UNIVERSIDADE FEDERAL DE MINAS GERAIS / MACQUARIE UNIVERSITY

Faculdade de Letras / Department of Linguistics

Programa de Pós-graduação em Estudos Linguísticos

Rodrigo Araújo e Castro

Systemic-Functional modeling of text complexity in Brazilian Portuguese

Belo Horizonte

2021

Rodrigo Araújo e Castro

Systemic-Functional modeling of text complexity

in Brazilian Portuguese

Versão final

Tese apresentada ao Programa de Pós-graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Doutor em Linguística Aplicada.

Área de Concentração: Linguística Aplicada

Linha de pesquisa: Estudos da Tradução – 3B

Orientador(a): Dr. Adriana Pagano (UFMG)

Coorientador(a): Dr. Ilka Afonso Reis (UFMG)

Coorientador(a): Dr. David Butt (Macquarie University)

Coorientador(a): Annabelle Lukin (Macquarie University)

Belo Horizonte

2021

FICHA CATALOGRÁFICA

Castro, Rodrigo Araújo e.

C355s Systemic-Functional modeling of text complexity in Brazilian Portuguese [manuscrito] / Rodrigo Araújo e Castro. – 2021.
106 f., enc.: il, tab., color.

Orientadora: Adriana Pagano.

Coorientadora: Ilka Afonso Reis.

Coorientador: David But.

Coorientadora: Annabelle Lukin.

Área de concentração: Linguística Aplicada.

Linha de Pesquisa: Estudos da Tradução.

Tese (doutorado) – Universidade Federal de Minas Gerais,

Faculdade de Letras.

Tese apresentada em cotutela com a Macquarie University.

Bibliografia: f. 80-86.
Apêndices: f. 88-90.

1. Tradução e interpretação – Teses. 2. Linguística aplicada – Teses. 3. Linguística – Processamento de dados – Teses. 4. Funcionalismo (Linguística) – Teses. 5. Linguística de corpus – Teses. I. Pagano, Adriana Silvina. II. Reis, Ilka Afonso. III. But, David. IV. Lukin, Annabelle. V. Universidade Federal de Minas Gerais. Faculdade de Letras. VI. Macquarie University. VI. Título.

CDD: 418.02

Ficha catalográfica elaborada bibliotecária Priscila Oliveira da Mata - CRB6-2706

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

FOLHA DE APROVAÇÃO

Systemic-Functional modeling of text complexity in Brazilian Portuguese

RODRIGO ARAUJO E CASTRO

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 11 de novembro de 2021, pela banca constituída pelos membros:

Prof(a). Adriana Silvina Pagano - Orientadora
UFMG

Prof(a). Ilka Afonso Reis - Coorientadora
UFMG

Prof(a). Giacomo Patrocinio Figueredo
UFOP

Prof(a). Thiago Castro Ferreira
UFMG

Prof(a). Igor Antônio Lourenço da Silva
UFU

Prof(a). Kícila Ferreguetti de Oliveira
UFMG

Documento assinado eletronicamente por **Ilka Afonso Reis, Professora do Magistério Superior**, em 12/11/2021, às 14:48, 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 12/11/2021, às 15:33, 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 **Kicila Ferreguetti de Oliveira, Professora Magistério Superior-Substituta**, em 12/11/2021, às 16:25, 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 **Giacomo Patrocinio Figueredo, Usuário Externo**, em 16/11/2021, às 22:16, 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 17/11/2021, às 08:57, 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 18/11/2021, às 17:46, 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 **1032826** e o código CRC **F0866CB9**.

Acknowledgments

I appreciate all the assistance provided by many people who assisted me in this thesis, some of them that will be named afterward.

Firstly, I would like to say thanks to my wife Izabella Rosa Malta, my beloved rose, who has assisted me greatly in many ways over this long journey, both emotionally and academically, that is coming to an end.

I also would like to say thanks to Macquarie and UFMG staff, especially Annabelle and Fabio, for all the support and to my supervisors, Adriana Pagano, David Butt, and Ilka Reis, and my former supervisor Canzhong Wu, who have helped as much as they could for me to learn a lot during this time and reach my final destination in this academic journey. It if were not for Canzhong's assistance, certainly my methodology would be as rich and so computationally driven as it is.

Finally, I would like to say thanks to all my friends who assisted in my thesis one way or another, such as Raquel Rossini, Adriana Alves, and, especially, Arthur de Melo Sá, who assisted me with revisions and translations.

Complexity is itself a complex notion, and complexity in language is particularly so.

-- M.A.K. Halliday, *Complementarities in Language*, 2008

Resumo

O estudo da complexidade textual é um passo fundamental para a modelagem de tarefas de simplificação textual, uma vez que simplificação se configura como uma redução na complexidade do texto. Nas últimas duas décadas, estudos em Processamento de Língua Natural (PLN) têm procurado identificar estratégias eficientes de simplificação. Embora algumas tentativas de abordar esta questão com a construção de modelos computacionais baseados em teorias da linguagem tenham fornecido insights potencialmente valiosos, estes ainda são insuficientes para lidar efetivamente com a tarefa. Com o objetivo de preencher esta lacuna e com base em uma teoria abrangente da linguagem -- a Linguística Funcional Sistêmica (LSF) (Halliday & Matthiessen, 2014) --, esta tese explora a complexidade da linguagem com o objetivo de obter evidências que possam informar as tarefas de simplificação textual visando a produção de textos mais acessíveis em português brasileiro. Para tanto, foi compilado SIM-Pt (Simplificado Português Brasileiro), um corpus paralelo monolingüe de segmentos textuais alinhados nos domínios da física, biologia e psicologia. Os segmentos foram organizados em dois conjuntos de dados associados: (1) dois conjuntos de segmentos extraídos de textos científicos encontrados na Web, compostos, respectivamente, de segmentos mais simples e mais complexos; e (2) dois conjuntos de segmentos criados manualmente com base nos segmentos extraídos de textos, mantendo-se níveis distintos de complexidade. Cada conjunto contém aproximadamente 200 segmentos de texto. As orações em cada segmento foram analisadas manualmente de acordo com seus significados Ideacionais, Interpessoais e Textuais, e padrões na lexicogramática foram obtidos com base em frequências sistêmicas e estruturais que pudessem fornecer variáveis estreitamente relacionadas a diferentes níveis de metaforicidade gramatical. Por meio do mapeamento da complexidade textual nos estratos da lexicogramática, semântica e

contexto, foi proposta uma relação entre complexidade textual e metáfora gramatical experiencial. Os resultados mostram que, do ponto de vista experiencial, em média maior grau de metáfora gramatical experiencial está correlacionado com maior complexidade textual. As principais evidências que sustentam esta afirmação sob a perspectiva da lexicogramática foram a frequência mais elevada de orações relacionais e existenciais, juntamente com orações na voz média e orações incrustadas, e a frequência mais elevada de mudanças de classe de palavra (especialmente nominalizações) e mudanças na escala de ordens (Ravelli, 1999). Os resultados desta tese contribuem para os estudos da simplificação textual no português brasileiro, tanto no campo da linguística aplicada como no campo da PNL.

Palavras-chave: Linguística Sistêmico-Funcional; Linguística Aplicada; Simplificação textual; Complexidade textual; Metáfora gramatical; Textos de ciência; Português brasileiro

Abstract

Investigating text complexity is a significant step towards modeling text simplification tasks, as text simplification is the reduction of the complexity of a text. In the last two decades, studies in Natural Language Processing (NLP) have attempted to discover efficient simplification strategies. Although some attempts to address this issue with the construction of computer models based on language theories have provided potentially valuable insights, they remain insufficient to effectively deal with the task. Purporting to fill this gap and drawing on a comprehensive theory of language -- Systemic Functional Linguistics (SFL) (Halliday & Matthiessen, 2014) --, this thesis explores text complexity with a view to gathering findings that may inform text simplification tasks aimed to produce more accessible texts in Brazilian Portuguese. To that end, SIM-Pt (Simplified Brazilian Portuguese), a monolingual parallel corpus of aligned text segments in the physics, biology, and psychology domains, was compiled. Text segments were organized into two paired datasets: (1) two sets of naturally occurring segments, made up of, respectively, simpler and more complex segments extracted from science texts found on the Web; and (2) two sets of manually constructed segments based on the naturally occurring segments, ensuring distinct complexity levels. Each set contains approximately 200 text segments. Clauses in segments were manually analyzed in terms of Ideational, Interpersonal, and Textual meanings, and lexicogrammatical patterns were obtained on the basis of systemic and structural frequencies that could yield variables closely related to different levels of grammatical metaphor. By examining text complexity within the strata of Lexicogrammar, Semantics, and Context, we proposed a relationship between text complexity

and experiential grammatical metaphor. The results show that, from the experiential viewpoint, a higher degree of experiential grammatical metaphor on average correlates with higher text complexity. The main pieces of evidence supporting this claim from the perspective of lexicogrammar were the higher frequency of relational and existential clauses in combination with middle voice and embedded clauses and the higher frequency of class shifts (especially nominalizations) and rank shifts (Ravelli, 1999). The findings of this thesis are expected to contribute to text simplification accounts for Brazilian Portuguese in both applied linguistics and NLP.

Keywords: Systemic Functional Linguistics; Applied Linguistics; Text simplification; Text complexity; Grammatical metaphor; Science texts; Brazilian Portuguese

List of Figures

List of Tables

List of Abbreviations for computational studies

NLP - Natural Language Processing

NILC - Interinstitutional Center for Computational Linguistics

SRT - Semantic Representation of Text

SRL - Semantic Role Labeler

Summary

Chapter 1 Introduction

1.1 Overview

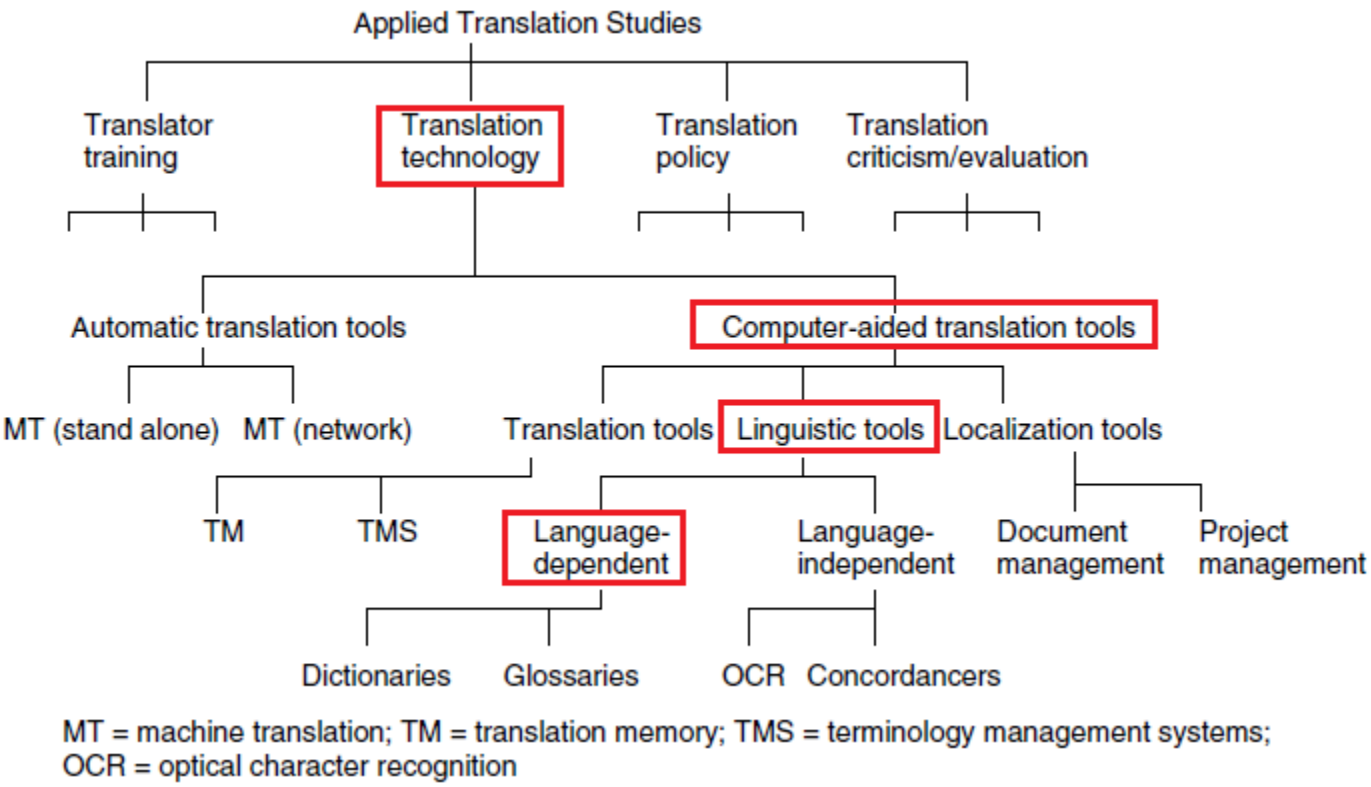
A sound theory of what makes a text complex is an essential step when investigating text simplification. This thesis supports Systemic Functional Linguistics, which is a broad enough linguistic perspective to provide a linguistic background for computational implementations aiming at making texts simpler or more complex. That is to say, the grammatical patterns can be used to train systems to perform text simplification tasks and to develop guidelines for simplifying science texts in Brazilian Portuguese. As a consequence, these representations may be adapted to the study of text simplification in other languages.

Taken into consideration that a "[c]ommon ground between Translation Studies and translation technology – and machine translation in particular – may be found within

functional approaches to translation." (Quah, 2006, p. 25). One of such theories is Systemic Functional Linguistics (SFL), on which this thesis draws, as shown in Figure 1, describing the extension of Holmes' map of Translation Studies proposed by Quah (2006).

Figure 1

Extension of Holmes' map of Translation Studies



Note: Adapted from Quah (2006, p. 42)

In Figure 1, the extension of Holmes' map of Translation Studies adds up a more extensive description of Applied Linguistics, which, according to Quah (2006), is not broadly described by Holmes (1988/2000: 182), which includes this as "Translation aids".

Drawing on SFL, to investigate text simplification, this thesis assumes that a segment is simpler in comparison to another one, if, upon performing operations to render the text comprehensible by a non-specialist audience, the segments present distinct configurations from the perspective of linguistics.

Simplification of text has received special interest, in particular amongst those working in Natural Language Processing (NLP)[1]. This line of inquiry into strategies of simplification is explored, for instance, in Shardlow (2014) and Paetzold & Specia (2017), and is directly related to the quest to provide more easily comprehended texts, particularly in the domains of science, of foreign language learning (Burstein, 2009; Petersen & Ostendorf, 2007), of literacy education (for students and people with low literacy levels: Candido et al., 2009), and for students with learning disabilities (Carroll et al., 1999; Max, 2005; Fajardo et al, 2013). Simplification is also significant for readers aiming to work across disciplines, especially those who wish to have access to science or other information outside their professional domain (Gonzalez-Dios, 2016).

The main challenge faced by researchers in NLP and other computational studies is the need for a reliable and appliable "linguistic profile with features indicating [text] complexity or simplicity". To be appliable, the account of complexity/simplicity needs to be sensitive to "the needs of particular audiences" (Pasqualini, 2018, p. 50). The users' needs vary by content required, by the registerial experience of the readers, and by the final purpose to which the simplified text is directed. Such a correlation between theory and practice is a challenge. This challenge grows for so-called 'low-resource' languages, as is the present case of Brazilian Portuguese.

Even though renowned research centers, such as the Interinstitutional Center for Computational Linguistics (NILC), at the University of São Paulo, have produced several tools, most of them available online, they are still insufficient to address the subtle issues of text simplification: the problem of clear linguistic guidelines to produce simpler texts. To supply the need for linguistic guidelines for Brazilian Portuguese, some NLP studies (e.g., Aluísio et al.

[1] NLP is defined as "range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis" (Liddy, 2001, p.1).

2008) have developed lists of text simplification strategies to be applied. These strategies are procedures designed by computer scientists to gauge text complexity (Damay et al, 2006, p. 34). Manipulating text according to these linguistic features - those indicating higher or lower text complexity - is expected to produce more readable and more comprehensible texts (Aluísio & Gasperin, 2010; Candido et al, 2009).

This thesis seeks to bridge computational and linguistic approaches to text simplification by investigating which linguistic features characterize varying levels of grammatical metaphoricity and text complexity from an SFL perspective on a corpus of science texts in Brazilian Portuguese. Also, this thesis addresses the need for studies on text complexity that bring more over to NLP from a language studies perspective. As a starting point, the perspective of those working in NLP can be described as: "[the] difficulty in comprehending written texts" is the inverse of text simplification. Text simplification is then regarded as "the process of transforming complex sentences into a set of equivalent simpler sentences with the goal of making the resulting text easier to read by some target group" (Damay et al, 2006, p. 34). Furthermore, NLP studies generally characterize text simplification as lexical and syntactic simplification. In response to this NLP perspective on the relationship between complexity and simplification, this thesis argues for the direct relationship between complexity and simplification, based on the notion of grammatical metaphor (Ravelli, 1999; Halliday & Matthiessen, 2014) and Steiner's (2005) hypothesis for associating explicitation and grammatical metaphoricity. But the bridge with linguistics and language-based views extends and deepens the factors behind complexity, and thereby, what counts as simplification.

Lexical simplification is defined in NLP circles as "the process of replacing complex words in a given sentence with simpler alternatives of equivalent meaning" (Paetzold & Specia,

2017, p. 549). Syntactic simplification is defined as “the process of reducing the grammatical complexity of a text while retaining its information content and meaning” (Siddharthan, 2003, p.3).

NLP’s accounts of text simplification (Damay et al., 2006) can be read in the light of the work of SFL linguists like Halliday & Martin (1993, p. 67). Halliday & Martin’s contention is that a language segment is simpler in comparison to another if, in order to be comprehensible by a non-specialist audience, the segments differ in their linguistic criteria configurations. In other words, a text produced by a specialist in a certain domain may not be comprehensible to another person without such expertise; therefore, the use of different configurations compared with the originally produced text may be necessary to increase text readability for a broader audience. The issues here involve the expertise of the proponents, their shared register, and what counts as a configuration.

Text complexity is a relevant topic for both Applied Linguistics and Computer Science. Regarding the former, researchers have sought to evaluate the impact of language features such as nominalization on translation tasks (Palumbo, 2008) and in “expanding students’ registerial repertoire” to improve their writing performance at university (Lee et al., 2019, p. 1). As for the latter, studies have shown that given enough software training (with specific training sets, namely corpora of annotated texts), text simplification can be performed automatically (Aluísio et al., 2008, among others) through simplification systems to improve text produced through machine translation (Aluísio et al., 2008; Shardlow, 2014; Štajner & Popović, 2016). Text simplification may be successful in assisting in the training of machine translation systems, allowing them to provide more diverse and more effective solutions to translation problems. This leads to more comprehensible texts. Some solutions, for example, may be the use of synonyms or

structures similar to those added in the translated language, allowing the production of simpler texts. These systems are potential resources to improve the population's literacy level (Gasperin et al., 2018). Text comprehension can come through lexical and grammatical strategies, thereby facilitating the reading process. These steps, focusing formally on lexis and grammar, can be extended by the registerial, textual, and lexicogrammatical approaches of functional linguists.

A particular approach in language studies posited by Michael Halliday, his "appliable linguistics" (Halliday, 1985; Matthiessen, 2012, p. 436), provides the chief motivation of this thesis. Appliable Linguistics is "a kind of linguistics where theory is designed to have the potential to be applied to solve problems that arise in communities around the world, involving both reflection and action" (Matthiessen, 2012, p. 436). As a "socially accountable linguistics", concerned with explaining functional variation in language with a "critical stance" (Matthiessen, 2012), Halliday's functional theory has the potential for adaptation in "different contexts and for diverse purposes" (Halliday, 1985, p. 3). As Applied linguistics assumes that "a theory is made manifest, or 'realized', in the processes of being applied", a linguistic theory needs to solve practical problems to reach its full potential (Knight & Mahbook, 2010, p. 5). For this reason, such theory should make a relevant difference to social situations, and cultural challenges. Halliday's Systemic Functional Theory (SFT) is regarded in this thesis as an appliable theory, with the potential to offer a basis for practical purposes, especially for those pursued in creating greater responsiveness for the application of computational linguistics.

Previous studies drawing on Systemic Functional Linguistics (SFL), such as Ravelli (1999) and Steiner (2005), showed an association between text complexity and a concept proposed by Halliday through his insights into the 'evolution' of patterns of exposition in the

history of sciences (from classical times in Greek to the present). The concept is grammatical metaphor. It is a semantic configuration, or reconfiguration, of versions of the world, configurations forced upon the writer by various semantic pressures. For instance, these pressures are to capture abstractions as ‘things’; to express things in terms of equivalence to other ideas, abstractions, definitions, and/or measurements; to give qualities and relations (of cause; time; or space) the central rather than the marginal role of the clause. Halliday called the products of these pressures ideational grammatical metaphors, or experiential grammatical metaphors. This meant that by dislocations of grammatical roles, our ordinary experience and our earliest versions of the world about us and ‘in us’, becomes strangely shifted and abstracted in texts. This can happen with even what seem to be simple words: “Stop (the) fighting” becomes more metaphorical in a politician’s seemingly simple words about Vietnam: “We need to make an end of this war”. In this case, the meaning of the clause “stop the fighting” is realized in the other clause as the nominal group “an end of this war”, distilling the meaning (Halliday & Matthiessen, 1993). Science and other specializations turn language into more and more specific forms. According to Halliday, this ‘turn’ in the grammar is like the swerve introduced in traditional notions of metaphor and lexical substitution. In the case of lexical metaphors, instead of congruent forms, non-congruent forms take part of clauses to ensure that the clauses are considered “texts”, according to SFL. The same applies to grammatical metaphors.

Ravelli (1999) argued that, through agnation, class shifts, and rank shifts, one may produce texts with varying degrees of grammatical metaphoricity, which may potentially lead to distinct levels of text complexity. Furthermore, Steiner (2005) associated the degree of grammatical metaphor with the level of explicitness in a text, concluding that agnation between congruent and metaphorical wordings can bring about lower or higher text complexity.

Rather than drawing on concepts posited by linguistic theories, such as that of grammatical metaphor, as aforementioned, research on NLP has tended to investigate text simplification by focusing mostly on lexical simplification (Paetzold & Specia, 2017), syntactic simplification (Seretan, 2012; Štajner et al, 2013; Siddharthan, 2003; Shardlow, 2014), paraphrasing (Damay et al., 2006), and text entailment (Dagan et al., 2013). Nevertheless, computational linguistics studies on text simplification for Brazilian Portuguese, such as Aluísio et al. (2008, 2010), Candido et al. (2009), Caseli et al. (2009), and Specia (2010), have suggested that a combination of linguistic and computational perspectives may allow the development of systems able to produce simplified sentences in Brazilian Portuguese. Contrary to expectations, a thorough review of the literature on computational and linguistic studies revealed a lack of studies on text complexity and text simplification, at least a lack of those drawing on input from linguistic theories.

In terms of NLP studies of Brazilian Portuguese, there is no consensus on which strategies assist to manipulate metaphoricity, and what may lead to varying degrees of text complexity. Studies by researchers at NILC (Interinstitutional Center for Computational Linguistics) suggest that if text complexity reduces, one can observe text simplification. In addition, only a few NLP studies, mainly conducted by NILC researchers, tackle text complexity in Brazilian Portuguese. A number of studies in NLP have postulated a convergence between lexical and syntactic manipulations in text complexity leading to text simplification. Different strategies have been found to be related to varying levels of text complexity, such as those reported by Siddharthan (2004, 2006), Hwang et al. (2015), and Štajner (2015) (see section 2.2.1). Thus, the apparently large number of different strategies used to manipulate text complexity and lead to text simplification is likely to be a result of the varied labels used to

describe those strategies. Some of the main text simplification strategies are actually the same or highly similar. However, they receive different names. For instance, dis-embedding of relative clauses and separation of subordinate clauses (Siddharthan, 2004, 2006), and sentence splitting (Štajner, 2015) are labeled differently but may be considered highly similar strategies. In this thesis, each strategy was interpreted in the light of the overall model of Systemic Functional Linguistics. This allowed the grammatical and lexical realizations to be related to the semantic and then cultural activities, which shaped the needs of the speakers and writers of the particular register of specialization.

In this light, the central task of this research is to investigate linguistic features that characterize text complexity from an SFL perspective on a corpus of science texts in Brazilian Portuguese. Text complexity is explored drawing on the concept of grammatical metaphor and the examination of degrees of metaphoricity. Language patterns are sought to characterize more metaphorical and less metaphorical wordings, assuming there is a connection between higher text complexity and higher metaphoricity levels. The aim is to identify which linguistic patterns are associated with varying levels of grammatical metaphoricity and text complexity in Brazilian Portuguese.[2]

A corpus having instances of different metaphoricity levels is a fundamental resource with which to explore text complexity. In NLP, corpora are used for machine learning; this is the case of the Newsela corpus for the English language (Xu et al., 2015). Inspired upon Newsela and with the specific aim of exploring text complexity in Brazilian Portuguese, the SIM-Pt corpus was compiled for the purposes of this thesis. Besides informing our study, SIM-Pt can be used for training systems of text complexity classification.

[2] As this thesis aims at indicating linguistic trends and patterns, statistical tests will not be used to validate the numbers. Further works, though, may need these tests to validate their conclusions.

SIM-Pt was designed as a monolingual parallel corpus of aligned science text segments (from physics, biology, and psychology texts), as ideational grammatical metaphor associates with education, science, bureaucracy, and law (Halliday, 1993; Halliday & Martin, 1993). For this reason, ideational grammatical metaphors can be realized differently, and, considering the nature of the texts from the corpus, these text types are the focus of this thesis.The corpus is organized in four datasets, two sets of simpler and two sets of more complex segments, with around 200 text segments each. These were manually analyzed clauses regarding IDEATIONAL, INTERPERSONAL, and TEXTUAL meanings and mapped lexicogrammatical patterns to investigate the degree of grammatical metaphoricity.

The systemic patterns were drawn based on descriptive statistics of the systemic choices, e.g., PROCESS, THEME, and CIRCUMSTANCE types. The structural patterns were semantic item combinations (PARTICIPANT, PROCESS, and CIRCUMSTANCE) and class and rank shifts indicating experiential grammatical metaphor, mainly, nominalizations (Halliday & Martin, 1993; Ravelli, 1999) and embedded clauses.

Drawing on the assumption that a survey of text complexity features is key to text simplification procedures, this thesis explores grammatical metaphor as an approach to examine text complexity which may provide valuable insights into text simplification. Manipulating text with distinct degrees in metaphoricity is posited as an instrument for judging and thereby manipulating text complexity level. Motivated by shortcomings in NLP studies of text simplification, which are not cognizant of metaphoricity as a factor in text complexity, this thesis seeks to probe experiential grammatical metaphor as an indicator of text complexity and strategy for future implementations aimed at simplifying science texts in Brazilian Portuguese.

Finally, in order to draw conclusions, these are the research questions this thesis seeks to address in detail in the Discussion:

I. How can Systemic Functional Linguistics map the association between textual complexity and experiential grammatical metaphor in science texts in Brazilian Portuguese?

II. According to Systemic Functional Linguistics (SFL), what evidence of varying metaphoricity degrees in terms of structure and system can be found in science texts in Brazilian Portuguese?

III. Which linguistic patterns discriminate between congruent and non-congruent clauses; in other words, different degrees of experiential grammatical metaphor in Brazilian Portuguese?

IV. Which linguistic patterns indicating experiential grammatical metaphor lead to varying degrees in text complexity in Brazilian Portuguese?

1.2 Thesis structure

This thesis is organized into 6 chapters, including this Introduction. Chapter 2 reviews the literature on text complexity, from the linguistic and computational perspectives, and introduces main concepts in SFL that base our assumption that manipulation of metaphoricity levels may increase or decrease text complexity and lead to a simpler or more complex text. Chapter 3 gives an account of the procedures undertaken for the compilation of the Sim-Pt, a text simplification corpus in Brazilian Portuguese, the steps to import and annotate the corpus, and to analyze the data using the software Sysfan to obtain frequencies of systemic and structural patterns. The systemic patterns refer to the systemic choices in the systems and the structural patterns refer to the distinct configurations of elements. Chapter 4 outlines the findings of this thesis, presenting

evidence of experiential grammatical metaphor as a key factor for determining the complexity level of segments, which are representative of the complexity level of texts, and, in turn, showing complexity levels as feasible text simplification criteria. Chapter 5 brings the discussion of the results considering the findings in Chapter 4 in the light of the theoretical framework presented in Chapter 2. Chapter 6, Conclusion, outlines the key aspects and the limitations of this study, along with suggestions for future studies. Finally, the References section lists the works cited in this thesis, followed by four Appendices.

