.e., descriptions on both source and target – *desc*).

To evaluate how well a neural end-to-end surface realizer can generate numerical referring expressions, we evaluate two forms of number representation: by digits or word form description, both in the meaning representations as well as in the verbalizations, totaling four different conditions. To exemplify, Table 10 depicts the four variations of a pair of meaning representations and their corresponding English verbalization.

Strategies	Numeric Referring Expressions			
	Area	Month	Year	
<i>no desc</i>	322.91	4	2021	In <u>April 2021</u> , <u>322.91</u> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE)
<i>desc src</i>	three hundred and twenty-two point nine one	four	two thousand and twenty-one	In <u>April 2021</u> , <u>322.91</u> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
<i>desc trg</i>	322.91	4	2021	In <u>April</u> <i>two thousand and twenty-one</i> , <i>three hundred and twenty-two point nine one</i> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
<i>desc</i>	three hundred and twenty-two point nine one	four	two thousand and twenty-one	In <u>April</u> <i>two thousand and twenty-one</i> , <i>three hundred and twenty-two point nine one</i> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).

Table 10 – The strategies and how they were used to represent the numeric referring expressions. They are *highlighted*.

Tables 11 and 12 depict the error rates and BLEU scores per numerical strategy and model for English and Portuguese, respectively. Across the different types of errors, we see that numerical ones are the most frequent in both languages. The results brought to our attention some hypotheses regarding the differences between the investigated languages (English and Brazilian Portuguese) as presented in Table 13. We infer that our approach cannot significantly improve the generated outputs due to the massive amount of training data of the LLMs’ English versions and a possible bias when referring to numerical expressions.

However, the Brazilian Portuguese language results improved, demonstrating that the approach reduced the number of hallucinations (Ji et al., 2023) and surpassed the lack of data in a Brazilian Portuguese-specific domain. For more information about the numerical referring expression approach, please refer to Appendix A Section.

Error Rate Full Results for English (EN)									
Strategies	LM	Number	Named Entity	Word	Context	Uncheckable	Other	Fluency	BLEU
No Desc	T5	0.48	0.05	0.08	0.03	0.03	0.08	0.05	0.69
	GPT2	0.65	0.28	0.03	0.53	0.08	0.60	0.95	0.14
	BART	0.50	0.00	0.00†	0.08	0.00†	0.00	0.15	0.61
	Avg.	0.54*	0.11	0.03	0.21	0.03	0.23	0.38	0.48
Desc Source	T5	0.45†	0.00†	0.00†	0.00†	0.05	0.00†	0.08	0.69
	GPT2	1.00	0.05	0.00†	0.03	0.00†	0.00†	0.23	0.41
	BART	0.48	0.00†	0.00†	0.05	0.00†	0.03	0.23	0.62
	Avg.	0.64	0.02*	0.00*	0.03*	0.02	0.01	0.18	0.57
Desc Target	T5	0.95	0.00†	0.00†	0.03	0.00†	0.00†	0.00†	0.87†
	GPT2	0.95	0.03	0.00†	0.13	0.13	0.00†	0.18	0.35
	BART	0.90	0.05	0.08	0.05	0.03	0.00†	0.18	0.60
	Avg.	0.93	0.03	0.03	0.07	0.05	0.00*	0.12	0.60*
Desc	T5	0.93	0.00†	0.03	0.03	0.00†	0.00†	0.00†	0.66
	GPT2	0.90	0.13	0.03	0.23	0.03	0.05	0.25	0.28
	BART	1.00	0.00†	0.03	0.03	0.00†	0.00†	0.05	0.58
	Avg.	0.94	0.04	0.03	0.09	0.01*	0.02	0.10*	0.50

Table 11 – English error rates and BLEU score for the 4 numerical strategies and 3 language models - Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.

Error Rate Full Results for Portuguese (PT)									
Strategies	LM	Number	Named Entity	Word	Context	Uncheckable	Other	Fluency	BLEU
No Desc	mT5	0.45	0.02	0.08	0.03	0.03	0.07	0.03	0.58
	GPT2-pt	0.24	0.04	0.03	0.03	0.01	0.14	0.09	0.60
	mBART	0.34	0.04	0.00†	0.09	0.01	0.07	0.10	0.51
	Avg.	0.34	0.04	0.04*	0.05*	0.02	0.09	0.07	0.56
Desc Source	mT5	0.37	0.01	0.08	0.05	0.01	0.04	0.02	0.59
	GPT2-pt	0.19†	0.01	0.02	0.12	0.01	0.08	0.03	0.61
	mBART	0.28	0.03	0.01	0.01	0.01	0.13	0.05	0.59
	Avg.	0.28*	0.01	0.04	0.06	0.01	0.09	0.03*	0.60
Desc Target	mT5	0.87	0.00†	0.04	0.00†	0.00†	0.01†	0.02†	0.65
	GPT2-pt	0.90	0.01	0.04	0.11	0.00†	0.08	0.05	0.64
	mBART	0.79	0.06	0.09	0.09	0.00†	0.10	0.09	0.61
	Avg.	0.85	0.02	0.06	0.07	0.00*	0.06	0.05	0.64
Desc	mT5	0.90	0.00†	0.05	0.02	0.00†	0.01	0.06	0.68†
	GPT2-pt	0.80	0.01	0.12	0.14	0.00†	0.03	0.15	0.67
	mBART	0.89	0.00	0.07	0.03	0.00†	0.03	0.15	0.65
	Avg.	0.87	0.00*	0.08	0.06	0.00*	0.02	0.12	0.67*

Table 12 – Portuguese error rates and BLEU score for the 4 numerical strategies and 3 language models. Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.

Incorrect Number Error Rate							
Strategies	Language Model	English (EN)			Portuguese (PT)		
		Deforestation Monthly	Deforestation Daily	Overall	Deforestation Monthly	Deforestation Daily	Overall
No Desc	T5/mT5	0.50	0.45†	0.48	0.55	0.18	0.36
	GPT2/GPT2-pt	0.65	0.65	0.65	0.00†	0.00†	0.00†
	BART/mBART	0.50	0.50	0.50	0.18	0.09	0.14
	Avg.	0.55*	0.53*	0.54*	0.24*	0.09	0.17*
Desc Source	T5/mT5	0.45†	0.45†	0.45†	0.73	0.07	0.40
	GPT2/GPT2-pt	1.00	1.00	1.00	0.27	0.00†	0.14
	BART/mBART	0.50	0.45†	0.48	0.45	0.00†	0.23
	Avg.	0.65	0.63	0.64	0.48	0.02*	0.25
Desc Target	T5/mT5	0.95	0.95	0.95	1.00	0.68	0.84
	GPT2/GPT2-pt	0.95	0.95	0.95	0.82	0.74	0.78
	BART/mBART	0.95	0.85	0.90	0.82	0.55	0.69
	Avg.	0.95	0.92	0.93	0.88	0.66	0.77
Desc	T5/mT5	1.00	0.85	0.93	1.00	0.68	0.84
	GPT2/GPT2-pt	0.90	0.90	0.90	0.82	0.52	0.67
	BART/mBART	1.00	1.00	1.00	1.00	0.69	0.84
	Avg.	0.97	0.92	0.94	0.94	0.63	0.78
Kappa Statistic		0.94	0.92	0.93	1.00	0.99	0.97

Table 13 – English and Portuguese “Incorrect Number” error rates results, grouped by number description strategies – higher values denote more errors. For comparison, we show only results for Monthly and Daily Deforestation, since these domains are present in both languages.

* Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.

5.3 Conclusion

The second study collaborates with the development of DaMata, a robot journalist based on a pipeline architecture of NLG, which generates and publishes multilingual daily and monthly news reports about deforestation in the Brazilian Amazon on Twitter. With the context enrichment of the REG component by leveraging structured data streams available for the domain, DaMata can provide more diverse information and reach wider audiences. Moreover, due to its rule-based nature, the pipeline framework ensures high-fidelity data publishing and the random choice of template level enhances lexical variety and audience engagement.

For future work, we plan to publish satellite images indicating the deforested areas alongside the generated texts. A further interesting feature to be implemented is a chatbot for the Twitter account – a similar solution to our third approach to context-enrichment (please refer to the following Chapter 6), as we have noticed Twitter recipients tend to react to the posts and seek to engage in conversations with the robot journalists for further information. We also intend to add more languages for our robot to output, intending to raise broader audience awareness about deforestation. Our robot journalism proposal can be used to report on additional domains, particularly sensitive issues that could benefit from ampler dissemination. For further improvement of our system’s performance, we plan to add data-driven techniques to the existing pipeline modules.

6 Enriching an NLG Model from a Human-Centered Approach

The content of this chapter follows two main topics:

Persona-based Corpus and Human-centered NLP model.

Summary of the Third Approach

In this third approach to context-enriched NLG models, we present our proposal in three sections to address the gaps we are trying to overcome. The first section explains the limitations of human-centered approaches to NLG, covering related work, presenting HCD methodologies, and how we apply them in our proposal. The second one introduces the lack of in-domain data, current developments in the field to solve this issue, the intersections of HCD to Text Simplification, and an approach related to the compilation of an enriched corpus with personas¹. Last, the third section correlates NLG with the two previous sections by highlighting a full overview of the proposal, thereby discussing the development of a context-enriched NLG model from an HCD perspective – human-centered NLP model². This thesis author is responsible for developing the personas and applying them to the corpus, conducting human-centered studies and evaluations, developing Python scripts for the corpus’ automatic validation, conducting training sessions about the human-centered approach employed in this study, and other similar tasks. Regarding the third section about the human-centered NLP model, this thesis’ author was responsible for most phases of the design, development, and testing, with the exception of some of the human-centered evaluations. Code and data are publicly available³.

¹ This thesis’ author is responsible for several tasks related to the personas’ development.

² This part of the study was conducted mostly by this thesis author.

³ <<https://github.com/Dia-Bete/DiaBeteBot>>

6.1 Addressing the Gap in HCD Approaches to NLG

In their studies about HCI, [Sharp, Preece and Rogers \(2019\)](#) draw attention to the need for a human-centered approach, more specifically to a *people-centered* approach to development; they extend the core principle of human- and user-centered design to people-centered design, which is understanding people, communities, and other stakeholders. Drawing on an extensive range of sources, the authors emphasize three main principles ([Sharp; Preece; Rogers, 2019](#), p. 48): (i) “[p]eople’s tasks and goals are the driving force behind the development of a product”, (ii) “[p]eople’s behavior and context of use are studied and the system is designed to support them”, and (iii) “[p]eople’s characteristics are captured and designed for”. This view is recurrent in domains that provide vital information and services that affect people’s everyday lives, such as governance and healthcare.

In the same vein, a recent study from [Štajner \(2021\)](#) presented a survey of *Automatic Text Simplification for Social Good* where the author reported some advances within the area and the limitations that research needs to address. Literacy is still a challenge in various populations, languages, and domains ([Štajner, 2021](#), p. 2638). The author emphasized the benefits of multidisciplinary collaboration in order to achieve better results for real end users. Moreover, this thesis addresses the challenge with a combination of HCD, NLG, TS, and context while bearing in mind a *people-centered* approach.

A similar perspective has been adopted by [Broderick et al. \(2014\)](#), who applied a UCD approach to the Health domain. The authors ([Broderick et al., 2014](#), p. 5–6) reported the development of mobile applications (mHealth) which were co-designed⁴, i.e., constructed alongside end-users in order to achieve straightforward communication between experts and lay users. The results of their research pointed out a way of maximizing the potential of mHealth applications in order to ensure health literacy⁵ for users of all proficiency levels. Their study was followed by

⁴ As [Nielsen \(2011, p. 1\)](#) states: “[c]o-design is collective creativity that is applied across the whole span of a design process. This means that co-design is a specific instance of co-creation”. She also adds that “[c]o-creation exploits that the users possess knowledge about their own needs and daily life and their ability to be creative” ([Nielsen, 2011, p. 1–2](#)).

⁵ According to the Affordable Care Act ([Act; Representatives, 2010](#)), health literacy is defined as “the degree to which an individual has the capacity to obtain, communicate, process, and

similar approaches as the works of (Vosbergen et al., 2015; Holden et al., 2017; Smith et al., 2017; Woods et al., 2017; Dew et al., 2018; Hughson et al., 2018; Trujillo; Senette; Buzzi, 2018; Shin; Holtz, 2019; Chiang et al., 2023). For more information about this kind of research, please refer to Noorbergen et al. (2021).

Despite the importance of the human-centered approach, there remains a paucity of evidence on combining this approach during the full cycle of the development of NLG models, specifically for multipurpose models, which are part of NLG and conversational AI applications. In their study, Bateman et al. (2019, p. 10) suggest that applying a human-centered artifact known as *persona* during the development of conversational AI systems could be pertinent to the field as a result of “making sure that what is said is appropriate both to the context and to the state of knowledge of the addressee”. This view is supported by Manishina (2016, p. 130) who writes that NLG is a human-driven field, where the interaction between the user and the NLG system is a fundamental part of the development of these kinds of systems where there is a strong need of “human likeness”. Manishina argues that the concern with intuitive and human-driven software is crucial during the conception of these language models.

An informative chatbot developed in Brazilian Portuguese within the Government domain to promote culture was TAIS (Virtual Assistant for the Culture Incentive Law – *Assistente Virtual da Lei de Incentivo à Cultura*)⁶, supported by LAPPIS (*Laboratório Avançado de Produção Pesquisa e Inovação em Software*)⁷. The development followed best practices of a UCD approach using personas⁸. This lab is coordinated by Ph.D. Prof. Carla Silva Rocha Aguiar⁹ who created a user-friendly template – Rasa Boilerplate¹⁰ for creating FAQ Chatbots for people without Programming skills (Lacerda, 2019) at UnB (*Universidade de Brasília* –

understand health information and services in order to make appropriate health decisions”.

⁶ TAIS (*Assistente Virtual da Lei de Incentivo à Cultura* – Virtual Assistant for the Culture Incentive Law) – <<https://lappis-unb.github.io/tais/>>.

⁷ LAPPIS (*Laboratório Avançado de Produção Pesquisa e Inovação em Software* – Laboratory for Advanced Software Production, Research, and Innovation) – <<https://github.com/lappis-unb>>.

⁸ <<https://github.com/lappis-unb/tais/wiki/Estudo-sobre-melhores-pr%C3%A1ticas-de-bots>>

⁹ Ph.D. Prof. Carla Silva Rocha Aguiar – <<http://lattes.cnpq.br/2831991076751452>>

¹⁰ <<https://github.com/lappis-unb/rasa-ptbr-boilerplate>>

University of Brasília).

In the same vein, conversational systems developed for specific domains have presented several directions to promote better interactions with the end user, as the work of [Yadav et al. \(2022\)](#) that uses summarization in order to generate question-type driven information. Next, some developments in conversational AI have presented user-centered solutions to health ([Molen et al., 2017](#); [Yui et al., 2023](#)) in order to discuss solutions regarding the existing drawbacks of the expert and non-expert relationship. In addition, a recent study conducted by [Majumder \(2023\)](#), highlights the necessity of building *user-centric* AI systems within an interactive explainability foundation (an intersection of knowledge, explanations, and interactions) and is the nearest study to the goals of this thesis.

6.1.1 Applying HAI and HCD Methodologies to NLG

In this study, we correlate HCD and UCD with Human-Centered AI (HAI), specifically some directions followed by eXplainable Artificial Intelligence (XAI)¹¹, in order to adjust our model to the key actors of this study. We investigate how to enrich an NLG model in different parts of their design, placing their end-users in a central position and bearing in mind how to engage users effectively with the outputs of this model. We explore human-computer interaction from the angle of natural language, more specifically from a human-centered NLP¹² ([Yang, 2023b](#)) perspective, which according to [Yang \(2023a, p. 10\)](#) “seeks to create NLP systems that are more user-friendly, accessible, and inclusive”.

In line with the human-centered NLP’s main concern, the generation of texts that target human interlocutors should reflect their language choices within a specific context. In order to achieve this goal from a human-centered perspective, we follow some guidelines from the human-centered XAI framework ([Ribera; Lapedriza, 2019](#)) presented in Figure 11, where the key actors from our study (developers/AI researchers, domain experts, final and lay users) are given the consideration in

¹¹ According to [Xu \(2019, p. 44\)](#), explainable AI “enables users to understand the algorithm and parameters used, which is intended to address the AI black-box problem”.

¹² The Human-centered NLP course ([Yang, 2023a](#)) at Stanford University led by Ph.D. Prof. Diyi Yang is resourceful regarding the intersections of HCI and NLP, i.e., Human-centered NLP. For more information, please refer to <http://web.stanford.edu/class/cs329x/>.

several phases of the development and design of the context-enriched NLG model. According to the framework, we can classify these groups as *Developers* and *AI researchers* (software developers, data analysts, and researchers involved in creating the AI system, i.e., the NLG model); *Domain experts* (specialists within the diabetes mellitus domain where the decisions made by the system belong to); and *Lay users* (final users of the NLG model).

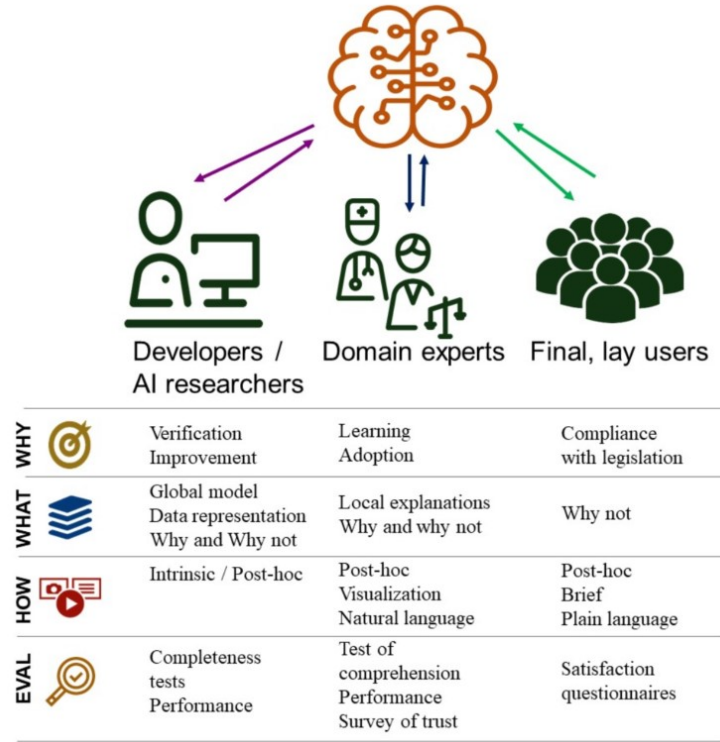


Figure 11 – Human-centered AI framework highlighting groups of users (Ribera; Lapedriza, 2019).

Instead of applying Ribera and Lapedriza (2019) framework to explanations to each step of the context-enriched NLG model generation (i.e., where the model pays attention to or the nuances of each iteration)¹³, we considered the output generated texts. Furthermore, the authors target their explanations to the previously mentioned groups as we did with the outputs of our NLG model. In this study, we

¹³ To understand the context of XAI in this thesis, please refer to Section 3.4.2. Furthermore, explanations and their nuances can be clarified by Arrieta et al. (2020) discussion about XAI.

considered two groups of domain experts: those who are directly involved with the NLG model’s development (i.e., Empoder@’s research team) and users who interact with the application of our NLG model, that is, the final users of the conversational question-answering system (CQA).

In this context, we validated the model’s outputs from three different perspectives. First, developers/AI researchers evaluated the outputs in terms of detecting system failures to ensure that the information fits end users. These researchers can understand programming language and NLG metrics, testing different scenarios to evaluate the generated texts. Second, domain experts validated the generated texts in relation to the context of the model (i.e., the need to adjust to the domain’s terminology) and performed a human evaluation considering some key features, such as fluency, adequacy, and grammatically (i.e., characteristics designed to facilitate users’ comprehension). Finally, with the model’s context enrichment and evaluation during each iteration of the NLG model development, the generated outputs are simplified versions designed to meet lay users’ needs and consider a human-centered perspective throughout the process.

Based on the demographic data collected from the Empoder@ Project (Chaves et al., 2017), the members of our research group; and the Ribera and Lapedriza (2019) framework, we added a complementary user-centered approach in order to enrich our model: *persona* – that refers to a “fictitious user described with basis in data” (Nielsen, 2019, p. 1) and used during all process of developing our third approach, presented in Figure 12. We addressed this need for a more user-driven generation by applying a group of personas that are more frequently encountered within a specific domain: expert and lay users. In addition, we reiterate the benefits of using personas presented in the Background Chapter 3, which is creating a common reference (e.g., the lay user can be referred to the Empoder@ lay user persona instead of an assumption of who this lay user could be) to guide our research team and engage them with the end-users of the diabetes mellitus domain. According to Harley (2015), “[f]raming a statement around a specific persona breaks the listeners out of self-referential thinking and removes the speaker’s reliance on opinions, shifting the discussion away from personal judgments toward that particular persona’s needs”. Recent studies regarding personas (Nielsen, 2019;

Salminen et al., 2020; Salminen et al., 2021; Yang, 2023a) reiterate the benefits of using them within an NLP context.

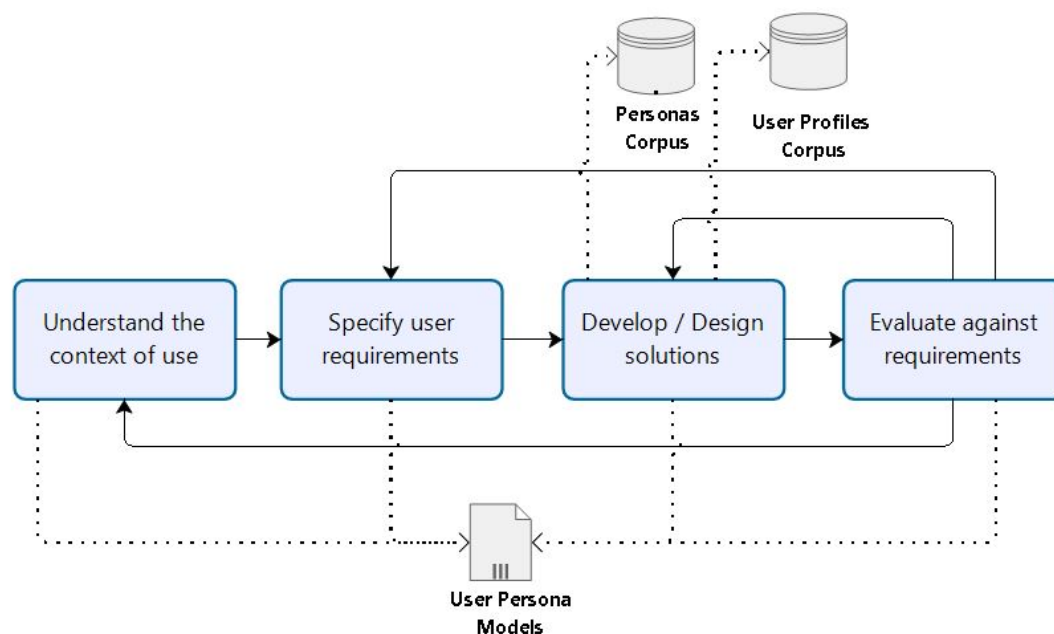


Figure 12 – User-centered design and how to integrate personas and corpora – adapted from Soegaard and Dam (2015).

Another advantage highlighted by Suojanen, Koskinen and Tuominen (2015b, p. 135–136), within a translation context¹⁴, when applying personas to translated texts was finding real-life examples of personas being used in translation work. The authors applied their created personas in Translation and Master’s courses where they emphasized the personas’ use as a methodology “to discuss translation strategies, the information structure of a text, typography, and many other features of translations from the perspective of this persona’s needs and preferences”.

Within this proposal, we applied our created personas (please refer to the next Subsections) from two complementary perspectives. The first one is during the process of developing the context-enriched NLG model: design, implementation, and development phases as well as the compilation of the corpus (as highlighted in Figure 12). And the second one addressed the texts generated from our context-

¹⁴ Please refer to Text Simplification Section 3.2 to TS approaches.

enriched model, by contrasting them with each persona. For instance: *Eris Carvalho*, a lay persona, is described as a middle-aged woman who does not know if she has diabetes, she also has problems finding information due to her sight and technology limitations (please refer to Figure 13).

According to Salminen et al. (2021, p. 1685), data-driven personas are created “from the behaviors and demographics of user segments”, in line with this, we developed a persona based on the demographic data and their persona’s demographic groups identified in the results of the studies conducted by the Empoder@ Project (Pagano, 2015; Chaves et al., 2017; Vieira et al., 2017). For more information, please refer to Pagano (2015), Chaves et al. (2019), Neves et al. (2021), Cortez et al. (2022), and Nunes et al. (2023). Consequently, during the human-centered evaluation of the context-enriched NLG model outputs, some samples were validated by domain experts from the Empoder@ project, who verified if the generated text met the persona. Additionally, Suojanen, Koskinen and Tuominen (2015b) 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”. For more information about data-driven personas, please refer to Suojanen, Koskinen and Tuominen (2015b) and Salminen et al. (2021).