Chapter 2 Theoretical framework

This chapter begins with a brief overview on science writing according to SFL, followed by a review of selected works on text complexity geared towards text simplification, covering both computational and linguistic approaches, to subsequently introduce Systemic Functional Linguistics (SFL), our main theoretical framework, especially the concept of grammatical metaphor and its potential application for text simplification.

2.1 Science writing

Systemic Functional Linguistics' approach to science writing is essential to understand how the theory considers science's contribution to the accumulation of knowledge, as science "is totally dependent on scientific language" (Halliday & Martin, 1993, p. 77), which implicates technical terms. "[T]echnical terms accumulate information, allowing […] to move from one explanation to the next" (Halliday & Martin, 1993, p. 182). This accumulation, referred to by the authors as "distillation" or condensation, is related to the use of experiential grammatical metaphors in the logogenesis of texts, which is associated to greater complexity. The comprehension of a more metaphorical text requires, necessarily, to decode knowledge through "decoding strategies" (p. 69), unpacking the meanings that were condensed by the author in the logogenetic unfolding of the text.

According to Halliday & Martin (1993, p. xi), science is an "inter-organistic practice, a linguistic/semiotic practice which has evolved functionally to do specialized kinds of theoretical and practical work in social institutions". By this, the authors mean that science is a practice

"codifying" (p. 132) knowledge in specific ways and integrates a range of institutions involved with some specialized area (e.g., physics). Packing can be extensive and increasingly complex and, at a certain point, allows only the "literate" in certain fields to understand a text. However, a science text can be popularized to reach a broader audience, as is done in popular science news or school textbooks. This leads to the conclusion that "science texts" can relate to several semiotic activities – from expounding (research articles) to reporting (popular science news reports) to explaining (reviews) and others. Besides this, a text in itself, as is the case of the research article, encompasses different semiotic activities – explaining (Introduction); Reporting (Methodology and Results); Expounding (Literature Review). This is referred to by Matthiessen (2015) as "hybridity":

> These types and subtypes [of socio-semiotic activities] shade into one another; there are various hybrid types involving overlaps, blends and neutralizations (cf. Halliday and Matthiessen 2006) — but this hybridity can be brought out and described precisely because we have identified prototypical types and subtypes (p. 9)

For this thesis we have opted to focus on the main activities implicated in science texts, namely, "explaining" and "categorizing", (cf. Figure 6 and Table 6).

As regards procedures that make texts more complex, particularly experiential grammatical metaphor, Halliday & Martin (1993) states that "[t]his function of technicality to distill or condense is seen when we try to unpack technical texts; that is, try to rewrite them without using any technical terms. This procedure reveals that there are degrees of distillation." (p. 163). In other words, the degree of "technicality" is one of the seven criteria to measure the complexity of a text (Halliday, 1993, p. 71). A text can be "unpacked"; this "unpacking"

operation can be executed by degrees, for instance, what Aluísio & Gasperin (2010) calls "natural" and "strong" simplification.

Text simplification strategies based on varying degrees of metaphorization can play an important role to reduce text complexity and produce similar texts that are more "accessible" and aimed at a broader audience (Ravelli, 2006). Assuming that text simplification derives from a decrease in text complexity and is also regarded as a type of paraphrase (Vila et al, 2010; Steiner, 2019), this chapter reviews approaches with applications on text complexity, text simplification, and paraphrase studies, as well as the literature on Systemic Functional Linguistics.

2.2 Computational linguistics and NLP studies on text simplification

This section explores the advantages and drawbacks of Harris' (1957, 1965) Transformational Grammar and Chomsky's (1965) Generative Grammar as computational linguistics in which several works are based on. These approaches are contrasted with Corpus Linguistics' approaches, regarding limitations and considerations on how to best approach text simplification. Definitions of text simplification, text entailment, and paraphrase by researchers in NLP are briefly discussed, as well as text simplification approaches. particularly the one developed for Brazilian Portuguese at NILC (Interinstitutional Center for Computational Linguistics). NILC text simplification guidelines (Specia, Aluísio & Pardo, 2008) are examined from a linguistic perspective in the light of SFL. Then, SFL authors, as well as dissertations and thesis affiliated to this theory are explored. Based on the definitions found and drawing on SFL, Ravelli's (1999) and Steiner's (2004, 2005) hypotheses regarding the association of experiential grammatical metaphor with varying degrees of text complexity are presented.

2.2.1 Text simplification

Studies in computational linguistics approach "text simplification" as an attempt to automatically deal with varying degrees of text complexity. This thesis aims at exploring text complexity levels in order to contribute to text simplification.

What we know about text simplification is largely based upon empirical studies aimed at how to employ computational methods to reduce text complexity or to produce texts with similar degree of complexity, which is largely subsumed under the topic of paraphrase for NLP researchers. The main approaches pursued towards text simplification draw either on traditional and generative grammar or have been developed in studies within the disciplines of corpus linguistics and computational linguistics.

Among the different concepts proposed in Traditional grammar, the notions of coordinate and subordinate clauses, as well as noun, relative and adverbial clauses have been used to describe the relationship between sentences with similar meaning but different structure. Some works in computational linguistics, such as Gonzalez-Dios (2016), have used these notions (coordinate and subordinate clauses, noun, relative and adverbial clauses) to study particular languages, in their case, Basque.

In contrast, Harris (1957), in his Transformational Grammar[3], proposed the concept of transformation as modifications between different linguistic constructions, which lead to different meanings. For this purpose, Harris (1957) compared similar structures with different

[3] Generative Grammar has been referred to by various names while it was being developed, such as "Transformational Grammar (TG), Transformational Generative Grammar, Standard Theory, Extended Standard Theory, Government and Binding Theory (GB), Principles and Parameters approach (P&P) and Minimalism (MP)" (Carnie, 2013, p. 6).

elements belonging to certain “classes” (e.g., noun, verb, adjective, conjunction) to investigate modifications in the co-occurrence of items. As an example, Harris compares “I like this” and “This I like” and “He learned a lesson” and “He learned his lesson” (Harris, 1965, pp. 540-546), and argues that the modification in the sequence of the elements in the first case and the type of modifier (from singular, indefinite article to masculine possessive singular pronoun) in the second case leads to a transformation in the first sentence into the second in each case. According to Madnani and Dorr (2013), Harris’ work led to various proposals on paraphrasing, which were used as the basis for other studies in NLP, such as Chomsky's (1965). One major disadvantage in relying on this approach pointed out by the authors is that “items that are distributionally similar may not necessarily end up being paraphrastic […] [as they] can occur in similar contexts but are not semantically equivalent” Madnani and Dorr (2013, pp. 348-349). Concerning Transformational Grammar, Carnie (2013) states that:

> sentences are generated by a subconscious set of procedures (like computer programs). These procedures are part of our minds (or of our cognitive abilities if you prefer). The goal of syntactic theory is to model these procedures. In other words, we are trying to figure out what we subconsciously know about the syntax of our language. (p .6)

In sum, Generative grammar uses sets of “formal grammatical rules” to model sentences (Carnie, 2013, p. 6). This modeling suggests that in a computational perspective relying on generative grammar to perform paraphrase, syntactic structures known as “surface structures” and “deep structures” can explain how new instances are generated.

Despite the fact that many works from NLP researchers have drawn on Traditional Grammar (e.g., Gonzalez-Dios., 2016) and on Generative Grammar (Štajner et al, 2014; Štajner,

2015, among others), researchers affiliated to other theories, e.g., Sinclair (affiliated to Corpus Linguistics), are opposed to this approach. Sinclair (2004) states that grammars such as Transformational and Generative grammars the well-formed sentences accepted by these types of grammars “fail to describe” the fact that multiple sentences can characterize one single choice of meaning (p. 170).

SFL, on the other hand, brings the concept of grammatical metaphor to attempt to explain this type of linguistic complexity. As grammatical metaphor allows the association between segments of text belonging to different languages or even to the same language, ideational grammatical metaphor, specifically, can be related to the conventional concept of paraphrase.

According to Steiner (2019, p. 743), what is conventionally called paraphrase is "a truth-condition preserving relationship among sets of propositions and the sentences expressing them". From an SFL perspective, it can be said that experiential meanings are to some extent analogous in a relation of paraphrase, whereas textual and interpersonal meanings vary. The notion of conventional paraphrase viewed as texts in which there is analogy in experiential meanings is taken in this thesis as a way to approach the notions of paraphrase and text simplification as discussed by researchers in NLP and computer science (e.g., Vila et al, 2010).

Text simplification is one of the concepts that can be approached using corpora and computational methods (e.g., sampling and parsing), as in Siddharthan (2004), Siddharthan (2006), Štajner (2015), and Hwang et al (2015). In addition, as SFL associates both paradigmatic and syntagmatic axes through realization (Halliday & Mattthiessen, 2014, p. 20), this approach can also be used to investigate text simplification.

The next section goes further into detail on the computational perspective of text simplification based on NLP.

2.2.2 Text simplification application in NLP tasks

Several works on text simplification - e.g., Daelemans et al (2004), Damay et al. (2006), Aluísio et al (2008), Aranzabe et al (2012), Seretan (2012), Barlacchi & Tonelli (2013), Klaper et al (2013), Stymne et al (2013), Štajner et al (2013), Shardlow (2014), and Paetzold & Specia (2017) - are related to the field of Natural Language Processing (NLP). NLP studies use either quantitative (number of words, sentences, or words per sentence) or hybrid methods (frequency tables, collocation, and n-grams).

In general, the main objective of text simplification is to improve text accessibility; however, it can be applied in a wide range of contexts, such as a reading assisting tool for users or as a pre-processing step for other NLP tasks.

In NLP studies, text simplification is defined as a "process of reducing the linguistic complexity of a text, while still retaining the original information content and meaning" (Siddharthan, 2003, p. 3). Bearing on this definition, authors regard a simplified text as a paraphrase or an entailment. Paraphrase is defined as "(approximate) sameness or equivalence of meaning between different wordings. [...] a vague and complex phenomenon with a broad range of manifestations that can involve lexical, syntactic, semantic and pragmatic knowledge" (Vila et al, 2010, p. 191).

Entailment is closely related to paraphrase. A distinction between the two notions is made in the literature by Nevěřilová (2014) as: "[w]hen people explain something, they proceed in two ways: they can express the same thing in other words or they can explicitly voice the implicit knowledge. The former phenomenon is called paraphrase, the latter makes part of textual entailment" (Nevěřilová, 2014, p. 293)

Rocha & Cardoso (2018) examine the distinction between both concepts as one pertaining to scope. As defined by Rocha & Cardoso (2018), "[a] text fragment entails another text fragment if, from the meaning of the former, one can infer the meaning of the latter. If such a relation is bidirectional, then we are in the presence of a paraphrase" (Rocha & Cardoso, 2018, p. 868).

Rocha & Cardoso (2018) view paraphrases as entailed texts but not all entailed texts are paraphrases; which may impact the development of text simplification systems.

Dagan et al (2013) point out that:

> textual entailment does not pertain only to the derivation of new information through reasoning, but also captures the variability of language expression, by which the same information may be stated in many different ways, or at different levels of abstraction. (p. 26)

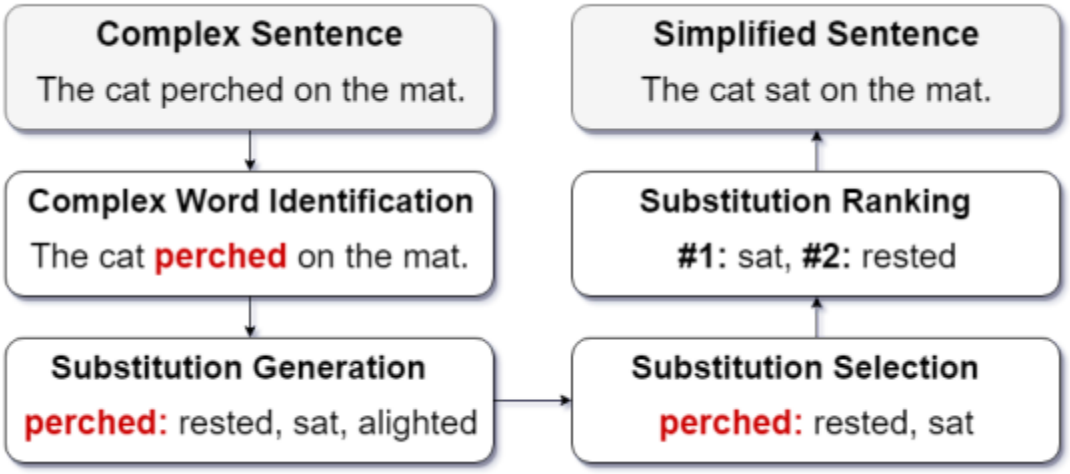
Dagan et al's (2013) remark, in contrast to the previous one, by Rocha & Cardoso (2018), takes into account language variability and different levels of abstraction (e.g., text, sentence, clause) as an issue in discussions of t similarity between different text fragments (or segments).

Text simplification can be performed through different methods with the use of lexical simplification, syntactic simplification, explanation generation, or all of these combined. Lexical simplification is defined as “the process of replacing complex words in a given sentence with simpler alternatives of equivalent meaning” (Paetzold & Specia, 2017, p. 549). Syntactic simplification is defined as “the process of reducing the grammatical complexity of a text while retaining its information content and meaning” (Siddharthan, 2003, p.3). Finally, explanation generation is defined by Shardlow (2014, p. 63) as the addition of extra information to contextualize a concept that is considered difficult by some audience and improve its understanding for this group.

According to Paetzold & Specia (2017, p. 550), most methods of lexical simplification share the following sequence of steps to convert a complex sentence into a simpler sentence, complex understood here as being more difficult to understand. Figure 2 lists the sequence of the following steps and illustrates them with an example: i) complex word identification; ii) substitution generation; iii) substitution selection; iv) and substitution ranking.

Figure 2

Lexical Simplification flowchart



Note: Reprinted from Paetzold & Specia (2017, p. 551).

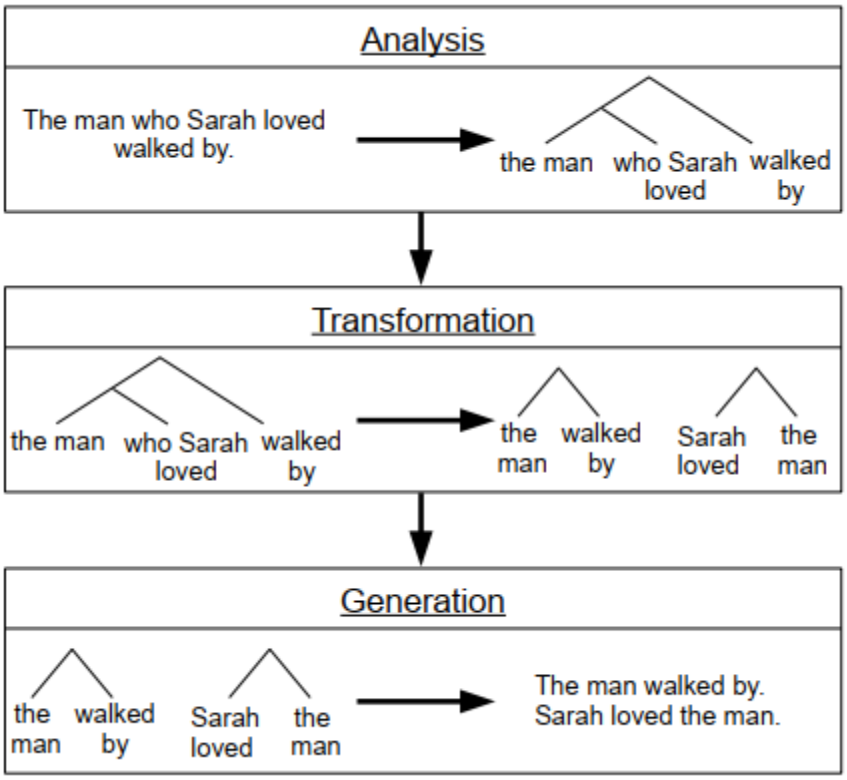
Figure 2 shows that, to simplify a sentence, firstly a software identifies what is deemed as complex words (words that can posit reading comprehension problems to the reader) to be replaced with simpler ones; subsequently, it generates potential candidates for substitution, selects the best candidates and ranks them to achieve the best possible results. These results establish how the simplified sentence is formed, as shown in the last box.

The second text simplification approach is syntactic simplification. According to Shardlow (2014, p. 62), syntactic simplification is often performed with the following three steps applied to the syntactic structure of a sentence: analysis, transformation, and generation (of a new text). Analysis is usually made with the use of a parser; transformation uses some strategies,

such as the splitting of long or complex sentences; and generation processes the results from transformation, correcting eventual mistakes, such as gender or number assignment. Figure 3 presents these steps with the sentence “The man who Sarah loved walked by.” (Shardlow, 2014, p. 62) as an example.

Figure 3

Syntactic Simplification flowchart



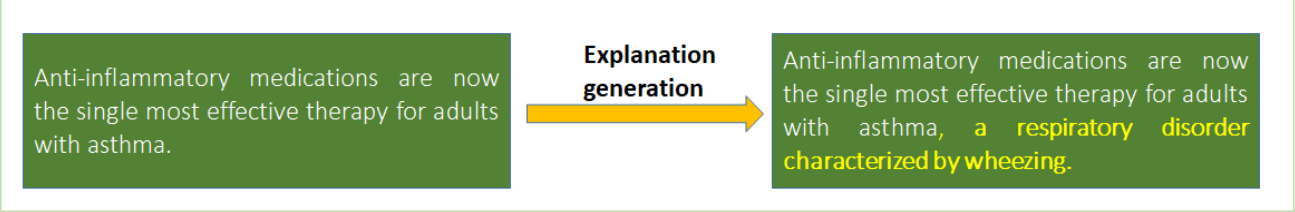
Note: Reprinted from Shardlow (2014, p. 62).

Figure 3 shows the intermediate stages in the syntactic simplification process applied to the sentence “The man who Sarah loved walked by.”, which results in the sentences: “The man walked by.” and “Sarah loved the man.”

The third text simplification approach is explanation generation, which is performed in texts with a high degree of technicality, as in the healthcare literature. In these texts, a technical term is often used without an explanation, considering that their intended audience is already familiar with the technical jargon. When these texts target readers without the technical expertise to deal with this jargon, readability and understandability decrease. To widen the range of readers for texts like these, a text simplification strategy that is applied is to recognize a technical term considered “difficult” and automatically add an explanation to the technical term retrieved from a lexicon. Figure 4 shows an example from Damay et al. (2006, p. 37) of this strategy being applied in a text from the medical field, based on a definition retrieved from SimText lexicon.

Figure 4

Example of Explanation generation in a text segment from the medical field



Note: Adapted from Damay et al. (2006, p. 37).

In Figure 4, as explained by Shardlow (2014), there is explanation generation (highlighted in yellow). The text was simplified through the automatic addition of an explanation to the medical concept “asthma” at the end of the sentence. This is meant to increase understandability, considering a broader audience; yet, if the explanation is still highly metaphorical, with the use of a nominal group highly technical (“respiratory disorder”) modified by a qualifier realized by an embedded clause (“characterized by wheezing”), this strategy is not as efficient as it can be. Therefore, an approach that manages to reduce text complexity must take metaphorization into account, unlike the traditional approaches reviewed in this thesis.

Other approaches use linguistic databases to inform text simplification systems. For instance, as seen in Ji & Eisenstein (2013), for text simplification, linguistic databases, mostly corpora containing pairs of instances, often provide input to be computationally analyzed. Some of these works retrieve some instances of non-simplified language and manually simplified language pairs, producing new datasets, such as *LSeval* (De Belder & Moens, 2012) and *LexMTurk4* (Horn et al., 2014). Other works (Kauchak, 2013; Zhu, Bernhard, & Gurevych, 2010; Kajiwara & Komachi, 2016) have produced datasets by automatically aligning sentences from online sources in English, like the Standard and Simple Wikipedia (Xu, Callison-Burch & Napoles, 2015). Others query linguistic databases to obtain their input; for instance, *WordNet* (Fellbaum, 1998) and *BabelNet* (Navigli & Ponzetto, 2010).

Text simplification studies have also been developed for several languages. Table 1 presents a brief overview of some studies, with some notes describing briefly the nature of these works.

Table 1

Text simplification in languages other than English

Year	Author(s)	Language	Title	Focus
2004	W. Daelemans, A. Höthker, EFTK Sang	Dutch	Automatic Sentence Simplification for Subtitling in Dutch and English	A comparison between machine learning and rule-based approaches for Dutch language simplification of subtitles. The rule-based approaches outperformed the machine learning because the corpus was too small to provide enough training for the machine and could not avoid “nonsensical errors” (p. 2)

2012	María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios	Basque	Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque	Automatic Basque language syntactic simplification carried out by an algorithm based on dependency trees drawing on the Constraint Grammar. After the analysis of relative and adverbial temporal clauses, the output of the simplification process consists of simpler sentences that must be corrected to ensure grammatical sentences.
2012	Violeta Seretan	French	Acquisition of Syntactic Simplification Rules for	The acquisition of a comprehensive list of

			French	rules for simplifying newspaper articles in French through a method that “leverages linguistic tools to provide annotators with statistically salient linguistic expressions that are potentially interesting from a rule inference point of view” (p. 4022). This method produces sets of rules comparable to those proposed by Dras (1999) and Specia et al. (2008), who investigated English and Brazilian Portuguese.
2012	H.B. Chen, H.H. Huang, H.H.	Chinese – English	A Simplification Translation Restoration Framework for	The implementation of a Chinese – English.

	Chen, C.T. Tan		Cross-domain SMT Applications	Simplification module for a “general purpose statistical machine translation (SMT)” system that translates domain specific knowledge After a medical summary translation, the integrated system outperforms SMT-based and other integrated systems.
2012	B.T. Hung, N. Le Minh, A. Shimazu	Vietnamese	Sentence Splitting for Vietnamese-English Machine Translation	The use of Vietnamese language sentence splitting to improve the output of automatic translation approaches avoids that syntactic differences between languages produce

				inefficient outputs in the target language.
2013	S. Štajner, B. Drndarevic, H. Saggion	Spanish	Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification	A news articles' corpus analysis of original and manually simplified texts in Spanish aimed to investigate relevant operations to implement in a text simplification system for people with cognitive disabilities. Findings suggest that summarization and paraphrases were the most relevant operations, followed by lexical and syntactic operations. Furthermore, complexity measures modifying average

				sentence length play an important role in simplifying texts for people with cognitive disabilities.
2013	G. Barlacchi, S. Tonelli	Italian	ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian	Aimed at children with low reading skills, ERNESTA, this Italian Syntactic Simplification system, analyzes texts resolving anaphoras to explicitate information and make children stories simpler. ERNESTA's impact covers different tasks, such as educational games and reading comprehension tasks.

2013	D. Klaper, S. Ebling, M. Volk	German	Building a German/Simple German Parallel Corpus for Automatic Text Simplification	A web corpus for the production of a German monolingual statistical l machine translation system to translate German into simple German.
2013	Sara Stymne, Jörg Tiedemann, Christian Hardmeier, Joakim Nivre	English – Swedish	Statistical Machine Translation with Readability Constraints	Aiming at enhanced readability by a specific audience, e.g., language learners, this English – Swedish. Simplification system produced simplified translations. These translations, though, can be influenced by several factors and lead to different types of simplifications.

2013	J.W. Chung, H.J. Min, J. Kim, J.C. Park	Korean	Enhancing Readability of Web Documents by Text Augmentation for Deaf People	This news article's simplification system of Korean for deaf readers converts complex sentences into simple sentences, presenting their relationship graphically.

Note: Adapted from Shardlow (2014, pp. 61-66)

Several tools and resources have been developed for Brazilian Portuguese at NILC and made available online[4]. Among the resources available on the NILC website, some are related to automated analysis of impaired speech and text, to word embeddings, to writing assistance and text simplification and evaluation, including corpora for different purposes, systems used for lexicon and semantic analysis, for syntactic analysis, and for semantic and discourse analysis. Other resources are associated with summarization tasks, pre-processing tasks and POS tagging, machine translation, and speech analysis.

A software developed to perform an automated analysis of impaired speech and text in dementia is called *Coh-Metrix-Dementia* and works on the browser Google Chrome. The word

[4] These tools and resources are available at http://www.nilc.icmc.usp.br/nilc/index.php/tools-and-resources

embeddings repository *NILC embeddings* stores and shares word vectors for NLP and machine learning tasks, and several evaluation datasets to assess linguistic models, that could be used to input text simplification tasks.

Another tool is a software created for executing syntactic analysis, which is called *Sucupira*. For semantic and discourse analysis, the following tools are available: *CST Parser* and *CST Tool* (for processing syntax trees); *DiZer 2.0*, for discourse parsing; *TextTilling* for Portuguese, for dividing texts into smaller sets that represent a certain topic; *RSTeval*, *RST*[5] *Toolkit* and *Segmentador RST* for working with Rhetorical Structure Theory; and *NLPNet - Semantic Role Labeler* (SRL).

Some tools were also developed for machine translation tasks, namely *Alinhadores* (for aligning texts at the sentence and word level), *PorTAl* (Machine Translation Portal – a portal to assist with the processing of multilingual texts), *Trapezio* (a translation post-editor), and *VisualLIHLA* (for aligning texts according to lexical criteria). Finally, the systems developed for treating speech data are *Petrus* (an automatic phonetic transcription system which is available online for Brazilian Portuguese), *Listener* (to assess the English pronunciation of Brazilian speakers), and *Aeiouadô*, a pronunciation dictionary for Brazilian Portuguese designed to be used by other computational tools, like those related to speech technologies. From these resources, *PorTal* could be used to do NLP tasks using machine translation for text simplification tasks.

The corpora available are *CorpusTCC*, *CorTrad*, *CSTNews*, *Lácio-Web*, *HPC*, *OpiSums-PT*, *PorSimples*, *RHETALHO*, *Corpus NILC,* and *TweetSentBR*. The systems developed for performing lexicon and semantic tasks are *ABNT Rules*, *e-Termos*, *Mini Gramática*, *Palavras*

[5] SRT stands for “Semantic Representation of Text”.

Compostas, *PortLex*, *Propbank.BR* (which is also a dataset), *VerbNet.BR*, *TEP*, *Unitex PB*, *WordNetBr*, a Neologism detection tool, and *LIWC* for Portuguese.

Some of the systems for assisting in text writing and for text simplification/evaluation tasks are *CATEAP*, *Simplifica* (developed on the basis of *PorSimples*), *Facilita Educational*, *SciPo*, *Scipo-Farmácia* in Portuguese and English, *MAZEA-web*, *Escrita Científica*, *CALEP-Web*, *CALeSe*, *AlCórpus* and *Coh-Metrix-Port* (an adaptation of *Coh-Metrix*[6] tool into Brazilian Portuguese).

NILC also developed datasets unrelated to word embeddings, namely *Mac-Morpho*, a corpus of Brazilian Portuguese texts with POS (part-of-speech) tags; *MilkQA*, a dataset of questions and answers for answer selection tasks; *ASSIN* (Semantic similarity and textual inference assessment), a dataset used to evaluate paraphrase and textual implication[7]; *PropBank.Br*, a collection of verb senses and verb mappings in English; *PorSimplesSent*, a corpus of aligned sentences in Portuguese for sentence readability assessment studies; and *SIMPLEX*, a lexical simplification database.

Other tools, developed for text summarization tasks, are *RSumm*, *GistSumm*, *ViSum*, *Summary coherence evaluation*, *NILC-WISE*, *Sentence ordering program*, and *GEI*. Besides, for pre-processing data some systems were developed, such as *Taggers*, *Stemmer*, *Lemmatizer* for Portuguese (for Brazilian Portuguese), *Senter* (which is pre-loaded with historical Portuguese abbreviations to segment these texts properly), and a POS tagger tool called *NLPNET – POS tagger*.

---

[6] Available at http://cohmetrix.com/

[7] According to Zaenen et al. (2005, p. 31), "local textual inferences come in three well-defined varieties (entailments, conventional implicatures/presuppositions, and conversational implicatures)". In this thesis, the notions of implicature/explicature and presupposition will not be contrasted since it is out of the scope of this study.

At NILC, Rocha & Cardoso (2018) explored the ASSIN corpus, a corpus for textual entailment and paraphrase in Brazilian and European Portuguese to perform automatic recognition of textual entailment and paraphrase. The authors employ supervised machine learning techniques[8] that analyze lexical, syntactic, and semantic features and assess the performance of the system. Based on their "outstanding" results, they conclude that this semantic approach can be considered promising, yet, combining data from European and Brazilian Portuguese can present extra challenges. The reason is that increasing the number of texts in Brazilian Portuguese decreases the performance for European Portuguese and vice versa, most likely because "syntactic and semantic differences between the two variants" have caused most of the errors in the system output.

Also, at NILC, aiming to map semantic similarity between texts to identify paraphrases, Ji & Eisenstein (2013) extracted latent (syntactic) representations of sentences to obtain features that were combined with features based on n-gram operations. The result was a 3% increase in performance compared to the prior state-of-the-art system. They also pointed out potential challenges faced by approaches aiming at identifying paraphrase. Among them, they highlight two main challenges:

> i) the "infinitely diverse set of possible linguistic realizations for any idea" (Bhagat and Hovy, 2013);
>
> ii) the low number of words in each sentence, which makes the "bags of words" used in standard approaches too sparse to produce results with great efficiency. (p. 891)

[8] Machine learning techniques can be unsupervised or supervised. In short, unsupervised techniques analyzed the data without any previous assumptions, drawing inferences by the use of the rules that are automatically detected by the machine. In contrast, supervised techniques depend on the user input, e.g., a preliminary annotation, from which to derive the patterns and replicate them to another part of the dataset.

For this reason, their approach was supervised, instead of unsupervised, using labeled data to improve the methods for finding paraphrase by introducing a new weighting scheme[9] that discriminates the most important features from the data and improves the performance in this task.

Aluísio & Gasperin (2010) describe the PorSimples project (Text Simplification in Brazilian Portuguese for the Inclusion and Digital Accessibility - Simplificação Textual do Português para Inclusão e Acessibilidade Digital), aimed to develop text simplification approaches for Portuguese through the use of text entailment. This approach assesses the readability of texts and integrates lexical and syntactic simplification to produce either a "natural" or a "strong" simplification[10], reducing slightly or greatly the complexity of a text. As a result, they developed three distinct systems: i) *SIMPLIFICA*, an "authoring system" to produce simplified texts; ii) *FACILITA*, an "assistive technology system" for assisting a less literate audience to read content on the web; and iii) Educational FACILITA, a "web content adaptation tool" for "assisting "low-literacy readers to perform detailed reading" (p. 50).

For the Portuguese language, studies such as Aluísio et al. (2008), Caseli et al. (2009), Candido et al. (2009), Gasperin et al. (2010), Specia (2010), and Aluísio et al. (2010) have contributed to the study of text simplification. Since the focus of this thesis is text simplification phenomenon in Brazilian Portuguese, these authors are taken into consideration when dealing with linguistic features specific to this language.

---

[9] A weighting scheme is a scheme that assigns different weights to distinct features to measure their relevance in face of some phenomenon that is being investigated.

[10] Aluísio & Gasperin (2010) state that a "natural" simplification is intended for an audience with "basic literacy level", who can comprehend a text that is simplified enough to be considered "natural", and the "strong" simplification for an audience with "rudimentary level", which requires a text "as simple as possible".

Works that explore text simplification in Brazilian Portuguese use parallel corpora of non-simplified and simplified texts and take into account the classification of functional literacy in Brazil proposed by INAF (National Indicator of Functional Literacy). These works on Portuguese reported that due to the lack of a segment corpus in Brazilian Portuguese, it was necessary to compile their own corpora to study. Nevertheless, most of these works - except Aluísio et al. (2010), which draws on Aluísio et al's (2008) methodology, use different strategies to draw conclusions on the basis of corpus analysis.

Summarizing the text simplification studies affiliated to NILC, the main contribution from Aluísio et al. (2008), in the scope of the PorSimples project, was the proposition of sets of steps associated with linguistic phenomena to simplify texts in Brazilian Portuguese. Caseli et al. (2009), Candido et al. (2009), Specia (2010), and Aluísio et al. (2010) used the set of strategies developed in this project to develop their works. Gasperin et al. (2010), in turn, used the strategies proposed by Siddharthan (2003), which suggests three necessary phases to simplify texts: identification of phenomena, application of simplification rules, and regeneration (recovery of the correct wordings) of segments that were left behind during the simplification process.

In 2008 NILC developed a text simplification manual (Specia, Aluísio & Pardo, 2008). Table 2 shows simplification operations, including examples translated into English for the sake of comprehension.

Table 2

Text simplification strategies from 2008 NILC manual

Targeted Structure	Simplification operation	Example of non-simplified excerpt	Example of simplified excerpt
1 - apposition	splitting sentences	Tarso Genro também visitou ontem o antecessor, **José Genoino.** *Tarso Genro also visited former president*[11] ***José Genoino*** *yesterday***.**	Tarso Genro também visitou ontem o antecessor. **O antecessor é José Genoino.** *Tarso Genro also visited the former president yesterday. The former president is* ***José Genoino***.
2 - relative clauses	splitting sentences	A sabatina**, que é aberta à participação dos assinantes da Folha**, será no Teatro Folha. *The interview,* ***which is open to all Folha de São Paulo subscribers****, will be at Folha*	A sabatina será no Teatro Folha. **A sabatina é aberta à participação dos assinantes da Folha.** *The interview will be at the Folha Theater.* ***The interview is open to all***

[11] Both Tarso Genro and José Genoíno were presidents of the political party Partido Trabalhista (Workers' party), usually referred to in the media as PT.

		Theater.	***Folha de São Paulo subscribers.***
3a - relative subordinate clauses	splitting sentences	O governo criou o Ministério de Assistência e Promoção Social, **o qual irá avaliar os programas já existentes.** *The government implemented the Ministry of Social Welfare and Development,* ***which will assess the existing programs.***	O governo criou o Ministério de Assistência e Promoção Social. **O Ministério de Assistência e Promoção Social irá avaliar os programas já existentes.** *The government implemented the Ministry of Social Welfare and Development.* ***The Ministry of Social Welfare and Development will assess the existing programs.***

3b - defining relative clauses	splitting sentences	O sol **que se filtra através das folhas** desenha no ar colunas de poeira. *The sun **shining through the leaves** makes dust columns swirl in the air.*	**O sol se filtra através das folhas**. O sol desenha no ar colunas de poeira. ***The sun shines through the leaves**. The sun makes dust columns swirl in the air.*
3c - causative subordinate clauses	- splitting sentences - replacing a discourse marker by a simpler or more frequent one - modifying syntactic order (to keep condition-effect)	Queiroz foi levado para o Pinel **porque** estaria muito exaltado. *Queiroz was taken to Pinel psychiatric hospital **because** he was reportedly mentally distressed.*	Queiroz estaria muito exaltado. **Com isso**, Queiroz foi levado para o Pinel. *Queiroz was reportedly mentally distressed. **Due to this**, he was taken to Pinel psychiatric hospital*

3d - comparative subordinate clauses	- splitting sentences - replacing a discourse marker by a simpler or more frequent one	Ele não foi localizado pela Agência Folha, **assim como** os outros deputados dissidentes. *He was not located by Folha Press; nor were the other dissident congressmen*	Ele não foi localizado pela Agência Folha. Os outros deputados dissidentes **também não foram localizados**. *He was not located by Folha Press. The other dissident congressmen **were not located either.***
3e - concessive subordinate clauses	- splitting sentences - replacing a discourse marker by a simpler or more frequent one	**Conquanto** Castello Branco **esteja** disposto a ajudar, há uma série de obstáculos legais e políticos. ***Although** Castello Branco is willing to help, there are several legal and political constraints.*	Castello Branco **está** disposto a ajudar. **Mas** há uma série de obstáculos legais e políticos. *Castello Branco is willing to help. **However,** there are several legal and political constraints.*
3f - conditional subordinate clauses	- modifying syntactic order (to keep condition-effect) - replacing a	Apóio o candidato do meu partido, **a menos que** seja o Malan. *I support the candidate of my*	**Se** o candidato **não** for Malan, **então** apoio o candidato do meu partido. *If the candidate is not Malan, then I will support*

	discourse marker by a simpler or more frequent one	*party **unless** it is Malan.*	*the candidate of my party.*
3g - consecutive subordinate clauses	- splitting sentences - replacing a discourse marker by a simpler or more frequent one	Estamos dispostos a ajudar um parceiro muito importante, **de forma que** o processo Alca seja benéfico. *We are willing to help a very important partner **so that** the Alca process is beneficial.*	Estamos dispostos a ajudar um parceiro muito importante. **Por isso**, o processo Alca será benéfico. *We are willing to help a very important partner. **Hence** the Alca process will be beneficial.*

3h - final subordinate clauses	- splitting sentences - replacing a discourse marker by a simpler or more frequent one	A gravidade dos fatos exige imediato posicionamento de vossa excelência **a fim de que** as medidas administrativas e judiciais pertinentes sejam adotadas. *The seriousness of the situation requires immediate response from Your Excellency* ***in order that*** *proper administrative and legal measures be adopted.*	A gravidade dos fatos exige imediato posicionamento de vossa excelência. **O objetivo é que** as medidas administrativas e judiciais pertinentes sejam adotadas. *The seriousness of the situation requires an immediate response from Your Excellency.* ***The objective is that*** *proper administrative and judicial measures be adopted.*
3i - proportional subordinate clauses	non-simplification[12]	**À medida que** as eleições se aproximam, fica mais difícil. *As elections approach, the situation turns out to be more challenging.*	

[12] There is no simplified version predicted by this criteria.

3j - conformative subordinate clauses	replacing a discourse marker by a simpler or more frequent one	**Conforme a Folha apurou**, os repasses totais do BB à DNA foram crescendo ano a ano. *As confirmed by Folha, the number of bank transfers from Bank of Brazil (BB) to DNA increased year by year.*	**A Folha apurou que** os repasses totais do BB à DNA foram crescendo ano a ano. *Folha has confirmed that the number of bank transfers from Bank of Brazil (BB) to DNA increased year by year.*
3k - temporal subordinate clauses	- splitting sentences - replacing a discourse marker by a simpler or more frequent one	Eu demiti o tesoureiro do sindicato **logo que** descobri os documentos falsificados. *I fired the labor union treasurer as soon as I found the counterfeit documents.*	Eu descobri os documentos falsificados. **Em seguida**, eu demiti o tesoureiro do sindicato. *I found the counterfeit documents. I fired the labor union treasurer immediately afterwards.*

4a – asyndetic coordinate clauses	splitting sentences	João subiu pela velha escada de madeira mal iluminada, **chegou a uma espécie de salão.** *João used that dimly lit old wooden staircase, arriving at a kind of hall.*	João subiu pela velha escada de madeira mal iluminada. **João chegou a uma espécie de salão.** *João used that dimly lit old wooden staircase. João arrived at a kind of hall.*
4b – syndetic coordinate clauses	splitting sentences	O presidente da Câmara se disse "surpreendido" com a renúncia de Valdemar **e** elogiou a "bravura" de Valdemar. *The Senate president claimed that he was "astonished" with Valdemar's resignation and praised Valdemar's "bravery".*	O presidente da Câmara se disse "surpreendido" com a renúncia de Valdemar. **O presidente** elogiou a "bravura" de Valdemar. *The Senate president claimed that he was "astonished" with Valdemar's resignation. The president praised Valdemar's "bravery".*
5 - non-finite verbs	non-simplification[13]	Reunidos na capital da Etiópia, os líderes dos países africanos recusaram a	

[13] There is no simplified version predicted by this criteria.

		proposta que havia sido oferecida pelo G4 em Londres. *Meeting in the capital of Ethiopia, the leaders of the African countries refused the proposal submitted by G4 in London.*	
6 - passive voice	converting passive voice into active voice	As transferências **foram feitas** pela empresa Boston Comercial e Participações por meio de uma conta CC-5. *The bank transfers were made by the Boston Commercial and Holdings company through a CC-5 account.*	A empresa Boston Comercial e Participações **fez** as transferências por meio de uma conta CC-5. *The Boston Commercial and Holdings company made the bank transfer through a CC-5 account.*

Note: Adapted from Specia, Aluísio & Pardo (2008, emphasis added). All the strategies and examples in Portuguese were translated into English by the author of this thesis.

The structures that are targeted in the authors' guidelines are (1) apposition, (2) relative clauses, (3) subordinate clauses, (4) coordinate clauses, (5) sentences with non-finite verbs, and (6) passive voice. The simplification operations in Table 2 are (a) splitting sentences, (b)

replacing a discourse marker by a simpler and/or more frequent one (to avoid the ambiguous ones), (c) converting passive into active voice, (d) inverting clause order, and (e) non-simplification. Another strategy that is not explicit in the report is the modification of the syntactic order to keep condition-effect order, though it takes place in 3c and 3f.

In Table 2, the strategies devised by the authors are applied to different syntactic structures to provide a simpler version of the excerpt.

In structure 1, the PARTICIPANT "José Genoino" (which refers to a famous politician in Brazil), is being used as the complement in a relational clause in the present tense with the structure "X be Y". X refers to "o antecessor" (*the predecessor*) and Y refers to "José Genoino". In this case, the clause complex was divided in two simplexes, through the "splitting sentences" strategy. The two sequences in in the simplified excerpt are linked by a cohesion device[14] called repetition (of the SUBJECT).

In structure 2, the relative clause "que é aberta à participação dos assinantes da Folha**"** **(***which is open to all Folha de São Paulo subscribers",* a famous newspaper from São Paulo, Brazil**)** is replaced by a new relational clause. The subject of this relational clause is the recovered PARTICIPANT "a sabatina" (a type of debate). The same takes place in 3a, with a relative subordinate clause (or dependent hypotactic clause), in which the relative pronoun "o qual" (which) is replaced by the recovered PARTICIPANT "O Ministério de Assistência e Promoção Social" (*Ministry of Social Welfare and Development*). In this case, both clauses from the simplified excerpt are also linked by the use of repetition of the subject.

[14] Further information on cohesion drawing on SFL is available in Halliday & Hasan (2013), an ebook version of the original work published in 1976.

A different operation is observed in 3a and 3b, in which the restrictive subordinate clause (embedded clause, according to SFL) becomes a new material clause, due to the nature of the process “filtrar” (*to filter*). In both cases, both clauses in the simplified excerpt are linked by the use of the cohesion device “repetition” (of the SUBJECT). From 3c to 3k, though, a discourse marker (in bold in Table 2) is replaced by a more frequent (thus, supposedly simpler) one. In 3f, though, a new strategy that was not stated before comes up, a modification of the syntactic order to follow the sequence condition-effect in a conditional clause, which would be hypotactic, drawing on SFL.

In 4a and 4b, a free clause in a paratactic relationship (either without or with conjunction, respectively) is agnated as a new simplex, instead of part of a complex.

In 5, the manual states that clauses with non-finite verbs, minor clauses to SFL, are not simplified, except for some specific cases not mentioned in Table 2. Finally, the last example shows that passive voice is replaced by an active voice, which takes place in the verbal group.

Another recurrent operation to make texts simpler is paraphrasing - regarded by Rocha & Cardoso (2018) as a type of text entailment. Paraphrasing is explored further by other authors, such as Rigat (2013) and Madnani and Dorr (2010).

Rigat (2013) categorizes paraphrases into two main types: “reformulative” and “non-reformulative”. While reformulative paraphrases produce modifications in a source text excerpt, the non-reformulative paraphrase behaves differently. In Computational Linguistics, these two types require different techniques (Barrón-Cedeño et al, 2013, p.181), which are not error-free (Rigat, 2013, p. 30). These techniques can be complex, since “a great variety of linguistic

operations give rise to paraphrases, [for this reason,] a single paraphrase may include multiple combined paraphrase phenomena" (Rigat, 2013, p. 2).

According to Rigat (2013), one of the main issues paraphrasing faces is the acquisition of an adequate data source, as it requires a corpus in which the same meanings are realized in different wordings and a large number of contexts. The availability of these meanings defines the quality and the number of obtainable paraphrases (p. 5-6). Also, it is not possible to obtain "general and comprehensive paraphrase corpora", only corpora covering "specific paraphrasing types or facets" (p. 2) may be compiled.

Some works aim at building paraphrase corpora; yet, this number is not high. An alternative that is often used is to rely on works from related research topics; for instance: text compression, aiming at obtaining "a summary paraphrase", text simplification, "a specific type of paraphrases where the complexity of the text is reduced", text entailment, aforementioned and defined by Rocha & Cardoso (2018), and plagiarism studies, which relies mostly on paraphrasing (p. 11). Then, corroborating the conclusions of Rigat (2013), for each of these purposes different corpora are required due to the dissimilarities between these research topics.

Different types of corpora can be used to generate paraphrases, such as documents retrieved from the web, monolingual and parallel corpora. Madnani and Dorr (2010) suggest that web documents can be used to build corpora, with a "non-trivial pre-processing phase" to deal with the noise"[15]. They also claim that a monolingual corpus may not be the best choice, because there are "no explicit clues available that indicate semantic equivalence" (p. 354), while in

[15]Although the definition of noise is not explicit in Madnani and Dorr (2010), Wu & Zhu (2004) regard noise as those errors introduced into data, intentionally or not. The two major sources of noise are contradictory examples and misclassifications, by the machine or by an expert.

parallel corpora “the sentence pairs are paraphrases almost by definition”, it is more feasible to find and compare the representation of the same meanings in different wordings.

One way to solve this issue is to provide more instances and more precision in the paraphrasis extraction through the use of “multiple sources of information, yield[ing] paraphrases with much higher accuracy" (Madnani and Dorr, 2010, p. 380).

On the basis of a corpus with segments from texts retrieved online and a manually constructed corpus compiled by linguists, this thesis seeks evidence on the association between experiential grammatical metaphor and varying degrees of text complexity, which may lead to text simplification. Although this manual analysis is limited to a reduced number of segments, the analysis benefits from a higher accuracy in detecting pieces of evidence.

To establish which text simplification strategies are aligned to the SFL framework, this thesis analyzed the strategies proposed by Siddharthan (2004), Siddharthan (2006), Štajner (2015), and Hwang et al (2015), which suggested a wide range of text simplification strategies that are followed by other authors revised in this thesis. These strategies are shown in Table 3.

Table 3

Text simplification strategies in NLP literature[16]

Strategy	**Siddharthan (2004)**	**Siddharthan (2006)**	**Štajner (2015)**	**Hwang et al (2015)**
Conversion from passive voice to active voice	✗	✗	✗	✓
Addition of explanation or examples	✗	✗	✗	✓
Replacement of more technical words with less technical ones (lexical simplification)	✓	✗	✗	✗
Syntactic order modification	✗	✗	✗	✓
Sentence splitting	✗	✗	✓	✗

[16]In the third line of this table, we refer to lexical simplification by the comparison between "more technical words" and "less technical words", but the texts from NLP literature refer to them as "difficult" and "easy" words.

Sentence elimination	✗	✓	✗	✗
Sentence compression	✗	✗	✗	✓
Dis-embedding of relative clauses	✓	✗	✗	✗
Separation of subordinate clauses	✓	✗	✓	✗

Table 3 shows a range of strategies for simplifying texts through automatic processes that can also be applied manually as well. These come from different sources from the literature, from which Table 3 presents only a few. They are adaptations of the procedures applied by, mainly, professional editors to simplify texts, such as in the Newsela corpus.

The interpretation of these strategies in the light of linguistics probably yields more effective results. The linguistic perspective, including those studies focused on SFL, is described in the next section, that reviews SFL as an appliable theory and how this thesis approaches text complexity levels to clarify text simplification.

2.3 Systemic Functional Linguistics

Systemic Functional Linguistics (SFL) assumes that language construes reality through choices made in a given context. Language is organized on the basis of dimensions and ordering principles, which are shown in Table 4.

Table 4

The dimensions in language and ordering principles

	Dimension	Principle	Orders
1.	stratification	realization	semantics ~ lexicogrammar ~ phonology ~ phonetics
2.	instantiation	instantiation	potential ~ subpotential/ instance type ~ instance
3.	metafunction	metafunction	ideational [logical ~ experiential] ~ interpersonal ~ textual
4.	system (paradigmatic order)	delicacy	grammar ~ lexis [lexicogrammar]
5.	structure (syntagmatic order)	rank	clause ~ group/phrase ~ word ~ morpheme [lexicogrammar]; tone group ~ foot ~ syllable ~ phoneme [phonology]

Note. Adapted from Halliday & Matthiessen (2014, p. 20)

Table 4 shows the 5 language dimensions posited by SFL, as well as their organizing principles and their orders. These are explained in turn in the following paragraphs.

The first principle is stratification, which organizes language by orders through realization between content and expression. Content expands into lexicogrammar and semantics and expression into phonology and phonetics. In other words, meanings are realized from the most abstract order (context) to the least abstract (phonetics). Context is realized by semantics,

semantics is realized by the lexicogrammar, lexicogrammar is realized by phonetics, and phonology is realized by phonetics.

The second principle is instantiation, which is shown in Table 5.

Table 5

Cline of Instantiation in context and language

	Potential	Subpotential	Instance type	Instance
context	context of culture (cultural potential)	institutional (subcultural) sites	situation types	contexts of situation
language	language system (meaning potential)	register	text types	texts (acts of meaning)

Note: Retrieved from Matthiessen et al (2010, p. 123).

Presenting Figure 7 (see section 2.3.3.1) as a table, Table 5 shows that there is a continuum (from left to right) between the poles of language potential and instance. While the potential includes all the possibilities enabled by the linguistic system, the linguistic instance describes each configuration realized by the system in a text. In between, the sub-potential or instance type categorizes the potential into subtypes. The CONTEXT OF CULTURE is associated with the potential and CONTEXT OF SITUATION with the instance, given the degree of specificity within this continuum. The institution or situation type is at the intermediate point, in between the potential and the instance.

Martin (2013) defined how instantiation is organized:

> The cline of instantiation is methodologically and theoretically important because it defines the domains of observation, analysis, description and theory in scientific

> engagement with language [...]. Systemic functional linguists (or more generally, semioticians) study the phenomenal realm of language (or more generally, semiotic systems) by observing, sampling and analysing instances at the instance pole of the cline of instantiation - texts in their contexts of situation. Based on the analysis of instances, they can move further up the cline of instantiation towards the potential pole by making generalizations about sets of texts sampled to be representative of some point higher up the cline of instantiation such as a text type or a register, or of the overall potential itself, the particular semiotic system being studied. (p. 116)

The third dimension is metafunction, organized in terms of IDEATIONAL (further divided into logical and experiential components), INTERPERSONAL, and TEXTUAL meanings. These metafunctions are sets of systems that are grouped together according to grammatical criteria. Each metafunction has a different viewpoint. The IDEATIONAL metafunction accounts for linguistic resources used to transform human experiences into meanings, (Halliday & Matthiessen, 2014, p. 30). Each flow of discourse has a logical structure (logical component) and a content (experiential component). The INTERPERSONAL metafunction accounts for meanings as an exchange, "as an interactive event involving speaker, or writer, and audience" (Halliday & Matthiessen, 2014, p. 134). The TEXTUAL metafunction, in turn, organizes language structure, working together with the other two metafunctions. In doing so, the TEXTUAL metafunction accounts for the way discourse functions in language, such as which information is highlighted or how a new piece of information restates one that was mentioned previously in a text.

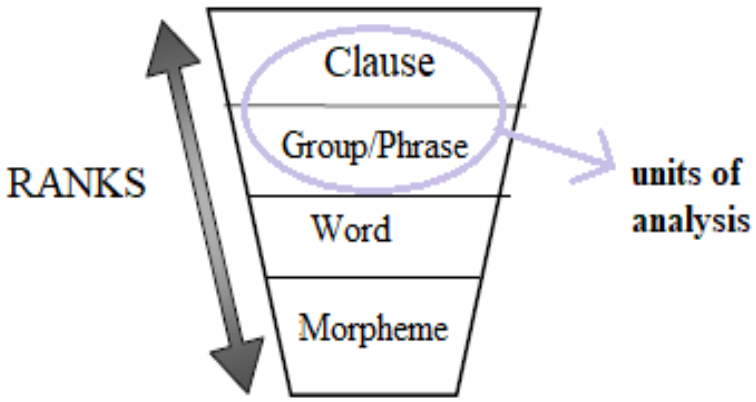
These three dimensions can be analyzed according to language's paradigmatic and syntagmatic orders in structure and system. Concerning the system, language is ordered by the

principle of delicacy, which represents the cline between grammar and lexis, regarded as lexicogrammar. Concerning the structure, written language composition in English follows the hierarchy clause ~ group/phrase ~ word ~ morpheme, associated with the lexicogrammar, and spoken language follows the hierarchy tone group ~ foot ~ syllable ~ phoneme, related to phonology.

Halliday and Matthiessen (2014, p. 115) state that the lexicogrammar of English is organized according to the following ranks: clause, group/phrase, word, and morpheme - each rank associated with the rank above through composition. Figueredo (2007, p. 256) states that both the systems in Brazilian Portuguese and in English are organized in four ranks: morpheme, word, group/phrase, and clause. Thus, Brazilian Portuguese is organized on a rank scale with the same number of levels as English, as shown in Figure 5.

Figure 5

Rank scale for Brazilian Portuguese



Note: Adapted from Figueredo (2007, p. 256).

Figure 5 shows that the lexicogrammar of Brazilian Portuguese is organized according to the following ranks: clause, group/phrase, word, and morpheme. The arrow indicates that an analyst may shunt between ranks to achieve results.

The unit of analysis in this thesis is the clause, even though the group is also analyzed in terms of semantic categorization (as PARTICIPANT, PROCESS, CIRCUMSTANCE). The reason for that is the fact that Halliday (2004, p. 36) highlights these units of analysis as relevant while comparing the "clausal variant" and the "nominal variant" to describe which text excerpts are more metaphorical, regarding IDEATIONAL meanings.

This thesis focuses on the clause and the group ranks (cf. Figure 5) intersected with metafunctions, which take into consideration that language functions relate in a way that each systemic choice leads to another and produces a certain systemic configuration. Considering that SFL is a theory of language in context, these systemic configurations can be explained in terms of meaning in context.

In terms of instantiation, context "extends along the cline of instantiation from the potential pole (context of culture) to the instance pole (context of situation) via the intermediate region of subpotential/instance type (institution/situation type)" (Matthiessen et al, 2010, p. 77).

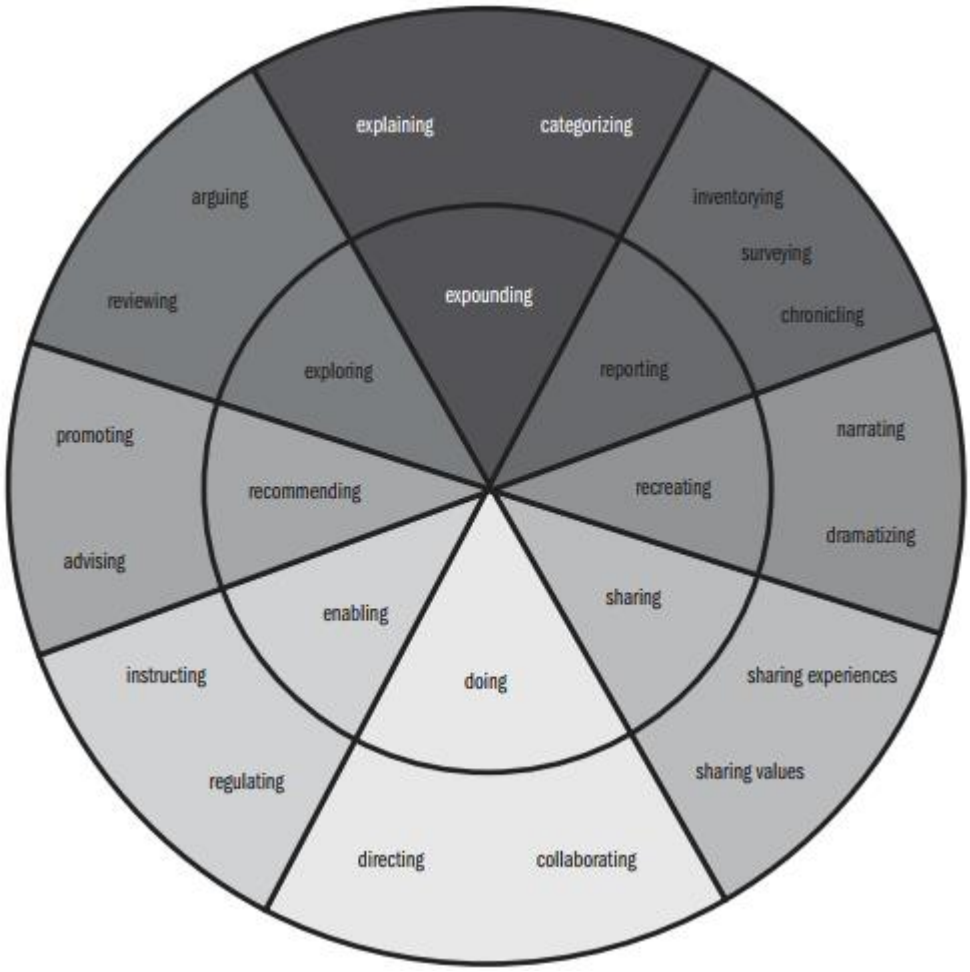
CONTEXT OF CULTURE is associated with all cultural elements which pertain to the meaning production and CONTEXT OF SITUATION is related to a specific situation in which a text is produced (Halliday & Matthiessen, 2014). CONTEXT OF SITUATION can be explained through the variables FIELD, TENOR, and MODE, which are associated with the IDEATIONAL, INTERPERSONAL, and TEXTUAL metafunctions, respectively. Consequently, a text, which is

produced in a certain context and situation, is “language functioning in context” (Halliday & Matthiessen, 2014, p. 3).