In the next subsections, we present the personas created for the diabetes self-care domain and their application within the context-enriched NLG model development scenario.

6.1.2 Final/Lay User Persona: Eris

Our first persona is a lay user in the context of the diabetes mellitus domain. We can see more characteristics of this persona in Figure 13. We highlight key features of this persona as biography, goals, and frustration, in the following subsections. Furthermore, we present the proficiency level within the domain, which is the need for simplified texts.

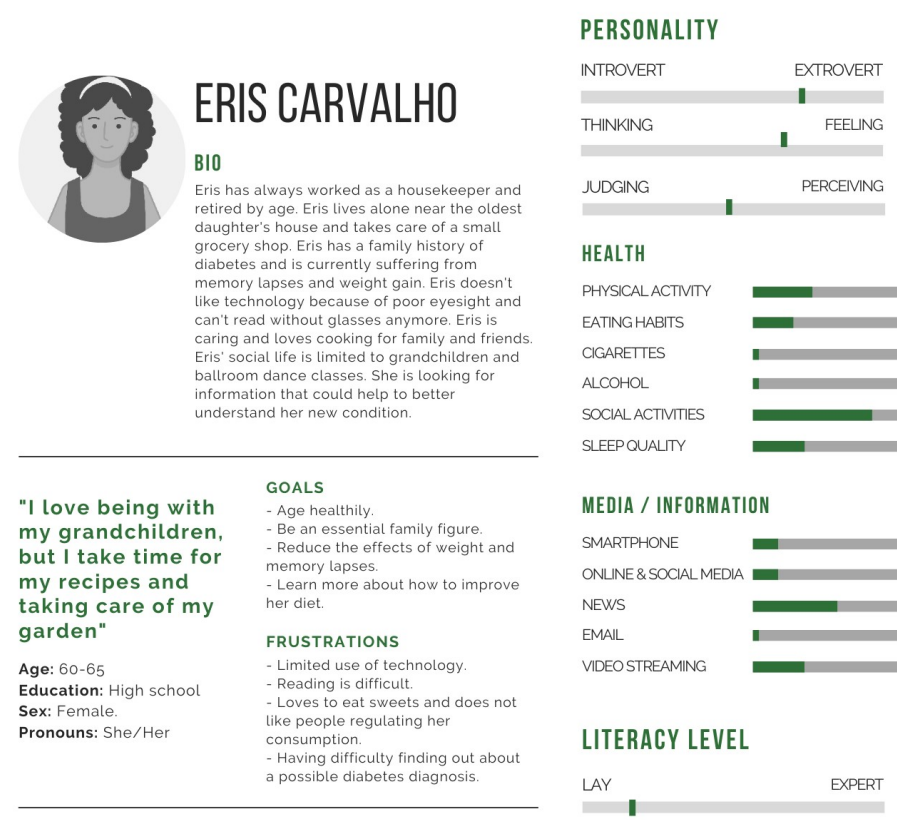


Figure 13 – Final/Lay-user persona: Eris (created with Canva (2023)).

6.1.2.1 Eris' Biography

Eris has always worked as a housekeeper and retired by age. Eris lives alone near the oldest 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 like technology because of poor eyesight and can't read without glasses anymore. Eris is caring and loves cooking for family and friends. Eris' social life is limited to grandchildren and ballroom dance classes. She is looking for information that could help to better understand her new condition. More information about Eris can be seen in Figure 13.

6.1.2.2 Eris' Goals

- Age healthily.
- Be an essential family figure.
- Reduce the effects of weight and memory lapses.
- Learn more about how to improve her diet.

6.1.2.3 Eris' 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.

6.1.3 Final/Lay User Persona: Otto

Our second persona is a software developer and lay user in the context of the diabetes mellitus domain. We highlight that even though this persona is an expert, he is not an expert within the Diabetes Mellitus domain. Please refer to this persona in Figure 14. Moreover, the proficiency level addresses mid-texts, an overlap between simplified and more technical texts in the domain.



Figure 14 – Final/Lay-user persona: Otto (created with Canva (2023)).

6.1.3.1 Otto's Biography

Otto lives in Belo Horizonte in a student's residence near the university, where he studies Computer Science and internships as a developer. He came across the diabetes project through his professor. Otto is familiar with the basic concepts of Machine Learning and was trying to work on 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 not accepting her new condition. Otto is looking for data to help him better understand why people with disabilities have not mobilized about this situation. He is also fond of reading and self-developing in different ways. During the weekends, Otto prefers to go hiking and running with a group of friends. More information about Otto can be seen in Figure 14.

6.1.3.2 Otto's Goals

- Help develop a system to prevent patients' illnesses.
- Social networking.
- Exercising discipline.
- Develop his own app.

6.1.3.3 Otto's Frustrations

- Managing stress and multitasking abilities.
- Experiment with a better way of taking care of different projects during the day and managing his studies.
- Working in different locations.
- Help his sister to manage her diabetes.

6.1.4 Domain Expert Persona: Alexis

Our third persona is a *domain expert* user in the context of the diabetes mellitus domain. Please refer to this persona in Figure 15. This persona's key features are highlighted: biography, goals, and frustrations. Moreover, the proficiency level requires more technical texts within the domain.



Figure 15 – Domain Expert Persona: Alexis (created with [Canva \(2023\)](#)).

6.1.4.1 Alexis' Biography

Alexis has worked with diabetes type 1 and 2 patients for the last five years and has become the project coordinator at a municipal hospital in Northeast Brazil. She also clinics at a private office 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 book vacations. Alexis is 20 weeks pregnant and recently discovered gestational diabetes. She realized the lack of published information about the subject, especially regarding diet, and has decided to investigate how to promote more information online. More information about Alexis can be seen in Figure 15.

6.1.4.2 Alexis' Goals

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

6.1.4.3 Alexis' 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.

6.2 Addressing the Lack of In-Domain Data

Considering NLG challenges that are associated with the technological shift to neural networks (NN), several studies indicated that this high-tech context has expanded and diversified the possibilities when addressing NLG models. Whereas traditional approaches such as statistical (e.g., phrase-based models) and rule-based performed well for the last decades, neural NLG (NNLG) models have been applying statistical models' learned lessons and have evolved to more engineered solutions. Moreover, a further comprehensive study would include neural network models more driven to HCI methodology in order to follow the current changing context.

In general, NLP models and applications need to overcome the lack of resource data (Koehn; Knowles, 2017; Junczys-Dowmunt et al., 2018) to both specific domains and languages. Commenting on the available data, Knowles, Bates and Håkansson (2018) argue that large volumes of corpora give you both coverage and ambiguity. As a consequence of this, generic NLP tools work pretty well on common topics but demonstrate a lack of good results on in-domain fields.

Many recent studies (Manishina, 2016; Chu; Wang, 2018; Moussallem; Wauer; Ngomo, 2018; Zhang et al., 2018b; Almansor; Hussain, 2019) have shown that domain adaptation models are categorized by their main models, i.e., data-driven (data-centric) or model-driven (model-centric). According to Chu and Wang (2018), data-driven models target training data selection or creating synthetic data, whereas model-driven models shift between in-domain and out-of-domain models. Here we are proposing a user-driven approach in order to enrich an NLG model, that is a mixture model that combines both domain-adaptation approaches.

According to Koehn and Knowles (2017) and Koehn (2020), one of the challenges of using in-domain data is related to the amount of data that is necessary to train NLG models. Lample et al. (2018) also highlight the need for large amounts of parallel sentences in order to reach a human-level parity. Another problem is relevant to the words that only appear in out-of-domain data due to the number of occurrences and the massive amount of data. In order to overcome this issue in the domain of diabetes mellitus self-care, where the scarcity of data is a fact, we propose the use of one of the available methods to train NLG models with out-of-domain data and then apply the in-domain data. However, Koehn (2020, p. 141) highlights that “we have to be careful that the out-of-domain data does not drown out the in-domain data”.

Recent studies in domain adaptation point to a mixture of methods in order to satisfy multiple domains as (Blackwood; Ballesteros; Ward, 2018), or to apply external knowledge features in a possible way to improve the NLG results as the studies of (Sennrich; Haddow, 2016; Arcan; Buitelaar, 2017; Ataman; Federico, 2018), and (Vilar, 2018). On the one hand, Li et al. (2016) encode personas as a combined set of individual characteristics within their model as a straightforward mechanism. In the same vein, Chu, Vijayaraghavan and Roy (2018) argue that training NLG models with external knowledge (e.g., persona embeddings) would benefit applications where custom texts are necessary. Moreover, they highlight the relevance of training and improvements in the model’s performance. Furthermore, Chu, Vijayaraghavan and Roy (2018) believe prior knowledge could be keen to benefit different domains.

On the other hand, Chatterjee et al. (2017) state that models are difficult

to control when adding external knowledge. In order to overcome this drawback, they introduced a method to compute particular terminology options and later guarantee an output suggestion appropriately. The authors justify their approach on account of the cost of neural networks training which was sustained by satisfactory performance when working with out-of-vocabulary terminology. Similarly, [Wang et al. \(2017\)](#) exploit sentence embedding and selection of similar sentences so the model would be close to in-domain data. They argue that the proposed methodology of sentence selection with additional training has resulted in fast convergence.

Other types of hybrid approaches could include unsupervised techniques, such as the works of [Zhang et al. \(2017\)](#), [Chu and Dabre \(2018\)](#), [Lample et al. \(2018\)](#), and [Marie and Fujita \(2018\)](#). Overall, these studies highlight an innovative methodology in order to train NLG systems counting only on monolingual corpora. While [Tars and Fishel \(2018, p. 1\)](#) hold the view that considering domains as distinct languages and using multilingual approaches would play a central role in building multi-domain models.

6.3 Addressing the Lay-Expert Communication Gap

Simplification aims to transform the 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 diverse 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; Gasperin, 2010](#); [Pagano, 2015](#); [Araujo, 2021](#); [Oliveira, 2022](#)).

Recent TS research has brought to attention the challenges behind the task itself ([Alva-Manchego; Scarton; Specia, 2020](#); [Al-Thanyyan; Azmi, 2021](#)). Even though TS is not a new task within NLG, there is still room for research concerning simplification in an inclusive way ([Stajner, 2021](#)). In relation to in-domain simplification which is one of the pillars of this thesis, several works have been pointing directions to a better outcome.

This study focuses on sentence simplification (SS) ([Alva-Manchego; Scarton; Specia, 2020](#)) within a hybrid approach (please refer to TS Background Section

3.2 for more information) 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 – CLEAR¹⁵. 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 conclude their work by emphasizing the endeavor while performing semantic annotation. The CLEAR corpus served as the basis for the development of a French biomedical text simplification corpus (Cardon; Grabar, 2020).

A Portuguese medical corpus in the domain of Parkinson’s disease was developed by Zilio et al. (2020) in conjunction with the development of a simplification web-based tool: MedSimples¹⁶. The corpus was created in order to evaluate MedSimples and contains a total of 570 annotated sentences and 2080 annotated instances. This tool was developed by the Federal University of Rio Grande do Sul (UFRGS) and received Google LARA (Latin America Research Awards) in 2019. MedSimples is a lexical simplification tool for promoting health literacy that works with general and specialized vocabularies. The tool is under development but has presented potential results and has extended the corpus to more domains, such as COVID-19 and Pediatrics.

Devaraj et al. (2021) revisit the medical domain by an automated simplification with a fine-tuning approach. The authors introduced an English dataset paired with technical and simplified texts derived from the Cochrane database (McIlwain et al., 2013)¹⁷. The results pointed out satisfactory results when training BERT (Devlin et al., 2019) and SciBERT (Beltagy; Lo; Cohan, 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; Alva-Manchego, 2022);

¹⁵ <<http://natalia.grabar.free.fr/resources.php>>

¹⁶ <<https://www.ufrgs.br/textecc/acessibilidade/page/cartilha/>>

¹⁷ <<https://consumers.cochrane.org/PLEACS>>

(ii) EASIER corpus composed of 260 annotated documents in Spanish (Alarcon; Moreno; Martínez, 2023); (iii) TESLEA, an automated corpus of medical text simplification fine-tuned with pre-trained models (Phatak et al., 2022); and some 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 MEDIQA¹⁸ at the 5th Clinical Natural Language Processing Workshop¹⁹, performed summarization and simplification tasks within medical clinical notes. The results presented new resources as multilingual benchmarks and discussions about the development of medical NLP (Abacha et al., 2023).

In the next section, a description of the procedures and methods used to compile a persona-based corpus will be presented.

6.4 Persona-Based Corpus

The use of the HCD approach has grown since 2015 (Vinyals; Le, 2015; Serban et al., 2018; Ram et al., 2018) with the development of NLP and NLG models that incorporate their engines: personas (Li et al., 2016; Wolf et al., 2019), user profiles (Cao; Cheung, 2019), and/or corpus built with the use of personas/profiles (Zhang et al., 2018a; Mazaré et al., 2018), during the training of these NLP applications. With the intent of providing a resource that could help the development of these kinds of systems and also surpass the lack of in-domain data, we developed a corpus named Dia(Bete)²⁰, i.e., a “collection of machine-readable texts compiled with a specific purpose that can be retrieved with particular computer software for linguistic research” (Lan, 2015, p. 465), 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.

¹⁸ MEDIQA is a series of shared tasks on Medical NLP – available at <<https://sites.google.com/view/mediqa2023/home>>

¹⁹ ACL–ClinicalNLP 2023 – <<https://clinical-nlp.github.io/2023/>>

²⁰ Dia(Bete) is the main project’s name at Empoder@ regarding the NLP developments to the Diabetes Mellitus domain. During the research group sessions, we also addressed the project by its nickname *Betinha* (Bete is a short version of the proper name *Elisabete* – very common in Brazil).

Additionally, due to the difficulty in finding an automatic solution that could provide domain-adapted versions of texts to end-user literacy, as can be seen in the example presented in Table 14, we first compiled the persona-based corpus manually following similar approaches (Grabar; Cardon, 2018; Oraby et al., 2019; Ramos; Monteiro; Paraboni, 2020) to use it in the training phase of the context-enriched NLG model.

Within the Empoder@ project, several studies tried to explore and solve the communication difficulty between patients and medical staff. By conducting an interdisciplinary approach, the project covered different areas, such as Applied Linguistics, Computer Science, Human-Computer Interaction, Nursing, Nutrition, and Medicine, to address the communication concerns related to language per se encountered regarding experts and non-experts in the Diabetes Mellitus domain (Pagano, 2015; Chaves et al., 2017; Chaves et al., 2019; Oliveira, 2022; Nunes et al., 2023).

Regarding simplification, we follow the project called PorSimples²¹ (Aluísio et al., 2008) developed by Ph.D. Prof. Sandra Maria Aluísio at the University of São Paulo (USP). This project was able to produce a substantial amount of NLP technologies related to Text Adaptation (TA) in relation to 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). Furthermore, Aluísio and Gasperin (2010, p. 46) highlight the main purpose of the PorSimples project:

“It aims at developing technologies to make access to information easier for low-literacy individuals, and possibly for people with other kinds of reading disabilities, by means of Automatic Summarization, Lexical Simplification, Syntactic Simplification, and Text Elaboration”.

PorSimples project intersects with the Empoder@ project concerning the generation of easier-to-read texts with the advent of shortening the gap within the

²¹ **PorSimples** – Simplification of Portuguese Texts for Digital Inclusion and Accessibility. For more details, please refer to <http://www.nilc.icmc.usp.br/nilc/index.php/porsimples>.

Persona	Texto de exemplo
Experto	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 .
Leigo	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 .
Persona	Example of text
Expert	In ketoacidosis, the symptoms are polyuria (increased urine), polydipsia (increased thirst), increased respiratory rate , nausea and vomiting, ketonic breath , cephalaea (headache), mental confusion, and abdominal pain .
Non-expert	In ketoacidosis, the symptoms are polyuria (increased desire to urinate), polydipsia (increased thirst), rapid breathing , nausea, vomiting, and “rotten apple” breath , cephalaea(headache), mental confusion, and pain in the region of the abdomen .

Table 14 – Samples of expert and non-expert texts in the Diabetes Mellitus domain (Original text in Brazilian Portuguese and the corresponding English translation) – Simplifications are **boldfaced**.

doctor-patient relationship. Through more accessible communication, the patient can better understand the information necessary for self-care. Language becomes a fundamental tool for accessing self-care and increases learning about their diabetes condition (Pagano, 2015; Oliveira, 2022).

6.4.1 Corpus and Annotation Process

This section mainly focuses on the development of the corpus and how an HCD method, i.e., the personas (expert and lay users), was able to guide the development, compilation, annotation, 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 mellitus) based on information from end-users in the data collected from the Empoder@ project in the context of the Brazilian Portuguese language.

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

Table 15 – Persona-based corpus compilation criteria.

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

6.4.2 Corpus

As presented in Personas’ Section 6.1.2, our corpus is oriented by the personas constructed with the data collected from the Empoder@ project. 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” (McEnergy; Hardie, 2011, p. 19), specifically, experts and non-experts from the diabetes mellitus domain. Besides, we need an intralingual corpus, which according to Jakobson (2000, p. 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. Finally, in order to achieve the aforementioned corpus characteristics, we follow corpus-based translation criteria (Fernandes, 2004; Fernandes, 2006) to compile the corpus as can be seen more closely in Table 15.

6.4.3 Compilation and Automatic Correction

One of our objectives was to create a corpus to satisfy the automatic question generation and the Conversational AI assistant development needs, as

described in the Conversational AI Section 3.3. In order to achieve this goal from an HCD perspective, we followed the corpus compilation pipeline shown in Figure 16. First, texts (comments/questions) were collected from discussion groups and forums from social media in the Brazilian Portuguese language that is related to the diabetes mellitus domain. This task was done by members of our research group, who were responsible for collecting the text mainly manually. Second, they used semi-automatically scripts created in Python to crawl texts from a web page. Third, we applied a semi-automatically pre-processing and cleaning phase, where we removed numerous noisy words from the collected corpus such as repeated punctuation, emojis, wrong capitalization, and misspelled words. Finally, the corpus was de-identified, post-edited, and annotated as will be discussed in the next sections.

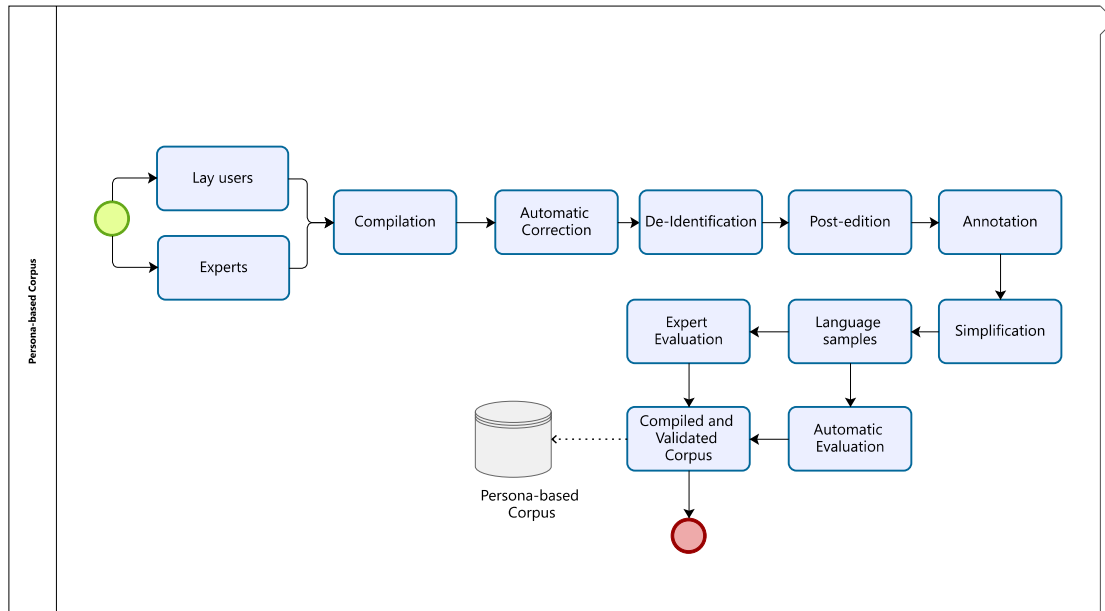


Figure 16 – Persona-based corpus compilation of lay and expert texts (created with Bizagi (2023)).

6.4.4 De-Identification

To maintain individuals' privacy, we follow Brazil's General Data Protection Law (Lei Geral de Proteção de Dados – LGPD)²² established in 2018, regarding privacy and data protection guidelines – an analogous regulation is the European Union's General Data Protection Regulation (GDPR)²³. LGPD provides comprehensive guidelines on how to process personal data, individual rights, and obligations. In our context, the de-identification process (also called *anonymization*) follows steps to prevent re-identification, reflecting the balance between privacy protection and data utilization within a sensitivity-information domain (Berman, 2016; Liu et al., 2017; Pagano et al., 2023). The de-identification constituted of 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 [age–5 or age+5], and relative mentions followed similar age's 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)²⁴;
4. Some paraphrasing (see Text Simplification Section 6.3 for more information).

6.4.5 Post-edition

During the post-editing phase, the original text was classified according to a linguistic proficiency scale created by LETRA linguists researchers. Then, some

²² Lei Geral de Proteção de Dados Pessoais (LGPD) – Lei nº 13.709/2018. Available at <https://www.gov.br/mds/pt-br/aceso-a-informacao/lgpd>

²³ European Union's General Data Protection Regulation (GDPR) – available at <https://gdpr-info.eu/recitals/no-159/>

²⁴ This kind of markers can be enough to identify the author (e.g., the repetition “GOOD MOOOOORNING!!!!” has fewer occurrences than its usual form “Good morning!” (Berman, 2016; Liu et al., 2017; Pagano et al., 2023).

improvements from the de-identification phase and the remaining punctuation issues were adjusted. The proficiency scale's creation followed Literacy and Simplification directions from:

- Functional Linguistic literature (Szmrecsanyi, 2004; Halliday; Matthiessen, 2014) regarding complexity – node counts, word counts, and a so-called ‘Index of Syntactic Complexity’;
- the guidelines of the Manual for Portuguese Simplification (Specia; Aluísio; Pardo, 2008; Aluísio et al., 2008) developed by NILC²⁵ during the PorSimples project (see Section 6.3 for more information);
- the National Indicator of Functional Literacy (INAF – *Indicador Nacional de Alfabetismo Funcional*)²⁶.

The INAF index categorizes literacy into five main categories for the Brazilian population, as presented as follows:

1. **Illiteracy**: the condition of those who cannot achieve simple tasks involving reading words and phrases, even though a portion of them can read family numbers (e.g., telephone numbers, house numbers, and prices);
2. **Literacy – rudimentary level**: the ability to locate explicit information, expressed in a literal way, in short texts, composed essentially of sentences or words that explore familiar everyday situations, such as advertisements, and it can compare, read and write familiar numbers (e.g., schedules, prices, bills/coins, telephones);
3. **Literacy – basic level**: the ability to locate one or more chunks of information in medium-length texts, solve basic operations with numbers, and also recognize the meaning involving situations of everyday domestic or social context;

²⁵ NILC – *Núcleo Interinstitucional de Linguística Computacional* (The Interinstitutional Center for Computational Linguistics) at <https://sites.google.com/view/nilc-usp/>.

²⁶ *Indicador Nacional de Alfabetismo Funcional* (National Indicator of Functional Literacy) – available at <https://alfabetismofuncional.org.br/>

Linguistic Literacy Scale	
1	very poorly structured comment/question
2	comment/question poorly structured and/or with many errors
3	comment/question reasonably well structured and with errors
4	well-structured comment/question with errors (especially related to the domain)
5	well-structured and error-free comment/question (or just one that is easy to identify)

Table 16 – Linguistic literacy criteria – based on NILC research (Specia; Aluísio; Pardo, 2008; Aluísio et al., 2008).

4. **Literacy – intermediate level:** the ability to find information expressed literally in medium-length texts (e.g., journalistic, scientific), solve mathematical problems involving percentage and proportion, interpret and prepare a synthesis of different texts, and make small inferences;
5. **Literacy – proficient level:** the ability to read long texts of high complexity (e.g., message, description, exposition, or argumentation), find multiple types of information related to tasks in different contexts, compare different texts, recognize effects of meaning and perform inference.

Additionally, the linguistic proficiency information is directly related to the designed personas and is expected to help the training of the NLG system in order to identify experts and non-expert users, as well as generate text accordingly. The researchers involved during the post-edition phase participated in an HCD workshop in order to use the personas to help and/or validate each occurrence of the corpus. Based on LETRA’s previous research and HCD background, the proficiency scale was designed by the research group following a Likert scale format and ranges from 1 to 5 as presented in Table 16.