Text, according to SFL, can be semantically described as the result of simultaneous co-selections within grammatical systems (Halliday & Matthiessen, 2014). Texts can be grouped under socio-semiotic activities. Eight activities, and their subtypes, represent human activities that involve language, shown in the central and the peripheral concentric circles of Figure 6 and described in Table 6, respectively.

Figure 6

Socio-semiotic activities represented as a topology with each function explained



Note: Retrieved from Halliday & Matthiessen (2014, p. 37)

Table 6

Socio-semiotic activities and their functions

Socio-semiotic activity	Function
Doing	Representing some social behavior, using gestures, gazes, and/or facial expressions
Expounding	Expounding world knowledge, by categorizing or explaining them
Reporting	Reporting facts or phenomena, by inventorying, surveying, or chronicling
Recreating	Recreating human life events, by dramatizing or narration
Sharing	Sharing personal values or experiences
Enabling	Enabling some course of action by instructing or regulating people's behavior
Recommending	Recommending some activity, either for the promotion of a product or a service or giving advice
Exploring	Exploring values of society, by arguing or reviewing

Note: Adapted from Halliday & Matthiessen (2014, pp. 35-36)

In Figure 6, the section associated with the expounding activity and its subtypes "explaining" and "categorizing" is focused on this thesis. The function of each main socio-semiotic activity from Figure 6 is described in Table 6.

This work investigates aligned text segments pertaining to the "cultural domain, or institution" (Halliday & Matthiessen, 2014, p. 33) science, associated with the "expounding" and "reporting" activities, described and highlighted in Figure 6. The text segments are mainly associated with the "expounding" activity on the account of their purpose being defining or classifying scientific concepts, as shown in the examples in Table 7.

Table 7

Corpus examples classified according to function within the expounding socio-semiotic activity

Pair	Nature	Set	Text segment	Function
053	Manually constructed	Set B	O daltonismo ou discromatopsia pode ser dividido em três tipos: Monocromacias, Dicromacias e Tricromacias Anómalas. *Daltonism or Dyschromatopsia can be classified into three groups: Monochromacy, Dichromacy, and Anomalous Trichromacy.*	Classifying

052	Manually constructed	Set B	O daltonismo é uma deficiência que na maioria dos casos é congênita. *Daltonism is a disability that is congenital in most cases.*	Defining

Note: Author's data. Examples retrieved from thesis' corpus.

Table 7 organizes two examples from the corpus across the columns "Pair" (the number of the segment pair), "Nature" (the text type in terms of naturally occurring or manually constructed), "Set" (the set each text belongs to), "Text Segment" (the content of the segment), and "Function" (one of the two types of "expounding" activity: classifying or defining). These examples were retrieved from a segment pair (column Pair), they were either Naturally occurring (retrieved from the web) or Manually constructed (written on the basis of a naturally occurring segment) and were classified as a subtype of the socio-semiotic activity "expounding". A translation into English is provided in italics.

Texts can be segmented according to the rank scale shown in Figure 5, namely (for English and Portuguese) as morphemes, words, groups/phrases, and clauses, which are compositional. In other words, to explain one rank, it is necessary to analyze the rank below and above as well (Halliday & Matthiessen, 2014**).** Also, when the clause is taken as the unit of analysis all metafunctions converge into "an integrated grammatical structure" (Halliday & Matthiessen, 2014, p. 10), the clause is the "point of origin" of several systems (Halliday & Matthiessen, 2014, p. 49), which can be subjected to a lexicogrammatical analysis. However, to

perform an analysis, it is often necessary to investigate agnate forms of a text segment. In other words, the same meaning can be realized as a clause or as a nominal group; therefore, the need to examine meanings at both rank levels.

Concerning the investigation of IDEATIONAL grammatical, Ravelli (1999) states that:

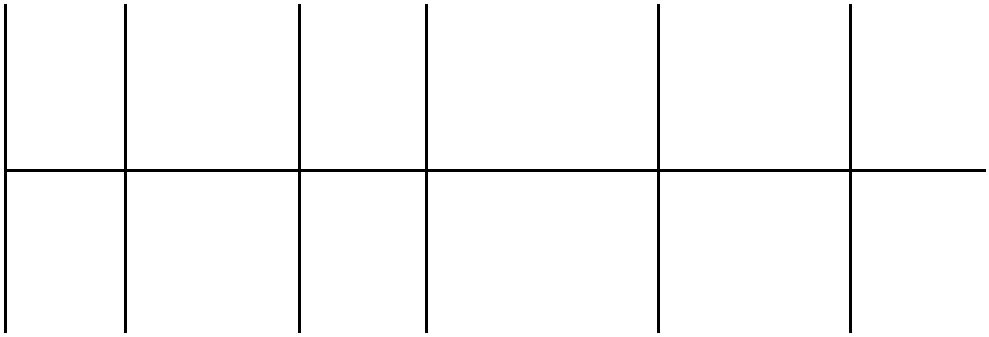
> Grammatical metaphor cannot be a feature at the rank of the clause, because although the entire clause may be metaphorical, often only parts of a clause are metaphorical. Thus, the grammatical metaphor would appear to be a feature at the rank of group/phrase – the constituents of the clause. Yet it is not the case that groups – such as nominal groups, for example – may be realized metaphorically: the group is the metaphorical realization of something else. (p. 99)

According to Ravelli (1999, p. 99), "it is extremely difficult to capture any descriptive generalisations about grammatical metaphor at the level of lexicogrammar". For this reason, it is necessary to define what exactly this concept represents and to approach it as a phenomenon that takes place between ranks and the lexicogrammatical and semantic strata. Thus, in Table 8, which presents the location of this research in the function-rank matrix, both the rank of the clause and the group are in bold.

Table 8

Location of this thesis' research topic in the function-rank matrix

Stratum	**Rank**	**Class**	**Ideational**		**Interpersonal**	**Textual**
			Logical	**Experiential**		
Lexicogrammar	clause			TRANSITIVITY	MOOD	THEME
	group	nominal				
		verbal				
		adverbial				
		prepositional phrase				



Note. Adapted from the function-rank matrix in Halliday & Matthiessen (2014, p.87)

2.3.1 Brief account of systemic descriptions in Brazilian Portuguese.

Table 9 brings a concise account of systemic functional descriptions in Brazilian Portuguese, covering systems related to a range of ranks, namely clauses and groups (nominal, verbal)/ phrases (prepositional), though through realization the word rank is also associated with the group/phrase.

Table 9

Studies describing Brazilian Portuguese from 2007 to 2020

Study	**System or features of a system described**
Araújo (2007)	System of PROJECTION
Figueredo (2007)	Nominal group, mainly the systems of THEME, MOOD, and TRANSITIVITY
Ferreguetti (2014)	Systems of PREDICATION and IDENTIFICATION
Pagano, Ferreguetti e Figueredo (2011)	Relational clauses
Pagano, Ferreguetti e Figueredo (2014)	Existential clauses
Ferreguetti (2014)	Existential clauses
Ferreguetti (2018)	Prepositional phrase functioning as a qualifier in the nominal group
Sá (2016)	Verb and the verbal group
Braga (2016)	CIRCUMSTANCES
Rosa (2017)	System of EXPERIENCE MODIFICATION, taking the verb as the entry condition
Paula (2017)	Verbal clauses
Alves (2017)	System of TRANSITIVITY
Alves (2018)	System of CONJUNCTION
Sá (2020)	System of TIME

As illustrated in Table 9, a range of studies in chronological order describing Brazilian Portuguese drawing on Systemic Functional Linguistics shows that while not as vastly as English, Brazilian Portuguese is currently partially described according to this theory.[17]

There are some differences between the grammatical system of English and Brazilian Portuguese, though. One major difference is that in the literature there is not enough evidence of lexicogrammatical structures realizing behavioral PROCESSES in Brazilian Portuguese (Figueredo, 2014, p. 273). Therefore, this thesis assumes that in the system of Brazilian Portuguese there is no behavioral clause, despite their shared historical roots in Indo-European languages[18].Yet, the description in those studies is by far incomplete compared to the English description. For this reason, this thesis assumed some principles from the English description and investigated them to obtain findings.

The next section describes the criteria for mapping the text complexity phenomenon drawing on SFL.

2.3.2 Criteria for mapping text complexity according to SFL

As previously stated, the text segments in this thesis pertain to the science domain. Halliday (1993, p. 71) identifies seven criteria for discussing difficulties associated with the comprehension and teaching of science literacy: i) interlocking definitions; ii) technical taxonomies; iii) special expressions; iv) lexical density; v) syntactic ambiguity; vi) grammatical metaphor, and vii) semantic discontinuity.

---

[17] Systemic Functional Linguistics was firstly described for Chinese (Halliday, 1956), before its first description for English (Halliday, 1985). Descriptions for other languages, then, partially derive from these, gradually receiving more contributions from the scientific community.

[18] It is beyond the scope of the thesis to do a full systemic grammar of Brazilian Portuguese, though it was necessary to validate the theoretical framework for English to apply it to Brazilian Portuguese, making explicit some key differences, such as the lack of evidence for behavioral clauses in the latter.

Interlocking definitions, which “translate common sense into specialized knowledge” (Halliday & Martin, 1993, p. 229), refer to the fact that some concepts depend on other concepts to be understood. For instance, to understand the diameter of the circle as equivalent to twice its radius, it is required some prior knowledge of what the radius of a circle is. Table 10 presents segment 52 to illustrate this criterion. The concept “daltonismo” (*daltonism*) depends on the meaning of “deficiência” (*deficiency*) to be comprehended, as “daltonismo” is portrayed as a type of deficiency, categorizing it. Yet, Halliday & Martin (1993, p. 80) warn that “[w]riters sometimes try to make the task simpler by adding further definitions, not realizing that in a construct of this kind the greater the number of things defined the harder it becomes to understand.”

Table 10 shows an example of interlocking definitions.

Table 10

Interlocking definitions example

Pair	Nature	Set	Text segment
052	Naturally occurring	Set A	O **daltonismo** é uma deficiência que na maioria dos casos é congênita. *Daltonism is a disability that is congenital in most cases.*

Note: Author's data. Example retrieved from the corpus, specifically from segment pair 52.

Technical taxonomies are "highly ordered constructions in which every term [e.g., concept] has a definite functional value", from which concepts derive their meaning (Halliday & Martin, 1993, p. 81). In Table 11, which presents the naturally occurring segment 53 as an example, the concepts in italics "monocromacias", "dicromacias" e "tricromacias Anómalas" are defined deriving on the concept "daltonismo" (*daltonism*), also called "discromatopsia" (*dyschromatopsia*).

Table 11 presents segment 53 to illustrate this criterion.

Table 11

Technical taxonomy example

Pair	Nature	Set	Text segment
053	Manually constructed	Set A	O **daltonismo** ou **discromatopsia** pode ser dividido em três tipos: Monocromacias, Dicromacias e Tricromacias Anómalas. *Daltonism or Dichromatopsy can be classified into three groups:*

		Monochromacy, Dichromacy, and Anomalous Trichromacy.

Note: Author's data. Example retrieved from the corpus, specifically from segment pair 53.

In Table 11, the concept "daltonismo" (*daltonism*) is classified into three types, which is realized by the material PROCESS "dividido" (*divided*). Thus, this concept is associated with the other three presented afterward.

Special expressions are specific grammar constructions used in certain fields of study, as in mathematics, to realize meanings. For instance, in the sentence "Your completed table should tell you what happens to the risk of getting lung cancer as smoking increases.", what can be inferred is that the interpretation of the table should inform you about something, in this particular case, how the risk of lung cancer is related to how much someone smokes. In this example, "Your completed table should tell" is considered a more metaphorical special expression than the version with all participants ("the interpretation of the table should inform you that …") because the whole expression has a single meaning. (Halliday, 1993, p. 82).

Lexical density is the measure of the frequency of lexical terms (*content words*) compared to the sum of all words from a text, both lexical and grammatical terms (Sokolova and Bobicev, 2018, p. 4); the higher the lexical density, the more meanings, and consequently the higher is the cognitive effort needed to interpret the text. In contrast, according to Halliday

(1993, p. 76) and Ravelli (2006, p. 55), lexical density is the "measure of the proportion of lexical items in a clause". Both measures can be compared to investigate the efficiency of these metrics, as performed in the exploratory study[19] for this thesis. Thus, this work does not consider the lexical density measurement on its own as an efficient method to discriminate between less and more metaphorical text segments.

Syntactic ambiguity is a grammatical phenomenon caused by the use of certain constructions, especially by nominalizations, in which semantic information that is necessary to comprehend the text is not realized. For example, in the sentence "Lung cancer death rates are clearly associated with increased smoking", this sentence could mean that lung cancer death rates *cause* increased smoking on individuals or that lung cancer death rates *are caused by* increased smoking on individuals. This syntactic ambiguity is caused by the expression "is/are associated with", which can have two meanings: "to cause" or "to be caused by something".

Thus, grammatical metaphor can take place because of a high lexical density and syntactic ambiguity. Grammatical metaphor is defined as the substitution process of a congruent linguistic segment with a non-congruent form, or vice versa, to realize approximate semantic content (Halliday, 1993, p. 87).[20] In other words, grammatical items in congruent form can be replaced by items in non-congruent form to modify the level of metaphoricity. The example in Table 12 shows an increase of metaphoricity comparing the segment from Set A and Set B.

[19] The exploratory study took into account the embedded clauses into the count, but the thesis did not, because the thesis sought to focus on text simplification through experiential grammatical metaphor. This study was presented in 2019 at ESFLC in Leiria, Portugal.

[20] Halliday (1993), Halliday & Martin (1993) often use the terms "non congruent" and "metaphorical" interchangeably.

Table 12

Grammatical metaphor example

Nature	Set	Segment
Manually constructed	Set A	Pessoas vulneráveis podem sofrer sintomas que são causados por **estressores**, seja bioquímicos ou sociais. *Vulnerable people can present symptoms that are caused by stressors, be they biochemical or social.*
Manually constructed	Set B	**Eventos estressantes como a perda de um emprego ou o término de uma relação amorosa** podem causar esses sintomas. *Stressful events such as job loss or the end of a relationship can cause these symptoms.*

Note: Author's data. Examples retrieved from the corpus, specifically from segment pair 71.

In Table 12, the nominal groups "Eventos estressantes como a perda de um emprego ou o término de uma relação amorosa" (*Stressful events such as job loss or the end of a relationship*) and "estressores" (*stressors*) from manually constructed segment pair 071 show that the metaphoricity increased. This shift is due to extra meanings within the first nominal group that

were added to the sentence, namely examples of which events qualify as stressful events: "estressantes como a perda de um emprego ou o término de uma relação amorosa (*"such as job loss or the end of a relationship"*). In contrast, information was omitted from the other segment; in this case who suffers the symptoms and the types of stressful factors (*estressores*).

Finally, semantic discontinuity is the resource used by some writers to realize meanings without providing all the essential information, demanding that the reader infer the missing information. The pairs provided in Table 13 show that the use of the verbal group "have resulted" in the original segment allows the omission of some information, specifically why the laws resulted in cleaner factories and cleaner countryside and why the light-colored moths increased in number.

Table 13

Semantic discontinuity example

Original segment	Rewritten segment
However, strong anti-pollution laws over the last twenty years have resulted in cleaner factories, cleaner countryside and an increase in the number of light-coloured pepper moths.	Over the last twenty years, **[the government have passed]** strong laws to stop **[people]** polluting; so the factories **[have become]** cleaner…

Note: Adapted from Halliday (1993, p. 91)

Comparing the original and rewritten segments, the extra information is highlighted in Table 13 in bold in the other segment since it was made explicit in the rewritten segment. These shifts increase the metaphoricity of the rewritten segment.

This segment from Table 13 was retrieved from the text shown below, also retrieved from Halliday (1993, p. 91):

> In the years since 1850, more and more factories were built in northern England. The soot from the factory smokestacks gradually blackened the light-coloured stones and tree trunks.
>
> Scientists continued to study the pepper moth during this time. They noticed the dark-coloured moth was becoming more common. By 1950, the dark moths were much more common than the light-coloured ones.
>
> However, strong anti-pollution laws over the last twenty years have resulted in cleaner factories, cleaner countryside and an increase in the number of light-coloured pepper moths.

Taking into consideration the segment in Table 13, the law of natural selection is briefly described in terms of cause and effect, without mentioning the PARTICIPANTS of the PROCESSES involved. In other words, since the anti-pollution laws are not mentioned in the first two paragraphs of the text, it is a piece of new information presented metaphorically, regarding experiential grammatical metaphor. This new information is the cause of "cleaner factories, cleaner countryside and an increase in the number of light-coloured pepper moths", as the PROCESS "have resulted" shows. This indicates that these laws avoided pollution in the factories and the countryside, but *not* in moths (as they are animals, not places). This is where the semantic discontinuity takes place: the fact that these environments became cleaner due to the anti-pollution laws have activated some natural mechanism that allows light-colored moths to increase in number, a classic example of natural selection.

Technical taxonomies and grammatical metaphor (explained in further detail in the next section) are also directly associated with the concept of technicality. According to Halliday & Martin (1993), technicality is defined as a "field-creating process", which enables the distillation (or condensation) of meanings in a text. In other words, with the text development, meanings realized in clauses accumulate by the use of nominalizations that reinstate information that was previously mentioned. Thus, technicality allows texts to become more compact and the nominalization of PROCESSES enables them to be taxonomized for the development of further explanations (p. 182). In the case of science texts, concerned with explaining phenomena, technicality is highly present, since "the more a field is concerned with explaining phenomena rather than just ordering them the greater the distillation offered by technicality" (p. 182).

Table 13 shows that nominal groups with participants being realized by nominalized clauses can be realized as simple nominal groups (without embedding) in a non-congruent form. The excerpts in bold show the equivalence between the first nominal group from each segment in congruent form and non-congruent form. This can be observed in Figure 10 through the examples in the ideational grammatical metaphor continuum[21].

The next section describes the literature on grammatical metaphor, describing this concept in detail, followed by a review of the literature on text simplification integrating both the computational and linguistics perspectives.

[21] The interpersonal grammatical metaphor, more commonly found in spoken texts, is not analyzed in this thesis, since the main objective of this study is to investigate the phenomenon of text simplification in written texts.

2.3.3 Grammatical metaphor

This subsection explores in further detail the literature on the topic of grammatical metaphor drawing on SFL, starting with its definition, explaining each type, and integrating this concept with text simplification.

2.3.3.1 Definition of grammatical metaphor

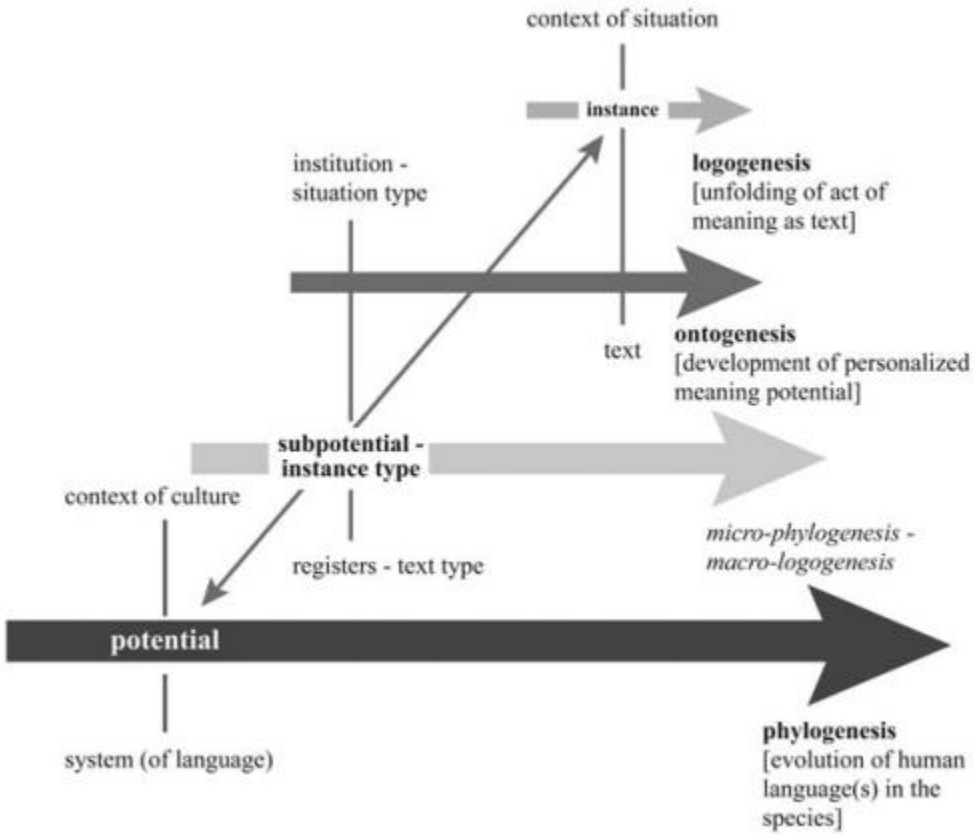
Grammatical metaphor is defined as the substitution process of a congruent linguistic segment with a non-congruent form, or vice versa, to realize approximate semantic content (Halliday, 1993, p. 87).[22] In this thesis, a congruent pattern is defined as "a regular pattern of relationship between the grammatical categories (clause, verbal group, etc.) and the semantic categories (FIGURE, PROCESS, etc.)" (Halliday & Martin, 1993, p. 188). In contrast, any pattern that does not follow this is regarded as non-congruent. Due to this grammatical mapping described by grammatical metaphor, it allows construing extra layers of meaning through different wordings (Halliday & Matthiessen, 2014, p. 699).

Grammatical metaphor can be approached based on the perspectives of logogenesis, ontogenesis, and phylogenesis (Webster, 2004, p. 30). These perspectives can be analyzed in conjunction with semogenesis, presented in Figure 7, associated with the cline of instantiation on the left.

Figure 7

Integration between semogenesis and the cline of instantiation

[22] Halliday (1993), Halliday & Martin (1993) often use the terms "non congruent" and "metaphorical" interchangeably.



Note: Reprinted from Matthiessen et al. (2010, p. 198)

Figure 7 presents the continuum between the poles of language potential, all the possibilities enabled by the linguistic system, and the linguistic instance, the configuration realized by the system in a text. In between, the sub-potential or instance type divides the potential into subtypes. The CONTEXT OF CULTURE is associated with the potential and CONTEXT OF SITUATION to the instance, given their degree of specificity within this continuum, and the institution or situation time is in the intermediate point, acting like a fork in which the potential is being classified. The concepts of logogenesis, ontogenesis, and phylogenesis are also presented along the continuum, from the instance pole to the potential pole, in association with, respectively, the unfolding of an act of meaning in a text, the development of personalized meaning potential, and the evolution of human language(s) in the species. These concepts can be

taken into account in combination with the umbrella term “semogenesis” to enable the conceptualization and study of various phenomena.

The logogenetic perspective represents “the creation of meaning as acts of meaning in the instantiation of the meaning potential in the course of the unfolding of text.” (Matthiessen et al, 2010, p. 151). In other words, while the text unfolds, meaning is creation through instantiation, from the potential to the instance.

The ontogenetic perspective represents “[t]he three phases of language development—protolanguage, the ‘child tongue’ before the child starts learning the mother tongue (phase I), the transition into adult language (phase II), and the period of learning adult language (phase III).” (Matthiessen et al, 2010, p. 133).

The phylogenetic perspective concerns “the history of the system in the species”; in other words, “the time frame of evolution in the species or a social group” (Matthiessen et al, 2010, p. 197). This concept is in contrast with ontogenesis and logogenesis in a way that the first is “the time frame of development in the individual” and the latter “the time frame of the unfolding of a text” (Matthiessen et al, 2010, p. 197). This way, each of these perspectives complement each other, allowing a meaning-creating process called semogenesis to take place.

The semogenetic perspective is a “[m]eaning-creating process, within different time frames: phylogenesis (the time frame of evolution in the species or a social group), ontogenesis (the time frame of development in the individual) and logogenesis (the time frame of the unfolding of a text)” (Matthiessen et al, 2010, p. 197). In sum, this is the process described in Figure 7.

2.3.3.2 Types of grammatical metaphor

According to Matthiessen (1995), Halliday & Matthiessen (1999), and Halliday and Matthiessen (2014), there are two kinds of grammatical metaphor, associated with their respective metafunction: interpersonal grammatical metaphors and ideational grammatical metaphors. Martin (1992) includes a third type, the textual grammatical metaphor, which is discussed briefly, though this type is not thoroughly analyzed in this thesis because it is not within the scope of this study.

2.3.3.2.1 Textual grammatical metaphor

On the use of textual resources for producing the same meaning in distinct manners, Martin (1992) points out that:

> discourse systems can be used to construe text as 'material' social reality. […] similarly, text reference [e.g., this; LR] identifies facts, not participants, and internal conjunction [e.g., finally, LR] orchestrates textual not activity sequences. (p. 416)

Table 14 shows an example of textual grammatical metaphor.

Table 14

Example of textual grammatical metaphor

Types of metaphorical realizations	Example 1	Example 2
Conjunctive relations	I think Governments are necessary at different levels for **a number of reasons**.	**For example**, they make laws, without which people would be killing themselves, and help keep our economic system in order.
Text reference	That **point** is just silly	**That's** ridiculous

Note. Examples retrieved from Martin (1992, pp. 416-417).

In Table 14, through conjunctive relations, a nominal group in a prepositional phrase is realized as a conjunction, and the text reference is not the same anymore, resulting in a different ideational metaphoricity degree.

By assuming that textual resources can also present congruent and metaphorical realizations, Martin's (1992) contrasts with most researchers. Therefore, this type of grammatical metaphor is not taken into consideration in this study.

2.3.3.2.2 Interpersonal grammatical metaphor

As interpersonal metaphors are associated with the interaction between speakers, they can be classified as metaphors of modality and metaphors of mood, in which the contrast between the congruent and the metaphorical version are, respectively, due to the choice of a different modality type (e.g., probability) or mood type (e.g., converting imperative into declarative mood).

Examples of metaphors of modality, as realizations of expressions of probability, are presented in Table 15.

Table 15

Examples of the metaphor of modality - expressions of probability

Category		Type of realization	Example
(1) subjective	(a) explicit	I think, I'm certain	I think Mary knows
	(b) implicit	will, must	Mary'll know

(2) objective	(a) implicit	probably, certainly	Mary probably knows
	(b) explicit	it's likely, it's certain	it's likely Mary knows

Note: Reprinted from Halliday & Matthiessen (2014, p. 689)

Table 15 presents some examples of metaphorical realizations of probability, such as 1a, subjective and explicit, realized by "I think" in "I think Mary knows". This can be contrasted with "Mary knows", which presents 100% certainty, not less than 100% in the example.

Examples of mood metaphors are shown in Table 16, contrasting the congruent and the metaphorical meanings of the examples provided.

Table 16

Examples of the metaphor of mood

Congruent meaning	Metaphorical meaning	Congruent example	Metaphorical example
(1) command	(a) warning	*Don't* buy a car.	*I wouldn't* buy a car *if I were you*.

	(b) advice	*Maybe I'll* buy a car.	*I've a good mind to* buy a car.
(2) offer	(a) threat	*She should* buy a car.	*She'd better* buy a car.

Note: Adapted from Halliday (1985, p. 365)

In Table 16, examples of metaphors of mood are presented. For instance, in 2a, "She'd better buy a car." is the metaphorical realization of "She should buy a car." by the use of a threat. The difference is the use of the expression "she'd better" instead of the modal verb "should", which is more congruent.

2.3.3.2.3 Interpersonal vs ideational grammatical metaphor

The main difference between interpersonal and ideational grammatical metaphor concerning the context in which they take place is the fact that interpersonal grammatical metaphor is found more frequently in "ordinary, spontaneous conversation that children meet in the home and neighbourhood" (p. 709), while ideational grammatical metaphor associates with education, science, bureaucracy, and law. For this reason, ideational grammatical metaphors are realized differently, and, considering the nature of the texts from the corpus, they are the focus of this thesis.

2.3.3.2.4 Ideational grammatical metaphor

Based on the perspectives of logogenesis, phylogenesis, ontogenesis and semogenesis described in section 2.3.3.1, Table 17 compares the realization of the meanings in two or more equivalent segments to present ideational grammatical metaphor in contrast with interpersonal metaphor.

Table 17

Representation of ideational grammatical metaphor through the comparison of congruent and non-congruent examples

Congruent form	Non-congruent form
The rates of people dying after they develop lung cancer clearly increase as the number of people smoking increases.	**Lung cancer death rates** are clearly associated with increased smoking.
The way the creature behaves when it feeds shows that it has become more responsive.	Increased responsiveness may be reflected in **feeding behaviour.**
More goods cannot be produced unless	**Higher productivity** means more

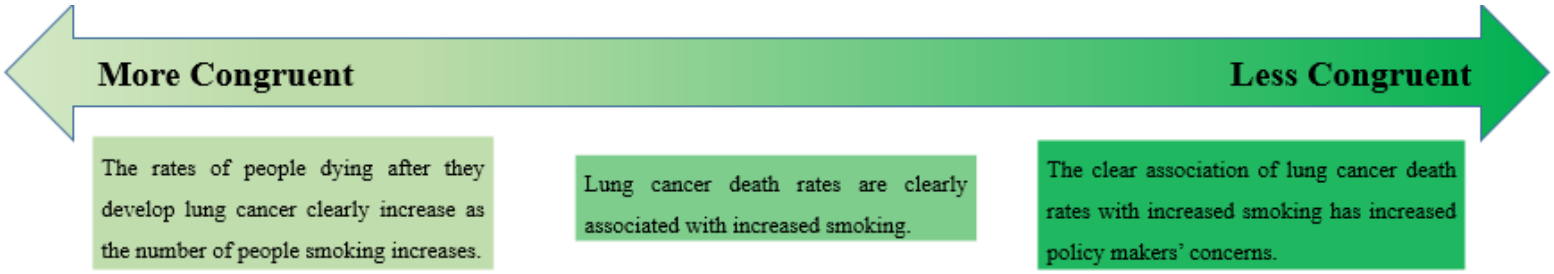
more supporting services are provided.	supporting services.

Note: The examples were retrieved from Halliday & Matthiessen (1999, p. 231, emphasis added). Emphasis was added to the first nominal group of each segment in the congruent (non-metaphorical) form and the equivalent in non-congruent (metaphorical) form.

According to Halliday (2004, p. 36), a "clausal variant" is more congruent, as opposed to a "nominal variant", which is more metaphorical. Three reasons are presented for that. Firstly, the history of language development (phylogenetic perspective) shows that a clausal variant is developed first in the system of the language, in the language user's life (the child) (ontogenetic perspective), and in a text progression (logogenetic perspective). Secondly, it was in the clausal variant (or congruent mode) that "human language first came into being; hence it determines our collective categorisation of the world we live in" (Halliday, 2004, p. 37). Lastly, a text typically starts with clausal constructions, which can be observed in the choice of THEME, and the ideas develop as nominal constructions in a thematic position.

Table 17 shows that nominal groups formed by nominalized clauses can be realized as simple nominal groups (without embedding) in a non-congruent form. The emphasis added shows the equivalence between the first nominal group from each segment in congruent form and in non-congruent form. This is illustrated by the examples in the ideational grammatical metaphor continuum in Figure 8.

Figure 8



Ideational grammatical metaphor continuum

Note: Adapted from Silva (2012, p. 41). The examples were retrieved from Halliday & Matthiessen (1999, p. 231) and Silva (2012, p. 43).

The ideational grammatical metaphor continuum in Figure 8 shows three sentences, classified between the poles "more congruent" (less metaphorical) and "less congruent" (more metaphorical). In the most congruent segment, the participant "the rates of people dying after they develop lung cancer", which is also the THEME of this clause, is realized as the PARTICIPANT and THEME "lung cancer death rates" in the second example. Consequently, the second example is more metaphorical than the first because the first PARTICIPANT is a clause in the first example and a nominal group in the second, indicating nominalization. Also, a new meaning was added to the clause, realized by the RHEME of the clause: "are clearly associated with increased smoking". In the last example, the PARTICIPANT and THEME "the clear association of lung cancer death rates with increased smoking" realizes the meanings of the second sentence as a whole. Besides, it adds a new meaning, realized by the RHEME "has increased policy makers' concern". In conclusion, the second example is less congruent (more metaphorical) than the first and more congruent (less metaphorical) than the third because the clause from the second example is equivalent to the first PARTICIPANT from the third example. The text development associated with

the choice of another THEME type presented in the segments shows that meanings were accumulated from the first example to the third, i.e., they were condensed to include new ideas.

This accumulation of content can be mapped through ideational metaphors. Halliday & Matthiessen (2014) define ideational metaphor as

> a 're-mapping' between sequences, figures, and elements in the semantics and clause nexuses, clauses, and groups in the grammar. In the congruent mode of realizations [...], a sequence is realized by a clause nexus and a figure is realized by a clause. In the metaphorical mode, the whole set of mappings seems to be shifted 'downwards': a sequence is realized by a clause, a figure is realized by a group, and an element is realized by a word. (pp. 713-714)

This ‘re-mapping’ allows the remapping between the congruent mode and the metaphorical mode or vice versa since the ideational metaphor is associated with patterns in the realization mode. These patterns are expanded in certain text types in which ideational metaphor takes place systematically, such as in scientific discourse (Halliday & Matthiessen (2014, p. 713).

As stated by Halliday and Matthiessen (1999, p. 235), considering a set of “metaphorically agnate wordings”, through phylogenetic, logogenetic, and ontogenetic analysis, it is possible to investigate whether they stand in a relationship involving a different metaphoricity from one of the instances. In other words, taking into account a set such as “the announcement [was made] of his probable resignation” and “he announced that he would probably resign” (segment 1 and segment 2, respectively, for comparison’s purpose), initially one could not assume which is more metaphorical and which is more congruent. However, in

terms of the history of the language (phylogenesis), children learning processes (ontogenesis) and textual organization (logogenesis), segment 1 precedes segment 2. Therefore, segment 1 is considered more congruent (or less metaphorical) than segment 2. A similar example can be observed in Figure 4.

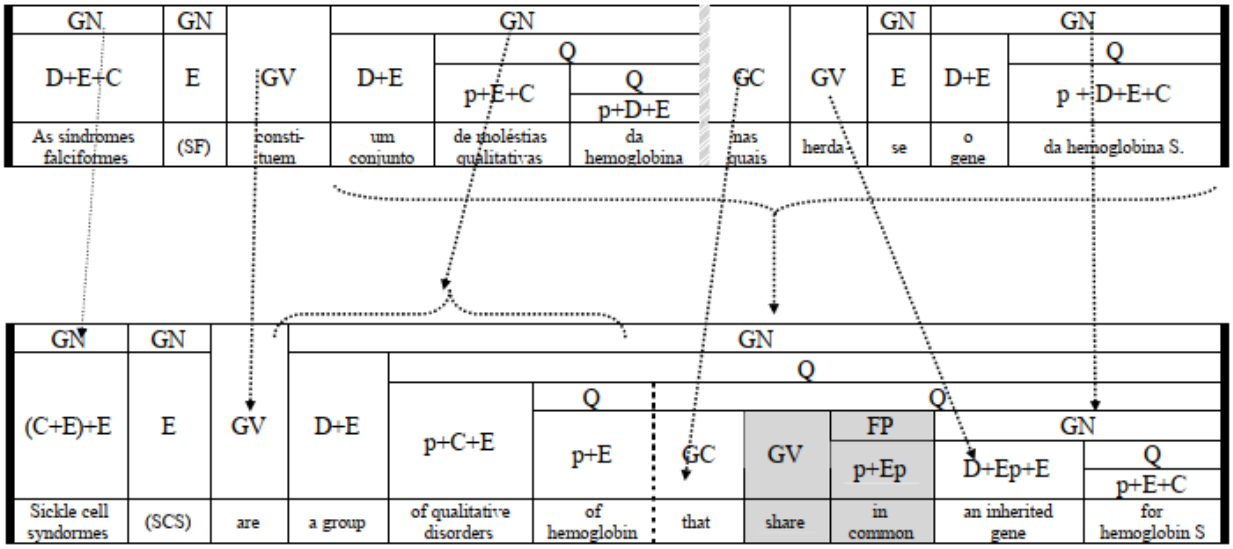
2.3.3.3 Further works exploring ideational grammatical metaphor

Drawing on Halliday & Matthiessen's (2014) concept of ideational metaphor to perform an empirical study to model peak performance in translation, Silva (2007) focused on ideational linguistic categories from the source texts and target texts. Subjects (four medicine expert researchers) carried out two translation tasks, which were monitored and recorded by a keylogging software (Translog©). Silva (2007) approached data according to three different viewpoints: i) an interface between translation studies and expertise and expert knowledge studies ii) an analysis of the rhetorical structure in source and target texts drawing on the Rhetorical Structure Theory (RST); iii) and an investigation of the subjects' (de)metaphorization processes between source and target texts during translation process.

Silva's (2007) third approach made use of an analysis of the metaphoricity of segments in the source text (in Brazilian Portuguese) and the target text (in English). Figure 9 introduces an example of this analysis, comparing two segments focusing on the definition of sickle cell syndromes.

Figure 9

Experiential analysis of an introduction on sickle cell disease on the group rank



Note. Adapted from Silva (2007). GC = Conjunctive group; GN = Nominal group; GV = Verbal group; C = Classifier; D = Deictic; E = Thing; Ep = Epithet; Q = Qualifier; p = preposition.

In the example provided in Figure 9, the verbal group “herda” (*to inherit*) from the source text was translated as an epithet from the nominal group “an inherited gene”. On the one hand, the analysis in Figure 2 shows that the target text is more metaphorical than the source text. To condense this information, the author of the target text deemed necessary to include extra information, by translating “herda” (*inherits*) as “share in common”, even though the meaning realized by the prepositional phrase “in common” was not present in the source text. This simplification strategy is also observed in Silva (2007, p. 206).

Thus, ideational metaphor plays an important role in scientific discourse (Halliday & Matthiessen (2014, p. 713). For this reason, this thesis analyzes samples retrieved from a segment corpus composed of pairs of segments retrieved from science texts.

Other studies focusing on grammatical metaphor are Bateman (1990) and Byrnes (2009). Bateman (1990) is mostly related to the field of computer science and Byrnes (2009) to linguistics, and, drawing on SFL, these studies suggest different applications of this concept.

Bateman (1990) approaches SFL to find a linguistic theory that can help identify translation equivalents for the task of machine translation, which began its development without the necessary support of linguistic theories to model language properly. Two types of grammatical metaphor explored in this paper are the rank shifting and the complexity metaphors, which can assist in the identification of congruent and non-congruent segments to be considered translation equivalents for the machine translation task (p. 14). This way, instead of selecting the most frequent ('topmost') nodes for the translation equivalents, the most appropriate expression in each case can be chosen from a set of possible expressions. Even some distinctions across languages can be described through the rank shifting and the complexity metaphors (p. 15), such as in the comparison of the English-German versions of a sentence provided in the paper, that is, the translation of "Mary cut her finger." as "Mary schnitt sich in den Finger". For this task, two monolingual grammars and lexicons work separately, dealing especially with the unique features of each language, as well as with the ranks and pairing of the congruent- non-congruent pairs of segments. Besides, with the use of grammatical metaphor, nominal groups and the modality in English and in German can be compared and the adequate translation equivalents selected, considering the linguistic differences between both languages. In sum, on the basis of the

linguistic representation of language drawing on the concept of grammatical metaphor from SFL, it was possible to design computational representations of the translation equivalents, to be selected by the algorithm for performing machine translation between English and German.

Byrnes (2009) investigates the German writing process development of 14 college students enrolled in levels 2, 3, and 4 from a curriculum project called "Developing Multiple Literacies"" between 1997 and 2000, aiming at enabling them to develop advanced writing skills in German. This project was based on the assessment of their texts according to the concept of grammatical metaphor, drawing on SFL (p. 3) as the theoretical framework. Data coding draws on Derewianka (2003) and Steiner (2002) to track three potential indicators of grammatical metaphor: derivations, rank-shifting, and agnations. These indicators were observed in sets of congruent and non-congruent realizations, as well as in similar lexical items. Also, all de-adjectival and de-verbal nominalizations functioning as grammatical metaphors were annotated. The results showed a relation between the grammar of the clause and the clause complex, by measuring lexical density, grammatical intricacy, and an expected increase in the rate of grammatical metaphors associated with the level of the students. This increase was shown through several measurements associated with the nominalizations, such as the number of tokens, number of clauses, number of "normal" (common) nouns, and grammatical metaphors. In sum, this study allowed to shed light on "the emergent capacity for meaning-making in FL writing by adult learners" (Byrnes, 2009, p. 3).

The next section shows the relationship between grammatical metaphor, text complexity, and text simplification.

2.3.4 Grammatical metaphor, text complexity, and text simplification

According to Halliday and Matthiessen (1999, p. 258), “the overall effect of the grammatical metaphor is that semantic relations between one element and another, and between one figure and another, become progressively less explicit as the degree of metaphoricity increases”

This can be illustrated by comparing the more congruent and more metaphorical variants of the following examples (Halliday and Matthiessen, 1999, pp. 258-259):

[i] most congruent (less metaphorical)

Glass cracks more quickly the harder **you** press on it.

[ii] less congruent (more metaphorical)

Cracks in glass grow faster the more pressure is put on. (pp. 258-259)

These examples show that by the agnation of the more congruent (or less metaphorical) example into a more metaphorical (or less congruent) form, some semantic information is lost in the process. In this case, the semantic information is the PARTICIPANT “you” from the clause “the harder you press on it”, which is implicit in the clause “the more pressure is put on”.

Furthermore, by analyzing the notion of ‘explicitness’, Steiner (2005, pp. 21-23) investigated and corroborated five hypotheses on informational density and grammatical metaphoricity, presented in Table 18:

Table 18

Five hypotheses on informational density and grammatical metaphoricity

(H1) The more informationally dense and the more grammatically metaphorical a stretch of text, the less it will be explicit grammatically (and cohesively).
(H2) The more informationally dense and the more grammatically metaphorical a stretch of text, the more the explicit grammatical and cohesive marking will be of the general nominal, rather than verbal type.
(H3) The more informationally dense and the more grammatically metaphorical a stretch of text, the higher the proportion of 'intermediate phrase types' (groups, phrases, rather than words or clauses) per clause.
(H4) The more informationally dense and the more grammatically metaphorical a stretch of text, the higher the proportion of phrases with a nominal head relative to phrases with a verbal head per clause.
(H5) The more informationally dense and the more grammatically metaphorical a stretch of text, the higher the number of grammatical features per unit.

Note: Reprinted from Steiner (2005, p. 21-23)

From these hypotheses in Table 18, H1 and H2 associate the degree of grammatical metaphor and the level of explicitness in a text. Consequently, these findings corroborate the criteria used for the creation of the manually constructed segments, which were initially based on text simplification techniques. Also, these findings confirm that the criteria for decreasing text complexity (e.g., making a text simpler) and agnating between congruent and metaphorical wordings can be contrasted successfully. Finally, they also allow other comparisons relevant to this thesis, as the measurement of grammatical density[23] and grammatical intricacy.

The next section reviews how text simplification studies can be approached through the linguistic perspective.

2.4 A linguistic view of text simplification

Drawing on SFL, this thesis approaches text simplification by assuming that a segment is simpler in comparison to another one, if, upon performing operations to render the text comprehensible by a non-specialist audience, the segments present distinct configurations from the perspective of linguistics, specifically SFL framework (Halliday & Martin, 1993). This view takes into consideration the assumption that in order to be more comprehensible, a text should be less complex in terms of its level of grammatical metaphoricity.

This thesis reviews NLP's approach to text simplification, which regards it as a type of paraphrase (which is a type of entailed text), obtained through a “process of reducing the

[23] Grammatical density is the density of grammatical items, in the same manner as the lexical density is the density of lexical items in a sentence or clause, depending on the author.

linguistic complexity of a text, while still retaining the original information content and meaning" (Siddharthan, 2003, p. 3).

According to Shardlow (2014, p. 58), both text readability and understandability are increased when the complexity of a text is reduced. Shardlow states that readability for a given audience is associated with features such as grammar complexity, sentence length, and knowledge of the text's subject. Understandability is the amount of information a text can provide to the user.

Ravelli (1996) defines an "accessible" text as:

> a text which does not presume a high level of reading knowledge (as in, say, a very academic textbook); which does not compromise the scientific integrity of the information needing to be conveyed; and which functions successfully as text, that is, as a cohesive and a coherent unit. (p. 371)

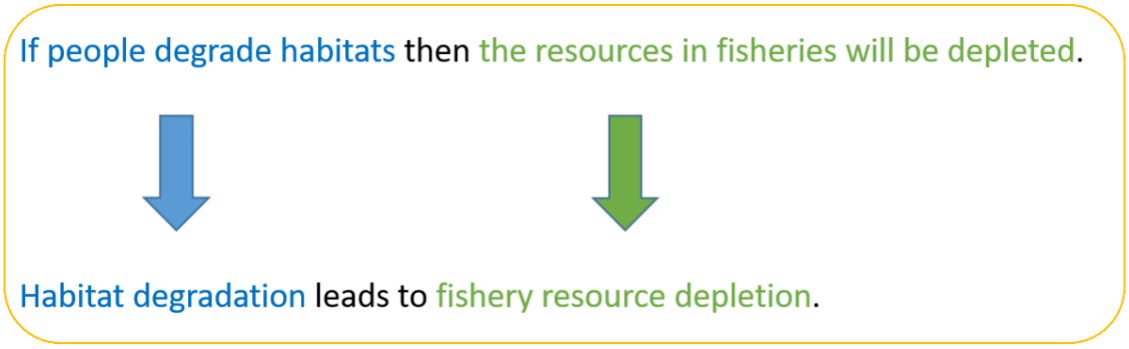
This definition of accessibility assumes that three conditions must be fulfilled for a text to be considered accessible. By comparing two or more texts and regarding one of them as more "accessible", Ravelli means that this text is less complex.

According to Ravelli (2006), one of the operations to increase complexity is the use of nominalization. Nominalization may increase text complexity because nominalization is "one of the key distinguishing features of written language", condensing meanings and providing extra information (Ravelli, 2006, p. 61). Drawing on Ravelli (1999, p. 58), Figure 10 illustrates the use of nominalization as an attempt to produce the more metaphorical clause "Habitat degradation

leads to fishery resource depletion" based on the clause complex "If people degrade habitats then the resources in fisheries will be depleted".

Figure 10

Example of nominalization process



Note: The example in this figure was retrieved from Ravelli (1999, p. 58)

The first example in Figure 10 shows a clause complex in which two clauses in a hypotactic relation are divided by a conjunction "then", being the second clause a consequence of the first. In the second example (below the arrows), there is only one sentence, but each PARTICIPANT is a nominalized form of a clause from the first example.

In the example, the excerpt "if people degrade habitats" is organized as a hypotactic clause with the conjunction ("if"), subject ("people"), process ("degrade"), and goal ("habitats"). In this excerpt, "people degrade habitats" was nominalized as the nominal group "habitat degradation", formed by CLASSIFIER and THING, and the clause "the resources in fisheries will be depleted" was nominalized as "fishery resource depletion", a nominal group with two CLASSIFIERS ("fishery" and "resource") and a THING ("depletion") (Ravelli, 2006, p. 58). Consequently, two clauses in the first example were nominalized in comparison with the second,

increasing the degree of metaphoricity. The nominalization process condenses the meanings of the clauses, as the two material clauses are realized as a relational clause indicating causal relation between the meanings realized in each nominal group. Thus, according to Ravelli (1999), nominalization is a special type of class shift and is related to two types of complexities – lexical and grammatical complexity – which can describe how compact or intricate is the "ideational information". These complexity measures can be obtained through lexical density and grammatical intricacy and can be associated with the metaphoricity level of texts and segments.

Ravelli (1996, p. 371) characterizes the less metaphorical version as "accessible" since it is considered a text that requires lower expertise to be comprehended without any loss of information. Ravelli's contention sustains this thesis proposal and the potential of its findings to contribute to applications within an Appliable Linguistics framework (Halliday, 2007), such as guidelines for text simplification in Brazilian Portuguese – which can be used to develop text simplification or text summarization systems.

2.5 Computational approach for text simplification contrasted to SFL approach for metaphoricity degree

Notably, there are differences between the computational approach, that proposes text simplification strategies to produce more accessible texts, and the linguistics approach, proposing criteria for producing texts with varying degrees of metaphoricity, which can be more accessible if metaphoricity decreases. Even though their view of what an accessible text is differs, the output may be the same, taking into account that the computational approach disregards grammatical metaphor. This distinction is explored in this section.

According to Ravelli (1999) and Halliday & Matthiessen (2014), the main ways of identifying grammatical metaphors are through derivation and agnation. For instance, taking into consideration two sentences: “The staff wanted to determine the cause of the accident at the dam.” and “The determination of the cause of the accident at the dam depended on the staff.”. In this case, if “determine” (a verb, that realizes a PROCESS) and “determination” (a noun, that is, part of a nominal group) are used in the same context (with approximately the same meanings), since they derive from the same root “determin”, they are regarded as related and; therefore, the noun “determination” can be a grammatical metaphor of “determine”.

Agnation, which is the main way for identifying grammatical metaphor, can be realized as two types: class shifts and rank shifts. In other words, conversion between a grammatical class and another (verb to noun, mostly) or between certain ranks (clause to noun, mostly). These are to be taken into account for the assessment of the most adequate strategies used in this thesis.

The strategies that were analyzed in the light of SFL for compiling SIM-Pt were from the NLP literature and a simplification manual developed by NILC, in São Paulo, Brazil. These strategies were compared with the SFL approach proposed in this thesis based on metaphoricity criteria in Table 19.

Table 19

Compatibility between computational approaches and the current approach from this thesis based on metaphoricity criteria

Strategy	**Computational approaches**	**Strategy compatible with SFL Approach**
Conversion from passive voice to active voice	✓	×
Addition of explanation or examples	✓	×
Replacement of more technical words with less technical ones (lexical simplification)	✓	✓
Syntactic order modification	✓	✓
Sentence splitting	✓	✓

Sentence elimination	✓	✗
Sentence compression	✓	✗
Dis-embedding of relative clauses	✓	✓
Separation of subordinate clauses	✓	✓

Based on Table 19, which shows which strategies for simplifying texts can accounted for as having a linguistic basis drawing on the grammatical metaphor concept, Table 20 presents possible reasons why the criteria in bold can be relevant for a text analysis following SFL and assuming an association between complexity and grammatical metaphor (Steiner, 2004).

Table 20

Possible reasons for the relevance of some text simplification strategies in the light of SFL concept of grammatical metaphor

Strategy considered useful for text simplification based on varying degrees of metaphoricity degree	**Reason**
Replacement of more technical words with less technical ones (lexical simplification)	The use of less frequent words increases the complexity of a text, which influences grammatical metaphor.
Syntactic order modification	The syntactic order is one of the variables that increase the complexity of a text.
Sentence splitting	The conversion of a clause complex into two or more clauses reduces the complexity, which reduces the level of metaphoricity.
Dis-embedding of relative clauses	By converting embedded clauses into free clauses, the figures are not at the rank of the group anymore, but at the rank of the clause, reducing the level of grammatical metaphor.

Separation of subordinate clauses	The conversion of a clause complex with a hypotactic clause into two or more clauses reduces the complexity, which reduces the level of metaphoricity.

These strategies from Table 20, except for non-simplification, were interpreted drawing on SFL and the results summarized in Table 21.

Table 21

Text simplified strategies interpreted in the light of SFL

Text simplification strategy	Equivalent in SFL	The rank in which the strategy applies	Affected system	Associated with grammatical metaphor
1) splitting sentences	converting a clause complex into two or more simplexes	clause	Logico-Semantic type	yes, because it affects grammatical intricacy and lexical density
2) replacing a discourse marker by a simpler and/or more	replacing a more frequent word with a less frequent one	word	Conjunction	no

frequent one				
3) converting passive into active voice	converting an effective active into an effective receptive verbal group	group	Voice	no
4) inverting clause order	modifying the logical structure of the clause complex, affecting theme	clause	Theme	yes, because it affects the information flow
6) syntactic order modification	modifying the logical structure of the clause, affecting theme	clause and group	Theme	yes, because it affects the information flow

Note: SFL stands for Systemic Functional Linguistics, drawing on Halliday & Matthiessen (2014)

Table 21 presents the SFL interpretation of the text simplification strategies proposed by Specia, Aluísio & Pardo (2008) and is illustrated by examples in Table 2. Table 21 also associates each strategy with a system from the SFL and with varying degrees of metaphoricity. From these five strategies, three of them can be related to grammatical metaphor, one of them (splitting sentences) because of effects on grammatical intricacy and lexical density, and two

(modification in syntactic structure and word order in a clause) due to an impact on the information flow.

Finally, by considering not only these text simplification strategies explained in Table 2, but also the identification of derivation and agnation (by class and rank shifts), the analysis can be performed.

2.6 Conclusion

Among the most popular linguistic approaches in computational linguistics, this thesis shows that Chomsky (1965) is not the most adequate theoretical framework to investigate text simplification. The principal limitation of Chomsky's (1965) approach compared to SFL is that the comparison between a simpler and a more complex clause is quite limited in terms of deep structure alone. In contrast, the grammatical metaphor allows the assessment of a greater number of factors from semantics and grammar, as agnation allows the differences between meanings to be tracked in each system. This way, it is a major disadvantage in using this theory for manipulating grammatical metaphoricity aiming at varying degrees of complexity, potentially resulting in text simplification.

In the light of SFL, one can infer that Dagan et al's (2013) definition of "text entailment", which takes into consideration language variability and its levels of abstraction, can be associated with grammatical metaphor. Therefore, if text entailment, which can extend or not to paraphrase if it is bidirectional, can be used to perform text simplification, then the same may be applicable for grammatical metaphor. This conclusion is confirmed by the results obtained by

Aluísio & Gasperin (2010), which used the “text entailment" approach to develop three different systems

Taking into account all these features characterizing text complexity and due to these inconsistent results associated with the metrics proposed in different Corpus Linguistics and NLP studies, this work does not consider the lexical density measurement as an efficient method by itself to discriminate between less and more metaphorical text segments, though it is useful together with grammatical intricacy, to investigate text complexity. Thus, the need for another measure to develop a linguistic representation of text complexity to input a future text simplification model with the use of statistical variables associated with linguistic categories. This measure could be the count of explicit grammatical functions according to SFL, namely the functions within the Figures (Halliday & Matthiessen, 2014, p. 30): PARTICIPANT, PROCESS, and CIRCUMSTANCE.

However, even though works such as Wu (2000) manage to integrate linguistics and computational methods efficiently, few works investigate an applied linguistics perspective on text complexity geared towards computer applications.

Based on the differences in the metaphoricity degrees within the segment pairs from the corpora, this thesis adopts an approach that integrates both NLP's and linguistics’ perspectives, drawing on SFL. These variables are associated with categories pertaining to the contextual variables of FIELD, TENOR and MODE, and IDEATIONAL, INTERPERSONAL, and TEXTUAL METAFUNCTIONS for the clause and nominal group ranks. Aware of SFL's problematization of the conventional concept of paraphrase, given that any variation in register results in new meaning construal and not a "similar" meaning construal in two texts, this thesis pursues an analysis of

metaphoricity in text complexity that may inform NLP studies and add to their approach to text simplification.

The next chapter describes the methodology to compile the SIM-Pt corpus, as well as the annotation of the selected segment pairs and the statistical analysis of the results intended to be carried out to produce a linguistic representation of text complexity in science texts aiming to input a future empirical model of text simplification in Brazilian Portuguese.

Chapter 3 Methodology

This Chapter is divided into two sections: Corpus compilation and Analysis procedures. "Corpus compilation" describes the steps followed to compile SIM-Pt, the corpus used in this study. "Analysis procedures" introduces the steps to import and annotate the corpus in Sysfan, to analyze annotations to obtain frequencies of systemic and structural patterns. The systemic patterns refer to the systemic choices in the systems (AGENCY, PROCESS, POLARITY, MOOD, THEME, and CONJUNCTION) and the structural patterns to the distinct ordering of elements. These two types of patterns provide evidence of ideational grammatical metaphor, more specifically experiential grammatical metaphor, which can be associated with a higher or lower degree of text complexity.

3.1 Corpus compilation

This section describes the compilation of SIM-Pt, a Brazilian Portuguese monolingual corpus which comprises text segment pairs. In this corpus, 50% of the segments are "naturally occurring" and 50% of them are "manually constructed".

According to Lester & O'Reilly (2018, pp. 99-100), naturally occurring data represent "naturally occurring activities", which take place regardless of the researcher. The use of this

type of data reduces bias, improving the quality of research. In this thesis, naturally occurring refers to segments retrieved from science texts found on websites.

A common practice in Natural Language Processing studies is augmenting the amount of data either by translating or manually constructing text. By the latter, "manually constructed", we refer to a set of texts, segments, or clauses written by a human to provide input to a supervised analysis of some kind[24]. An example of a corpus of manually constructed texts is the Newsela corpus[25], a collection of news articles produced by professional editors and leveled in terms of complexity.

The manually constructed segments in SIM-Pt, are rewritings of naturally occurring segments collected from the web. Pairs of a naturally occurring and a manually constructed segment will be referred to henceforth as "matching pairs" concerning experiential meaning, as they were deemed as being related to each other through metaphoricity shifts (c.f. section 2.2.1, Halliday & Matthiessen, 1999).

The next subsection describes the SIM-Pt corpus design, providing an overview of all steps carried out in this methodology.

3.1.1 SIM-Pt corpus design

This section is concerned with the corpus design used for carrying out this study, each step is explained in detail.