6.4.6 Annotation

The corpus annotation is part of a broader study conducted by our research group dedicated to the Brazilian Portuguese Named Entity Recognition (NER) and Relation Extraction (RE) within the Diabetes Mellitus domain (Pavanelli, 2022). The annotation had five main phases: setup, guidelines, entity types design, and

annotation process. A summary of the main characteristics that intersect with the purpose of the persona-based corpus is described as follows.

First, we conducted two parallel annotations with the Label Studio tool²⁷ after each iteration of the corpus compilation. In the first annotation, we considered the pre-processed version of the corpus where researchers annotated the intent and slot-filling according to some relevant features of the domain: date, time, food, glucose, symptom, complication, and medication. We can see an example of a post in Brazilian Portuguese in Table 17 and the corresponding annotation presented in Figure 17.

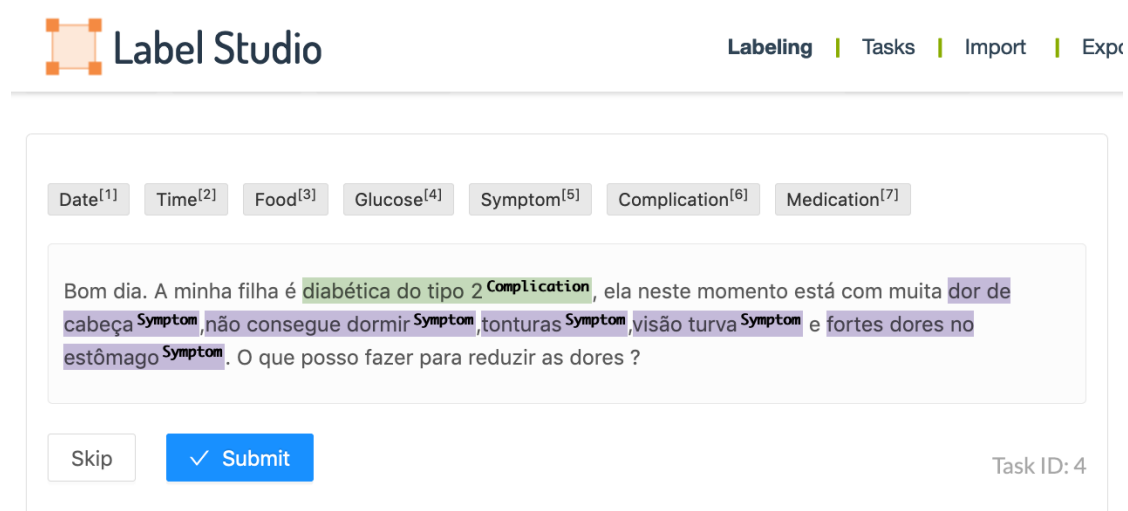


Figure 17 – Intent and slot-filling annotation of social media post in the Diabetes Mellitus domain (Brazilian Portuguese).

²⁷ <<https://app.labelstud.io/>>

Language	Question
Brazilian Portuguese	<i>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?</i>
English	Good Morning. My daughter is type 2 diabetic, she is currently having episodes of intense headaches, she can't sleep, dizziness, blurred vision, and severe stomach pains. What can I do to help her alleviate her headache?

Table 17 – Annotation sample from our corpus in Brazilian Portuguese and its translation to English.

In the second annotation, we selected the original version of the corpus, considering all emojis and extra punctuation, in order to annotate the posts in relation to a sentimental 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; Wessler; Avioli, 1972) and according to their level of knowledge regarding the diabetes mellitus domain.

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 18 – Entities description and examples according to Figure 17 – adapted from Pavanelli (2022, p. 26).

Entities were annotated according to the ontology proposed by Abacha and Zweigenbaum (2015) and some examples of these entity types are presented in Table 18 accordingly. A summary of the Dia(Bete)²⁸ corpus entities is presented in Figure 18 as well as their relations in numbers in Figure 19.

²⁸ Please refer to NER Bete corpus at <https://github.com/pavalucas/Bete>

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 18 – 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 19 – Bete Relations: Number of occurrences, percentage, average, and standard deviation annotation per document sorted in decreasing count order – adapted from Pavanelli (2022).

6.4.7 Simplification

This study focuses on sentence simplification (SS) (Alva-Manchego; Scarton; Specia, 2020) within a hybrid approach (please refer to TS Background Section 3.2 for more information) in order to develop a simplification booklet following the guidelines of Szmrecsanyi (2004) concerning syntactic complexity, the steps of the Manual for Portuguese Simplification (Specia; Aluísio; Pardo, 2008) developed by NILC, and the INAF index (please see more information in the Post-edition Section 6.4.5).

Following the documentation provided by the PorSimples project, our research team developed a booklet for project researchers at the Empoder@ project and students of the graduation course in Letters at UFMG in order to guide them within the simplification task carried out in the persona-based corpus. In addition to the personas presented in Section 6.4, we followed the simplification booklet to simplify the corpus (please refer to the Appendix B for the full simplification booklet) and Gasperin, Maziero and Aluisio (2010) guidelines presented in Table 19. The booklet guidelines are summarized as follows:

#	Cases	Ph	Rules
1	Passive voice clauses	1	Identify passive voice by syntactic tags.
		2	Transform passive into active voice.
2	Apposition	1	Identify apposition by syntactic tags; Identify the anchor of the appositive clause.
		2	Remove apposition from the original sentence; Create new sentence containing anchors subject, “é” (is) as a verb, and the appositive clause as predicate.
3	Asyndetic coord. clauses	1	Identify pairing of clauses by syntactic tags.
		2	Split original sentence into one sentence per clause.
4	Additives coord. clauses	1	Identify pairing of clauses by syntactic tags; Identify additive conjunctions by PoS tags.
		2	Split original sentence into one sentence per clause.
5	Adversative coord. clauses.	1	Identify adversative conjunctions by token matching.
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by “Mas” (But).
6	Correlated coord. clauses	1	Identify correlation conjunctions by token matching.
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by “Também” (Also).

Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from Gasperin, Maziero and Aluisio (2010).

#	Cases	Ph	Rules
7	Result coord. clauses	1	Identify result conjunctions by token matching and PoS tags.
		2	Split original sentence; Create a sentence for the main clause; Create a for the coordinate clause starting by “ <i>Como resultado</i> ” (As a result).
8	Explanatory coord. clauses	1	Identify explanation conjunctions by token matching and PoS tags.
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the coordinate clause starting by “ <i>Isto ocorre porque</i> ” (This happens because).
9	Causal sub. clauses	1	Identify causal conjunctions by token matching.
		2	Split original sentence; Create a sentence for the subordinate clause; Create a sentence for the main clause starting by “ <i>Com isso</i> ,” (Given this).
10	Comparative sub. clauses	1	Identify pairing of clauses by syntactic tags; Identify additive conjunctions by PoS tags.
		2	Split original sentence into one sentence per clause.
11	Comparative sub. clauses	1	Identify comparison conjunctions by token matching.
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the subordinate clause starting by “ <i>Também</i> ” (Also).
12	Concessive sub. clauses	1	Identify concession conjunctions by token matching.
		2	Split original sentence; Create a sentence for the subordinate clause; Create a sentence for the main clause starting by “ <i>Mas</i> ” (But).
13	Consecutive sub. clauses	1	Identify consecutive conjunctions by token matching.
		2	Split original sentence; Create a sentence for the main clause; Create a sentence for the subordinate clause starting by “ <i>Assim</i> ,” (Thus).
14	Purpose sub.clauses	1	Identify purpose conjunctions by token matching and PoS tags.

Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from [Gasperin, Maziero and Aluisio \(2010\)](#).

#	Cases	Ph	Rules
		2	Split original sentence; Create a sentence for the main clause; Create a for the subordinate clause starting by “ <i>O objetivo é</i> ” (The purpose is).
15	Conformative sub. clauses	1	Identify conformative conjunctions by token matching. Rearrange clauses into subordinate/main order,
		2	linked by “ <i>confirma que</i> ” (confirms that), or “ <i>que</i> ” (that) when the subordinate clause contains a verb.
16	Temporal sub. clauses	1	Identify temporal markers by token matching.
		2	Split original sentence; Create a sentence for the subordinate clause; Create a sentence for the main clause starting by “ <i>Então,</i> ” (Then).
17	Proportional sub. clauses	1	Identify proportional conjunctions by token matching.
		2	Do not simplify.
18	Non-finite sub. clauses	1	Identify non-finite clauses by syntactic tags.
		2	Do not simplify.
19	Restrictive relative clauses	1	Identify relative clause by syntactic tags and relative pronoun; Identify the anchor of the appositive clause.
		2	Remove relative clause from original sentence; Create new sentence containing anchor as subject and the relative clause as predicate.
20	Non-restr. relative clauses	1	Identify relative clause by syntactic tags, relative pronoun, and punctuation; Identify the anchor of the appositive clause.
		2	Remove the relative clause from the original sentence; Create a new sentence containing anchor as subject and the relative clause as predicate.
21	No Subj-Verb -Obj order	1	Identify non-SVO order by syntactic tags:
		2	for now, we only look for sentences starting with a VP. Transform into SVO order
22	Adverbial phrases	1	Identify adverbial phrases by syntactic tags; Identify if adverbial phrases are long (carry post-modification).

Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from [Gasperin, Maziero and Aluisio \(2010\)](#).

#	Cases	Ph	Rules
		2	Remove long adverbial phrases (one or more, if consecutive) from original sentence; Create a sentence for the adverbial phrases starting by “ <i>Isso ocorre</i> ” (This happens).

Table 19 – Simplification guidelines – phenomena identification and simplification steps – adapted from [Gasperin, Maziero and Aluisio \(2010\)](#).

- identifying relative clauses, sentences with non-finite verbs, and discourse markers inside a sentence;
- identifying negative connectives and checking if they can be replaced by positive ones;
- performing the noun-to-verb conversion;
- rewriting sentences as individual ones following the subject-verb-object ordering;
- replacing technical words with simpler and/or more frequent ones.

After the compilation, de-identification, post-edition, and classification phases, Applied Linguistics, Pharmacy, Medicine, and Nutrition expert members of our research group were responsible for creating an answer for each utterance of the corpus (comment/question). Moreover, the researchers included a simplified version of the comment/answer in order to enrich our corpus and drive the compilation to end-users. The simplification performed by the Medical domain experts considered

preserving the same meaning for the lay user (simplified) version within the domain, even if this version could increase the number of characters of the expert’s version. In addition, some simplified versions have additional information (e.g., explanations or a warning in case of sensitive questions) in order to preserve the main idea of the source text, a concern shared by medical domain simplification studies (Grabar; Cardon, 2018; Cardon; Grabar, 2018; Shardlow; Nawaz, 2019; Bercken; Sips; Lofi, 2019; Cardon; Grabar, 2020; Koptient; Grabar, 2020). The alignment of analogous segments of the language of expert and lay discourse is the basis for enriching NLG models through an HCD approach. Please refer to a sample of the simplification spreadsheet information in Table 20.

Original Post	Omitted due to ethical considerations (please refer to the De-identification Section 6.4.4).
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?
Translated Post-edited Post	Hello! How will I know if the device for measuring blood glucose is faulty?
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.
Simplified 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.
Translated Answer	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 in pharmacies.

Table 20 – Sample of the simplification spreadsheet in Brazilian Portuguese and the corresponding translation (some columns were omitted due to ethical considerations).

Translated Simplified Answer	We recommend you calibrate the glucose meter (a device that measures the amount of sugar in the blood) whenever you open a new package of testing strips. You can calibrate the glucose meter in pharmacies.
Linguistics Literacy (1 to 5)	4 – well-structured comment/question with errors (especially related to the domain) – <i>comentário/pergunta bem estruturado(a) e com erros (especialmente relacionados ao domínio)</i>
Diabetes Literacy (1 to 5)	3 – comment/question reasonably well structured and with errors – <i>comentário/pergunta razoavelmente bem estruturado(a) e com erros</i>
Acceptance Level (1 to 5)	5 – Acceptance – <i>Aceitação</i>
Classification	6.3 – Glucose Meter <i>Glicosímetro</i>
Category	<i>comida-e-dieta</i> – food-and-diet

Table 20 – Sample of the simplification spreadsheet in Brazilian Portuguese and the corresponding translation (some columns were omitted due to ethical considerations).

6.4.8 Expert Evaluation

The validation of the corpus compilation was done at the end of each corpus compilation iteration (for more information, please refer to Figure 16), where the verification was performed with the use of personas in order to identify possible persona mismatches (as explained in Subsection 6.1.1). The initial validation plan was aligned with the number of collected posts. After achieving the expected number of posts (i.e., posts of each forum, which are 1,000 in total), an expert validation was performed by different areas of the Empoder@ project, i.e., Applied Linguistics, Medicine, Pharmacy, and Nutrition expert members.

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 Model²⁹ as aforementioned in Section 6.4.6. Additionally, the corpus has been investigated regarding question-answering entailment by Ferreira et al. (2021) during his Post-doc period. Finally, a Sentiment Analysis study was conducted by Gumiel et al. (2021) during his Post-doc research.

6.5 Revisiting the NLG Model from an HCD Approach

For our third approach to context-enriched NLG models, we chose the healthcare domain, Diabetes Mellitus. In order to achieve more suitable results, we turned to HAI and HCD methods to model the language according to the different specificity of use: persona, which is in the context of this research, an artifact built in order to guide the design and the enrichment or development of NLG models. Moreover, the personas were developed from representations of researchers and participants of the Empoder@ project – Health Sciences consultants and applied linguists (for more information, see Section 6.1.2) and were added following eXplainable AI framework guidelines (Ribera; Lapedriza, 2019).

As a starting point, we first integrated personas according to the investigated context: experts and non-experts. Secondly, we compared and classified multi-domain NLG models to apply a human-centered methodology, specifically, hybrid models for NLG that consider the use of personas or personas corpora indicated by Wolf et al. (2019), Mazaré et al. (2018), Curry et al. (2018), Dinan et al. (2020), Lin et al. (2020), Paranjape (2022); and Majumder (2023). We based our approach on the lessons learned from these models and then performed set-up modifications to the Brazilian Portuguese language. Thirdly, we studied the effects of adding personas and external knowledge (a compilation of the personas created in Section 6.1.2 and the persona-based corpus compiled in Section 6.4, respectively), in order to provide more consistent data considering limited or nonexistent domain corpora,

²⁹ <<https://github.com/pavalucas/Bete>>

as suggested by the studies of [Li et al. \(2016\)](#) and [Ram et al. \(2018\)](#). The whole NLP model analysis can be summarized in Figure 20.

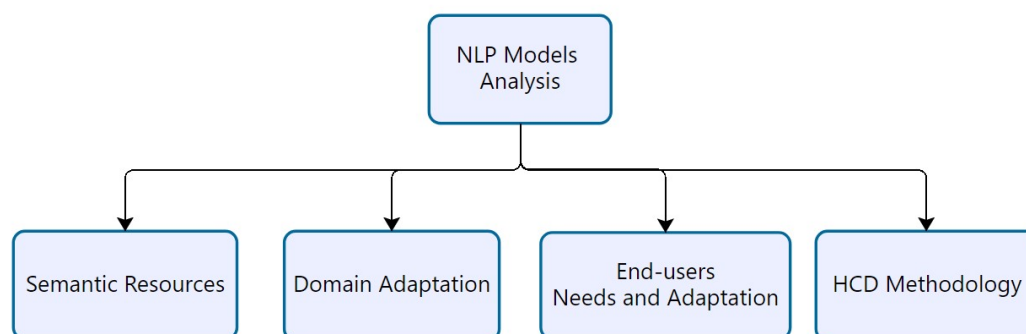


Figure 20 – NLP models analysis – semantic resources, end-users, domain adaptation requirements, and HCD methodology (created with [Bizagi \(2023\)](#)).

Following the human-centered NLG model pipeline presented in Figure 21, we first compiled the statistical and neural networks models results and then compared with super, semi, and/or unsupervised approaches ([Oraby et al., 2018](#); [Oraby, 2019](#); [Harrison et al., 2019](#); [Gopalakrishnan et al., 2019](#); [Liang et al., 2020](#)). Then, we selected the best models and added semantic resources and/or personas as external knowledge features in order to provide consistent data, considering limited or non-existent domain corpora. Finally, we compiled a summary of the investigated models that are summarized in Table 21 in order to interpret the results.

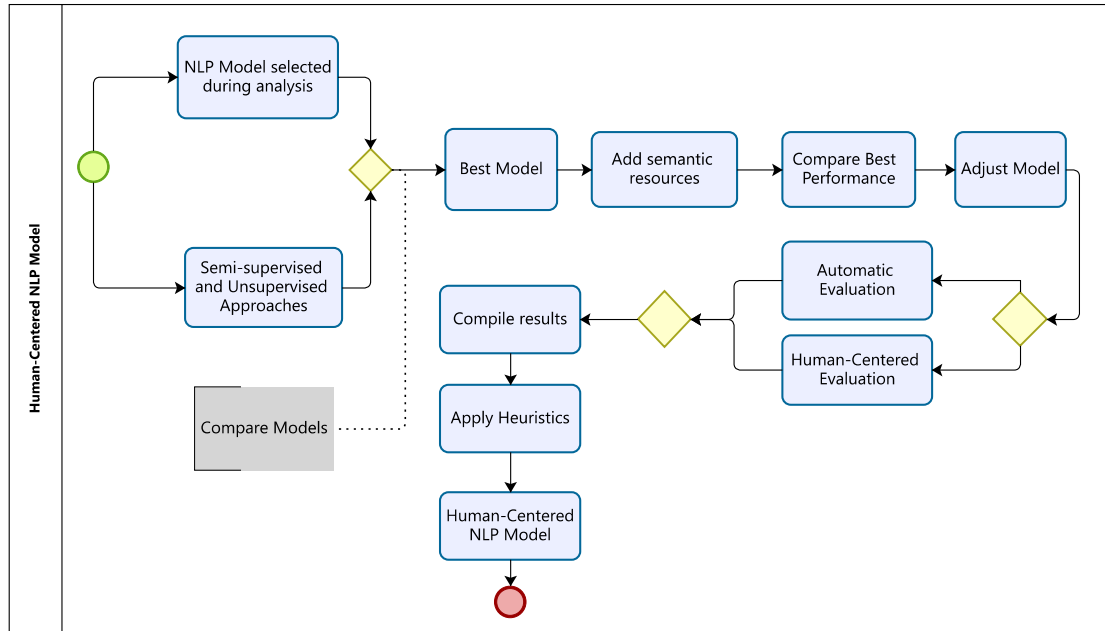


Figure 21 – Third approach on enriching NLG models – Human-Centered NLG model pipeline (created with Bizagi (2023)).

NLG Model	Personas	Ext.Res.	Corpora	HCD
Ferreira et al. (2018a)		X	X	
Cao and Cheung (2019)		X	X	
Zhang et al. (2018a)		X	X	
Mazaré et al. (2018)		X	X	
Ram et al. (2018)	X	X	X	
Curry et al. (2018)	X	X	X	
Wolf et al. (2019)		X	X	
Dinan et al. (2020)		X	X	
Lin et al. (2020)		X	X	
Liang et al. (2020)		X	X	
Oraby et al. (2018)		X	X	
Harrison et al. (2019)		X	X	
Song et al. (2021)	X	X	X	
Lim et al. (2022)	X	X	X	
Paranjape (2022)	X	X	X	
Majumder (2023)	X	X	X	X

Table 21 – NLP models analyzed according to the use of personas, external resources, corpora (persona or user profiles), and HCD approach.

6.5.1 Hybrid Approach

Regarding the procedures, this research has as the principle of a survey of NLP models, in particular, NLG models with statistical and neural networks approaches that consider multi-domain adaptation scenarios. As highlighted in previous sections, a hybrid approach is considered a more suitable architecture to enrich an NLG model. This approach fits the complexity of different features within this study, which are: (i) external knowledge; (ii) semantic resources; (iii) end-user needs and experience; (iv) representative of different levels of literacy and text generation; and (v) with an application to a conversational AI question-answering context.

The main difficulty of this third approach was the related work phase and its extensive literature review since this approach correlates different tasks within NLG and the nuances of the domains, specifically the healthcare domain. To clarify these challenges, the tasks we had to unveil within NLG were question-answering, text simplification, paraphrasing, machine translation, conversational AI, explainable AI, human-centered AI, persona-based models, human-centered NLP, and context enrichment. Next, we covered a sensitive domain to context enrichment where information has a very high cost, whether concerning the selection of data or their annotation and validation.

Within two years, we had weekly research group meetings during the COVID-19 pandemic. During these meetings, the discussed topics included annotation, translation, design, and software development (NLG models, web page, mobile app, and chatbot). The group consisted of linguists, software developers, and students and/or professionals from different areas, such as Computer Science, Letters, Information Science, Medicine, Nursing, and Nutrition.

The first development phase was building the persona-based corpus in parallel to the NLG models investigation. Next, the second phase involved decisions regarding the best model to fit this third approach need, i.e., developing a human-centered NLP model. Finally, after deciding on the model's architecture, we started investigating the most suitable tools to address the conversational AI QA system. These steps will be addressed in the following sections.

6.5.2 Context

Accordingly, we first integrated personas in line with the investigated context: experts and non-experts, complex and simplified texts (the developed persona-based corpus in Section 6.4), during the compilation, annotation, and validation of the persona-based corpus. This corpus was compiled and annotated manually and automatically. After achieving a total of 2,000 pairs (question and answer), we performed the 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 of the Empoder@ project. The result was a monolingual corpus aligned with expert and non-expert discourse in the diabetes mellitus domain. The resulting corpus was used in the next step of the context enrichment of the NLG model and can be summarized in Figure 22.

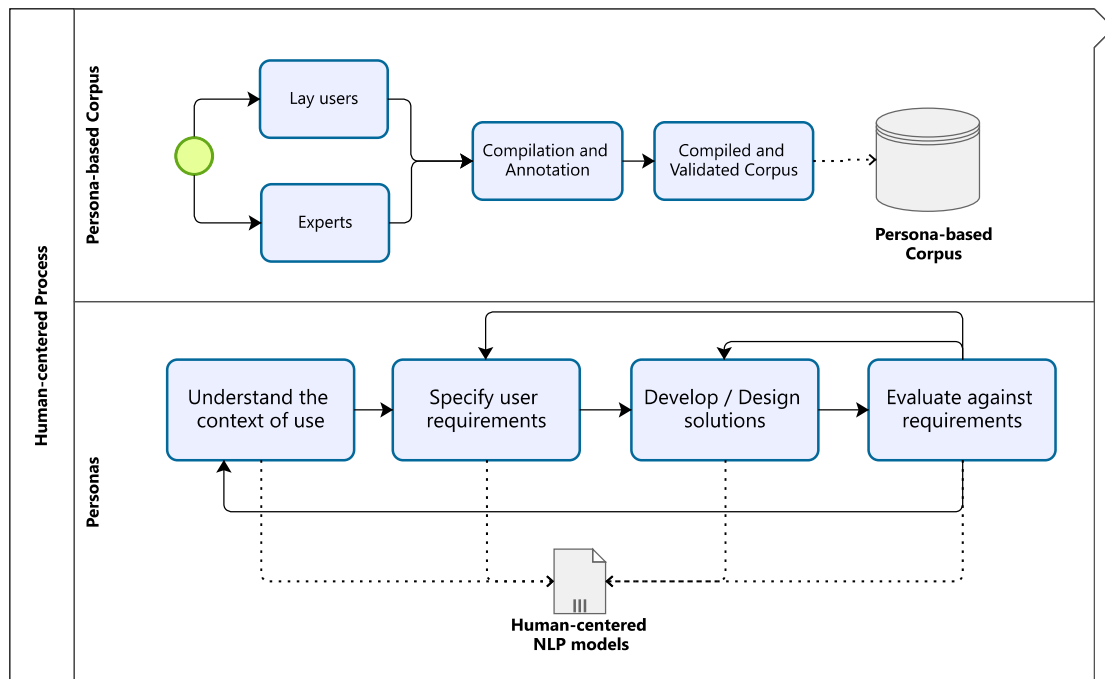


Figure 22 – Applying HCD methodology to the context enrichment of an NLG model (created with Bizagi (2023)).

6.5.3 Fine-tuning

We validated our approach by testing different models as we have proposed. After the analysis, we reached a conclusion based on the main requirements of the **(RQ2)** “*How are NLG models designed to provide an adequate production of meanings and a better understanding within critically sensitive domains, such as healthcare and governance?*”, and conducting an extensive literature review, the most suitable solution to answer this thesis question was enriching an NLP model with a fine-tuning approach. Then, the selection of pre-trained language models (PLM) considered the nuances of the domain of the limited number of public datasets due to the cost of annotation and the ethical considerations regarding privacy. [Jurafsky and Martin \(2023, p. 210\)](#) clarify the pre-training process which concerns “learning some sort of representation of meaning for words or sentences by processing very large amounts of text.., and then we call the resulting models pre-trained language models”.

In addition, [Wang et al. \(2022\)](#) point out the advantages of fine-tuning a model than training from scratch (please refer to Figure 23 for a visual difference between the two paradigms), more specifically within dialogue systems (i.e., conversational AI and virtual assistants), as well as low-resource domains to downstream NLP tasks, by which PLMs can exploit unannotated data for self-supervised and improve the performance of the results. This approach has attracted a lot of attention with BERT ([Devlin et al., 2019](#)), and especially with the use of Generative Pre-Training (GPT) ([Radford et al., 2018a](#); [Radford et al., 2019](#)) models, additionally to fine-tuning approaches ([Zhang et al., 2020](#)).

At first, we considered training a model from scratch, but due to the time-consuming and mostly to the computationally cost, considering the large language models with many parameters like GPT-2, we opted for pre-training the models from their checkpoints, i.e., their intermediate saved versions during the training process. For more information about fine-tuning, please refer to a dedicated chapter of [Jurafsky and Martin \(2023\)](#) book or pre-training surveys ([Wang et al., 2021b](#); [Paaß; Giesselbach, 2023](#)).

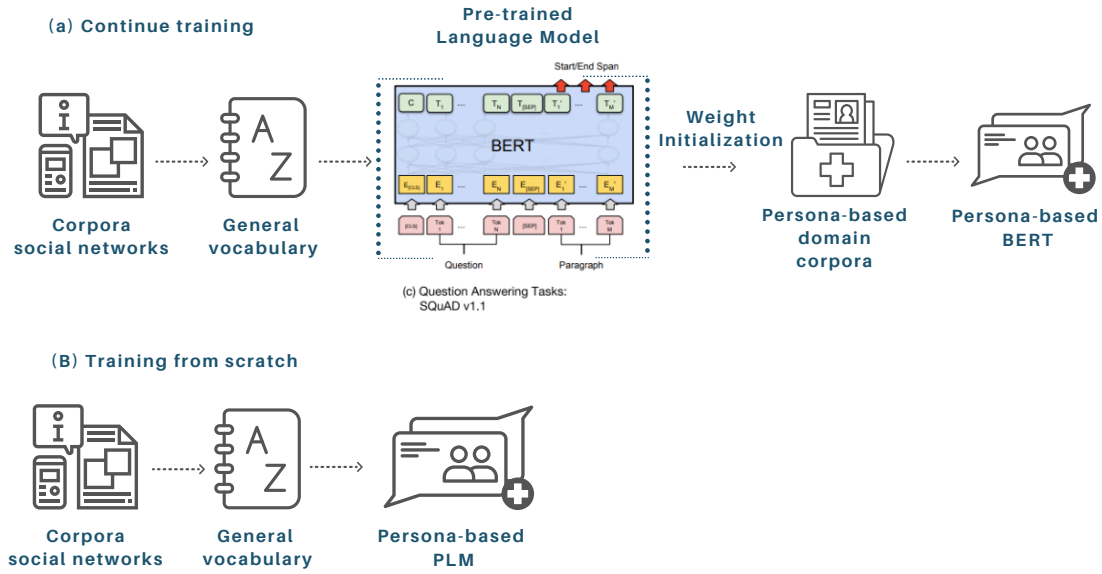


Figure 23 – Training paradigms within PLMs – adapted from Cai et al. (2023, p. 4).

6.5.3.1 Baseline Models

We performed fine-tuning in competitive baseline models as shown in Table 22. The fine-tuning concerned two complementary tasks of the human-centered NLP model: text classification – Natural Language Inference (NLI) – and question answering – conversational AI question answering. The analysis of the pre-trained models (BERT and GPT-2) and their extensions focus on their individual characteristics, training data, applications, and performance metrics, emphasizing their respective advantages in specific domains, and are listed as follows:

Name	Configuration	Task
BERT (mBERT)	bert-base-multilingual-cased	NLI, CQA
BERTimbau	neuralmind/bert-base-portuguese-cased	NLI, CQA
BioBERT	dmis-lab/biobert-base-cased-v1.2	NLI, CQA
BioBERTpt(all)	pucpr/biobertpt-all	NLI, CQA
BioBERTpt(clin)	pucpr/biobertpt-clin	NLI, CQA
BioBERTpt(bio)	pucpr/biobertpt-bio	NLI, CQA
ClinicalBERT	emilyalsentzer/Bio_ClinicalBERT	NLI, CQA
GPT-2	gpt2	NLI, CQA
GPT-2 Portuguese-2	pierreguillou/gpt2-small-portuguese	NLI, CQA

Table 22 – Fine-tuned models approach followed by their respective configuration and tasks (NLI – Natural Language Inference and CQA – Conversational AI Question Answering).

- BERT (Bidirectional Encoder Representations from Transformers) ([Devlin et al., 2019](#)) revolutionized NLP by introducing bidirectional context to better understand word meanings in context – Transformer architecture ([Vaswani et al., 2017](#)). BERT excels in tasks like natural language inference, sentiment analysis, and question answering. We use the general-domain and multilingual version – mBERT – in order to test our approach.
- BERTimbau ([Souza; Nogueira; Lotufo, 2020](#)) extends BERT’s capabilities specifically for the Portuguese language by pretraining on brWaC corpus ([Filho et al., 2018](#)) and achieves improved performance in language-related tasks.
- BioBERT ([Lee et al., 2020](#)) is specialized in biomedical text processing in the English biomedical domain. The model is pre-trained in domain-specific corpora (pre-trained in PubMed abstracts and PMC full-text articles), enabling better performance in tasks like biomedical named entity recognition, relation extraction, and question-answering.
- BioBERTpt ([Schneider et al., 2020](#)) is a domain-specific model designed for clinical and biomedical NLP tasks, particularly in the Portuguese language. The model is pre-trained on biomedical corpora and available in three versions: BioBERTpt(clin), BioBERTpt(bio), and BioBERTpt(all). The advantage’s use of BioBERTpt lies in its ability to handle specialized medical jargon and

shows the benefits of domain transfer learning in clinical and biomedical tasks.

- ClinicalBERT (Alsentzer et al., 2019) is a BERT-inspired model designed for clinical text – generic clinical text and discharge summaries. The main advantage of ClinicalBERT is its ability to capture context-specific clinical semantics due to the use of domain contextual embeddings – trained over MIMIC corpus (Johnson et al., 2016), which leads to more context-aware clinical information for NLP tasks.
- GPT-2 (Generative Pre-trained Transformer 2) (Radford et al., 2019) extends OpenAI GPT Model – GPT-1 (Radford et al., 2018b) and uses its own developed dataset WebText. The model generates coherent and contextually relevant text, and its diversity encompasses a broad number of NLP tasks from Reading Comprehension to Machine Translation. GPT-2 results present a high level of fluency and diversity in generated texts.
- GPorTuguese-2 (Schneider et al., 2021) is a variant of GPT-2, specifically fine-tuned for clinical and biomedical texts written in the Portuguese language. The model benefits from transfer learning in order to provide better clinical information results, making it a valuable tool for clinical NLP tasks.

6.5.3.2 Fine-tuned Models Settings

About the model parameters, we followed most of the settings from each of the fine-tuned models (please refer to Table 22). We trained each model with 8 epochs with a dropout of 0.1. Moreover, we set the early stopping of the neural networks to 5 and the beam size to 1. We applied a maximum output limit generation of 30. Furthermore, we set the batch and state sizes to 8, and 512, respectively. Additionally, we set persona word embeddings and hidden size to 768D.

6.5.4 Classification – Natural Language Inference

After deciding about the fine-tuning approach, we followed Li et al. (2016) in relation to adding a Natural Language Inference (NLI) step in order to presume

the persona behind the question that the model is trying to answer. After users ask a question, the model infers by the input question which persona is more suitable to the user. After classifying the question, the next step in the pipeline is the conversational AI question-answering (CQA) model, responsible for answering the user’s question. The CQA model takes into account the user’s input question and the inferred persona. The fine-tuned CQA model selects an answer according to the persona and gets back to the user. If the model is not able to select an answer that fits the parameters, then the NoSQL database (i.e., built with the Elasticsearch³⁰ platform) step of the pipeline is able to select an answer from the persona-based corpus.

Literacy	Persona	Literacy
1	0	simple
2	0	simple
3	1	technical
4	1	technical
5	1	technical

Table 23 – Persona model’s settings based on the literacy scale presented in Section 16.

Model	ID F1-Score	Desc. F1-Score	ID Accuracy	Desc. Accuracy
BERT	0.5002	0.6831	0.4576	0.7797
mBERT	0.4145	0.7156	0.4145	0.7119
BERTimbau	0.2012	0.2579	0.2712	0.3051
BioBERT	0.6416	0.7129	0.6102	0.6949
BioBERTpt	0.6585	0.6976	0.6949	0.7119
BioBERTpt(clin)	0.6435	0.6269	0.6271	0.5932
BioBERTpt(bio)	0.5959	0.6532	0.5593	0.6271
ClinicalBERT	0.6831	0.7212	0.7797	0.7966

Table 24 – Persona-based corpus NLI: F1-Score and Accuracy considering two forms of adding persona context – ID and Description.

³⁰ <<https://www.elastic.co/>>

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

Table 25 – Persona-based corpus statistics considering the original corpus (annotated by Empoder@ research group) and the nutrition corpus (annotated by nutrition students based on the medicine annotation).

The second perspective concerning the NLI phase was to classify the input question on a 1 to 5 scale (please refer to the Annotation Guidelines Section 6.4.6 for more information) and then get the persona description according to the classifier model as exemplified in Table 23. Next, the CQA model receives a triple of information that includes the question, the inferred persona (lay or expert persona), and the persona description. Results showed that the first approach presented better classifications according to the test set of the persona-based corpus as presented in Tables 24 and 25.

6.5.5 Question-Answering & Conversational AI

According to Budler, Gosak and Stiglic (2023), “[h]ealthbots have the potential to transform care by focusing it on the user; yet they are still in their development and need further work on development, automation, and adoption to have a population-level health effect”. This third context enrichment approach benefits from the development of BeteQA (Ferreira et al., 2021) concerning the accumulated knowledge related to the LLMs used, the best metrics, and the drawbacks.

During the development of this third context-enrichment approach that started in October 2019, innovative technologies emerged concerning Conversational AI, some of them discontinued, and others followed the steady development of industry and academic research. We narrowed down some of them in order to find the most suitable platform for our approach, they are listed in Table 26. Accordingly,

we added corresponding NLU technologies of each framework (Abdellatif et al., 2021).

Framework	NLU	
IBM Watson	Watson Conversation	<https://www.ibm.com/watson>
Dialogflow	Google Dialogflow	<https://cloud.google.com/dialogflow>
Rasa	Rasa	<https://rasa.com/>
Microsoft LUIS	LUIS	<https://dev.botframework.com/>
Pandorabots	Pandorabots	<https://home.pandorabots.com/home.html>
Mycroft	Mycroft	<https://mycroft.ai/>
Botpress	Botpress	<https://botpress.com/>
Meta Wit.ai	Wit.ai	<https://wit.ai/>

Table 26 – Conversational AI Frameworks/Platforms.

6.5.6 Conversational AI

In order to develop the chatbot application – Betinha, we first investigated which technology would be the best approach. First, the development started with the Dialogflow framework (Google, 2023) following the lessons learned from the COVID-19 chatbot – Ana (Ferreira et al., 2020). Next, we investigated which of the frameworks would provide the possibility to customize an NLP model. After the analysis, the chatbot was developed in Python³¹ programming language and the open-source Conversational AI framework called RASA (Bocklisch et al., 2017)³². One of the advantages of RASA is the feature of deployment on different platforms, such as WhatsApp³³, Slack³⁴, Telegram³⁵, and so on. In order to reach a higher number of users, we implemented Betinha to the deployment to WhatsApp due to its common use in Brazil.

Due to the characteristics of this third approach, we conclude that Rasa’s framework (Bocklisch et al., 2017) suits better this study’s main purpose which is enriching an NLG model. Rasa is an open-source conversational AI toolkit with a licensed version called Rasa Pro (Rasa’s enterprise solution). Additionally, Rasa Pro

³¹ [<https://www.python.org/>](https://www.python.org/)

³² [<https://github.com/RasaHQ/rasa>](https://github.com/RasaHQ/rasa)

³³ [<https://about.meta.com/brand/resources/whatsapp/whatsapp-brand/>](https://about.meta.com/brand/resources/whatsapp/whatsapp-brand/)

³⁴ [<https://slack.com/>](https://slack.com/)

³⁵ [<https://telegram.org/>](https://telegram.org/)

offers Rasa X³⁶ – a low-code user interface that supports conversational AI teams. The free Rasa version can be extended according to the business rules needs which adds an advantage when compared to black-box platforms, such as Dialogflow or IBM Watson.

To achieve the current implementation, we first developed the Knowledge base actions³⁷ feature from Rasa. Next, we implemented the knowledge base with our corpus in a JSON file and then used the set-up Elasticsearch database. Due to its experimental approach, we decided to replace Rasa’s knowledgebase feature with an external API approach. The decision was based on a business rule, that is to provide a more intuitive and easy-maintenance implementation. We chose an open-source and user-friendly Python API – FastAPI³⁸. Besides, the idea behind the external API was to create a FastAPI connection with Rasa’s framework.

We implemented the API completely independent of Rasa’s own framework in order to add new LLM implementations. We have also developed unit tests to validate each part of the approach: FastAPI, Rasa, and then, FastAPI–Rasa integration. We focused the development on an intuitive approach in order to extend to new models (pre-trained or not). The models were implemented based on QA Entailment models (Ferreira et al., 2021) developed during the Empoder@ project, and serve as a complementary feature to our approach, more specifically, to the CQA fine-tuned model. This was not planned in the first version of the chatbot but was added to the model as a consistent part of the human-centered NLP model. The API approach was a very product-driven decision, in order to make the chatbot extensible to more models and possible continuous updates. Examples of unit tests, both to the RASA API (interface layer) and to the FastAPI (business layer). Please refer to Figure 24 regarding a general overview of the developed Conversational AI system.

³⁶ <<https://rasa.com/docs/rasa-enterprise/>>

³⁷ <<https://rasa.com/docs/rasa/action-server/knowledge-bases/>>

³⁸ <<https://fastapi.tiangolo.com/>>

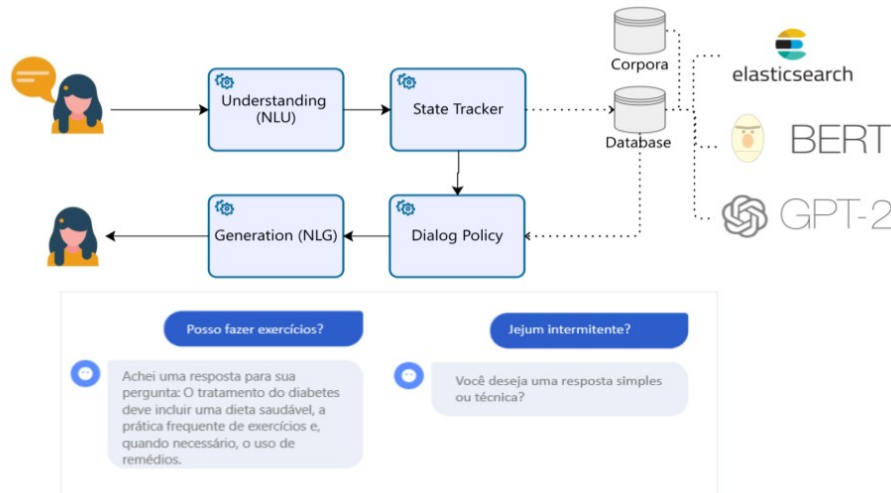


Figure 24 – Main Components of the Dia(Bete) Conversational AI System.

One of the benefits of using an API to build the QA step of the pipeline is leveraging a core part of Rasa’s framework for natural language understanding (NLU), which is the DIET implementation (Bunk et al., 2020). We started the development with Rasa on version 2.0 and then updated it to 3.0. During this period, some tools used at the beginning of the experiment were discontinued as BotSociety³⁹, a design-conversational interface responsible for exporting a pre-configured version to different chatbot frameworks.

³⁹ <<https://botsociety.io/>>

```
1  version: "3.1"
2
3  stories:
4
5  - story: welcome
6    steps:
7      - intent: greet
8      - action: utter_greet
9
10 - story: say_goodbye
11   steps:
12     - intent: goodbye
13     - action: utter_goodbye
14
15 - story: main_answer_question_and_say_goodbye
16   steps:
17     - intent: answer_question
18     - action: action_answer_question
19     - intent: goodbye
20     - action: utter_goodbye
```

Figure 25 – Betinha stories developed with Rasa framework.

In Rasa, we are able to develop “stories”, that guide the NLU to what happens, but the main benefit of the stories is that we do not need to create a huge amount of examples, just a few, just the necessary to make Rasa works. The Rasa NLU is capable of learning through interaction and that is an interesting characteristic of the HCD approach we are modeling. Please refer to Figure 25 with the main stories of Betinha.

6.5.7 Evaluation

For validation purposes, we applied the context-enriched NLG model to an initial sample to examine the data based on quality metrics better suited to the investigated model. In this respect, we focus on two types: (i) automatic – BLEU (Papineni et al., 2002), SARI (Xu et al., 2016); and (ii) human. The automatic evaluation considered string similarity metrics, these metrics are often used on Machine Translation (MT) systems. We used BLEU (Bilingual Evaluation Understudy) as a common sense in NLG literature. Regarding simplification metrics, we investigated common metrics within the area: SARI (System output Against Refe-

rences and Input sentence) – focused on paraphrasing (Xu et al., 2016) and SAMSA (Simplification Automatic Evaluation Measure through Semantic Annotation) – focused on sentence splitting (Sulem; Abend; Rappoport, 2018), and opted to run only SARI (Xu et al., 2016) due to the research timeline. Finally, we computed F1-Score and Accuracy as additional metrics to BLEU and SARI.

Task	Model	Configuration	F1-Score	Accuracy
NLI	BERT	bert-base-cased	0.6831	0.7797
NLI	mBERT	bert-base-multilingual-cased	0.7156	0.7119
NLI	BERTimbau	neuralmind/bert-base-portuguese-cased	0.2579	0.3051
NLI	BioBERT	dmis-lab/biobert-base-cased-v1.2	0.7129	0.6949
NLI	BioBERTpt	pucpr/biobertpt-all	0.6976	0.7119
NLI	BioBERTpt(clin)	pucpr/biobertpt-clin	0.6269	0.5932
NLI	BioBERTpt(bio)	pucpr/biobertpt-bio	0.6532	0.6271
NLI	ClinicalBERT	emilyalsentzer/Bio_ClinicalBERT	0.7212	0.7966

Table 27 – Pre-trained NLI models – F1-Score and Accuracy metrics in NLI task. The best results are **boldfaced**.

6.5.7.1 NLI Evaluation

The NLI models presented in Table 22 were fine-tuned with the persona-based corpus developed in Section 6.4. The F1 and Accuracy scores regarding the NLI task are presented in Table 27. The third approach was trained from checkpoints, i.e., from an existing set-up (e.g., weights) as a pre-requisite of this kind of approach. Some models that were trained using the same technique, such as BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), BioBERTpt (Schneider et al., 2020), and more recently CardioBERTpt (Schneider et al., 2023). Regarding the NLI model parameters, we trained these models following most of the settings, our set-up modifications of each pre-trained are shown in Section 6.5.3.2.

We classified the texts into two main personas (expert and lay), converting the Literacy scales (i.e., Diabetes and Linguistics, please refer to the Post-edition Section – Table 16, for more information) into two labels: simple (items 1 to 2) and technical (items 1 to 5). Regarding human evaluation, we chose to evaluate only with automatic metrics – specifically for the NLI task, since the classification

is a straightforward task and an additional cost at this stage. We performed an exploratory test in order to verify some samples of the generation during this step.

6.5.7.2 CQA Evaluation

In order to evaluate the generated texts, we applied BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016) in all pre-trained models results. Due to the CQA evaluation comprising two main goals: text generation in a question-answering context and text simplification in a conversational scenario, the primary metric was the combination of best scores of BLEU and SARI, with the prevalence of SARI, presented in Table 28.

Task	Model	Configuration	BLEU	SARI
CQA	BERT	bert-base-cased	<u>2.4122</u>	<u>52.5325</u>
CQA	BERTimbau	neuralmind/bert-base-portuguese-cased	2.7125	50.0948
CQA	BioBERT	dmis-lab/biobert-base-cased-v1.2	1.7709	45.7023
CQA	BioBERTpt	pucpr/biobertpt-all	1.3967	51.0576
CQA	BioBERTpt(clin)	pucpr/biobertpt-clin	2.5390	50.5933
CQA	BioBERTpt(bio)	pucpr/biobertpt-bio	1.5499	48.7800
CQA	ClinicalBERT	emilyalsentzer/Bio_ClinicalBERT	1.9248	53.8805
CQA	GPT2	gpt2	0.0003	40.8725
CQA	GPorTuguese-2	pierreguillou/gpt2-small-portuguese	0.0005	39.3900

Table 28 – Pre-trained CQA models – BLEU and SARI metrics in CQA task. The best results are **boldfaced**.

6.5.7.3 HCD & Human Evaluations

Due to the use of a human-centered approach, we decided to use HCI assessment as the primary human evaluation in this study. One of the considered human metrics was the set of personas created for this research, according to users' characteristics in addition to the information about the NLG models and investigated domains. In fact, the human evaluation was performed within the context of experts and non-experts among the research group participants. After the automatic metrics, we conducted an exploratory evaluation and validation of the results in which heuristics were examined in the text generation against the user need (Suojanen; Koskinen; Tuominen, 2015b; Cunha, 2016; Cunha; Fernandes, 2016). The exploratory evaluation followed an HCD approach in which an HCD

expert performs a usability test, following lessons learned from Cunha (2016) and validating the generated text with the according personas. This HCD evaluation was in parallel to the exploratory tests performed in the Rasa implementation of Betinha.

6.5.8 Results & Discussion

6.5.8.1 Metrics

Baselines were analyzed in different tasks, each of them with particular approaches. The NLI evaluation is presented in Table 27 where the best results are boldfaced. It is noteworthy that the language in the classification task was not a decisive factor as we can infer from the results, that the best outcome corresponds to the multilingual model (ClinicalBERT). Regarding CQA evaluation, we had a difference in results when compared to the classification ones. We disregarded the mBERT from the CQA analysis due to its unsuitable results, i.e., they were next to zero, which can imply a problem with the model's calibration or a small corpus for training (Gao; Galley; Li, 2019). For the CQA task, fine-tuning an open-domain model or a task-based one like ClinicalBERT or BERTimbau demonstrated a slight advantage.

6.5.8.2 Strategies

The NLI results in Table 27 considered only the two labels strategy to the personas: simple and technical (lay and expert, respectively). For the CQA results in Table 28, we only presented the results for the best strategy, that is, considering the persona as a classification task and embedding the persona as context without additional tags. The second strategy, which is adding the descriptions of the personas is considered after conducting an ablation study with different persona strategies.

6.5.8.3 Pre-trained Models

Regarding the selected models we can check the differences between NLI and CQA tasks in relation to the best scores for each task. Regarding the NLI task,

the best model was ClinicalBERT with an F1-Score of 0.7212 and an accuracy of 0.7966. The model surpasses by a considerable difference from the other investigated models. Considering Bertimbau, the classification was the lowest, indicating that for the NLI task, the language is not as important as the domain. BioBertpt had an average score among its versions, lacking the ability to classify in-domain personas.

Considering the CQA task, ClinicalBERT is the best model again if you consider BLEU and SARI combined – BLEU of 1.9248 and SARI of 53.8805. In the CQA task, BERT had the second-best score combination with a BLEU score of 2.4122 and SARI of 52.5325, showing an analogous performance in the CQA task. GPT-2 and its Portuguese version for clinical texts GPorTuguese-2 presented lower results even with fluent texts in the output. We can infer from the results that the domain is still determinant when applied to the diabetes mellitus persona-based corpus.

6.5.9 Conclusion

This study introduced an innovative corpus due to the combination of characteristics: simplification (lay and expert texts), persona-based (lay and expert personas), medical domain (Diabetes Mellitus), and Portuguese language. We conducted the experiment of the in-domain persona-based corpus within an HCD approach in the context of its application, i.e., a chatbot or conversational AI QA application, called Dia(Bete) – aka Betinha. The CQA agent is the application of the resulting human-centered NLP model, that was trained in state-of-the-art NLG inference and CQA models which benefits from the gained knowledge HCD approach of the persona-based corpus. The best F1-Score and Accuracy highlighted the preference for the clinical NLP models regarding the NLI task. The combination of BLEU and SARI best scores presented the prevalence of the domain over the language considering the CQA task. Overall, the best results were the in-domain NLP model for both NLI and CQA tasks.