[24] The difference between supervised and unsupervised analysis/learning is that, in the supervised one, it is necessary to perform human analysis previously to provide input for the machine, which can learn from the input in a "supervised" way. In turn, the machine can also learn in an automatic, unsupervised, data-driven way, provided that examples are available to draw rules from some phenomena, e.g., text simplification

[25] The Newsela corpus is available at https://newsela.com/data/

The corpus is a compilation of science texts. As defined by Halliday & Martin (1993, p. xi), science is an "inter-organistic practice, a linguistic/ semiotic practice which has evolved functionally to do specialized kinds of theoretical and practical work in social institutions". Science texts can relate to all semiotic activities, as one of the possible text types to report science knowledge are papers, in which each section has a different purpose. Thus, they can be associated with different socio-semiotic activities. Based on this definition and these features, this thesis refers to the texts compiled for our research, in the domain of physics, biology, and psychology, as "science texts". The main socio-semiotic activities pertaining to the segments extracted from the texts in our corpus are "expounding" and "reporting" (Halliday & Martin, 2014, p. 33).

The steps for compiling the corpus are presented in Table 22.

Table 22

Corpus compilation steps

Phase	Step	Task
Review of text simplification methods	1	Identification of the text simplification methods used in the computer science literature
Collection of naturally occurring texts	2	The manual query for science websites based on these techniques

	3	The manual query for texts in the websites
	4	The manual query for segments in the texts
	5	Manual segment extraction onto an electronic spreadsheet
Initial complexity assessment	6	Segment classification in terms of complexity
Segments complexity reassessment	7	Segment assessment by students
	8	Segment reclassification based on student feedback
Construction of further segments	9	Manual construction of segments

As shown in Table 22, the corpus compilation process consists of five phases: i) Review of text simplification methods; ii) Collection of naturally occurring texts; iii) Segments complexity reassessment; iv) Segments complexity reassessment; v) Construction of further segments. Each phase is further detailed in one or more steps. The next section describes each step in more detail.

3.1.2 Corpus compilation steps

This section describes all the steps for compiling the SIM-Pt corpus, consisting of segments retrieved from science websites and segments rewritten on the basis of these segments.

3.1.2.1 Review of text simplification methods

The text simplification strategies used as criteria to ascertain whether a segment could be considered a simplified version of its counterpart were retrieved from the literature concerning topics such as text simplification and text summarization (cf. section 2.2.1). The strategies found were the following[26]:

- Conversion from passive voice to active voice
- Addition of explanation or examples
- Replacement of more technical words with less technical ones
- Syntactic order modification
- Sentence splitting
- Sentence removal
- Sentence compression
- Dis-embedding of relative clauses
- Subordinate clauses splitting

[26] These strategies were characterized in this way by the computational linguistics literature and use terms that do not draw on Systemic Functional Linguistics. These labels were kept at this point but the strategies were analyzed and interpreted in the light of Systemic Functional Linguistics.

Table 23 shows an example of each strategy in English[27].

Table 23

Examples of text simplification strategies[28]

Text simplification strategy	Example	Simplified example
Conversion from passive voice to active voice	During the 13th century, gingerbread **was brought** to Sweden by German immigrants.	German immigrants **brought** it to Sweden during the 13th century.
Addition of explanation or examples	Humidity is the amount of water vapor in the air	Humidity (adjective: humid) refers to water vapor in the air, **but not to liquid droplets in fog, clouds, or rain.**
Replacement of more technical words with less technical ones (lexical simplification)	...extracted the £75 on-the-spot cash fine which has outraged him and other clamped **motorists**...	... extracted the £75 on-the-spot cash fine. It has shocked him and other clamped **drivers**...

[27] In general, the studies in Brazilian Portuguese (e.g., those authored by Specia), tend to follow trends set by those studies in English, thus the examples will be shown here only in English.

[28] In the third line of this table, we refer to lexical simplification by the comparison between "more technical words" and "less technical words", but the texts from the computer science literature refer to them as "difficult" and "easy" words.

Syntactic order modification	**During the 13th century**, gingerbread was brought to Sweden by German immigrants.	German immigrants brought it to Sweden **during the 13th century**.
Sentence splitting	The mayor, who recently got a divorce, is getting married again.	The mayor recently got a divorce. The mayor is getting married again.
Sentence removal	Also contributing to the firmness in copper, the analyst noted, was a report by Chicago purchasing agents, which precedes the full purchasing agents report that is due out today and gives an indication of what the full report might hold.	Also contributing to the firmness in copper, the analyst noted, was a report by Chicago purchasing agents. The Chicago report precedes the full purchasing agents report. The Chicago report gives an indication of what the full report might hold. **The full report is due out today.**

Sentence compression	Falaj irrigation is an ancient system dating back thousands of years and is used widely in **Oman, the UAE, China, Iran and other countries**	The ancient falaj system of irrigation is still in use in **some areas.**
Dis-embedding of relative clauses	...extracted the £75 on-the-spot cash fine which has outraged him and other clamped motorists...	... extracted the £75 on-the-spot cash fine. **It has shocked him and other clamped drivers**...
Subordinate clauses splitting	**While** the law generally supports clampers operating on private land, Mr Agar claims CCS's sign was not prominent enough to be a proper warning...	The law generally supports clampers operating on private land. **But** Mr Agar claims CCS's sign was not prominent enough to be a proper warning…

Note. The examples were collected from the following studies: Sentence splitting (Štajner, 2015, emphasis added); Sentence removal (Siddharthan, 2006, emphasis added); Replacement of more technical words with less technical ones (Siddharthan, 2004, emphasis added); Conversion from passive voice to active voice, Addition of explanation or examples, Syntactic order modification, Sentence compression (Hwang et al., 2015, emphasis added); and Dis-embedding of relative clauses (Siddharthan, 2004, emphasis added). Bold was added by the author of this thesis to emphasize contrasts.

3.1.2.2 Manual query for science websites

This step was carried out in parallel with the steps from section 3.1.2.3, as it was necessary to store the segments and the metadata (the website information) for each search.

The steps to select these science websites were the following:

- Identification of some terms discussed in science textbooks in Brazilian Portuguese on nuclear energy from CDTN website[29], e.g., "fissão nuclear" (*nuclear fission*);
- Identification of other concepts from texts belonging to other branches of physics as well as from biology and psychology, with higher availability than texts from other areas, e.g., "osmose" (*osmosis*), "divisão celular" (*cellular division*);
- Web query in search engines seeking the chosen domains and terms related to science teaching, such as "ciência para crianças" (*science for children*), "definição escolar" (*school definition*) and "o que é" (*what is*) plus term (e.g., "o que é divisão celular" - *what is cell division*), among others;
- Web query in search engines in search of other terms on the recurrently chosen websites on the first searches, e.g., "osmose" (*osmosis*).

After performing these steps to find potential segments, many websites that contained segments related to the selected domains were queried. A more detailed search resulted in 193[30]

[29] CDTN stands for the Centre for the Development of Nuclear Technology ("Centro de Desenvolvimento da Tecnologia Nuclear"), located at UFMG (Federal University of Minas Gerais).

[30] The 196 websites initially found were narrowed down to 193 due to accessibility issues at the time of the research to ensure data availability as long as possible. Reasons for the accessibility issues were: i) the website address changed during the corpus collection; ii) it was not possible to retrieve the files from university websites iii) or from book previews. In these cases, just the excerpts were retrieved.

websites from which segments were extracted. The websites from which the highest number of segments are listed in Table 24. Complete data is available in Appendix A.

Table 24

Number of segments extracted from texts in websites

Website address	Number of segments
https://www.msdmanuals.com	43
https://mundoeducacao.bol.uol.com.br	25
https://www.todamateria.com.br	12
https://pt.wikipedia.org	11
http://escolakids.uol.com.br	5
http://alunosonline.uol.com.br	4
http://brasilescola.uol.com.br	4
https://www.estudopratico.com.br	4
http://www.ufrgs.br	3

http://www.scielo.br		2

Note. In this table, all occurrences with 1 hit were excluded to highlight the most frequent text sources, as well as most of the website addresses with 2 or 3 occurrences. The full list of websites mentioned in this table is available in Appendix A and the list of websites used to access all texts are available in Appendix B. Due to time constraints and the need to focus on general aspects of the sources, it is not within the scope of this thesis to give further information on these websites.

Table 24 shows that the most frequent source for the segments was the website "msdmanuals", which provides translations from English reference texts on the domain of biology and psychology aimed at healthcare professionals (e.g., doctors) and laypeople (students and consumers).[31] These translations are curated by experts in each field to ensure their reliability, as these texts are accessed by a broad audience in Brazil. On the website's home page, a menu is provided to choose from one of three versions, either "Professional version", "Consumer version" or "Veterinary version". In either professional or consumer version, in each article a button points to the other version (see Figure 11 and Figure 12).

Some websites, such as "mundoeducacao", "todamateria" and "escolakids" provide information on different subjects. Yet, their webpages lack buttons pointing to another version of the same text. Figure 11 and Figure 12 show snapshots of a segment from the "msdmanuals" website and Figure 13 shows a snapshot of a segment from the "todamateria" website.

[31] This feature is highly valuable for investigating both text simplification and metaphoricity degree since it increases the chances of finding matching pairs of segments in terms of experiential meaning.

Figure 11

Snapshot of a segment retrieved from "mdsmanuals" website - version for healthcare professionals

Note. Retrieved from: https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/transtornos-de-personalidade/vis%C3%A3o-geral-dos-transtornos-de-personalidade

Figure 12

Snapshot of a segment retrieved from "msdmanuals" website -- version for consumers.

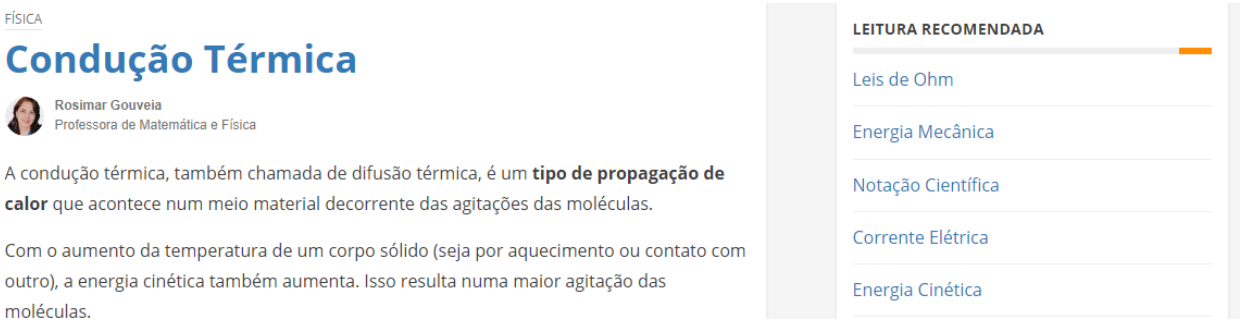
Note. Retrieved from https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/transtornos-de-personalidade/considera%C3%A7%C3%B5es-gerais-sobre-transtornos-de-personalidade#v6580311_pt.

Figure 11 and Figure 12 illustrate, respectively, a snapshot of the search for "schizophrenia" at the website "msdmanuals", from which segment pair 81 was retrieved. Figure 11 shows the website version for healthcare professionals and Figure 12 the version for lay-people (patients or consumers).

Figure 11 and Figure 12 show a hyperlink in red “Clique aqui para a versão para a família” (localized in English as "View consumer version") and “Clique aqui para a versão para profissionais” (localized in English as "View Professional version"). These hyperlinks point to aligned website pages, so a user who accesses one page can navigate from one version to another.

Figure 13

Snapshot of a segment retrieved from “todamateria” website



Note. Retrieved from: https://www.todamateria.com.br/conducao-termica/

Some websites, though, lack a hyperlink to another version of the same text. Figure 13 shows a snapshot of a text on Thermal conduction (“Condução Térmica”), which does not present this hyperlink. The only hyperlinks this website contains are hyperlinks to other pages with similar contents under “Leitura Recomendada” (Recommended reading), such as Energia Mecânica (Mechanical Energy).

Another type of website is one belonging to institutions. Websites such as “http://www.ufrgs.br” and “http://www.scielo.br” are hosted and managed by government institutions and are aimed at current or prospective college students/professors. For this reason,

some texts retrieved from such websites, e.g., papers or dissertations/theses, are assumed to have a higher degree of complexity.

Furthermore, these websites' mapping were the initial criteria for identifying which segment from each pair was more complex than its counterpart.

3.1.2.3 Manual query for segments in texts from the selected websites

This step was performed simultaneously with the search for websites, as it was necessary to store the texts, the segments, and the websites' addresses for each search. This information assisted in setting up the environment for the software used.

Taking into consideration the text simplification strategies in Section 2.10.1, a subset of 200 segment pairs was manually extracted. These segments were retrieved from texts from the domains of physics, biology, and psychology, targeting readers with two different levels of science literacy. Those labels (consumers and professionals) were associated with the source texts from which the segments were retrieved, considering that most of them mentioned this information explicitly as part of their URLs. For instance, segments retrieved from the URL "http://mundoeducacao.bol.uol.com.br/fisica/forca-centripeta.htm" were associated with physics ("física").

Table 25 shows the distribution of segments according to each domain

Table 25

Distribution of segments according to each field

Domain	Absolute frequency of segment pairs	Relative frequency of segment pairs
Physics	50	25.12 %
Biology	95	47.74 %
Psychology	54	27.14 %
Total	199	100%

Note. Elaborated by the author for this research

Table 25 shows the frequency of segments from the domains of physics, biology, and psychology, as well as the greater availability of texts from biology. 50 segment pairs from these domains (biology and psychology) were sampled to make sure that the results would not be biased. This amount (50) was equal to the minimum number of texts from the domain of physics. The sampling process was performed using the R[32] software environment **(**R CRAN, 2021), by using the sample function to select 50 texts from all texts belonging to the physics domain and 50 from the biology domain.

[32] This script can be found in <https://github.com/rodrigoacastro/Castro2020>, more specifically in <https://github.com/rodrigoacastro/Castro2020/tree/master/sampling_and_websites_analysis>.

After the sampling process, the selected segments were organized in an electronic spreadsheet in *.xlsx* format that was later imported into the database manager Filemaker Pro 17 Advanced for analysis' purposes[33]. The balanced distribution is presented in Table 26.

Table 26

A balanced distribution of segments according to each domain

Domain	Absolute frequency of segment pairs
Physics	50
Biology	50
Psychology	50
Total	150

Table 26 presents the balanced corpus' distribution, with 50 pairs of physics, biology, and psychology texts (and segments) collected from the web. This way, it was possible to prevent bias from occurring due to a higher number of segments belonging to one of the domains. This sampling to avoid bias was carried out after the reassessment of the segments described in 3.1.7.

[33] Usually, in order to import data to software, documents in *.txt* or *.csv* format (without any kind of formatting) are required. Yet, not only FileMaker Pro is capable of handling *.xls* or *.xlsx* files, it also facilitates the process of selecting the columns of interest and processing them properly as part of the created database.

Table 27 presents two aligned segment pairs retrieved from the web, highlighting in bold the aligned nominal groups, e.g., “fissão nuclear” (*nuclear fission*) and “agorafobia” (*agoraphobia*).

Table 27

Examples with aligned segments retrieved from scientific websites

Type	Segment number	Set A	Set B
Naturally occurring	001	Fissão nuclear é **o processo de divisão do núcleo atômico instável em outros núcleos mais estáveis**. *Nuclear fission is the process of splitting an unstable atomic nucleus into other more stable nuclei.*	Dito de forma resumida, a fissão nuclear é **a quebra de núcleos grandes, formando núcleos menores** e liberando grande quantidade de energia. *In sum, nuclear fission is the fission of large nuclei, forming smaller nuclei and releasing a great amount of energy.*

Naturally occurring	085	Agorafobia é **o medo e a ansiedade antecipatória acerca de ficar aprisionado em situações ou lugares sem uma maneira de escapar facilmente** e sem ajuda **caso ansiedade intensa ocorra**. *Agoraphobia is anticipatory anxiety and fear of being imprisoned in situations or places without easy exit or help in case of intense anxiety.*	Agorafobia é **ansiedade de estar preso em situações ou lugares dos quais não há nenhuma forma de escapar facilmente se surgirem ansiedade ou pânico**. *Agoraphobia is the anxiety of being imprisoned in situations or places in which there is no way of easily escaping in case of anxiety or panic.*

Note. The aligned nominal groups from the segments are in bold. Emphasis added.

The aligned nominal groups highlighted in bold in Table 27 show the common wording in both segments, a criterion to align them and investigate them concerning text complexity, aiming at exploring the phenomenon of text simplification.

After that, all texts collected from the web were extracted manually from the websites, were assigned an identifier, and were saved as *.txt* or *.pdf* files, depending on their availability. After that, with the use of BASH Shell scripts[34], the pdf files were converted to *.txt* format and saved in a folder to be imported into the database. Then they were associated with the respective segment they referred to, organized in spreadsheets (.xlsx files).

[34] The script used for this is also available in: https://github.com/rodrigoacastro/Castro2020

Figure 14 shows a snapshot of one of the texts in *.txt* format.

Figure 14

Snapshot of text source number 7

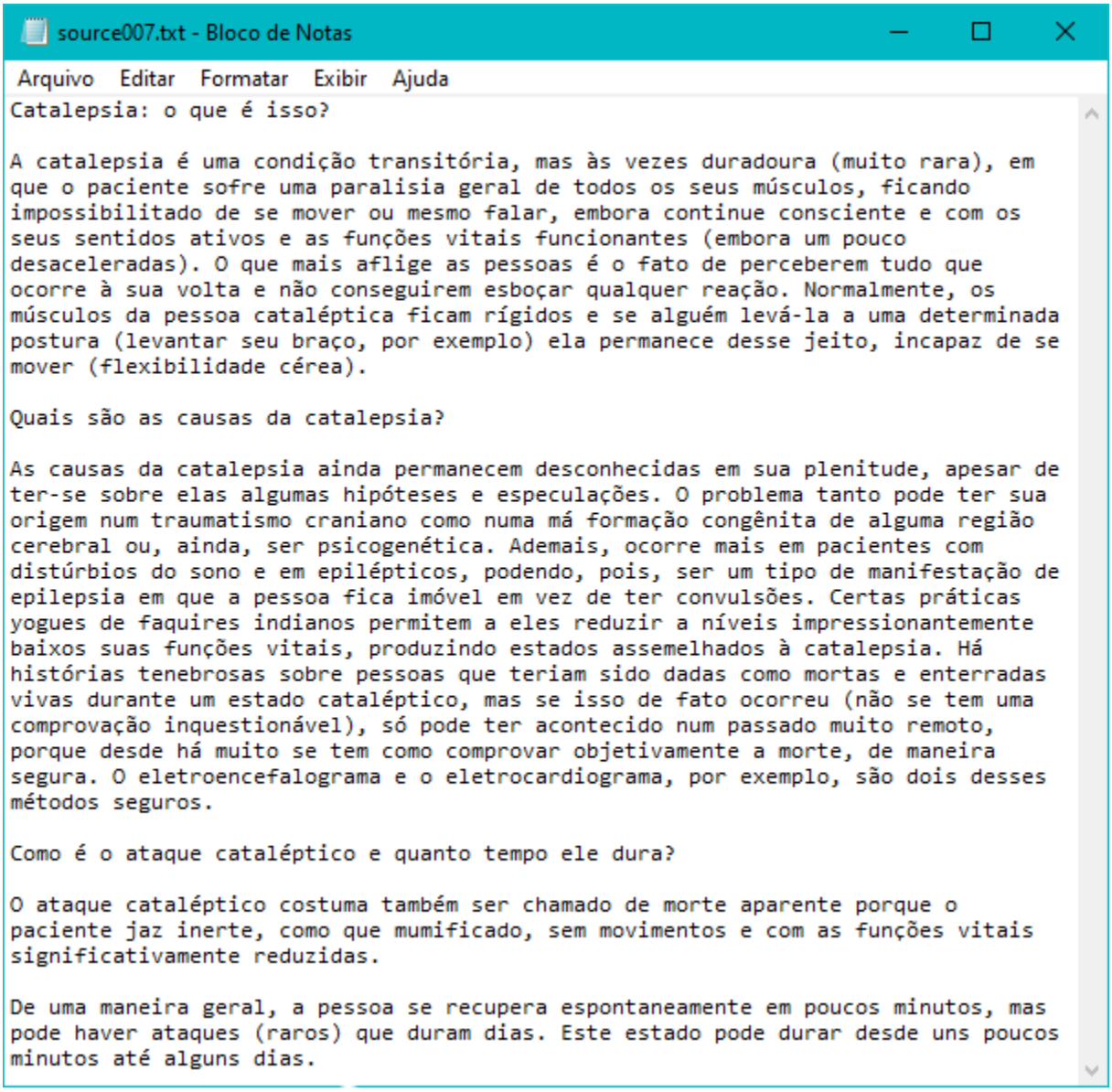
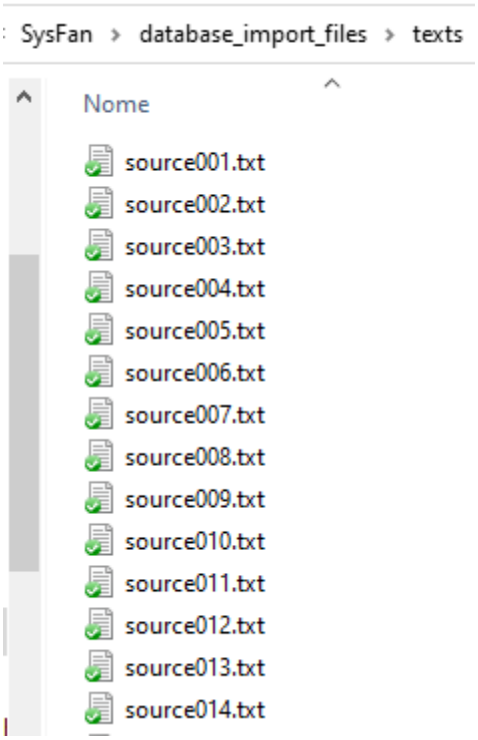


Figure 15 shows the folder which stores the 193 source texts in UTF-8 codification, required for the software Sysfan to recognize all characters and to avoid corrupting their content.

Figure 15

Folder storing all 193 source texts



Note: For the sake of brevity, only 14 from the 193 source texts were shown.

The next step was developing a coding system that was implemented automatically in FileMaker Pro 17 Advanced software to preserve the correspondence between text sources text sets, texts, segments, and clauses. This system consisted of assigning numbers to the pairs, segments, clause complexes, and clauses. The pairs were numbered between 1 and 150, the segments between 1 and 600 (Set A - 1 to 150, set B - 151 to 300, Set C - 301 to 450, Set D - 451

to 600), the clause complex within the segment according to its order (1, 2 or 3, for instance), and the clause within the complex also according to its order (1 or 2, for instance). An example of this coding is presented in Table 28, showing the full code in the last column, which does not include the pair number.

Table 28

Example of the coding system

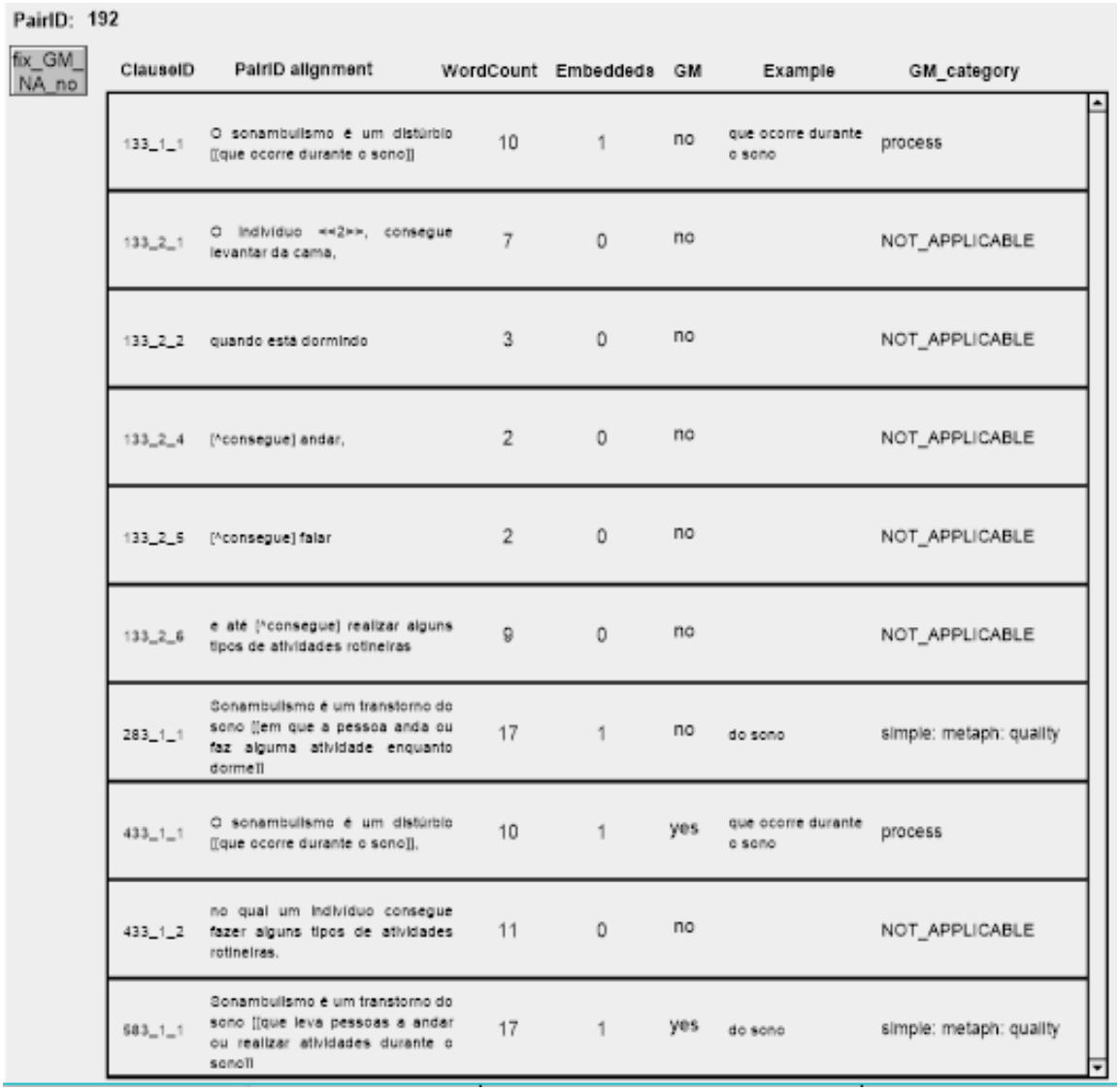
Segment	Clause complex	Clause	Full code
134	1	1	134_1_1

Table 28 shows how each clause was coded using the aforementioned system. The pair number was not included in the full code, but they were used to track the clauses during the analysis.

This way, after organizing the data in spreadsheets and developing the coding system, the texts in UTF-8 encoding could be imported into FileMaker Pro 17 Advanced software to be analyzed drawing on SFL categories and other automatic measurements, such as type/token ratio. Figure 16 shows a Sysfan snapshot with this coding in the column “ClauseID”.