To investigate the accuracy of the generated texts, we performed a human-centered evaluation using the chatbot from a shell window exemplified in Figure 26, in contrast with the corresponding persona. This analysis consisted of manually selecting utterances from the test set developed during the persona-based corpus

```

(venv) kubuntu:~/projects/BeteBot$ make shell
cd bot/ && rasa shell --endpoints endpoints.yml
2023-08-11 18:19:16 INFO      root - Starting Rasa server on
http://0.0.0.0:5005
2023-08-11 18:19:18 INFO      rasa.core.processor - Loading model model
s/20230811-180803-old-landfall.tar.gz...
2023-08-11 18:19:19 INFO      rasa.nlu.utils.spacy_utils - Trying to
load SpaCy model with name 'pt_core_news_sm'.
2023-08-11 18:19:21 INFO      rasa.nlu.utils.spacy_utils - Trying to
load SpaCy model with name 'pt_core_news_sm'.
2023-08-11 18:19:41 INFO      root - Rasa server is up and running.
Bot loaded. Type a message and press enter (use '/stop' to exit):
Your input -> Eu tenho diabetes e estou sem fazer exercícios, posso
começar mesmo assim?
Betinha -> Achei uma resposta para sua pergunta:
O tratamento do diabetes deve incluir uma dieta saudável, a prática
frequente de exercícios e, quando necessário, o uso de remédios.
Normalmente, a intenção é que os valores da glicemia (açúcar no sangue)
abaixem devagar e se mantenham relativamente estáveis ao longo do dia.
Quando a pessoa com diabetes tem sintomas no momento do diagnóstico
normalmente significa que a glicemia dela estava muito alta, então pode
ser necessário um tratamento que abaixe esses níveis mais rapidamente.

```

Figure 26 – Rasa shell during the HCD evaluation.

compilation and distinct ones from the QA dataset developed by [Ferreira et al. \(2021\)](#). The results presented a different perspective from the automatic evaluation with the prevalence of fluency on GPT-2 models' outputs. This outcome reinforces the application of a human-centered approach to a broader overview concerning accuracy and fluency within context-sensitive domains.

6.5.9.1 Limitations and Future Work

Due to the specificities of the corpus used in this research, we were not able to compare the fine-tuned models with different benchmarks. We plan to consider clinical corpora for future work as they are the nearest to the simple-technical discourse we addressed during this research. In relation to simplification, we plan to investigate this study's resulting models with simplification datasets in order to evaluate the personas both in corpora of the same language or the same domain. Regarding the analyzed models, we aim to test our approaches in the

new models, such as Llama2 (Touvron et al., 2023), BLOOM5⁴⁰, and GPT-J6⁴¹, choosing open-source versions of the recent LLMs development whenever available.

⁴⁰ BLOOM: BigScience Large Open-science Open-access Multilingual Language Model – <https://huggingface.co/bigscience/bloom>

⁴¹ https://huggingface.co/docs/transformers/model_doc/gptj

7 Conclusion

This study presents the combination of methods from natural language processing (i.e., natural language generation and conversational AI), human-centered AI, user-centered design, human-centered NLP, and employing Applied Linguistics (i.e., corpus and linguistic proficiency), as well as Semantic Web resources. Furthermore, this research uses a hybrid approach, that is the application of statistical and deep learning techniques (neural networks), to produce contextualized texts. We draw on classical approaches, as the pipeline architecture and contrast to end-to-end implementations (Ferreira et al., 2019), to determine the most suitable strategy in order to improve user comprehension as well as user experience. Our study addresses the target user needs in information-driven, decision-making, and critically sensitive domains, such as deforestation (governance) and diabetes mellitus domain (healthcare).

The study carries out a mixed-methods investigation, in which qualitative and quantitative approaches are collected, examined, and validated (Saldanha; O'Brien, 2014, p. 23), i.e., they are oriented to the solution of specific problems for a practical application that is, the support of NLG for multiple systems in sensitive domains. The present study is composed of three studies in order to encounter solutions to the production of texts using low-resource corpora, through a hybrid approach where the comprehension of these automatically generated texts is a central concern.

In summary, we started our investigation in the context of referring expression generation (please refer to Chapter 4 for more information). First, we enriched a state-of-the-art REG Model (Ferreira et al., 2018a) by adding external resources, such as ontologies and knowledge-base information, in order to provide referring expressions closer to the ones produced by humans as well as by helping the model to better generalize to unseen entities. Based on automatic and human evaluations in addition to an ablation study, we concluded that the enriched version of the NeuralREG model was able to produce more relevant referring expressions.

Besides, these generated texts were consequently easier to read, which is an essential criterion to provide end-users with a better experience.

Next, we presented our second approach to the enrichment of NLG models in the context of the governance domain: deforestation (please refer to Chapter 5 for more information). Furthermore, we introduced how we developed DaMata, a robot journalist that creates daily and monthly reports about deforestation in the Brazil Legal Amazon. After presenting DaMata, its main functionalities, and the benefits of developing an NLG tool with a social-good purpose, we explained the ongoing experiment to produce more user-driven texts within DaMata’s domain. In 2020, just after the qualification exam, we started an experiment regarding the preferences of geographical entities (e.g., Brazilian cities or conservation areas) within the Brazilian Legal Amazon. However, at the beginning of the experiment, more specifically during the data collection phase, we came across the opportunity to address the numerical referring expressions hallucinations issue in conjunction with DCU and USP. We have decided to address the hallucination problem and started an investigation into strategies regarding number tokenization to overcome this issue. Since this is an ongoing project, we decided to add the submitted paper in the Appendix A (please refer to the paper for more information). In addition, we believe there is still room for research about addressing the REG part of DaMata with NeuralREG’s extension presented in Chapter 4.

Finally, we showed the last and third approach of context-enriched NLG Models. We enriched an NLG model since its conception by employing human-centered approaches: HAI and HCD. We first created personas, that is a “fictitious user described with a basis in data” (Nielsen, 2019, p. 1) in the context of the healthcare domain. These personas were created based on data collected from the Empoder@ project and are used in the following stages of the NLG model’s context enrichment. Therefore, we applied the verified personas during the compilation, de-identification, annotation, validation, and evaluation of the persona-based corpus (see more in Section 6.4). This corpus is compiled semi-automatically in order to assist the training of the NLG model for the diabetes mellitus domain. Later, we compared NLG models that could be enriched from a human-centered perspective by including semantic resources and/or personas as external knowledge features in

order to provide consistent data, considering limited or non-existent domain corpora. In this respect, we examined our approach with two types of metrics: automatic and human-centered. The human metric took into account the personas created for this research, according to users' characteristics, in addition to the information about the NLG models and investigated domain. After applying automatic and human evaluations, an exploratory evaluation and validation of the results are done in order to contrast the text generation against a human-centered perspective, using personas and heuristics (Suojanen; Koskinen; Tuominen, 2015b; Cunha, 2016; Cunha; Fernandes, 2016).

Accordingly, we applied this NLG model to a conversational AI question-answering system that would assist users regarding the diabetes mellitus domain by answering questions about diagnostic and prognostic issues from a human-centered approach varying from simplified to complicated responses. Moreover, we believe providing end-users with relevant information could increase end-users' self-care management and consequently encourage healthy behaviors.

7.1 Contributions

In this section, we describe our contributions to this thesis based on the [research questions](#).

7.1.1 External Knowledge

RQ1: *What strategies add external knowledge to NLG models within a specific domain with a central concern of generating understandable texts?*

During our first context enrichment approach in Chapter 4, we were able to confirm the advantages of using external resources and how they could be a good strategy for producing more diverse texts. However, we believe with some improvements to our extended model, we are able to achieve more fluent texts and foster better engagement from end-users. With the conducted experiment discussed in Chapter 5, we addressed the production of more understandable texts in the deforestation domain. Within a REG pipeline architecture, we added extra

information collected from Semantic Web and Wiki pages in order to generate more accurate texts. During the experiments, numerical referring expressions posed a problem that was addressed with different kinds of tokenization strategies. Besides, the results pointed out issues specific to the language in use. We believe refactoring the strategies used with a more language-driven approach would benefit the results. The last approach in Chapter 6 corroborates the benefits of using an HCD approach in order to enrich a sensitive domain such as the healthcare domain – Diabetes Mellitus. We compared our approach by fine-tuning baseline models and confirmed more meaningful results with the developed solution. In addition, the HCD approach brought to an interdisciplinary research team the benefits of balancing expertise levels in order to a common outcome.

7.1.2 Interpersonal Relations

RQ2: *How are NLG models designed to provide an adequate production of meanings and a better understanding within critically sensitive domains, such as healthcare and governance?*

We addressed the production of meanings from context enrichment after finishing the last experiments of the three studies: REG, robot journalism, and HCD. In the last approach within the diabetes mellitus domain, we confirmed our hypothesis of applying context from a knowledge perspective (using Semantic Web, NoSQL databases, and different embeddings) and the HCD approach (personas, persona-based corpus, and HCD evaluation) corroborates a more adequate generation of texts considering sensitive domains. After the NLG models analysis presented in Table 21, we selected a fine-tuning technique as the less expensive approach to a domain with low-resource conditions, such as data (healthcare domain), and language (Brazilian Portuguese).

7.1.3 Enriching NLG Models with HCD Approaches

RQ3: *How a human-centered approach can be applied to NLG models to improve users' comprehension and experience?*

We addressed user comprehension, literacy, and experience from an HCD

approach, which presented advantages and fine outcomes to a conversational question-answering AI system in specific domains. The NLG models were guided by context with known techniques and exploratory ones. We proposed a persona-based model in the context of the healthcare domain, which achieved benefits from an HCD perspective to both humans – the research team in general, and the generated results – more comprehensive (simplified or not) texts. We showed competitive results with human-centered context enrichment, giving directions to more adequate implementations within the intersections of humans and NLP, i.e., human-centered NLP. We showed that an HCD approach can benefit language generation not only from the perspective of machines but mostly humans, who are the key actors within NLP – an AI-in-loop¹ approach, a progression of the human-in-the-loop term. Finally, this research brought future directions more specifically to low-resource contexts, where promoting literacy is critical. In addition, the persona approach for a low-resource context can open up new directions to research in languages other than English and promote the production of similar resources (corpora and/or human-centered NLP models).

7.2 Future Work

7.2.1 External Knowledge

The context from syntactic and semantic resources – Semantic Web, DB-Pedia, Wikipedia, ontologies, Open Data, and so on – gave the REG task more possibilities to contextualize outputs. We plan to add a different external resource to the extended REG model: profiles, to increase the performance and generate more diverse outputs within the deforestation domain. In relation to the healthcare domain, we plan to extend the use of different corpora and ontologies from the clinical domain, since we demonstrated the improvements with clinical data. Following this, augmenting training data from persona-based English corpora (Mazaré et al., 2018) by translating to Portuguese could explore out-of-domain data

¹ A discussion about the *Human-in-the-loop* term and its replacement with *AI-in-the-loop* and *humans in charge* is available at <https://hai.stanford.edu/events/2022-hai-fall-conference-ai-loop-humans-charge>.

and consequently experiment with the human-centered NLP approach. Another possibility of extending our corpora is in relation to the persona information, lay and expert users, by adding corpora from clinical data and aligning with respective personas: (*lay* $\rightarrow$ *patient*) and (*expert* $\rightarrow$ *doctor*).

7.2.2 Interpersonal Relations

This work explored an explainable AI framework in order to promote interpersonal relationships. Other possibilities of using XAI can be by extending the generation to more detailed explanations (Majumder, 2023), that is, giving users more detailed feedback about the generated text by connecting this feedback with the respective persona. In the same vein, the level of explanation of the generated text could be summarized as healthcare in-domain summarization tasks (Abacha et al., 2023).

7.2.3 Enriching NLG Models with HCD Approaches

The intersections between HCI and NLP have significantly grown these last years, adding a human-centered perspective to the development of language models. There are works in conversational AI that could be explored in order to address the persona-based clinical data in terms of the evaluation of language models with human-centered metrics. We explored the personas in contrast to the generated data as a human evaluation metric and this work can be extended to known usability scales within HCI. We can envisage human-centered NLP as a pillar within NLP focused on guiding language models in the direction of humans.

7.3 Ethical Considerations

As elucidated in the human evaluation's sections, such as the human evaluation of the extension of the NeuralREG model (see Section 4), the corpus de-identification, annotation, and post-edition (please refer to Section 6.4, more specifically to De-Identification Section 6.4.4), and the human-centered expert evaluation (see Section 3.4.2), all subjects are members of the research group and

were responsible for evaluating with an equal amount of occurrences. Code and data are publicly available. No consent from data subjects was required as this data is purely factual, containing no personal data due to the de-identification process conducted during this research.

References

- Abacha, A. B.; Yim, W.-w.; Fan, Y.; Lin, T. 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*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. p. 2291–2302. Available on: <<https://aclanthology.org/2023.eacl-main.168>>.
- Abacha, A. B.; Zweigenbaum, P. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information processing & management*, Elsevier, v. 51, n. 5, p. 570–594, 2015.
- Abdellatif, A.; Badran, K.; Costa, D. E.; Shihab, E. A comparison of natural language understanding platforms for chatbots in software engineering. *IEEE Transactions on Software Engineering*, IEEE, v. 48, n. 8, p. 3087–3102, 2021.
- Act, A. C.; Representatives, U. H. o. *Patient Protection and Affordable Care Act 42 USC § 18001 et seq.* 2010.
- Al-Thanyyan, S. S.; Azmi, A. M. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 54, n. 2, p. 1–36, 2021.
- Alarcon, R.; Moreno, L.; Martínez, P. Easier corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, Public Library of Science San Francisco, CA USA, v. 18, n. 4, p. e0283622, 2023.
- Almansor, E. H.; Hussain, F. K. Survey on intelligent chatbots: State-of-the-art and future research directions. In: Springer. *Conference on Complex, Intelligent, and Software Intensive Systems*. 2019. p. 534–543.
- Alsentzer, E. et al. Publicly available clinical BERT embeddings. In: *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019. p. 72–78. Available on: <<https://aclanthology.org/W19-1909>>.
- Aluísio, S.; Gasperin, C. 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*. 2010. p. 46–53.

- Aluísio, S. M.; Specia, L.; Pardo, T. A.; Maziero, E. G.; Fortes, R. P. Towards brazilian portuguese automatic text simplification systems. In: *Proceedings of the eighth ACM symposium on Document engineering*. 2008. p. 240–248.
- Alva-Manchego, F.; Scarton, C.; Specia, L. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info . . . , v. 46, n. 1, p. 135–187, 2020.
- Araujo, R. *Systemic-Functional modeling of text complexity in Brazilian Portuguese*. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2021.
- Arcan, M.; Buitelaar, P. Translating domain-specific expressions in knowledge bases with neural machine translation. *CoRR*, *abs/1709.02184*, 2017.
- Arrieta, A. B. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, Elsevier, v. 58, p. 82–115, 2020.
- Ataman, D.; Federico, M. An evaluation of two vocabulary reduction methods for neural machine translation. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. 2018. p. 97–110.
- Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2015.
- Baker, M. Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, John Benjamins, v. 7, n. 2, p. 223–243, 1995.
- Bateman, J.; McDonald, D.; Hiippala, T.; Couto-Vale, D.; Costetchi, E. Systemic functional linguistics and computation: New directions, new challenges. *The Cambridge handbook of systemic functional linguistics*, p. 561–586, 2019.
- Bateman, J.; Zock, M. Natural language generation. In: *The Oxford Handbook of Computational Linguistics. 2nd edition.* : Oxford Handbooks Online, 2016.
- Beltagy, I.; Lo, K.; Cohan, A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.
- Bercken, L. Van den; Sips, R.-J.; Lofi, C. Evaluating neural text simplification in the medical domain. In: *The World Wide Web Conference*. 2019. p. 3286–3292.

- Berman, J. J. Chapter 5 - identifying and deidentifying data. In: Berman, J. J. (Ed.). *Data Simplification*. Boston: Morgan Kaufmann, 2016. p. 189–231. ISBN 978-0-12-803781-2. Available on: <https://www.sciencedirect.com/science/article/pii/B9780128037812000059>.
- Bizagi. *Bizagi Modeler*. 2023. <https://www.bizagi.com/>.
- Blackwood, G.; Ballesteros, M.; Ward, T. Multilingual neural machine translation with task-specific attention. *arXiv preprint arXiv:1806.03280*, 2018.
- Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*, 2017.
- Broderick, J. et al. *Designing health literate mobile apps*. : Institute of Medicine of the National Academies Washington, DC, 2014.
- Brown, T. et al. Language models are few-shot learners. *Advances in neural information processing systems*, v. 33, p. 1877–1901, 2020.
- Budler, L. C.; Gosak, L.; Stiglic, G. Review of artificial intelligence-based question-answering systems in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 13, n. 2, p. e1487, 2023.
- Bunk, T.; Varshneya, D.; Vlasov, V.; Nichol, A. Diet: Lightweight language understanding for dialogue systems. *arXiv preprint arXiv:2004.09936*, 2020.
- Cai, L. et al. Incorporating domain knowledge for biomedical text analysis into deep learning: A survey. *Journal of Biomedical Informatics*, Elsevier, p. 104418, 2023.
- Campos, J.; Teixeira, A.; Ferreira, T.; Cozman, F.; Pagano, A. Towards fully automated news reporting in brazilian portuguese. In: SBC. *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*. 2020. p. 543–554.
- Canva. *Canva Design*. 2023. <https://www.canva.com/design/>.
- Cao, M.; Cheung, J. C. K. 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)*. 2019. p. 3154–3163.
- Cardon, R.; Grabar, N. 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*. 2018. p. 83–93.

- Cardon, R.; Grabar, N. French biomedical text simplification: When small and precise helps. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020. p. 710–716.
- Carlson, M. The robotic reporter. *Digital Journalism*, Routledge, v. 3, n. 3, p. 416–431, 2015. Available on: <<https://doi.org/10.1080/21670811.2014.976412>>.
- Carpenter, R. *Xtensio*. 1997. Available on: <<http://www.cleverbot.com/>>.
- Carroll, J. M. Human–computer interaction. *The Encyclopedia of Human-Computer Interaction, 2nd Ed.*, The Interaction Design Foundation, 2009.
- Caseli, H. d. M.; Nunes, M. d. G. V. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. São Carlos: BPLN, 2023. ISBN 978-65-00-80693-9. Available on: <<https://brasileiraspln.com/livro-pln/>>.
- Chatterjee, R. et al. Guiding neural machine translation decoding with external knowledge. In: *Proceedings of the Second Conference on Machine Translation*. 2017. p. 157–168.
- Chaves, F. A.; Cecilio, S. G.; Reis, I. A.; Pagano, A. S.; Torres, H. de C. Translation and cross-cultural adaptation of the behavior change protocol for educational practices in diabetes mellitus. *Revista latino-americana de enfermagem*, Escola de Enfermagem de Ribeirao Preto, Universidade de Sao Paulo, v. 27, 2019.
- Chaves, F. F.; Reis, I. A.; Pagano, A. S.; Torres, H. d. C. Translation, cross-cultural adaptation and validation of the diabetes empowerment scale–short form. *Revista de saude publica*, SciELO Public Health, v. 51, p. 16, 2017.
- Chiang, Y.-T. et al. Developing the “healthcare ceo app” for patients with type 1 diabetes transitioning from adolescence to young adulthood: A mixed-methods study. *Nursing Open*, Wiley Online Library, v. 10, n. 3, p. 1755–1766, 2023.
- Chu, C.; Dabre, R. Multilingual and multi-domain adaptation for neural machine translation. In: *Proceedings of the 24th Annual Meeting of the Association for Natural Language Processing (NLP 2018)*. 2018. p. 909–912.
- Chu, C.; Wang, R. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*, 2018.
- Chu, E.; Vijayaraghavan, P.; Roy, D. Learning personas from dialogue with attentive memory networks. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018. p. 2638–2646.

- Colby, K. M.; Weber, S.; Hilf, F. D. Artificial paranoia. *Artificial Intelligence*, Elsevier, v. 2, n. 1, p. 1–25, 1971.
- Cooper, A.; Reimann, R.; Cronin, D.; Noessel, C. *About face: the essentials of interaction design*. : John Wiley & Sons, 2014.
- Cortez, D. N.; Souza, D. A. S.; Oliveira-Cortez, A.; Reis, I. A.; Torres, H. de C. Efeito individual das atitudes e empoderamento para o autocuidado em diabetes mellitus. *Saúde Coletiva (Barueri)*, v. 12, n. 80, p. 11302–11317, 2022.
- Cripwell, L.; Legrand, J.; Gardent, C. Document-level planning for text simplification. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, 2023. p. 993–1006. Available on: <<https://aclanthology.org/2023.eacl-main.70>>.
- Cunha, R. *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]*. Dissertação (Mestrado) — Federal University of Santa Catarina, Florianópolis/SC - Brazil, 8 2016.
- Cunha, R.; Fernandes, L. P. Avaliação comparativa de ferramentas web de apoio à tradução através de critérios de usabilidade e ergonomia cognitiva. *Tradução em Revista*, v. 21, p. 67–90, 2016.
- Cunha, R.; Ferreira, T. C.; Pagano, A.; Alves, F. Referring to what you know and do not know: Making referring expression generation models generalize to unseen entities. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020. p. 2261–2272. Available on: <<https://aclanthology.org/2020.coling-main.205>>.
- Curry, A. C. et al. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking. *Alexa Prize Proceedings*, 2018.
- Curry, A. C.; Rieser, V. Inclusive conversational ai: The case of female voice assistants. *Inclusive Conversational AI*, 2020.
- Dale, R.; Reiter, E. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, Elsevier, v. 19, n. 2, p. 233–263, 1995.

- Deemter, K. van. Designing algorithms for referring with proper names. In: *Proceedings of the 9th International Natural Language Generation conference*. 2016. p. 31–35.
- Devaraj, A.; Wallace, B. C.; Marshall, I. J.; Li, J. J. Paragraph-level simplification of medical texts. In: NIH Public Access. *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. 2021. v. 2021, p. 4972.
- Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. p. 4171–4186. Available on: <https://aclanthology.org/N19-1423>.
- Dew, K. N.; Turner, A. M.; Choi, Y. K.; Bosold, A.; Kirchhoff, K. Development of machine translation technology for assisting health communication: A systematic review. *Journal of biomedical informatics*, Elsevier, v. 85, p. 56–67, 2018.
- Dimitrakis, E.; Sgontzos, K.; Tzitzikas, Y. A survey on question answering systems over linked data and documents. *Journal of Intelligent Information Systems*, Springer, p. 1–27, 2019.
- Dinan, E. et al. The second conversational intelligence challenge (convai2). In: *The NeurIPS'18 Competition*. : Springer, 2020. p. 187–208.
- Diniz, C. G. et al. Deter-b: The new amazon near real-time deforestation detection system. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, v. 8, n. 7, p. 3619–3628, July 2015. ISSN 2151-1535.
- Eisenstein, J. *Natural language processing*. : MIT Press, 2018.
- Erdem, E. et al. Neural natural language generation: A survey on multilinguality, multimodality, controllability and learning. *Journal of Artificial Intelligence Research*, v. 73, p. 1131–1207, 2022.
- Escobar, H. Deforestation in the brazilian amazon is still rising sharply. *Science*, American Association for the Advancement of Science, v. 369, n. 6504, p. 613–613, 2020. ISSN 0036-8075. Available on: <https://science.sciencemag.org/content/369/6504/613>.
- Fadhil, A.; Schiavo, G. Designing for health chatbots. *arXiv preprint arXiv:1902.09022*, 2019.

Fang, H. et al. Sounding board—university of washington’s alexa prize submission. *Alexa prize proceedings*, 2017.

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

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

Ferreira, T. C. et al. 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)*. Held Online: INCOMA Ltd., 2021. p. 234–243. Available on: <https://aclanthology.org/2021.ranlp-1.28>.

Ferreira, T. C.; Krahmer, E.; Wubben, S. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016. p. 568–577.

Ferreira, T. C.; Krahmer, E.; Wubben, S. Generating flexible proper name references in text: Data, models and evaluation. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, 2017. (EACL’17), p. 655–664. Available on: <http://www.aclweb.org/anthology/E17-1062>.

Ferreira, T. C.; Lee, C. van der; Miltenburg, E. van; Krahmer, E. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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)*. 2019. p. 552–562.

Ferreira, T. C. et al. Ana: A brazilian chatbot assistant about covid-19. In: *Proceedings of Conference ACL 2020 Workshop on Natural Language Processing for COVID-19 (NLP-COVID)*. 2020.

Ferreira, T. C.; Moussallem, D.; Kádár, Á.; Wubben, S.; Krahmer, E. Neuralreg: An end-to-end approach to referring expression generation. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018. p. 1959–1969.

Ferreira, T. C.; Moussallem, D.; Krahmer, E.; Wubben, S. Enriching the webnlg corpus. In: *Proceedings of the 11th International Conference on Natural Language Generation*. 2018. p. 171–176.

- Fikri, A.; Takamura, H.; Okumura, M. Stylistically user-specific generation. In: *Proceedings of the 11th International Conference on Natural Language Generation*. 2018. p. 89–98.
- Filho, J. A. W.; Wilkens, R.; Idiart, M.; Villavicencio, A. The brWaC corpus: A new open resource for Brazilian Portuguese. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Available on: <<https://aclanthology.org/L18-1686>>.
- Finch, S. E. et al. Emora: An inquisitive social chatbot who cares for you. *arXiv preprint arXiv:2009.04617*, 2020.
- Gao, J.; Galley, M.; Li, L. Neural approaches to conversational ai. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 2019. p. 1371–1374.
- Gardent, C.; Shimorina, A.; Narayan, S.; Perez-Beltrachini, L. Creating training corpora for nlg micro-planners. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017. p. 179–188.
- Gardent, C.; Shimorina, A.; Narayan, S.; Perez-Beltrachini, L. The webnlg challenge: Generating text from rdf data. In: *Proceedings of the 10th International Conference on Natural Language Generation*. 2017. p. 124–133.
- Garibay, O. O. et al. Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, Taylor & Francis, v. 39, n. 3, p. 391–437, 2023.
- Gasperin, C.; Maziero, E.; Aluisio, S. M. Challenging choices for text simplification. In: Springer. *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9*. 2010. p. 40–50.
- Gatt, A.; Krahmer, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, v. 61, p. 65–170, 2018.
- Gatt, A.; Krahmer, E. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, v. 61, p. 65–170, 2018.

Goldberg, Y. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, Morgan & Claypool Publishers, v. 10, n. 1, p. 1–309, 2017.

Gonzalez, A. *Towards Human-Centered Natural Language Processing*. Tese (Doutorado) — University of Copenhagen, 2021.

Google. *Dialogflow. Lifelike conversational AI with state-of-the-art virtual agents*. 2023. Available on: <<https://cloud.google.com/dialogflow>>.

Gopalakrishnan, K. et al. Topical-chat: Towards knowledge-grounded open-domain conversations. In: *INTERSPEECH*. 2019. p. 1891–1895.

Grabar, N.; Cardon, R. CLEAR – simple corpus for medical French. In: *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*. Tilburg, the Netherlands: Association for Computational Linguistics, 2018. p. 3–9. Available on: <<https://aclanthology.org/W18-7002>>.

Graefe, A. *Guide to Automated Journalism*. New York, 2016.

Gumiel, Y. B.; Lee, I.; Soares, T. A.; Ferreira, T. C.; Pagano, A. Sentiment analysis in portuguese texts from online health community forums: data, model and evaluation. In: SBC. *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. 2021. p. 64–72.

Gunning, D. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, v. 2, p. 2, 2017.

Gunning, D.; Aha, D. Darpa’s explainable artificial intelligence (xai) program. *AI Magazine*, v. 40, n. 2, p. 44–58, Jun. 2019. Available on: <<https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2850>>.

Halliday, M. A. K.; Matthiessen, C. Construing experience through meaning: A language-based approach to cognition. *Cassell, London*, 1999.

Halliday, M. A. K.; Matthiessen, C. M. *Halliday’s introduction to functional grammar*. : Routledge, 2014.

Han, T. et al. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

Harley, A. Personas make users memorable for product team members. *Nielsen Norman Group*, 2015.

- Harrison, V.; Reed, L.; Oraby, S.; Walker, M. Maximizing stylistic control and semantic accuracy in nlg: Personality variation and discourse contrast. In: *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*. 2019. p. 1–12.
- Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural computation*, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.
- Holden, R. J.; Kulanthaivel, A.; Purkayastha, S.; Goggins, K. M.; Kripalani, S. Know thy ehealth user: Development of biopsychosocial personas from a study of older adults with heart failure. *International journal of medical informatics*, Elsevier, v. 108, p. 158–167, 2017.
- Holley, K. L.; Becker, S. *AI First Healthcare*. : O'Reilly Media, Inc., 2021.
- Holtzman, A. et al. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087*, 2018.
- Horn, M.; Li, X.; Chen, L.; Kafle, S. A multi-talent healthcare ai bot platform. *Studies in Computational Intelligence*, Springer, 2019.
- Hughson, J.-a. et al. Health professionals' views on health literacy issues for culturally and linguistically diverse women in maternity care: barriers, enablers and the need for an integrated approach. *Australian Health Review*, CSIRO, v. 42, n. 1, p. 10–20, 2018.
- ISO. Iso 9241-151:2008. ergonomics of human-system interaction - part 151: Guidance on world wide web user interfaces. *International Standardization Organization (ISO)*. Switzerland, 2008.
- ISO. Iso 9241-210: 2019. ergonomics of human-system interaction - part 210: Human-centred design for interactive systems. *International Standardization Organization (ISO)*. Switzerland, 2019.
- Jakobson, R. On linguistic aspects of translation. *On translation*, Cambridge (Mass.), v. 3, p. 30–39, 2000.
- Ji, Z. et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, ACM New York, NY, v. 55, n. 12, p. 1–38, 2023.
- Johnson, A. E. et al. Mimic-iii, a freely accessible critical care database. *Scientific data*, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2016.
- Junczys-Dowmunt, M.; Grundkiewicz, R.; Guha, S.; Heafield, K. Approaching neural grammatical error correction as a low-resource machine translation task. *arXiv preprint arXiv:1804.05940*, 2018.

- Jurafsky, D.; Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd edition. : Pearson London, 2023.
- Knowles, B.; Bates, O.; Håkansson, M. This changes sustainable hci. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018. p. 1–12.
- Koehn, P. *Neural Machine Translation*. : Cambridge University Press, 2020.
- Koehn, P.; Knowles, R. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.
- Koptient, A.; Grabar, N. 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*. 2020.
- Krahmer, E.; Deemter, K. van. Computational generation of referring expressions: An updated survey. In: *The Oxford Handbook of Reference*. : Oxford University Press, 2019.
- Krahmer, E.; Theune, M. Efficient context-sensitive generation of referring expressions. In: *Information sharing: Givenness and newness in language processing*. : CSLI publications, 2002. p. 223–264.
- Kübler-Ross, E.; Wessler, S.; Avioli, L. V. On death and dying. *Jama*, American Medical Association, v. 221, n. 2, p. 174–179, 1972.
- Lacerda, A. R. T. d. *Rasa-ptbr-boilerplate: FLOSS project that enables brazilian portuguese chatbot development by non-experts*. 2019.
- Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; Ranzato, M. Phrase-based & neural unsupervised machine translation. *arXiv preprint arXiv:1804.07755*, 2018.
- Lan, L. Corpus. In: *Routledge encyclopedia of translation technology*. : Routledge, 2015. p. 465–479.
- Läubli, S. et al. A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, v. 67, p. 653–672, 2020.
- Lazar, J.; Feng, J. H.; Hochheiser, H. *Research methods in human-computer interaction*. : Morgan Kaufmann, 2017.

- Lee, C. van der; Gatt, A.; Miltenburg, E. van; Wubben, S.; Krahmer, E. Best practices for the human evaluation of automatically generated text. In: *Proceedings of the 12th International Conference on Natural Language Generation*. 2019. p. 355–368.
- Lee, C. van der; Krahmer, E.; Wubben, S. PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In: *Proceedings of INLG*. Santiago de Compostela, Spain: , 2017. (INLG'2017).
- Lee, J. et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Oxford University Press, v. 36, n. 4, p. 1234–1240, 2020.
- Leppänen, L.; Munezero, M.; Granroth-Wilding, M.; Toivonen, H. Data-driven news generation for automated journalism. In: *Proceedings of INLG*. 2017.
- LeRouge, C.; Ma, J.; Sneha, S.; Tolle, K. User profiles and personas in the design and development of consumer health technologies. *International journal of medical informatics*, Elsevier, v. 82, n. 11, p. e251–e268, 2013.
- Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady*. 1966. v. 10, p. 707–710.
- Lewis, M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.
- Lewis, M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. p. 7871–7880.
- Li, J. et al. A persona-based neural conversation model. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016. p. 994–1003.
- Liang, K. et al. Gunrock 2.0: A user adaptive social conversational system. *Alexa Prize Proceedings*, 2020.
- Lim, J. et al. You truly understand what I need : Intellectual and friendly dialog agents grounding persona and knowledge. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022. p. 1053–1066. Available on: <<https://aclanthology.org/2022.findings-emnlp.75>>.

- Lin, Z. et al. Xpersona: Evaluating multilingual personalized chatbot. *arXiv preprint arXiv:2003.07568*, 2020.
- Liu, Z.; Tang, B.; Wang, X.; Chen, Q. De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of biomedical informatics*, Elsevier, v. 75, p. S34–S42, 2017.
- Luan, Y.; Brockett, C.; Dolan, B.; Gao, J.; Galley, M. Multi-task learning for speaker-role adaptation in neural conversation models. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2017. p. 605–614.
- Majumder, B. P. *User-centric Natural Language Processing*. Tese (Doutorado) — UC San Diego, 2023.
- Majumder, B. P.; Rao, S.; Galley, M.; McAuley, J. Ask what’s missing and what’s useful: Improving clarification question generation using global knowledge. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021. p. 4300–4312.
- Manishina, E. *Data-driven natural language generation using statistical machine translation and discriminative learning*. Tese (Doutorado) — Université d’Avignon, 2016.
- Manning, C. D. Computational linguistics and deep learning. *Computational Linguistics*, MIT Press, v. 41, n. 4, p. 701–707, 2015.
- Marie, B.; Fujita, A. Unsupervised neural machine translation initialized by unsupervised statistical machine translation. *arXiv preprint arXiv:1810.12703*, 2018.
- Mazaré, P.-E.; Humeau, S.; Raison, M.; Bordes, A. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*, 2018.
- McEnery, T.; Hardie, A. *Corpus linguistics: Method, theory and practice*. : Cambridge University Press, 2011.
- McIlwain, C. et al. *Standards for the reporting of Plain Language Summaries in new Cochrane Intervention Reviews (PLEACS)*. 2013.
- Meyes, R.; Lu, M.; Puiseau, C. W. de; Meisen, T. Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*, 2019.

- Mille, S. et al. The first multilingual surface realisation shared task (sr'18): Overview and evaluation results. In: *Proceedings of the First Workshop on Multilingual Surface Realisation*. 2018. p. 1–12.
- Molen, P. van der et al. Identifying user preferences for a digital educational solution for young seniors with diabetes. *Diabetes Spectrum*, Am Diabetes Assoc, v. 30, n. 3, p. 182–187, 2017.
- Montal, T.; Reich, Z. I, robot. you, journalist. who is the author? *Digital Journalism*, Routledge, v. 5, n. 7, p. 829–849, 2017. Available on: <<https://doi.org/10.1080/21670811.2016.1209083>>.
- Moryossef, A.; Goldberg, Y.; Dagan, I. Step-by-step: Separating planning from realization in neural data-to-text generation. 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)*. 2019. p. 2267–2277.
- Moussallem, D.; Wauer, M.; Ngomo, A.-C. N. Machine translation using semantic web technologies: A survey. *Journal of Web Semantics*, Elsevier, v. 51, p. 1–19, 2018.
- Neves, N. W. T. et al. Knowledge of self-care practices in diabetes: compasso. *Research, Society and Development*, v. 10, n. 5, p. e41410515062–e41410515062, 2021.
- Nielsen, L. Personas in co-creation and co-design. In: *Proceedings of the 11th Human-Computer Interaction Research Symposium*. 2011. p. 38–40.
- Nielsen, L. *Personas-user focused design. 2nd edition.* : Springer, 2019.
- Noorbergen, T. J.; Adam, M. T.; Roxburgh, M.; Teubner, T. Co-design in mhealth systems development: insights from a systematic literature review. *AIS Transactions on Human-Computer Interaction*, v. 13, n. 2, p. 175–205, 2021.
- Norman, D. *Human-Centered Design*. The Interaction Design Foundation, 2023. Available on: <<https://www.interaction-design.org/literature/topics/human-centered-design>>.
- Norman, D. A.; Draper, S. W. *User centered system design; new perspectives on human-computer interaction.* : L. Erlbaum Associates Inc., 1986.
- North, K.; Ranasinghe, T.; Shardlow, M.; Zampieri, M. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*, 2023.

- North, K.; Zampieri, M.; Ranasinghe, T. ALEXSIS-PT: A new resource for Portuguese lexical simplification. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, 2022. p. 6057–6062. Available on: <<https://aclanthology.org/2022.coling-1.529>>.
- Novikova, J.; Dušek, O.; Rieser, V. The e2e dataset: New challenges for end-to-end generation. In: *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 2017. p. 201–206.
- Novikova, J.; Dušek, O.; Rieser, V. Rankme: Reliable human ratings for natural language generation. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018. p. 72–78.
- Nunes, L. B.; Santos, J. C. d.; Reis, I. A.; Torres, H. d. C. Avaliação do programa comportamental em diabetes mellitus tipo 2: ensaio clínico randomizado. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 28, p. 851–862, 2023.
- Oliveira, F. S. *O discurso do autocuidado em saúde: uma descrição de gêneros na covariação experto-leigo*. Tese (Doutorado) — Universidade Federal de Minas Gerais, 2022.
- OpenAI. *OpenAI: Introducing ChatGPT*. 2022. Available on: <<https://openai.com/blog/chatgpt>>.
- OpenAI. *GPT-4 Technical Report*. 2023.
- Oraby, S. *Stylistic Control for Neural Natural Language Generation*. Tese (Doutorado) — UC Santa Cruz, 2019.
- Oraby, S.; Harrison, V.; Ebrahimi, A.; Walker, M. 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*. 2019. p. 5938–5951.
- Oraby, S. et al. Controlling personality-based stylistic variation with neural natural language generators. In: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*. 2018. p. 180–190.
- Paaß, G.; Giesselbach, S. *Foundation Models for Natural Language Processing: Pre-trained Language Models Integrating Media*. : Springer Nature, 2023.

- Paes A.; Vianna, V.; Rodrigues, J. Modelos de linguagem. In: _____. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. book chapter 2. ISBN 978-65-00-80693-9. Available on: <https://brasileiraspln.com/livro-pln/1a-edicao/parte7/cap15/cap15.html>.
- Paetzold, G. H.; Specia, L. A survey on lexical simplification. *Journal of Artificial Intelligence Research*, v. 60, p. 549–593, 2017.
- Pagano, A.; Moro, C.; Schneider, E. T. R.; Cintho, L. M. M.; Gumiel, Y. Pln na saúde. In: _____. *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. BPLN, 2023. book chapter 21. ISBN 978-65-00-80693-9. Available on: <https://brasileiraspln.com/livro-pln/1a-edicao/parte9/cap21/cap21.html>.
- Pagano, A. S. A linguagem na construção das práticas educativas nas ciências da saúde. *Heloísa de Carvalho Torres, Ilka Afonso Reis, and Adriana Silvina Pagano. Empoderamento do pesquisador nas ciências da saúde. Belo Horizonte: FALE/UFGM*, p. 19–36, 2015.
- Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In: Association for Computational Linguistics. *Proceedings of the 40th annual meeting on association for computational linguistics*. 2002. p. 311–318.
- Paranjape, A. P. *Neural Systems for Informative Conversations*. Tese (Doutorado) — Stanford University, 2022.
- Patel, S.; Nargunde, R.; Verma, S.; Dhage, S. Summarization and simplification of medical articles using natural language processing. In: IEEE. *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. 2022. p. 1–6.
- Pavanelli, L. A. *An End-to-End Model for Joint Entity and Relation Extraction in Portuguese*. Dissertação (Mestrado) — PUC-Rio, 2022.
- Pearl, C. *Designing voice user interfaces: principles of conversational experiences*. : "O'Reilly Media, Inc.", 2017.
- Phatak, A.; Savage, D. W.; Ohle, R.; Smith, J.; Mago, V. Medical text simplification using reinforcement learning (teslea): Deep learning-based text simplification approach. *JMIR Medical Informatics*, JMIR Publications Toronto, Canada, v. 10, n. 11, p. e38095, 2022.
- Pichl, J. et al. Alquist: The alexa prize socialbot. *arXiv preprint arXiv:1804.06705*, 2018.

- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving language understanding by generative pre-training. *OpenAI*, San Francisco, 2018.
- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. et al. Improving language understanding by generative pre-training. *OpenAI*, 2018.
- Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog*, v. 1, n. 8, p. 9, 2019.
- Raffel, C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, JMLRORG, v. 21, n. 1, p. 5485–5551, 2020.
- Ram, A. et al. Conversational ai: The science behind the alexa prize. *Alexa Prize Proceedings*, 2017.
- Ram, A. et al. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*, 2018.
- Ramos, R. M.; Monteiro, D. S.; Paraboni, I. Personality-dependent content selection in natural language generation systems. *Journal of the Brazilian Computer Society*, SpringerOpen, v. 26, n. 1, p. 1–21, 2020.
- Reiter, E.; Dale, R. *Building Natural Language Generation Systems*. : Cambridge University Press, 2000. (Studies in Natural Language Processing).
- Ribera, M.; Lapedriza, A. Can we do better explanations? a proposal of user-centered explainable ai. In: *IUI Workshops*. 2019.
- Riedl, M. O. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, Wiley Online Library, v. 1, n. 1, p. 33–36, 2019.
- Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; Saenko, K. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.
- Rubin, J.; Chisnell, D. *Handbook of usability testing: how to plan, design and conduct effective tests*. : John Wiley & Sons, 2008.
- Saggion, H. Syntactic simplification. In: _____. *Automatic Text Simplification*. Cham: Springer International Publishing, 2017. p. 33–45. ISBN 978-3-031-02166-4. Available on: <https://doi.org/10.1007/978-3-031-02166-4_4>.

- Saggion, H. et al. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In: *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. Abu Dhabi, United Arab Emirates (Virtual): Association for Computational Linguistics, 2022. p. 271–283. Available on: <https://aclanthology.org/2022.tsar-1.31>.
- Saldanha, G.; O'Brien, S. *Research methodologies in translation studies*. : Routledge, 2014.
- Salminen, J.; Guan, K.; Jung, S.-G.; Jansen, B. J. A survey of 15 years of data-driven persona development. *International Journal of Human–Computer Interaction*, Taylor & Francis, v. 37, n. 18, p. 1685–1708, 2021.
- Salminen, J.; Rao, R. G.; Jung, S.-g.; Chowdhury, S. A.; Jansen, B. J. Enriching social media personas with personality traits: A deep learning approach using the big five classes. In: Springer. *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*. 2020. p. 101–120.
- Santhanam, S.; Shaikh, S. A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions. *arXiv preprint arXiv:1906.00500*, 2019.
- Schneider, E. T. R. et al. Cardiobertpt: Transformer-based models for cardiology language representation in portuguese. In: IEEE. *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. 2023. p. 378–381.
- Schneider, E. T. R.; Souza, J. V. A. de; Gumiel, Y. B.; Moro, C.; Paraiso, E. C. A gpt-2 language model for biomedical texts in portuguese. In: IEEE. *2021 IEEE 34th international symposium on computer-based medical systems (CBMS)*. 2021. p. 474–479.
- Schneider, E. T. R. et al. Biobertpt-a portuguese neural language model for clinical named entity recognition. In: *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 2020. p. 65–72.
- Schwartz, H. A. et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, Public Library of Science, v. 8, n. 9, p. e73791, 2013.
- See, A.; Liu, P. J.; Manning, C. D. Get to the point: Summarization with pointer-generator networks. In: *Proceedings of the 55th Annual Meeting of the*

- Association for Computational Linguistics (Volume 1: Long Papers)*. 2017. p. 1073–1083.
- Sennrich, R.; Haddow, B. Linguistic input features improve neural machine translation. *arXiv preprint arXiv:1606.02892*, 2016.
- Serban, I. V.; Lowe, R.; Henderson, P.; Charlin, L.; Pineau, J. A survey of available corpora for building data-driven dialogue systems. *Dialogue & Discourse*, v. 9, n. 1, p. 1–49, 2018.
- Shardlow, M. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications(IJACSA), Special Issue on Natural Language Processing 2014*, The Science and Information Organization, v. 4, n. 1, 2014. Available on: <<http://dx.doi.org/10.14569/SpecialIssue.2014.040109>>.
- Shardlow, M.; Alva-Manchego, F. 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*. 2022.
- Shardlow, M.; Nawaz, R. Neural text simplification of clinical letters with a domain specific phrase table. Association for Computatioal Linguistics (ACL), 2019.
- Sharp, H.; Preece, J.; Rogers, Y. *Interaction Design: Beyond Human-Computer Interaction. 5th edition.* : Wiley, 2019.
- Shin, J. Y.; Holtz, B. E. Towards better transitions for children with diabetes: user experiences on a mobile health app. In: *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 2019. p. 623–628.
- Siddharthan, A. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, John Benjamins, v. 165, n. 2, p. 259–298, 2014.
- Siddharthan, A.; Nenkova, A.; McKeown, K. Information status distinctions and referring expressions: An empirical study of references to people in news summaries. *Computational Linguistics*, MIT Press, v. 37, n. 4, p. 811–842, 2011.
- Smith, W. et al. Designing an app for pregnancy care for a culturally and linguistically diverse community. In: *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. 2017. p. 337–346.
- Soegaard, M.; Dam, R. F. *The encyclopedia of human-computer interaction, 2nd Ed.* : The Interaction Design Foundation, 2015.

- Song, H.; Wang, Y.; Zhang, K.; Zhang, W.-N.; Liu, T. BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, 2021. p. 167–177. Available on: <https://aclanthology.org/2021.acl-long.14>.
- Souza, F.; Nogueira, R.; Lotufo, R. Bertimbau: pretrained bert models for brazilian portuguese. In: Springer. *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I* 9. 2020. p. 403–417.
- Specia, L.; Aluísio, S. M.; Pardo, T. A. Manual de simplificação sintática para o português. *Txosten teknikoa NILC-TR-08-06*, Sao Carlos-SP, 2008. Available on: http://www.nilc.icmc.usp.br/nilc/download/NILC_TR_08_06.pdf.
- Štajner, S. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, p. 2637–2652, 2021.
- Štajner, S. Automatic text simplification for social good: Progress and challenges. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, 2021. p. 2637–2652. Available on: <https://aclanthology.org/2021.findings-acl.233>.
- Štajner, S. et al. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, Frontiers, v. 5, p. 991242, 2022.
- Stephanidis, C. Design for all. *The Encyclopedia of Human-Computer Interaction*, 2nd Ed., The Interaction Design Foundation, 2015.
- Stephanidis, C. et al. Seven hci grand challenges. *International Journal of Human-Computer Interaction*, Taylor & Francis, v. 35, n. 14, p. 1229–1269, 2019.
- Sulem, E.; Abend, O.; Rappoport, A. Semantic structural evaluation for text simplification. *arXiv preprint arXiv:1810.05022*, 2018.
- Suojanen, T.; Koskinen, K.; Tuominen, T. Usability as a focus of multiprofessional collaboration: A teaching case study on user-centered translation. *connexions: international professional communication journal*, New Mexico Tech, v. 3, n. 2, p. 147–166, 2015.
- Suojanen, T.; Koskinen, K.; Tuominen, T. *User-centered translation*. : Routledge, 2015.

- Szmrecsanyi, B. On operationalizing syntactic complexity. In: *Le poids des mots. Proceedings of the 7th international conference on textual data statistical analysis. Louvain-la-Neuve*. 2004. v. 2, p. 1032–1039.
- Tars, S.; Fishel, M. Multi-domain neural machine translation. *arXiv preprint arXiv:1805.02282*, 2018.
- Teixeira, A. L. R. et al. DaMata: A robot-journalist covering the Brazilian Amazon deforestation. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, 2020. p. 103–106. Available on: <https://aclanthology.org/2020.inlg-1.15>.
- Theune, M.; Klabbers, E.; Pijper, J. R. D.; Krahmer, E.; Odijk, J. From data to speech: a general approach. *Natural Language Engineering*, Cambridge University Press, v. 7, n. 1, p. 47–86, 2001.
- Touvron, H. et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Trujillo, A.; Senette, C.; Buzzi, M. C. Persona design for just-in-time adaptive and persuasive interfaces in menopause self-care. In: Springer. *International Conference of Design, User Experience, and Usability*. 2018. p. 94–109.
- Vaswani, A. et al. Attention is all you need. *Advances in neural information processing systems*, Curran Associates, Inc., v. 30, 2017. Available on: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Vieira, G. d. L. C.; Pagano, A. S.; Reis, I. A.; Rodrigues, J. S. N.; Torres, H. d. C. Translation, cultural adaptation and validation of the diabetes attitudes scale-third version into brazilian portuguese 1. *Revista Latino-Americana de Enfermagem*, SciELO Brasil, v. 25, 2017. Available on: <https://www.scielo.br/j/rlae/a/JvN46pZWtpkMDG5LcQKRjJR/?lang=en>.
- Vilar, D. Learning hidden unit contribution for adapting neural machine translation models. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. 2018. p. 500–505.
- Vinyals, O.; Le, Q. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 2015.
- Vosbergen, S. et al. Using personas to tailor educational messages to the preferences of coronary heart disease patients. *Journal of biomedical informatics*, Elsevier, v. 53, p. 100–112, 2015.