Figure 16

Sysfan snapshot - automatic alignment of segment pair 192



PairID: 192

fix_GM_NA_no

ClauseID	PairID alignment	WordCount	Embeddeds	GM	Example	GM_category
133_1_1	O sonambulismo é um distúrbio [[que ocorre durante o sono]]	10	1	no	que ocorre durante o sono	process
133_2_1	O indivíduo <<2>>, consegue levantar da cama,	7	0	no		NOT_APPLICABLE
133_2_2	quando está dormindo	3	0	no		NOT_APPLICABLE
133_2_4	[^consegue] andar,	2	0	no		NOT_APPLICABLE
133_2_5	[^consegue] falar	2	0	no		NOT_APPLICABLE
133_2_6	e até [^consegue] realizar alguns tipos de atividades rotineiras	9	0	no		NOT_APPLICABLE
283_1_1	Sonambulismo é um transtorno do sono [[em que a pessoa anda ou faz alguma atividade enquanto dorme]]	17	1	no	do sono	simple: metaph: quality
433_1_1	O sonambulismo é um distúrbio [[que ocorre durante o sono]],	10	1	yes	que ocorre durante o sono	process
433_1_2	no qual um indivíduo consegue fazer alguns tipos de atividades rotineiras.	11	0	no		NOT_APPLICABLE
583_1_1	Sonambulismo é um transtorno do sono [[que leva pessoas a andar ou realizar atividades durante o sono]]	17	1	yes	do sono	simple: metaph: quality

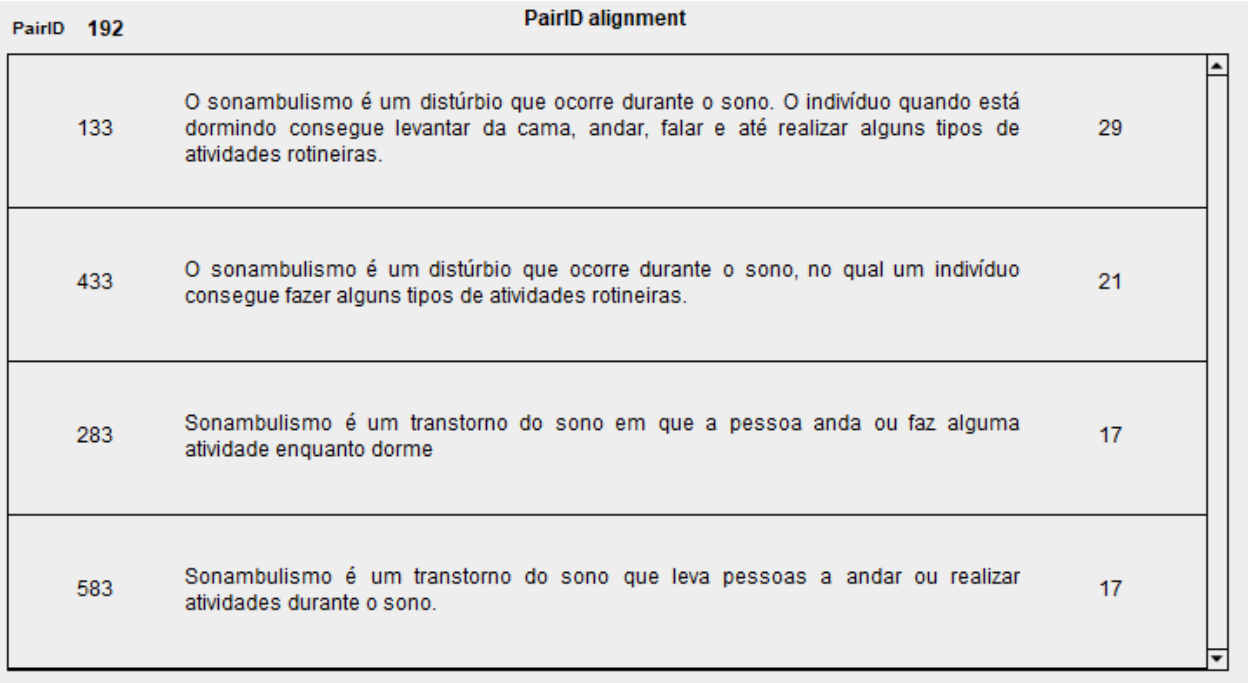
In this example, for instance, at the first three lines showing 133.1.1, 133.1.2, and 133.1.3, we split it into three using the dot (.) as a divider. In these, 133 is the segment number,

the clause complex number is 1 because there is only one clause complex in the segment, and the clause number can be 1, 2, or 3.

The same alignment is shown in a less detailed manner in Figure 17

Figure 17

Sysfan snapshot of segment alignment



PairID 192

PairID alignment

133	O sonambulismo é um distúrbio que ocorre durante o sono. O indivíduo quando está dormindo consegue levantar da cama, andar, falar e até realizar alguns tipos de atividades rotineiras.	29
433	O sonambulismo é um distúrbio que ocorre durante o sono, no qual um indivíduo consegue fazer alguns tipos de atividades rotineiras.	21
283	Sonambulismo é um transtorno do sono em que a pessoa anda ou faz alguma atividade enquanto dorme	17
583	Sonambulismo é um transtorno do sono que leva pessoas a andar ou realizar atividades durante o sono.	17

Note: This example regards segment 192 as well.

The next section describes the manual extraction of all the segments onto an electronic spreadsheet.

3.1.4 Manual segment extraction onto an electronic spreadsheet

This step consisted of creating an empty template and filling it, as presented in Figure 18.

Figure 18

Snapshot of template for manual segment extraction

SEGMENT NUMBER	SIMPLER	MORE COMPLEX	Topic
001			
002			
003			
004			

The columns of the template in Figure 18 describe the segment pair number, the segment considered “simpler” (less technical), the one considered “more complex” (more technical), and the segment’s domain (“topic” column).

The next step was manually copying and pasting these segments, as the selection requirements necessary for the analysis of syntactic and semantic elements. This way, it was not feasible to do it automatically with the use of software[35].

3.1.5 Segment classification in terms of complexity

In this section, the template shown in Figure 18 was filled out and the segments from each pair were classified, one being considered "simpler" (more technical) and another "more complex" (more technical). The criteria for this classification were those from studies on Natural Language Processing (NLP), listed and described in section 3.1.2.1.

Figure 19 shows a snapshot of the updated spreadsheet with the segments and their domain ("Topic" column)[36].

[35] Some papers from computer science literature (such as Indig et al, 2020) have used softwares to search through the web and collect text from websites, that is, what they refer to as "crawling". The software able to do this task is called "crawler". This software was not used in this case because it was not the intention to crawl texts automatically, due to strict selection criteria used. In other studies, though, crawlers can be very useful.

[36] The domain column was initially named "Topic", but in this thesis only the term "Domain" is used to refer to the text field (physics, biology or psychology).

Figure 19

Snapshot of the updated spreadsheet

SEGMENT NUMBER	SIMPLER	MORE COMPLEX	Topic
001	Dito de forma resumida, a fissão nuclear é a quebra de núcleos grandes, formando núcleos menores e liberando grande quantidade de energia.	Fissão nuclear é o processo de divisão do núcleo atômico instável em outros núcleos mais estáveis.	Physics
002	Em 1938, o físico alemão Otto Hahn e seus colaboradores realizaram essas experiências de bombardeamento do urânio, e a física austríaca Lise Meitner (1878-1968) explicou esse fenômeno, dizendo que o núcleo do átomo de urânio era instável e ao ser bombardeado com nêutrons moderados ele se rompe praticamente ao meio, originando dois núcleos médios e liberando dois ou três nêutrons, além da liberação de uma grande quantidade de energia.	Esse processo foi descoberto em 1939, por Otto Hahn (1879-1968) e Fritz Strassmann (1902-1980). O processo ocorre em decorrência da incidência do nêutron sobre o núcleo atômico. Ao bombardear de forma acelerada o átomo que tem um núcleo fissionável, ele parte-se em dois.	Physics
003	Forças de contato: [ocorrem] Quando há contato direto entre dois corpos.	Forças de contato: [são] aquelas que agem sobre os corpos somente na medida que quem aplica a força está necessariamente em contato com	Physics
004	Forças de campo: [ocorrem] Quando a atuação da força ocorre a distância.	Forças de campo: [são] aquelas que agem sobre os corpos sendo que o corpo que exerce a força não se encontra em contato os outros,	Physics

Note. The columns represent the segment pair number, the segment considered simpler, the one considered more complex, and the segment's domain ("topic" column).

The next sections describe the student assessment and the classification, followed by the manual construction of segments.

3.1.6 Segment assessment

Concerning the reclassification of the segments, 8 students who were taking a Translation Studies course during the second semester of 2018 at UFMG were asked to classify the text segments according to their degree of complexity. For this task, they were trained in the SFL theory during translation classes, especially lessons on text types, metafunctions, stratification, instantiation and grammatical metaphor. Then, they received instructions via email (see Table 29) as well as an electronic spreadsheet with the designated segments as a template, which was filled (Figure 20). At the time, 175 segments were reclassified, a number that was narrowed down to 150 later on, for corpus balancing purposes.

The instructions sent by email for each student are presented in Table 29.

Table 29

Instructions given to the students who assessed the segment pairs

Email containing the instructions	English Translation
Prezados alunos, Segue em anexo a planilha para a tarefa da avaliação 2, com data de entrega dia 9/11. Na planilha vocês encontrarão segmentos sobre um mesmo tópico (paráfrases) extraídos de textos distintos. Vocês deverão: 1. avaliar os segmentos alocados a vocês (ver relação abaixo) e dizer qual deles é de mais fácil leitura e	Dear students, Please find attached a spreadsheet for course assignment 2, which is due 9/11. On the spreadsheet, you will find segments about a common topic (paraphrases) retrieved from different texts. You are requested to 1. Assess the segments assigned to each of you (see list below) and determine which of them is more readable and comprehensible - there is validation for the column where you

compreensão - na planilha vocês selecionam a opção na coluna correspondente	are expected to select one of the values

Figure 20 shows the classification template in Brazilian Portuguese.

Figure 20

Snapshot of original classification template

	SEGMENTO A	SEGMENTO B	QUAL SEGMENTO É MAIS FÁCIL DE LER E COMPREENDER? CLICKAR NA SETA NO CANTO INFERIOR
134	Por exemplo: na geladeira os alimentos são resfriados dessa forma.	Um exemplo disso ocorre, por exemplo, no resfriamento dos alimentos dentro da geladeira.	

In this example from Figure 20, pair 134 is shown without any translation. The first column after segment B is related to comprehension of the segments, while the others, omitted from this figure, were not part of the scope of this work. The translation of the question is "Which segment is easier to read and to comprehend? Click in the arrow in the lower corner".

Based on this template, each of the 8 students was assigned some segments to classify. Table 30 shows the segment pair assigned to each student, summing up 175 segments for 8 students for this course activity.

Table 30

Segment pair distribution per student

Student	Segment pairs per student
Student 1	1 to 22
Student 2	23 to 44
Student 3	45 to 66
Student 4	67 to 88
Student 5	89 to 110
Student 6	111 to 132
Student 7	133 to 154
Student 8	155 to 175

Note. The students have given their permission, as part of the course, to have their activity used as part of this research, and numbers were assigned to each student to protect their identity.

Each student was asked to perform the activity for 20 or 21 segments, indicating which segment they considered more complex either Segment A or Segment B. All answers were carefully collected and consolidated to allow a reliable comparison between the initial evaluation and their assessment. This comparison showed that their assessment matched approximately 75% of the initial evaluation, which, given their previous training for the task and, was considered relevant to use this to reclassify the segments. This assessment of the segments was taken into consideration as the main reclassification criteria. This reassessment was done by collecting all assessments (each in a different *.xlsx* file) in one file and consolidating the main classification, as described in the next section.

3.1.7 Segment reclassification based on student feedback

Student agreement on the initial classification in Table 30 was used to improve the corpus. Table 31 shows the agreement between the initial classification and the students' assessment regarding the complexity of each segment, by choosing which they considered "simpler" (less technical).

Table 31

Students' agreement with initial text complexity classification

Segment	Absolute frequency	Relative frequency
A	129	**73.71**
B	46	26.29
Total	175	100

Note: The first column describes which segment was considered "simpler" (less technical).

As the agreement between the initial and the later assessment was 73.21 % (in bold), in some cases the segment order within the same pair was modified to match the student assessment. This way, the final version of the corpus was produced. In this final version, aiming at achieving a balanced corpus based on the configuration decided earlier on (the same number of segments per domain), approximately 25 segments were filtered out. This segment removal was performed after the author's assessment according to the student feedback.

3.1.8 Manual construction of segments

This section consisted of describing the procedures for manually constructing segments matching the naturally occurring segment pairs, ensuring that each segment within the pair has a different degree of grammatical metaphoricity. Researchers from LETRA that performed the text collection for this thesis[37] were responsible for this step to produce more examples that would be comparable with the first ones. The researchers also evaluated how manually constructed segments could be similar to naturally constructed ones regarding differences in grammatical metaphor degrees. Besides, the researchers verified all the segments to make sure the segments were chosen and classified correctly.

In this phase, the second subset of 150 aligned segments was manually constructed based on the segment pairs from the first subset and the same criteria; that is, the text simplification strategies from the literature. These new segments are associated with the same domains (physics, biology, and psychology). Also, if a less metaphorical (less "complex") example was used to produce segment A, and a more metaphorical (more "complex") segment from the same pair was used to construct segment B, consequently D will be more metaphorical than C. This assumption is based on a process of rewriting, which necessarily requires metaphorization; as a consequence, this would change the level of complexity of the text. This hypothesis will be investigated in this thesis, with the evidence from the corpus.

Table 32 presents two aligned segments from the manually constructed segments, associated with the segments shown in Table 27.

[37] These researchers were the author of this thesis and Sofia Pereira Sepúlveda <http://lattes.cnpq.br/4475513273164464>.

Table 32

Examples with manually constructed aligned segments

Type	Set	Segment pair	***Examples*** with aligned text segments	
Manually constructed		001	Fissão nuclear é o processo de cisão de núcleos instáveis em núcleos mais estáveis, com grande liberação energética. *Nuclear fission is the process of splitting unstable nuclei into more stable nuclei, with great energy release.*	Fissão nuclear é a quebra de núcleos grandes em núcleos menores mais estáveis. *Nuclear fission is the splitting of large nuclei into smaller and more stable nuclei.*
Manually constructed		085	Agorafobia é o quadro de medo e ansiedade de ficar aprisionado em um lugar ou situação caso a pessoa apresente quadro de ansiedade.	Agorafobia é o medo de estar preso em um lugar ou situação se a pessoa ficar ansiosa. *Agoraphobia is the fear of being stuck in a place or situation if the*

			Agoraphobia is a patient's condition characterized by the anticipatory anxiety and fear of being imprisoned in a place or situation if the individual suffers from anxiety symptoms.	*individual becomes anxious.*

Note. The aligned nominal groups from the segments were highlighted in bold. The translation of the segments into English is in italics.

The steps for manual construction of new segments were the following:

- Mapping of the topics from the segment pairs from the first set of texts;
- Elaboration of manually constructed segments based on the segments compiled from the web following the same criteria;
- Revision of the manually constructed segments to ensure their correspondence to the segments compiled from the web.

These procedures were necessary to validate the criteria used to collect the segments from the websites, as the same criteria as the web collection of the first set of texts were used to ensure

their alignment with the segments from the first text set. Also, with the use of constructed segments, it was possible to effectively double the number of segments analyzed, improving the results as well. Finally, the constructed segments were revised to avoid any typographic errors or any other errors that would introduce bias in the data analysis.

These procedures also followed research practices in machine learning and machine translation, which employ manually constructed texts as input for their algorithms' training. Aluísio et al (2010), for instance, use manually created segments produced in the scope of the PorSimples project to train their readability level identifiers, which will be later used in the actual tasks.

The next subsection describes the methodology used to analyze segments and clauses.

3.2 Analysis procedures

This section introduces the procedures carried out in the analysis, as well as the annotation tool and the manual analysis performed to obtain the results.

Table 33 lists the corpus analysis tools in steps.

Table 33

Corpus analysis steps

Phase	Step	Task
Database preparation	1	Database design in Sysfan
Manual importing text data into Sysfan	2	Importing text files in Sysfan for storage purposes
	3	Importing of the segment's spreadsheets in Sysfan
Manual identification of complexes and clauses	4	Identification of clause complexes in the segments
	5	Identification of clauses in the clause complexes
Automatic segment alignment	6	Automatic alignment of the segments belonging to the same pairs
Segments' Analysis	7	Analyzing segments to identify simplification strategies

	8	Analyzing clauses in terms of metafunctional categories
Frequencies' extraction and exporting	7	Extracting frequencies of annotated categories;
	8	Exporting data onto a spreadsheet
Frequencies' extraction and exporting	9	Extracting frequencies of the categories with R script and exporting them in *.csv* files
Manual analysis	10	Manual analysis of the shifts
	11	Analyzing the structural and systemic patterns and experiential metaphoricity

The next section describes the annotation tool, Sysfan, and the adjustments made to carry out the analysis, followed by the steps of the analysis.

3.2.1 Analysis

As previously described, the database template was adapted from the original setup provided by Wu (2000), and the following data management process steps were executed:

1. Database design in Sysfan

2. Importing text files in Sysfan for storage purposes;

3. Importing segments spreadsheets in Sysfan;

4. Identification of clause complexes in the segments;

5. Identification of clauses in the clause complexes;

6. Automatic alignment of the segments belonging to the same pairs;

7. Analyzing segments to identify simplification strategies

8. Analyzing clauses in terms of metafunctional categories

9. Extracting frequencies of annotated categories;

10. Exporting data onto a spreadsheet

11. Counting the frequency of the categories from the spreadsheet with the use of an *R* script

12. Manual analysis of the shifts

13. Analyzing the structural and systemic patterns and experiential metaphoricity

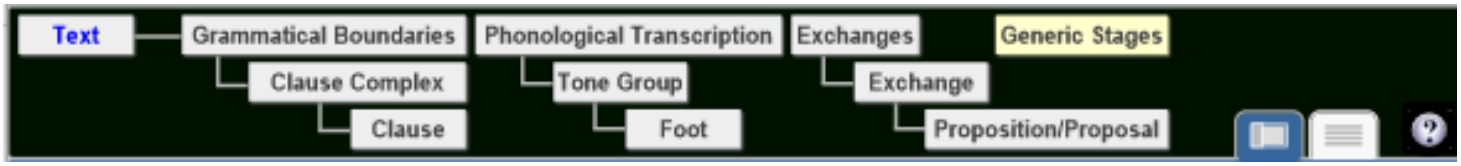
The next subsections describe each of these steps in detail.

3.2.1.1 Database design in Sysfan

3.2.1.1.1 Annotation tool - Sysfan

Sysfan is a Filemaker[38] plugin that introduces an interface specifically designed to do text analysis drawing on Systemic Functional Linguistics. In this interface, the user can import all the texts, segment them into clause complexes, then into clauses, and analyze texts according to the three metafunctions (IDEATIONAL, INTERPERSONAL, and TEXTUAL), as well as their graphological components. In other words, the interface allows the users to import texts and categorize them by context and type of text, then segment them into clause complexes and clauses to be classified according to the SFL system or structure categories. Finally, as this analysis is dependable on Filemaker Pro, is it also highly customizable, allowing new variables and new elements to be imported into the database and to the interface, respectively. Figure 21 shows a snapshot of Sysfan showing the analysis schema that was adapted to be used as a model for the database that stores the data from this thesis.

Figure 21



Snapshot of Sysfan analysis schema

[38] FileMaker Pro is a database management system that can both store and manage databases, as well as create apps with a user-friendly interface to interact with the data and perform analyses and computations. For more information, check Appendix C.

Figure 21, from Sysfan's initial page, shows a snapshot of the analysis schema[39] with several layouts used to store and analyze data, namely "Text", "Grammatical Boundaries", "Clause Complex" and "Clause". While "Text" is used to store texts, "Grammatical Boundaries" allows the user to segment texts into clause complexes, "Clause Complex" to analyze clause complexes, and to segment them into clauses, and "Clauses" allows the analysis of clauses. It is important to observe that all layers are connected through IDs of some sort, linking, for instance, text 1, its clause complexes, and its clauses.

As the original design of Sysfan was not suited to this research, the next section describes the adaptations made to the original template to fit the needs of this thesis.

3.2.1.1.2 Sysfan customization

To perform data analysis, it was necessary to adapt Sysfan's original setup provided by Wu (2000).

The main adaptations were:

1. Adapting the database to represent the relationship between the tables containing the distinct forms of textual data to be analyzed (texts, text segments, clause complexes, and clauses);

[39] This schema also provides shortcuts to access the different layouts available for analysis.

2. Updating word lists in the software, adding terms in Brazilian Portuguese, so lexical density and grammatical accuracy could be correctly calculated, as well as the average lexical density and the average grammatical accuracy[40];

3. Adapting the system of PROCESS of English according to Figueredo's (2007, 2011) description of Brazilian Portuguese;

4. Adding/deleting variables, fields in the database, and buttons to accommodate the needs of this research and improve its efficiency, such as the speed of data annotation;

5. Adding/editing functions from Filemaker to fit the interests of this research;

6. Adding layouts to process different kinds of results, not predicted in the original one;

7. Adjusting the export schemas selected to adequately select the data and the results that needed to be exported for further manual analysis or textual processing and statistical modeling within R.

While adaptations 2 and 3 were mostly related to the language of the texts analyzed (Brazilian Portuguese instead of English), all the others were associated with the requirements of this research. In other words, without these adaptations, it would not have been possible to execute this research and obtain results.

As a result, the updated layout follows in Figure 22.

Figure 22

[40] The average lexical density is the sum of all lexical density values for each set divided by the number of segments of such a set. The same applies for the average grammatical accuracy. The mean is calculated by adding two numbers and dividing this result by two.These measures are calculated automatically by Sysfan.

Snapshot of adapted Sysfan

schema

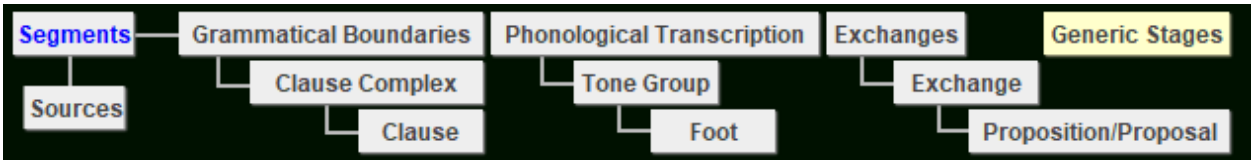


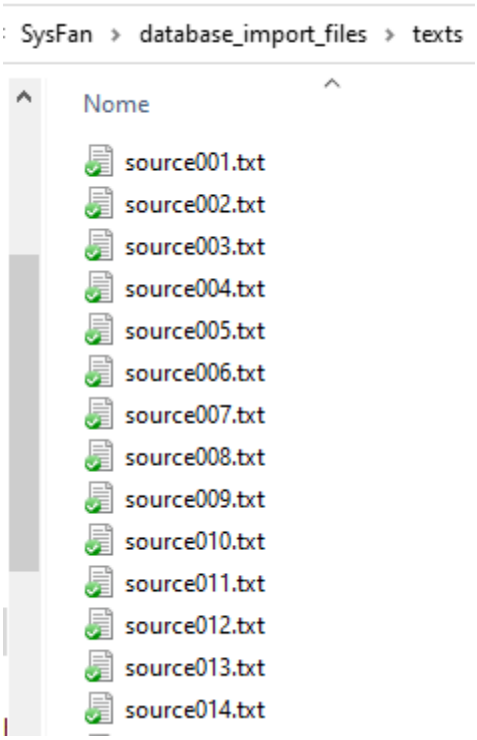
Figure 22, from Sysfan initial page, shows a snapshot of the updated schema with different layouts, namely “Sources”, “Segments”, “Grammatical Boundaries”, “Clause Complex”, and “Clause”. The shortcut “Sources” is used to store the source texts; “Segments” stores all the segments; “Grammatical Boundaries” is used to segment texts into clause complexes; “Clause Complex” allows the user to analyze clause complexes and to segment them into clauses; and “Clauses” allows the analysis of clauses.

3.2.1.2 Importing text files in Sysfan for storage purposes

To analyze the segments, as described in section 3.1, the segments from the SIM-Pt corpus were previously organized in electronic spreadsheets, and the texts were stored in a folder called “texts”. This way, they could be imported into Sysfan and divided into clause complexes and clauses to be analyzed in the Sysfan interface to find patterns indicating grammatical metaphor as evidence of text simplification. The first step, the text importing, is shown in Figure 23, Figure 24, and Figure 25.

Figure 23

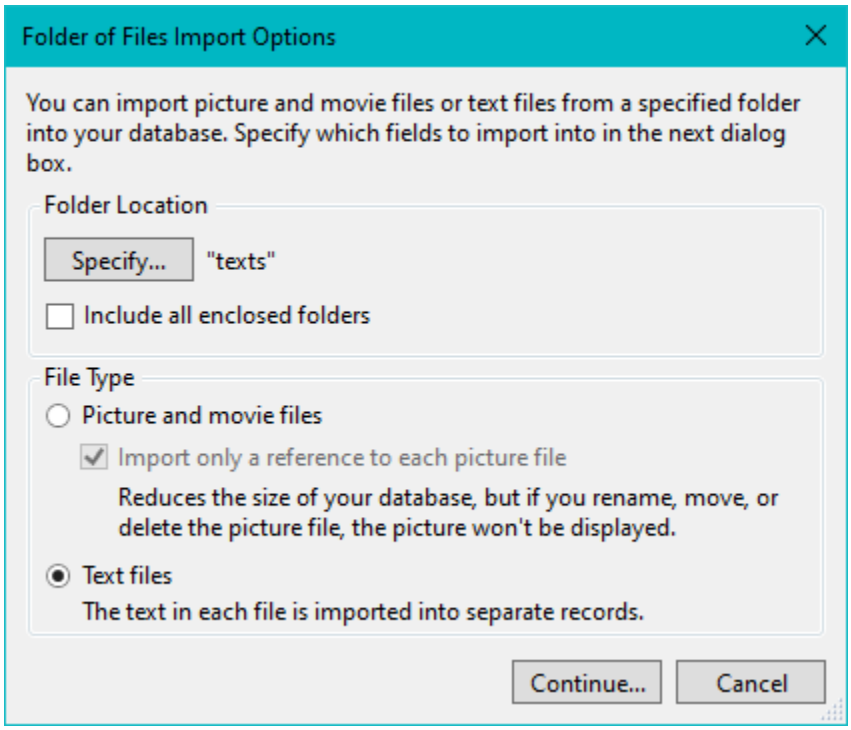
Snapshot of the folder containing all 193 source texts.



Note. These are the first 14 of the 193 source texts in *.txt* format and UTF-8 (without BOM) format.

The folder in Figure 23 with the content of all source texts retrieved from the websites was imported into Sysfan to allow the recovery of each segment context during the analysis. The importing process is illustrated in Figure 24, which shows the texts' importing process into the Sysfan interface.

Figure 24



Sysfan snapshot of text importing process - folder selection

Figure 24 shows that the folder “texts” was selected so that the text files could be imported into Sysfan. The next step was to select the information to be imported into Sysfan fields, namely “Text” and “URL”, as shown in Figure 25.

Figure 25

Sysfan snapshot of text importing process - field selection

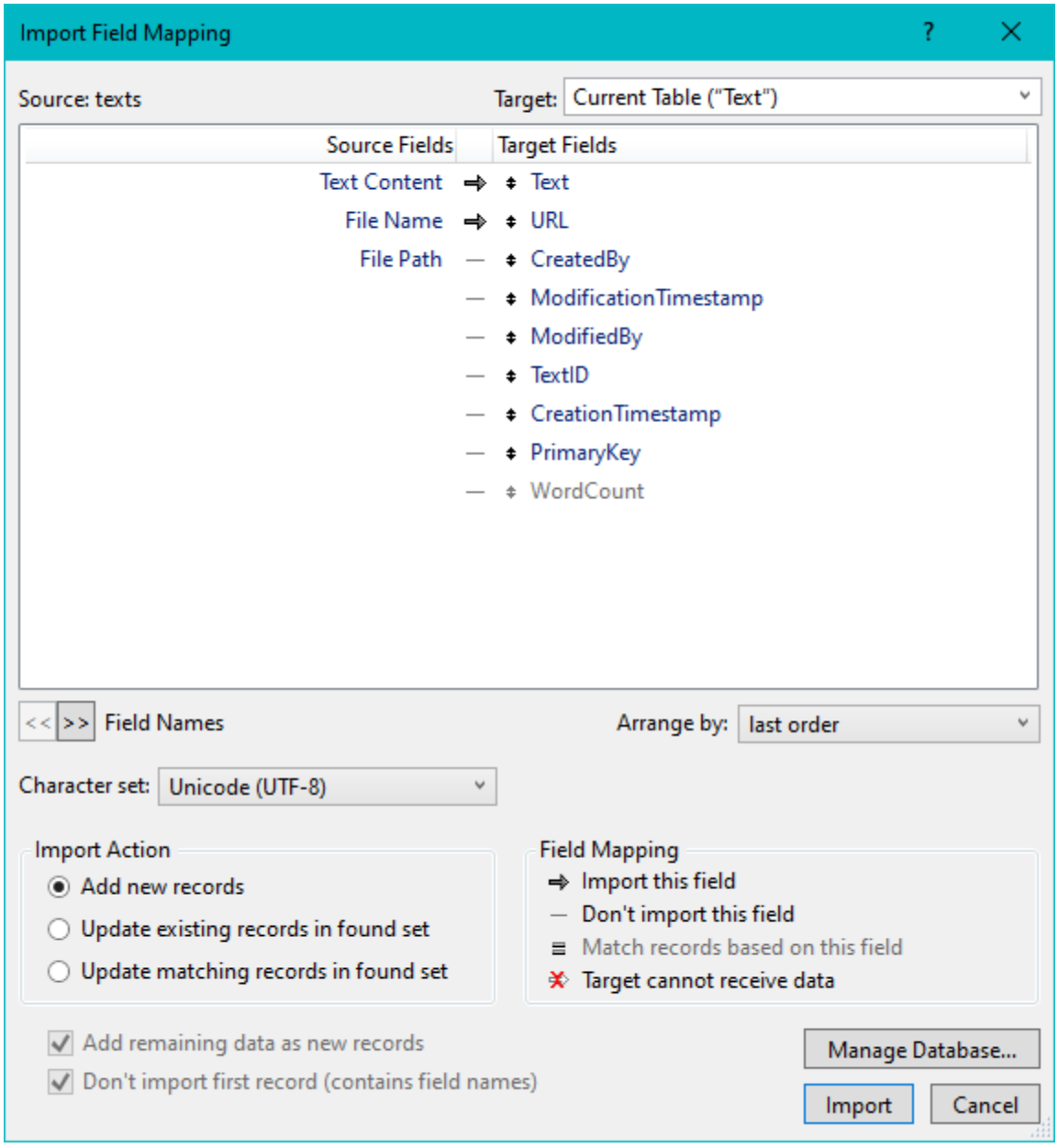


Figure 25 shows the “source fields” from the spreadsheet, specifically the content and the filename of the text files being matched to the “target fields” from the Sysfan interface. The options selected update the entries that already exist and save the remaining entries as new records.