- W3C. *Resource Description Framework (RDF)*. 2014. Available on: <https://www.w3.org/RDF/>.
- Wang, H.; Li, J.; Wu, H.; Hovy, E.; Sun, Y. Pre-trained language models and their applications. *Engineering*, Elsevier, 2022.
- Wang, R.; Finch, A.; Utiyama, M.; Sumita, E. Sentence embedding for neural machine translation domain adaptation. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2017. p. 560–566.
- Wang, W.; Feng, S.; Chen, L.; Wang, D.; Zhang, Y. Learning to improve persona consistency in conversation generation with information augmentation. *Knowledge-Based Systems*, Elsevier, v. 228, p. 107246, 2021.
- Wang, Z. J.; Choi, D.; Xu, S.; Yang, D. Putting humans in the natural language processing loop: A survey. In: *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, 2021. p. 47–52. Available on: <https://aclanthology.org/2021.hcinlp-1.8>.
- Weizenbaum, J. Eliza - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, ACM New York, NY, USA, v. 9, n. 1, p. 36–45, 1966.
- Wolf, T.; Sanh, V.; Chaumond, J.; Delangue, C. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*, 2019.
- Woods, L.; Cummings, E.; Duff, J.; Walker, K. The development and use of personas in a user-centred mhealth design project. In: *Proceedings of the 29th Australian Conference on Computer-Human Interaction*. 2017. p. 560–565.
- Xu, W. Toward human-centered ai: a perspective from human-computer interaction. *interactions*, ACM New York, NY, USA, v. 26, n. 4, p. 42–46, 2019.
- Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; Callison-Burch, C. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 4, p. 401–415, 2016.
- Yadav, S.; Gupta, D.; Abacha, A. B.; Demner-Fushman, D. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, Elsevier, v. 128, p. 104040, 2022.

Yang, D. *CS329X: Human-Centered NLP – Human-Centered NLP*. 2023. Accessed: 13 July 2023). Available on: <http://web.stanford.edu/class/cs329x/slides/s1_introduction.pdf>.

Yang, D. *Human-Centered Natural Language Processing Will Produce More Inclusive Technologies*. Stanford HAI, 2023. (Accessed 9 June 2023). Available on: <<https://hai.stanford.edu/news/diyi-yang-human-centered-natural-language-processing-will-produce-more-inclusive-technologies>>.

Yee, O. K. O. Natural language processing. In: *Routledge Encyclopedia of Translation Technology*. : Routledge, 2015. p. 601–615.

Yui, S.; Ramea, K.; Plochowietz, A.; Cromarty, B.; Pirmoradi, N. *Data-driven User Model Development of a Conversational AI Chatbot to Provide Assistance to Unemployed People during the COVID-19 Pandemic*. : OSF Preprints, 2023.

Yunxiang, L.; Zihan, L.; Kai, Z.; Ruilong, D.; You, Z. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

Zanettin, F. Translation and corpus design. *SYNAPS - A Journal of Professional Communication*, NHH, 2011.

Zhang, M.; Liu, Y.; Luan, H.; Sun, M. Adversarial training for unsupervised bilingual lexicon induction. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017. p. 1959–1970.

Zhang, S. et al. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*, 2018.

Zhang, W.-N.; Zhu, Q.; Wang, Y.; Zhao, Y.; Liu, T. Neural personalized response generation as domain adaptation. *World Wide Web*, Springer, v. 22, n. 4, p. 1427–1446, 2018.

Zhang, Y. et al. Dialogpt: Large-scale generative pre-training for conversational response generation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. p. 270–278.

Zhou, L.; Gao, J.; Li, D.; Shum, H.-Y. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, MIT Press, v. 46, n. 1, p. 53–93, 2020.

Zilio, L. et al. 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)*. Marseille, France: European Language Resources Association, 2020. p. 70–76. ISBN 979-10-95546-45-0. Available on: <https://aclanthology.org/2020.readi-1.11>.

Appendix

APPENDIX A – DaMata's Under-review Paper

Imaginary Numbers! Evaluating Numerical Referring Expressions by Neural End-to-End Surface Realization Systems

Anonymous ACL submission

Abstract

Neural end-to-end surface realizers output more fluent texts than classical architectures. However, they tend to suffer from adequacy problems, in particular *hallucinations* in numerical referring expression generation. This poses a problem to language generation in sensitive domains, as is the case of robot journalism covering COVID-19 and Amazon deforestation. We propose an approach whereby numerical referring expressions are converted from digits to plain word form descriptions prior to being fed to state-of-the-art Large Language Models. We conduct automatic and human evaluations to report the best strategy to numerical superficial realization. Code and data will be made publicly available.

1 Introduction

With the phenomenon of deep learning and its success in other text generation tasks, such as machine translation (Akhbardeh et al., 2021), Surface Realization approaches of *data-to-text* systems have moved from modular pipeline architectures (Reiter and Dale, 2000) to end-to-end ones, which convert a plain meaning representation into text with no explicit intermediate representations (Wen et al., 2015; Dušek and Jurčfček, 2016; Lebre et al., 2016; Gehrmann et al., 2018). Such end-to-end surface realizers do not require annotated intermediate representations and can be easily trained with parallel data to natural language output. While early neural data-to-text systems required a high amount of parallel training data, state-of-the-art architectures, known as Large Language Models (LLMs) (Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020a), have become available with some implicit linguistic awareness - obtained via pre-training - and require fine-tuning the desired generation task with far less parallel data, making the development process even easier. Plus than that, recent large language models do not even to be trained, obtain-

ing good results in a zero-/few-shot setting through prompt commands.

With respect to linguistic output, neural end-to-end surface realizers appear to generate more fluent text than classical pipeline architectures, but are more likely to suffer from adequacy problems, in particular, *hallucinations* (Ji et al., 2023), whereby the system produces text that contains information which is not present in the input representation. A particular hallucination problem that modern approaches seem to struggle with, unlike classical architectures, is numerical referring expression generation.

End-to-end surface realizers appear to have difficulties in generating the correct numbers contained in an input, even though they are able to ensure fluency in the texts. For instance, let’s hypothesize the case where a surface realizer produces the outcome “The country registered **458,098** cases of COVID-19” whereas the gold-standard reference points to “The country registered **408,098** cases of COVID-19”. Albeit there is only a single-digit difference between both texts (which can be overlooked by popular automatic quality metrics), the difference represents an arithmetic change of 50,000 and may lead readers to make drastic errors given the sensitivity of the context.

To the best of our knowledge, this problem has never been investigated in surface realization systems, despite having been addressed in other generation tasks such as text normalization (Zhang et al., 2019; Sproat, 2022), question-answering (Chen et al., 2021; Kim et al., 2022), and *text-to-speech* (Nikulásdóttir and Guðnason, 2019), which struggle to synthesize texts with numerical referring expressions represented by digits. To circumvent the problem, input texts are normalized by converting numerical referring expressions from digits to plain word form descriptions prior to being fed to a *text-to-speech* system (Nikulásdóttir and Guðnason, 2019).

To better investigate the phenomenon in the context

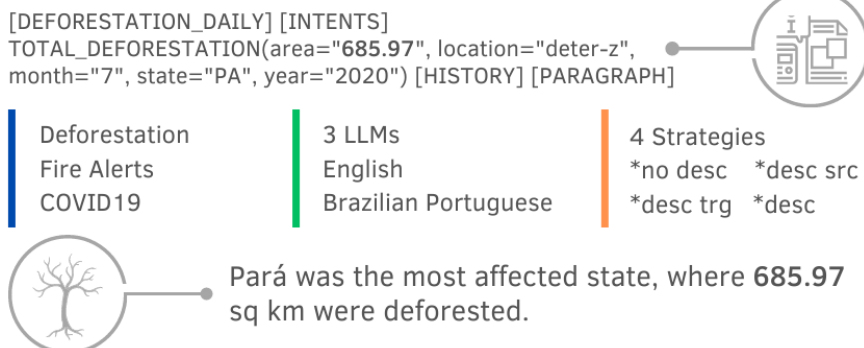


Figure 1: Summary of our proposed approach to solve the challenges when dealing with Numeric Referring Expressions.

of end-to-end surface realizers, this study raises two questions:

- (RQ1) How well do state of the art neural end-to-end surface realizers generate numerical referring expressions?
- (RQ2) Are numerical referring expressions better verbalized when represented by digits or text?

To answer these questions, we conducted automatic and human evaluations with three state-of-the-art (SOTA) Large Language Models (LLMs) as end-to-end surface realization approaches: GPT-2 (Radford et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2020b). Each one of these English LLMs and their multilingual counterparts was used to verbalize English and Brazilian Portuguese news about Amazon Deforestation, Fire Alerts, and COVID-19 cases. To evaluate the best strategy to represent numeric expressions, each of the 3 LLM surface realizers was trained in 4 versions with:

1. numbers represented by digits in the input meaning representation and the reference texts (*no desc*);
2. numbers described in their spell-out forms (i.e. *fifty-five*) in the input meaning representation but as digits (i.e. 55) in the target references (*desc src*);
3. numbers represented by digits (i.e. 55) in the meaning representations but in their spell-out forms (i.e. *fifty-five*) in the target references (*desc trg*);

	Portuguese			English		
	Train	Dev	Test	Train	Dev	Test
Daily Deforest.	2,031	252	242	10	10	10
Month Deforest.	162	10	11	159	8	10
Daily Fire	471	54	54	-	-	-
COVID-19	532	61	54	-	-	-
Total	3,196	377	361	169	18	20

Table 1: Data Statistics.

INPUT	
[DEFORESTATION_MONTH][INTENTS]	TOTAL_DEFORESTATION (area="322.91", location="deter-amz", month="4", year="2021") [HISTORY] [PARAGRAPH]
PORTUGUESE OUTPUT	
O Instituto Nacional de Pesquisa Espaciais (INPE) informou que foram desmatados 322.91 km² na Amazônia Legal, em abril de 2021.	
ENGLISH OUTPUT	
The National Institute for Space Research (INPE) detected 322.91 sq km of deforestation in the Legal Amazon in April 2021.	

Figure 2: Example of Portuguese and English Meaning Representation inputs and their corresponding outputs.

4. numbers in their spell-out forms (i.e. *fifty-five*) in both meaning representations and target references (*desc*).

Results show that surface realization approaches based on the BART and T5 architectures are less prone to generating erroneous numerical referring expressions, whereas the simple trick of representing numbers in text instead of digits leads to significantly lower numbers of errors in this kind of reference as shown in Table 2 of the corresponding example in Figure 2.

2 Data

For training and evaluating the models, we used automatic-generated reports by *BotsDoBem*, a group of Twitter robot-journalists such as CoronaReporter¹ and DaMata², which publish news in Brazilian Portuguese and English. For Brazilian Portuguese, the data comprises monthly and daily reports on Deforestation in the Legal Amazon area of Brazil (Rosa Teixeira et al., 2020), daily reports about Fires in the Brazilian Biomes, as well as COVID-19 cases in the country (Campos et al., 2020). For English, the data comprises Daily and Monthly Reports on deforestation in the Legal Amazon. Although automatically generated, these texts contain a high number of numerical referring expressions, being suitable to our goal of evaluating how well a neural end-to-end surface realizer generates numerical referring expressions. Table 1 introduces the number of instances per language and domain, split into training, development, and test sets. Each instance in the corpus consists of a meaning representation and a corresponding gold-standard verbalization in Brazilian Portuguese or English representing the sentence of a report. For both languages, the verbalizations were automatically generated by the pipeline system described in Rosa Teixeira et al. (2020) and Campos et al. (2020).

Figure 2 depicts how each instance of the English and Portuguese datasets looks like. The meaning representation starts with a tag representing the report domain, followed by a tag that marks the beginning of the sentence intents (e.g. INTENTS). Each intent in the meaning representation follows the *intent-attribute-value* schema. Finally, the tag [HISTORY] marks where the verbalization of the previous sentences in the paragraph of the target will be depicted. In the example, the tag [PARAGRAPH] means that the target sentence is at the beginning of the paragraph. The datasets and code are publicly available.

3 Generating Numerical Referring Expressions

The main purpose of this study was to evaluate how well neural end-to-end surface realizers can generate numerical referring expressions. For other language generation tasks, such as text synthesis, we know that neural models have difficulties in

handling numbers when represented by digits. For instance, text synthesis approaches are not able to synthesize a proper speech when the textual input contains numerical referring expressions represented by digits. In order to overcome the problem, texts to be synthesized go through several normalization methods, such as converting digits (e.g. 98) into their word form description (e.g. *ninety-eight*). To evaluate how well a neural end-to-end surface realizer can generate numerical referring expressions, we evaluate two forms of number representation: by digits or word form description, both in the meaning representations as well as in the verbalizations, totaling 4 different conditions: (1) numbers represented by digits in the meaning representation and the reference texts (*no desc*); (2) numbers are described in the input meaning representation in spell-out form and digits in the target references (*desc src*); (3) numbers represented by digits in the meaning representations and spell-out form descriptions in the target references (*desc trg*); and (4) numbers described in spell-out form in both the input meaning representations and target references (*desc*). To exemplify, Table 2 depicts the 4 variations of a pair of meaning representation and its corresponding English verbalization.

To convert a digit representation into a text numerical representation, we used the *num2words* Python library³, which works for both English and Brazilian Portuguese.

4 Base Surface Realization Model

To address our first research question (RQ1), we evaluate the performance of three state-of-the-art LLMs in generating numerical references. Specifically, we fine-tuned and compared GPT-2, BART, and T5 for English domains. Additionally, for Portuguese, we fine-tuned GPT-2, as well as mBART-50 and mT5, which are the multilingual versions of BART and T5, respectively. The model training process involved 30 epochs, a learning rate of 1e-5, a batch size of 1, 5 early stops, and a maximum token length of 300.

This study focuses on evaluating the accuracy of large language models in generating numerical references. By comparing different models and considering factors such as data preprocessing, tokenization techniques, and domain-specific context, we aim to gain insights into their performance. The

¹<https://twitter.com/CoronaReporter>

²<https://twitter.com/DaMataReporter>

³<https://pypi.org/project/num2words/>

Strategies	Area	Numeric Referring Expressions		
		Month	Year	
<i>no desc</i>	322.91	4	2021	In April <u>2021</u> , <u>322.91</u> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
<i>desc src</i>	three hundred and twenty-two point nine one	four	two thousand and twenty-one	In April <u>2021</u> , <u>322.91</u> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
<i>desc trg</i>	322.91	4	2021	In April <i>two thousand and twenty-one</i> , <i>three hundred and twenty-two point nine one</i> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).
<i>desc</i>	three hundred and twenty-two point nine one	four	two thousand and twenty-one	In April <i>two thousand and twenty-one</i> , <i>three hundred and twenty-two point nine one</i> sq km of the Legal Amazon were deforested, according to data from the National Institute for Space Research (INPE).

Table 2: The strategies and how they were used to represent the numeric referring expressions. They are highlighted. These strategies follow the example presented in figure 2.

evaluation process involves assessing the models' ability to handle various numerical expressions and their adaptability to different languages.

5 Evaluation

5.1 Models

For each language, we fine-tuned and evaluated 3 large language models according to the 4 numerical strategies, totaling 12 approaches per language. For English, we used GPT-2, BART and T5 language models, whereas for Portuguese, we used its Portuguese version - GPT-2-pt, and the multilingual versions of BART and T5 mBART and mT5, respectively.

5.2 Automatic Evaluation

We computed the BLEU score (Papineni et al., 2002) of the system to analyse the fluency of the generated texts automatically, and analyse whether errors in numerical referring expressions can reflect on its result.

5.3 Human Evaluation

To answer our research questions (RQ1) and (RQ2), we performed a human evaluation against the outcomes of our evaluated approaches.

Method We perform the human evaluation following the methodology of Thomson and Reiter (2020), which aims to quantify the quality of automatically generated texts according to the following taxonomy of errors: *Incorrect Number*, *Incorrect Named Entity*, *Incorrect Word*, *Context Error*, *Not Checkable* and *Other*. Besides these categories, a *Fluency Error* category was incorporated into

the evaluation, which allowed raters to assess the output for problems whether the text flow is not acceptable. We are primarily interested in the dimensions concerning the number errors i.e. *Incorrect Number* and *Incorrect Word*. We also drew on best practices with respect to error analysis and reporting as described in van Miltenburg et al. (2021).

Data preparation and Annotation process Errors were annotated in a binary categorical fashion, using a Google spreadsheet where annotators were instructed to use “y” (yes) or “n” (no) to rate whether the entries had errors according to the aforementioned dimensions. Data for rating was prepared by distributing entries in a collaborative spreadsheet, where each row deployed the original Meaning Representation (MR), the generated hypothesis, and the rating categories. Conditional formatting was used in the rating spreadsheets as a way to visually aid the raters and thus ease the error annotation process. Information about the language models used to generate the entries as well as the number representation strategy was not disclosed to annotators / omitted. Annotators were instructed to comment on difficult cases for further group discussion. The annotation guidelines are summarised below with additional details in the supplement:

- Entries were distributed in a collaborative spreadsheet.
- Each row consisted of the original Meaning Representation (MR), the generated hypothesis, and the rating categories.

Error Rate Full Results for English (EN)									
Strategies	LM	Number	Named Entity	Word	Context	Uncheckable	Other	Fluency	BLEU
No Desc	T5	0.48	0.05	0.08	0.03	0.03	0.08	0.05	0.69
	GPT2	0.65	0.28	0.03	0.53	0.08	0.60	0.95	0.14
	BART	0.50	0.00	0.00†	0.08	0.00†	0.00	0.15	0.61
	Avg.	0.54*	0.11	0.03	0.21	0.03	0.23	0.38	0.48
Desc Source	T5	0.45†	0.00†	0.00†	0.00†	0.05	0.00†	0.08	0.69
	GPT2	1.00	0.05	0.00†	0.03	0.00†	0.00†	0.23	0.41
	BART	0.48	0.00†	0.00†	0.05	0.00†	0.03	0.23	0.62
	Avg.	0.64	0.02*	0.00*	0.03*	0.02	0.01	0.18	0.57
Desc Target	T5	0.95	0.00†	0.00†	0.03	0.00†	0.00†	0.00†	0.87†
	GPT2	0.95	0.03	0.00†	0.13	0.13	0.00†	0.18	0.35
	BART	0.90	0.05	0.08	0.05	0.03	0.00†	0.18	0.60
	Avg.	0.93	0.03	0.03	0.07	0.05	0.00*	0.12	0.60*
Desc	T5	0.93	0.00†	0.03	0.03	0.00†	0.00†	0.00†	0.66
	GPT2	0.90	0.13	0.03	0.23	0.03	0.05	0.25	0.28
	BART	1.00	0.00†	0.03	0.03	0.00†	0.00†	0.05	0.58
	Avg.	0.94	0.04	0.03	0.09	0.01*	0.02	0.10*	0.50

Table 3: English error rates and BLEU score for the 4 numerical strategies and 3 language models - Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.

Error Rate Full Results for Portuguese (PT)									
Strategies	LM	Number	Named Entity	Word	Context	Uncheckable	Other	Fluency	BLEU
No Desc	mT5	0.45	0.02	0.08	0.03	0.03	0.07	0.03	0.58
	GPT2-pt	0.24	0.04	0.03	0.03	0.01	0.14	0.09	0.60
	mBART	0.34	0.04	0.00†	0.09	0.01	0.07	0.10	0.51
	Avg.	0.34	0.04	0.04*	0.05*	0.02	0.09	0.07	0.56
Desc Source	mT5	0.37	0.01	0.08	0.05	0.01	0.04	0.02	0.59
	GPT2-pt	0.19†	0.01	0.02	0.12	0.01	0.08	0.03	0.61
	mBART	0.28	0.03	0.01	0.01	0.01	0.13	0.05	0.59
	Avg.	0.28*	0.01	0.04	0.06	0.01	0.09	0.03*	0.60
Desc Target	mT5	0.87	0.00†	0.04	0.00†	0.00†	0.01†	0.02†	0.65
	GPT2-pt	0.90	0.01	0.04	0.11	0.00†	0.08	0.05	0.64
	mBART	0.79	0.06	0.09	0.09	0.00†	0.10	0.09	0.61
	Avg.	0.85	0.02	0.06	0.07	0.00*	0.06	0.05	0.64
Desc	mT5	0.90	0.00†	0.05	0.02	0.00†	0.01	0.06	0.68†
	GPT2-pt	0.80	0.01	0.12	0.14	0.00†	0.03	0.15	0.67
	mBART	0.89	0.00	0.07	0.03	0.00†	0.03	0.15	0.65
	Avg.	0.87	0.00*	0.08	0.06	0.00*	0.02	0.12	0.67*

Table 4: Portuguese error rates and BLEU score for the 4 numerical strategies and 3 language models. Higher error rates denote more errors. Higher BLEU scores denote greater Fluency. * Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.

- 284 - LLMs used to generate the entries were omitted in the spreadsheet.
- 285
- 286 - The spreadsheet was formatted to highlight
- 287 the options (y - red; n - green) aiming to
- 288 aid/ease the process with visual cues.
- 289 - Difficult cases were commented on to be further
- 290 discussed within the group of raters, fostering
- 291 improvements in the guidelines.
- 292 - Once, the annotation was finished, the spreadsheets
- 293 were exported in .csv files for result computation.
- 294
- 295 Overall, we selected a stratified sample (20% of
- 296 total output) of 852 instances from the Brazilian

Portuguese output (per strategy and per model) that were distributed among 3 annotators (each of them is proficient in the language and the domain and a member of the research team) for evaluation. A duplicate of the batch was redistributed among the same 3 raters to allow for double annotation. For English, all 240 outputs (per strategy per model) were annotated independently by 2 raters (each of them is proficient in the language and the domain and is also a member of the research team) one per batch, allowing for double annotation. We initially piloted the annotation process for both languages with a sample for 50 instances from each dataset, in order to resolve any ambiguities in the annotation guidelines, before proceeding with the full

297

298

299

300

301

302

303

304

305

306

307

308

309

310

311

Incorrect Number Error Rate							
Strategies	Language Model	English (EN)			Portuguese (PT)		
		Deforestation Monthly	Deforestation Daily	Overall	Deforestation Monthly	Deforestation Daily	Overall
No Desc	T5/mT5	0.50	0.45†	0.48	0.55	0.18	0.36
	GPT2/GPT2-pt	0.65	0.65	0.65	0.00†	0.00†	0.00†
	BART/mBART	0.50	0.50	0.50	0.18	0.09	0.14
	Avg.	0.55*	0.53*	0.54*	0.24*	0.09	0.17*
Desc Source	T5/mT5	0.45†	0.45†	0.45†	0.73	0.07	0.40
	GPT2/GPT2-pt	1.00	1.00	1.00	0.27	0.00†	0.14
	BART/mBART	0.50	0.45†	0.48	0.45	0.00†	0.23
	Avg.	0.65	0.63	0.64	0.48	0.02*	0.25
Desc Target	T5/mT5	0.95	0.95	0.95	1.00	0.68	0.84
	GPT2/GPT2-pt	0.95	0.95	0.95	0.82	0.74	0.78
	BART/mBART	0.95	0.85	0.90	0.82	0.55	0.69
	Avg.	0.95	0.92	0.93	0.88	0.66	0.77
Desc	T5/mT5	1.00	0.85	0.93	1.00	0.68	0.84
	GPT2/GPT2-pt	0.90	0.90	0.90	0.82	0.52	0.67
	BART/mBART	1.00	1.00	1.00	1.00	0.69	0.84
	Avg.	0.97	0.92	0.94	0.94	0.63	0.78
Kappa Statistic		0.94	0.92	0.93	1.00	0.99	0.97

Table 5: English and Portuguese “Incorrect Number” error rates results, grouped by number description strategies - higher values denote more errors. For comparison, we show only results for Monthly and Daily Deforestation, since these domains are present in both languages.

* Lowest error rate among strategies averages; † Lowest error rate among model and strategy combinations.

annotation task.

It is worth noting that for the Portuguese dataset annotators evaluated different entries in the first and second batches, allowing for inter-rater agreement assessment. To reduce bias during double annotation, access to corresponding entries in different batches was not allowed. With respect to both datasets, in line with Thomson and Reiter (2020) methodology, we removed disagreement as a result of raters not following annotations guidelines.

6 Results

Table 3 and 4 depict the error rates and BLEU scores results of per numerical strategy and model for English and Portuguese, respectively. Across the different types of errors, we see that numerical ones are the most frequent in both languages. However, when focusing in on the numerical error rates, results diverge across language, strategy, and models.

In English, average results per strategy show that representing numerical references with texts instead of digits did not have a positive impact, since the No Desc strategy, where numbers are represented with digits in the input meaning representation and in the automatically-generated texts, presented the lowest error rate. On the other hand, when inspecting the results per model, T5, when numbers in the input source are provided in spell-out form instead of digits (Desc Source),

presents the lowest error rate, followed by BART using the same strategy, and T5 with digit numerical references in the meaning representations and generated texts (e.g. No Desc). In the automatic evaluation, the strategy where numbers are described with digits in the source MR and with text in the target output (e.g. Desc Target) depicted the highest BLEU score with T5 in this strategy as the best model. The Kappa coefficient for inter-rater agreement regarding Incorrect Number error for both languages was up to 0.90 according to Table 5 (0.93 and 0.97, English and Portuguese respectively) which indicates a reasonable consensus between human evaluations.

In contrast to English, describing Portuguese numerical referring expressions with texts instead of digits in the input meaning representations to the models (e.g. Desc Source) presented the lowest error rate. The model with the lowest number of errors was GPT2-pt in the Desc Source strategy. Based on the automatic evaluation, the strategy where numbers are described with texts in the source MR and the outputted texts (e.g. Desc) depicted the highest BLEU score with mT5 in this strategy as the best model.

It is important to note that the Brazilian Portuguese approaches were evaluated in more domains than the English ones due to the differences in both datasets. Hence, in order to compare the numerical error rate of the models across languages, Table 5

depicts the numerical error rate of the approaches in the daily and month Amazon deforestation domains, which share the same meaning representations in English and Portuguese. In both languages, the No Desc was the best strategy. Across domains, error rates between month and daily deforestation were similar in English. However, for Portuguese, inputs in the form of daily reports introduced fewer numerical errors than monthly reports, likely due to the fact that the Portuguese models were trained with higher volumes of daily training sentences.

7 Discussion

This study evaluates how well state-of-the-art neural surface realizers generate numerical referring expressions. Three large language models (GPT-2, BART, and T5) in their English and multilingual/Portuguese versions and 4 strategies of handling numerical numbers, alternating the representation of numbers in meaning representations and verbalizations by digits and texts, were evaluated following the guidelines of Thomson and Reiter (2020) as well as the BLEU score, a standard automatic evaluation metric. Finally, we revisit the research questions outlined in Section 1:

(RQ1) How well do state of the art neural end-to-end surface realizers generate numerical referring expressions? A human evaluation was performed to annotate different error categories, such as numerical, named entities, context, word, uncheckable, other, and fluency errors. Results depicted across languages, models, and numerical strategies show the numerical error rate as the highest among the errors. Hence, with respect to this research question, there is clear evidence that state-of-the-art large language models struggle to generate adequate and faithful numerical referring expressions.

(RQ2) Are numerical referring expressions better verbalized when represented by digits or text? Results demonstrated that Brazilian Portuguese approaches which convert input meaning representations with numerical references in spell-out form into output text with numerical referring expressions presented as digits (Desc Source) perform better. On the other hand, for English, representing numerical references in spell-out form did not help regardless of whether it was present in the source meaning representa-

tion (Desc Source), in the target text (Desc Target) or both (Desc). As depicted in Table 1, the low results for English could be due to the scarcity of training resources in this language, and more work needs to be done to set the point.

Related Work Numeracy, the ability to comprehend, utilize, and attribute meaning to numerical information, is a crucial skill, while literacy focuses on the same aspects for written or spoken language (Spithourakis and Riedel, 2018). Language models trained on datasets containing numerical data have shown promise in capturing certain levels of numeracy (Wallace et al., 2019; Spithourakis and Riedel, 2018). However, significant advancements in this area, particularly in simple arithmetic, are yet to be achieved. Some researchers suggest that this could be attributed to factors such as data preprocessing techniques, tokenization methods, and word embeddings (Wallace et al., 2019; Muffo et al., 2022). Preprocessing techniques play a vital role in preparing numerical data for language models. In some cases, numbers in the corpus are modified or written in different forms as inputs to the models. During tokenization, where words are split into smaller units, all characters, including numbers, are treated as strings. Sub-word tokenizers, such as Byte-Pair Encoding (BPE) (Birnch et al., 2016), Word-Piece (Schuster and Nakajima, 2012), Sentence-Piece (an implementation of BPE and Unigram subword segmentation algorithms) (Kudo, 2018; Kudo and Richardson, 2018), and character-level representation (Dos Santos and Zadrozny, 2014), are commonly used in language models for this purpose.

Sub-word tokenizers operate by splitting words into smaller units to form the basis for training language models. They have the advantage of handling out-of-vocabulary words. WordPiece and BPE are widely used in LLM models such as BERT, GPT-2, T5, and BART. However, one limitation of WordPiece and BPE is that they can generate words that were not seen during training due to their tokenization technique. BPE’s deterministic symbol replacement approach lacks the ability to provide probabilistic segmentations, hindering subword regularization techniques that rely on segmentation probabilities. This can lead to errors in combining subword units during number and language generation (Kudo, 2018).

In other fields such as text synthesis, to circumvent the problem input texts are normalized by convert-

Expected output

The most affected state and municipality were respectively Pará (177.84 sq km) and Altamira, in the state of Pará (51.07 sq km).

Deforestation monthly intents

TOTAL_DEFORESTATION(area="177.84", location="deter-amz", month="4", state="PA", year="2021") [SEP]

TOTAL_DEFORESTATION(area="51.07", city="Altamira", location="deter-amz", month="4", state="PA", year="2021")

T5 nodesc

The most affected state and municipality were respectively Pará (**177.84 sq km**) and Altamira, in the state of Pará (**51.07 sq km**).

T5 descstrg

The state with the most deforestation in the month was Pará (**one hundred and seventy-seven point eight four sq km**), and the most devastated municipality was Altamira / Pará, where deforestation amounted to **fifty-one point zero seven sq km**.

T5 descsrc

The state with the most deforestation in the month was Pará (**177.84 sq km**), and the most devastated municipality was Altamira / Pará, where deforestation amounted to **51.07 sq km**.

T5 desc

The state with the most deforestation in the month was Pará (**one hundred and seventy-seven point eight four sq km**), and the most devastated municipality was Altamira / Pará, where deforestation amounted to **fifty-one point zero seven sq km**.

Table 6: Sample from T5 outputs for English considering all 4 strategies. T5 performed as the best model for English. The numeric referring expressions are **bolded**.

ing numerical referring expressions from digits to plain word form descriptions prior to being fed to a *text-to-speech* system (Nikulásdóttir and Guðnason, 2019), a method which mainly influenced this work.

Limitations. For future work, we plan to conduct experiments on the GPT3 and GPT4⁴ results of our ongoing analysis. However, since these models are neither free, nor reproducible due to limited or no information concerning model size, architecture, training parameters, and data set creation, we also plan to evaluate related open-source variations such as BLOOM⁵ and GPT-J⁶.

Ethics Statement. As highlighted in the human evaluation subsection 5.3, all annotators are members of the research group and were responsible for performing the evaluation with an equal amount of occurrences, hence ethical approval for conducting research with human subjects was not required. All data is publicly available (see Data subsection 2 for more information). No consent from data subjects was not required as this data is purely factual, containing no personal data, and hence compliant with the EU's General Data Protection Regulation (GDPR)⁷.

⁴<https://openai.com/blog/chatgpt>

⁵BLOOM: BigScience Large Open-science Open-access Multilingual Language Model - <https://huggingface.co/bigscience/bloom>

⁶https://huggingface.co/docs/transformers/model_doc/gptj

⁷<https://gdpr-info.eu/recitals/no-159/>

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. *Findings of the 2021 conference on machine translation (WMT21)*. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Rico Sennrich Birch, Barry Haddow, and Alexandra. 2016. *Das Hunderttage-Stadion : Entstehungsgeschichte des Bad Nauheimer Kunsteisstadions unter Colonel Paul R. Knight*. *Acl*, pages 1715–1725.

João Campos, André Teixeira, Thiago Ferreira, Fábio Cozman, and Adriana Pagano. 2020. *Towards fully automated news reporting in brazilian portuguese*. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 543–554. SBC.

Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Cicero Nogueira Dos Santos and Bianca Zadrozny. 2014. Learning character-level representations for part-of-speech tagging. *31st International Conference on Machine Learning, ICML 2014*, 5(2011):3830–3838.

Ondřej Dušek and Filip Jurčiček. 2016. *Sequence-to-sequence generation for spoken dialogue via deep syn-*

- tax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51, Berlin, Germany. Association for Computational Linguistics.
- Sebastian Gehrmann, Falcon Dai, Henry Elder, and Alexander Rush. 2018. [End-to-end content and plan selection for data-to-text generation](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 46–56, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Jeonghwan Kim, Junmo Kang, Kyung-min Kim, Giwon Hong, and Sung-Hyon Myaeng. 2022. Exploiting numerical-contextual knowledge to improve numerical reasoning in question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1811–1821.
- Taku Kudo. 2018. [Subword regularization: Improving neural network translation models with multiple subword candidates](#). *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:66–75.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). *EMNLP 2018 - Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Proceedings*, pages 66–71.
- Rémi Lebre, David Grangier, and Michael Auli. 2016. [Neural text generation from structured data with application to the biography domain](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, Austin, Texas. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Matteo Muffo, Aldo Cocco, and Enrico Bertino. 2022. [Evaluating Transformer Language Models on Arithmetic Operations Using Number Decomposition](#). 2022 *Language Resources and Evaluation Conference, LREC 2022*, pages 291–297.
- Anna Björk Nikulásdóttir and Jón Guðnason. 2019. [Bootstrapping a Text Normalization System for an Inflected Language. Numbers as a Test Case](#). In *Proc. Interspeech 2019*, pages 4455–4459.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Journal of Machine Learning Research*, 21(140):1–67.
- Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, U.K.
- André Luiz Rosa Teixeira, João Campos, Rossana Cunha, Thiago Castro Ferreira, Adriana Pagano, and Fabio Cozman. 2020. [DaMata: A robot-journalist covering the Brazilian Amazon deforestation](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 103–106, Dublin, Ireland. Association for Computational Linguistics.
- Mike Schuster and Kaisuke Nakajima. 2012. [Japanese and korean voice search](#). In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.
- Georgios P. Spithourakis and Sebastian Riedel. 2018. [Numeracy for language models: Evaluating and improving their ability to predict numbers](#). *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 1:2104–2115.
- Richard Sproat. 2022. Boring problems are sometimes the most interesting. *Computational Linguistics*, 48(2):483–490.
- Craig Thomson and Ehud Reiter. 2020. [A gold standard methodology for evaluating accuracy in data-to-text systems](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, Dublin, Ireland. Association for Computational Linguistics.
- Emiel van Miltenburg, Miruna Clinciu, Ondřej Dušek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. 2021. [Underreporting of errors in NLG output, and what to do about it](#). In *Proceedings of the 14th International Conference on Natural*

- 646 *Language Generation*, pages 140–153, Aberdeen, Scot-
647 land, UK. Association for Computational Linguistics.
- 648 Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh,
649 and Matt Gardner. 2019. [Do NLP models know num-
650 bers? Probing numeracy in embeddings](#). *EMNLP-
651 IJCNLP 2019 - 2019 Conference on Empirical Methods
652 in Natural Language Processing and 9th International
653 Joint Conference on Natural Language Processing, Pro-
654 ceedings of the Conference*, pages 5307–5315.
- 655 Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-
656 Hao Su, David Vandyke, and Steve Young. 2015. [Se-
657 mantically conditioned LSTM-based natural language
658 generation for spoken dialogue systems](#). In *Proceedings
659 of the 2015 Conference on Empirical Methods in Nat-
660 ural Language Processing*, pages 1711–1721, Lisbon,
661 Portugal. Association for Computational Linguistics.
- 662 Hao Zhang, Richard Sproat, Axel H Ng, Felix Stahlberg,
663 Xiaochang Peng, Kyle Gorman, and Brian Roark. 2019.
664 Neural models of text normalization for speech applica-
665 tions. *Computational Linguistics*, 45(2):293–337.

A Supplementary Material

A.1 Annotation Guidelines

After the most common error cases were identified and the treatment for the most difficult cases was agreed upon, annotators followed common guidelines for the rest of the evaluation process, as described in the following list:

- **Incorrect Number:** Has incorrect numerical values (e.g.: model verbalizes an area value of "354" as "345"); Numerical values not verbalized in numerical form in the final texts were considered incorrect (e.g.: "three hundred fifty-four" instead of "354");
 - **Incorrect Named Entity:** verbalizes entities incorrectly or verbalizes entities that do not exist;
 - **Incorrect Word:** occurrence of spelling errors;
 - **Context Error:** verbalizes some communicative intent incorrectly (e.g: verbalizes last month's deforestation variation instead of total area deforestation);
 - **Not checkable:** adds information that is not present in the input semantic representation in the verbalized text;
 - **Other:** other types of verbalization errors;
 - **Fluency:** the hypothesis verbalizes a not fluent text.
- Entries were distributed in a collaborative spreadsheet.
 - Each row consisted of the original Meaning Representation (MR), the generated hypothesis, and the rating categories.
 - LLMs used to generate the entries were omitted in the spreadsheet.
 - The spreadsheet was formatted to highlight the options (y - red; n - green) aiming to aid/ease the process with visual cues.
 - Difficult cases were commented on to be further discussed within the group of raters, fostering improvements in the guidelines.
 - Once, the annotation was finished, the spreadsheets were exported in .csv files for result computation.

Language	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error	Not Checkable	Other	Fluency Problem
Input	area="322.91"	city="Novo Progresso, Itaituba"	-	-	-	-	-
English	The National Institute for Space Research (INPE) estimated that deforestation of the Legal Amazon amounted to 2,322.91 sq km in April two thousand and twenty, which is a one hundred and twenty-six percent increase from the previous month.	"The National Institute for Space Research (INPE) reported that deforestation amounted to twenty-one point seven five sq km in the state of Pará , in the state of Pará , in February two thousand and twenty."	"The main class of deforestation was clear-cut deforestation, which removes all vegetetation of the soil, responsible for 317.93 sq km of deforested area."	"The most affected state and municipality were respectively Pará (177.84 sq km) and Altamira / Pará, in the state of Pará."	"The main cause of deforestation was the destruction of the soil, which leaves the soil clear of vegetation."	"The National Institute for Space Research (INPE) in Pará, where the most affected municipality was Novo Pro "	"The National Institute for Space Research (INPE) reported that deforestation amounted to 21.75 sq km in the state of Pará, in the state "
Input	cases="4091801" deaths="125584"	uc="PARQUE NACIONAL DO JAMANXIM"	-	-	area="0.32" day="22" month="8"	-	-
Portuguese	São registrados, no total, 135.584 mortes e 4.093.801 casos de #COVID19 no Brasil.	O INPE gerou alerta para devastação (0,19 km²) causada pelo desmatamento com solo exposto, que remove totalmente a vegetação da floresta, no dia 10 de agosto de 2020 na PARQUE NACIO	A cidade mais atingida foi SANTAQUITÉRIA , em CEARÁ , que registrou 22 focos de incêndio.	O Instituto Nacional de Pesquisas Espaciais(INPE) registrou um total de quinhentos e sessenta e nove focos de queimadas no território brasileiro, no dia onze de outubro de dois mil e vinte, o território brasileiro foi atingido.	O INPE gerou alerta para devastação (0,22 km2) causada pelo desmatamento com solo exposto, que remove totalmente a vegetação da floresta, no dia 22 de agosto de 2020 na RESERVA EXTRATIVISTA VERDE PARA SEMPRE / Pará - no mês já são 2 dias com alertas e 0,32 km2 desmatar.	A A A A A A A A BIOLÓGICA NASCENTES DA SERRA DO CACHIMBO somou dois vírgula sete três km2 de área desmatada no mês de novembro de dois mil e vinte.	Com um total de mil quinhentos e setenta e oito vírgula oito sete km2, o desmatamento com solo exposto , deixando a terra sem vegetação, a principal causa de destruição da Amazônia Legal no mês foi o desmatamento com solo exposto, deixando a terra sem vegetação.

Table 7: Examples of categories of error in human evaluation for English and Brazilian Portuguese.

APPENDIX B – Protocolo de Simplificação (Simplification Protocol)

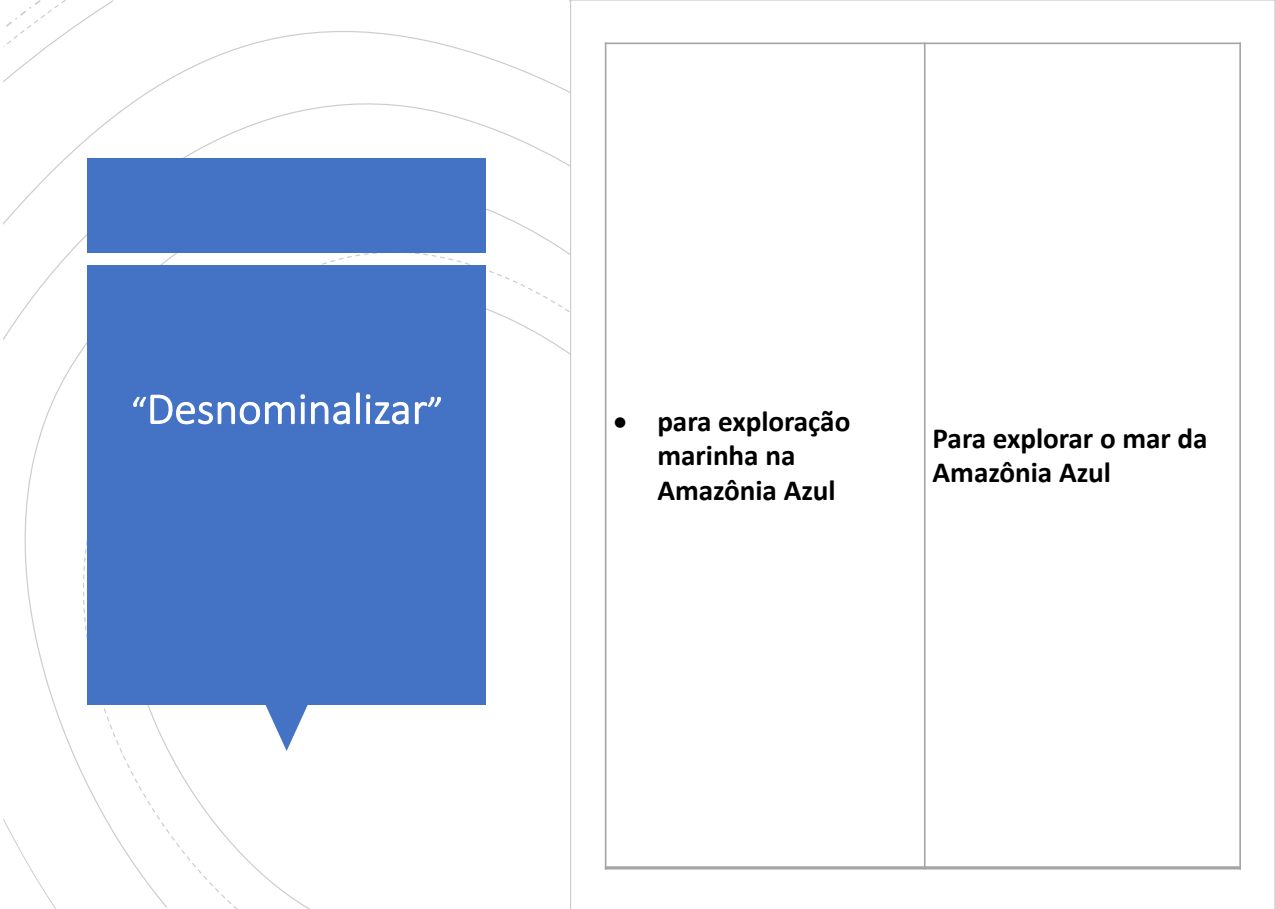
Identificar orações
finitas, não finitas e
nominalizações
dentro de uma
sentença

Areia e cascalho são os recursos que apresentam maior potencial para exploração marinha na Amazônia Azul, excedendo, em volume, o valor de qualquer outro recurso não vivo, excluindo óleo e gás, sem contar metais nobres como ilmenita, monazita, zirconita e rutilo, que ocorrem praticamente em toda a faixa litorânea da Amazônia Azul.

- para exploração marinha na Amazônia Azul
- Areia e cascalho são recursos
- [Areia e cascalho] apresentam maior potencial
- [Areia e cascalho] excedem, em volume, o valor de qualquer outro recurso não vivo,
- óleo e gás estão excluídos,
- metais nobres como ilmenita, monazita, zirconita e rutilo são recursos
- metais nobres como ilmenita, monazita, zirconita e rutilo ocorrem praticamente em toda a faixa litorânea da Amazônia Azul.

Identificar polaridade negativa e verificar se pode ser tornada positiva

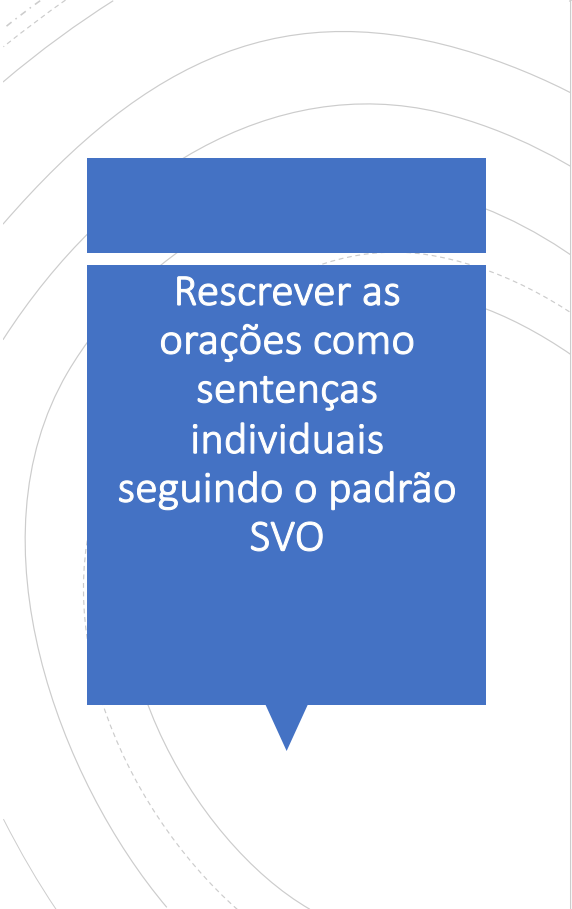
- | | |
|--|---|
| <ul style="list-style-type: none">• excluindo óleo e gás• sem contar metais nobres como ilmenita, monazita, zirconita e rutilo, | <ul style="list-style-type: none">• Os recursos não vivos de maior potencial são óleo e gás• Há metais nobres como ilmenita, monazita, zirconita e rutilo, |
|--|---|



“Desnominalizar”

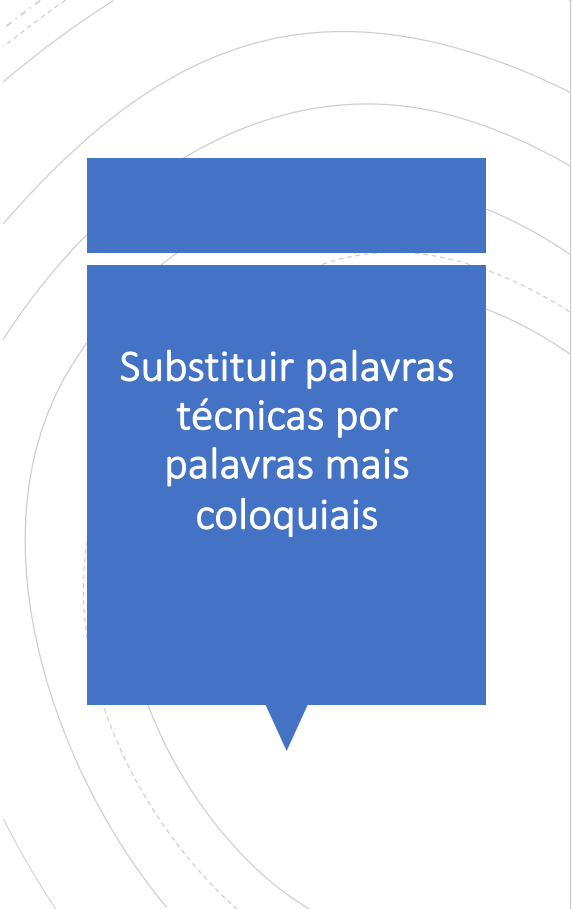
- para exploração marinha na Amazônia Azul

Para explorar o mar da Amazônia Azul



Rescrever as
orações como
sentenças
individuais
seguindo o padrão
SVO

Os recursos não vivos de maior potencial são óleo e gás. Depois deles, seguem areia e cascalho; há ainda metais nobres como ilmenita, monazita, zirconita e rutilo. Esses metais estão praticamente em toda a faixa litorânea.



Substituir palavras
técnicas por
palavras mais
coloquiais

Os recursos não vivos de maior potencial são óleo e gás. Depois deles, seguem areia e cascalho; há ainda metais nobres como ilmenita, monazita, zirconita e rutilo. Esses metais estão praticamente em toda a faixa litorânea.

Os recursos minerais mais importantes são óleo e gás. Depois deles, seguem areia e cascalho e metais nobres como ilmenita, monazita, zirconita e rutilo. Esses metais estão em quase todo o litoral.

Perguntas norteadoras

- Quantas sentenças há no texto?
- Quantas orações há em cada sentença?
- Quais são os Sujeitos e os Verbos de cada uma das orações?
- Há nominalizações? Quais os agentes do verbo nominalizado?
- Que palavras podem ser substituídas por sinônimos/hiperônimos (palavras mais coloquiais)?
- Qual o número total de palavras da versão simplificada?