3.2.1.3 Importing segments spreadsheets in Sysfan

Importing consisted of selecting the data spreadsheet and importing each column from the spreadsheet as a field from the Sysfan interface. An extract of the imported spreadsheet is shown in Figure 26, the spreadsheet selection process in Figure 27, and the segments’ importing in Figure 28.

Figure 26

Excel snapshot - Example of 20 first observations from the 600 segments imported

TextID	Segment_nature	Segment_set	PairID	Segme	Topic	URL	URL_ID
1	naturally occurring	SetA	3	Forças de	Physics	https://m	3
2	naturally occurring	SetA	4	Forças de	Physics	https://m	3
3	naturally occurring	SetA	6	A fecunda	Biology	https://co	5
4	naturally occurring	SetA	7	Qualquer	Biology	https://m	6
5	naturally occurring	SetA	12	O vitiligo	Biology	https://m	9
6	naturally occurring	SetA	13	A teoria m	Biology	https://m	9
7	naturally occurring	SetA	15	Quando a	Biology	http://bra	10
8	naturally occurring	SetA	16	Pareidolia	Psycholog	https://w	11
9	naturally occurring	SetA	17	Asteroide	Physics	https://m	12
10	naturally occurring	SetA	19	Por defini	Biology	http://pat	14
11	naturally occurring	SetA	20	Heterótro	Biology	http://ww	15
12	naturally occurring	SetA	21	Seres autc	Biology	https://w	16
13	naturally occurring	SetA	23	Acatisia é	Biology	https://w	18
14	naturally occurring	SetA	25	Clinicame	Biology	https://w	18
15	naturally occurring	SetA	26	A mitose e	Biology	http://esc	20
16	naturally occurring	SetA	27	Esse proce	Biology	http://esc	20
17	naturally occurring	SetA	28	A radioter	Biology	https://w	21
18	naturally occurring	SetA	35	As proteín	Biology	http://esc	24
19	naturally occurring	SetA	37	As proteín	Biology	http://esc	24
20	naturally occurring	SetA	38	Nas prote	Biology	https://m	25

Figure 26 shows that the fields TextID, Segment_nature[41], PairID, Segment, Topic, URL, and URL_ID describe the information imported in Sysfan. Table 34 contains a description of each field. It is important to notice that some of the field names, such as Text_ID, were not modified during the analysis, to avoid bugs with the software, which was already working adequately.

Table 34

Description of each imported field

Field	Description

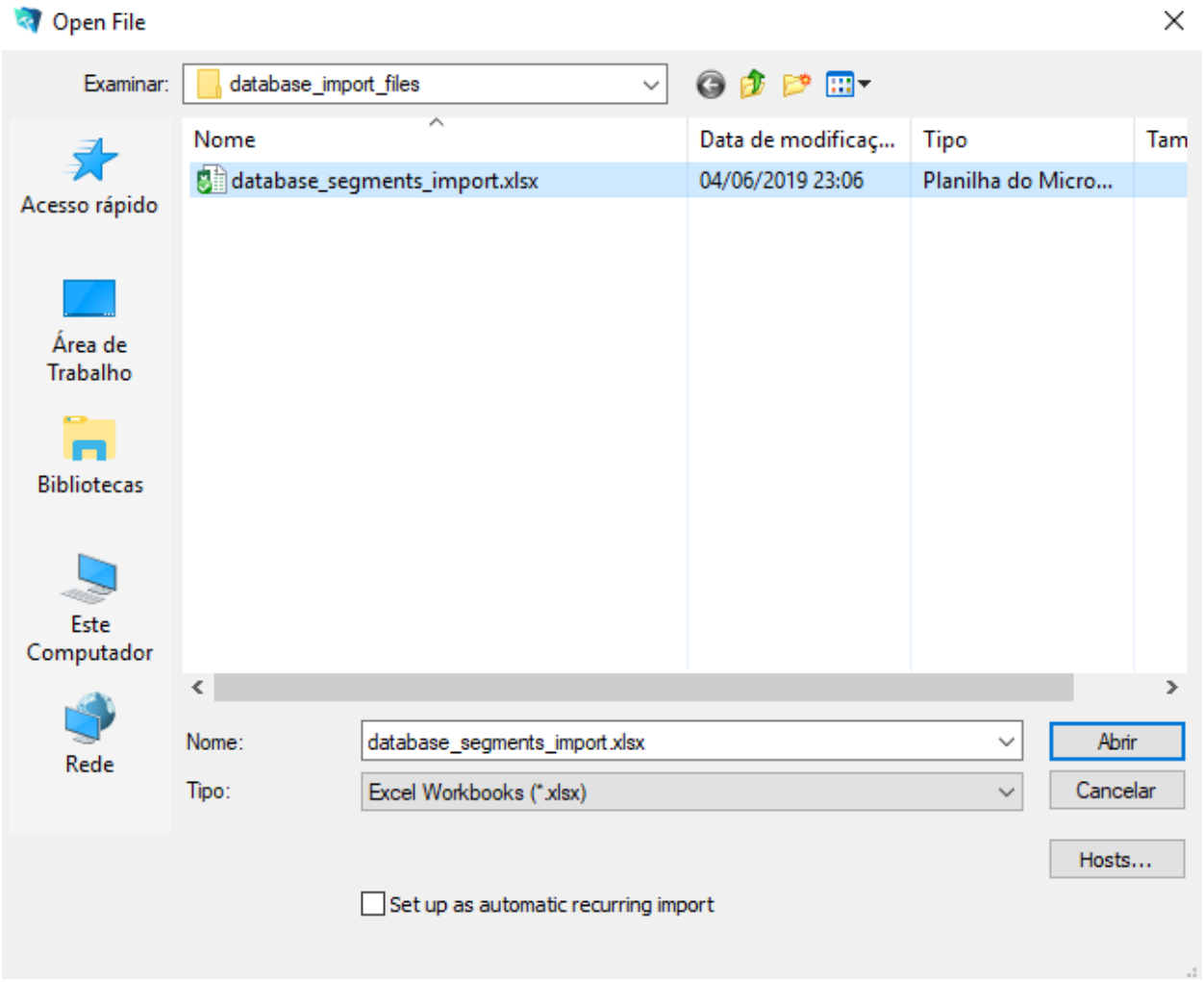
[41] As “nature” is not a technical term from SFT, the tables from this thesis will use the label “type” instead of “nature” for this information.

Text_ID	The ID of each segment
Segment_nature	Type of segment (naturally occurring or manually constructed)
PairID	The ID of the segment pair
Segment	Content of the segment
Topic	Segment topic, either Physics, Psychology or Biology
URL	Text source website URL (address)
URL_ID	The ID of the text source URL

The first part of the spreadsheet importing process is shown in Figure 27, describing how to choose a spreadsheet to read the information fields. In this case, the file "database_segments_import.xlsx" was chosen.

Figure 27

Sysfan snapshot of spreadsheet selection



The second part of the importing process is shown in Figure 28, namely how to import the data of each spreadsheet column as a field to read the information fields.

Figure 28

Sysfan snapshot of the spreadsheet importing process

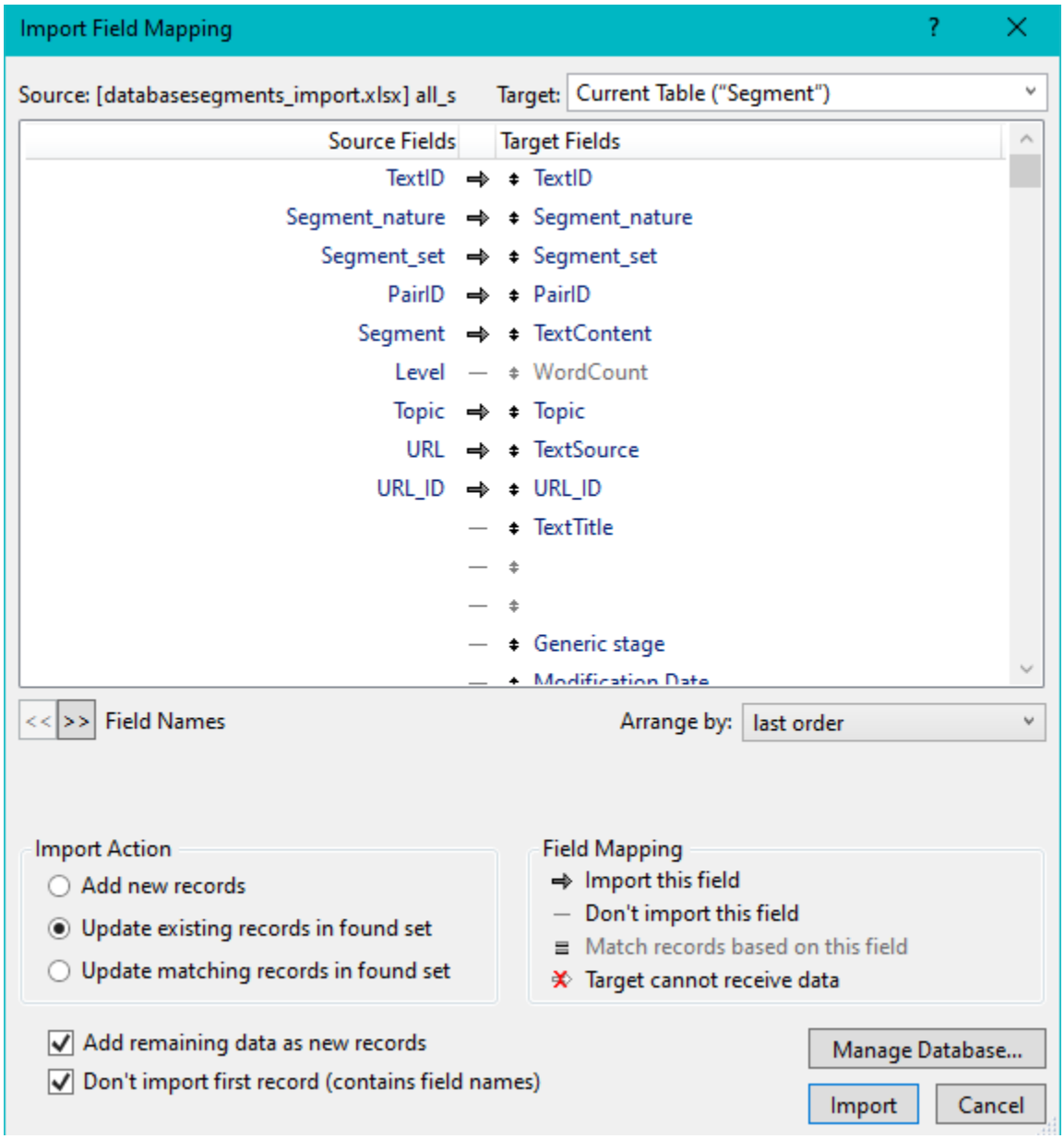


Figure 28 shows that the "source fields" from Table 34, retrieved from the spreadsheet (cf. Figure 26) are being matched to the "target fields" from the Sysfan interface. The options selected, such as "Update existing records in found set" (to update the entries that already exist) set the behavior of the importing process. It is important to notice that all fields but "Level" were

imported, as this classification was discarded during the analysis. This choice was made because i) the studies considering more than two levels of text simplification, such as Xu et al (2015), using Newsela corpus, designed their corpus for this purpose, which was not the case in this thesis; and ii) it was not possible to find instances to fill more than two levels of text simplification, as SIM-Pt was designed to distinguish two levels specifically: one more congruent and one more metaphorical. For these reasons, the field "Level" was not included in Figure 28 and Table 35.

The division of a segment into clauses complexes and then into clauses is exemplified in Table 35 with the naturally occurring segments from segment pair 192.

Table 35

Segmentation of naturally occurring segment 192

Segment	Clause complexes	Clauses
O sonambulismo é um distúrbio que ocorre durante o sono. O indivíduo quando está dormindo consegue levantar da cama, andar, falar e até realizar alguns tipos de atividades rotineiras.	O sonambulismo é um distúrbio que ocorre durante o sono	O sonambulismo é um distúrbio [[que ocorre durante o sono]].
Somnambulism is a disorder that takes place while in a stage of sleep. The asleep individual can get up from the bed, walk, talk, and even perform some routine activities.	*Somnambulism is a disorder that takes place while in a stage of sleep.*	*Somnambulism is a disorder that takes place while in a stage of sleep.*
	O indivíduo quando está dormindo consegue	O indivíduo [...] consegue

	levantar da cama, andar, falar e até realizar alguns tipos de atividades rotineiras.	levantar da cama *The individual [...] manages to get up from his/her bed*
	The asleep can get up from the bed, walk, talk, and even perform some routine activities.	quando está dormindo *When [the individual] is sleeping*
		(consegue) andar *(is able to) talk*
		(consegue) falar *(can) speak*
		e até (consegue) realizar alguns tipos de atividades rotineiras. *and (can) even perform some routine activities*

Note. The marker "[[" and "]]" indicates embedded clauses and "(" and ")" non realized items. The English translation is in italics.

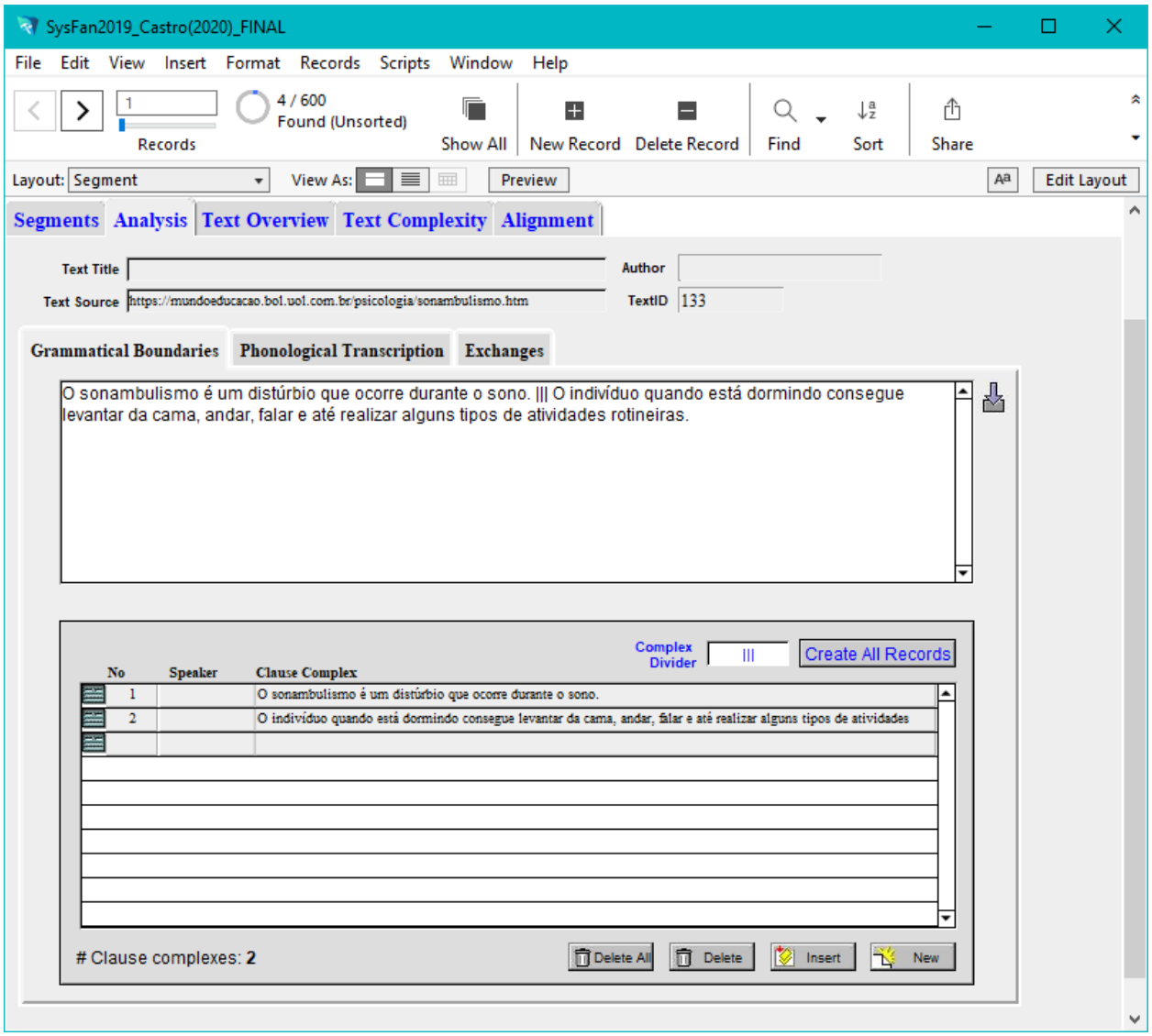
Table 35 shows the division of the segments belonging to pair 192 into Clause complexes and Clauses. The terms in parentheses were recognized as being implicit, so they were made explicit using the parenthesis notation, as in the last row: "e até (consegue) realizar" - translatable to "and (can) even perform".

3.2.1.4 Identification of clause complexes in segments

Figure 29 shows the Sysfan interface, in which the division of the segment takes place by adding the complex divider “III” and pressing “Create all records”.

Figure 29

Sysfan snapshot of the spreadsheet importing process



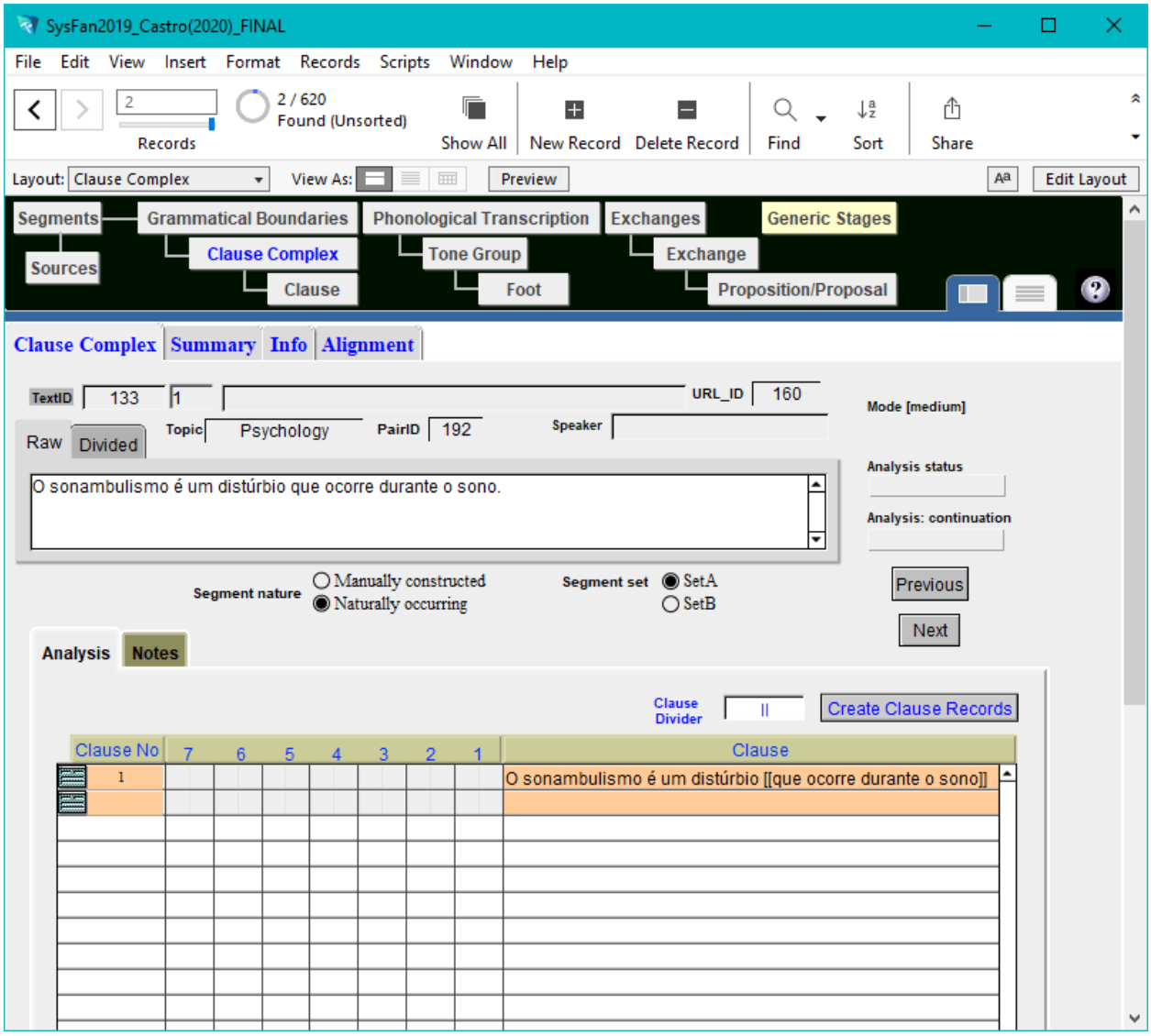
In Figure 29, the segment in the field “Grammatical boundaries” is divided into the clause complexes shown in the field “Clause complex” through the addition of the complex divider “|||”, which can be different if the user chooses to, and uses the button “Create all records”.

3.2.1.5 Identification of clauses in clause complexes

Figure 30 shows the Sysfan interface, in which the segmentation of the first clause complex takes place by adding the complex divider “||” and pressing “Create Clause records”.

Figure 30

Sysfan snapshot of the first clause complex segmentation



Following this step, Figure 31 shows the segmentation of the second clause complex into clauses.

Figure 31

Sysfan snapshot of the second clause complex segmentation

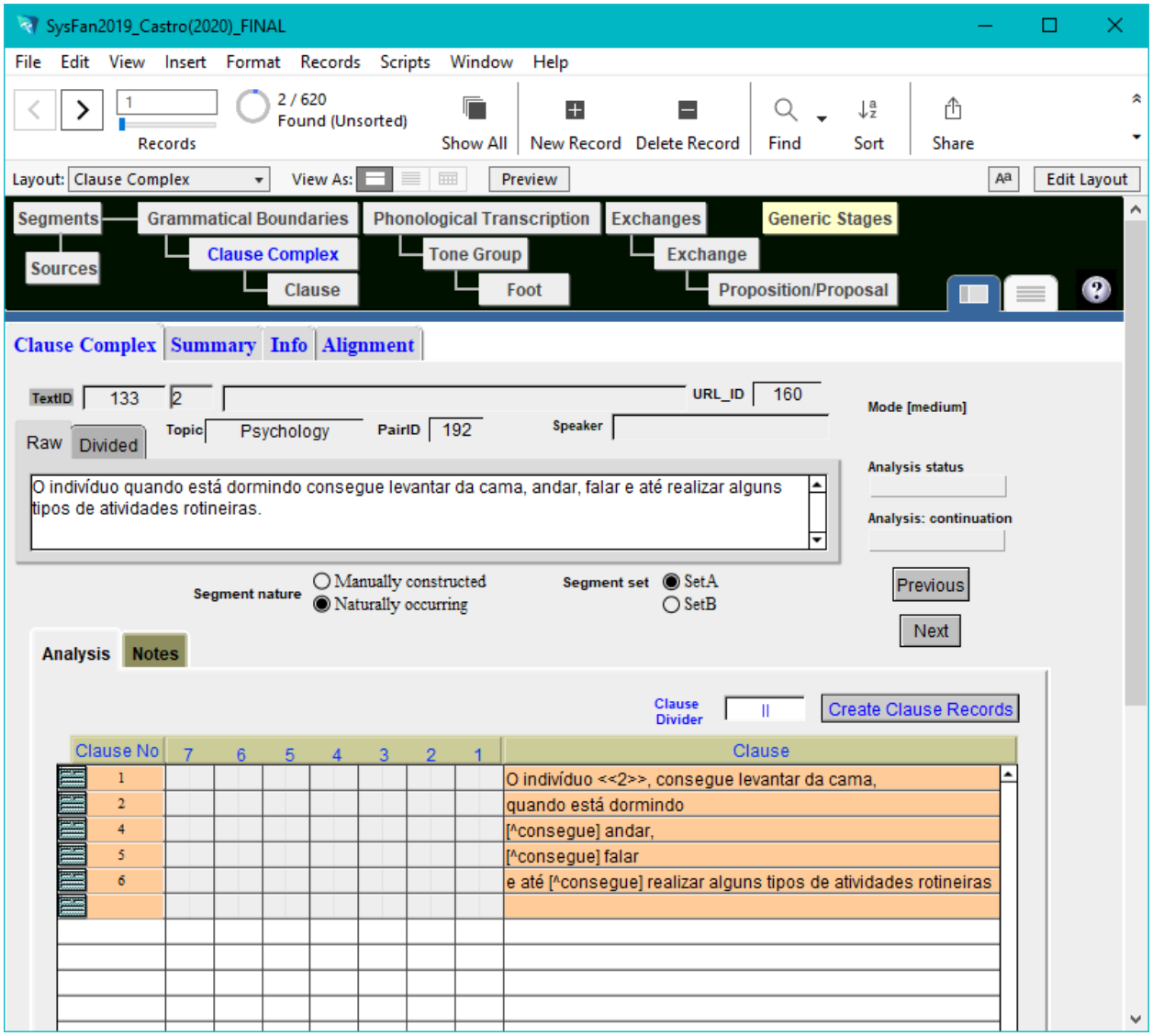


Figure 31 illustrates the segmentation of the second clause complex in five clauses, including a relative clause “quando está dormindo” (when he is sleeping), which takes the

notation “<<2>>” in its main clause to mark its position. This notation was created especially for this purpose, that is, to show the position of this clause that was classified separately in the next empty slot.

3.2.1.6 Automatic alignment of segments belonging to the same pairs

Table 36 shows the alignment of segment pair 192 in terms of clauses. An automatic alignment was performed automatically in Sysfan, organizing the segments in Set A, Set B, Set C, and Set D.

Table 36

Alignment of segment pair 192

Type	Set	Clauses	English Translation
Naturally occurring	Set A	O sonambulismo é um distúrbio [[que ocorre durante o sono]].	Somnambulism is a disorder that takes place while in a stage of sleep.
Naturally occurring	Set A	O indivíduo [...] consegue levantar da cama	The individual [...] can get up from bed,
Naturally occurring	Set A	quando está dormindo	in his sleep
Naturally occurring	Set A	[consegue] andar	[can] walk
Naturally occurring	Set A	[consegue] falar	[can] talk

Naturally occurring	Set A	e até [consegue] realizar alguns tipos de atividades rotineiras.	and even perform some routine activities.
Naturally occurring	Set B	Sonambulismo é um transtorno do sono [[em que a pessoa anda ou faz alguma atividade enquanto dorme]]	Somnambulism is a sleep disorder [in which the individual walks or performs some activity while s/he sleeps
Manually constructed	Set C	O sonambulismo é um distúrbio [[que ocorre durante o sono]],	Somnambulism is a disorder [[that takes place while in a stage of sleep]]
Manually constructed	Set C	no qual um indivíduo consegue fazer alguns tipos de atividades rotineiras.	in which an individual can perform some routine activities.
Manually constructed	Set D	Sonambulismo é um transtorno do sono [[que leva pessoas a andar ou realizar atividades durante o sono]].	Somnambulism is a sleep disorder [[that induces people to walk or perform activities during sleep]].

Note. The marker "[[" and "]]" indicates embedded clauses.

Table 36 presents in table format the alignment automatically performed in Sysfan, considering that the embedded clauses were marked between "[[" and "]]" and non-realized terms are between brackets. In Figure 17 and Figure 36, this alignment is shown in Sysfan, taking into account the categories described in the next section.

3.2.1.7 Analyzing segments to identify simplification strategies

The investigation began with the research of text simplification strategies from NLP, provided in Chapter 3, section 3.1.2.1. Then each strategy was interpreted according to a linguistic framework, filtering which strategies/categories could be explained on linguistic terms

taking SFL into account, focusing specifically on grammatical metaphor. This new subset of text simplification categories was used to investigate which segments contained clause complexes and clauses that would provide evidence of ideational grammatical metaphor. From the strategies considered in section 3.1.2.1, those that were chosen to ascertain which one provided sufficient evidence of metaphoricity are presented as a template in Figure 32.

Figure 32

Sysfan snapshot - analysis of the classification according to text simplification criteria

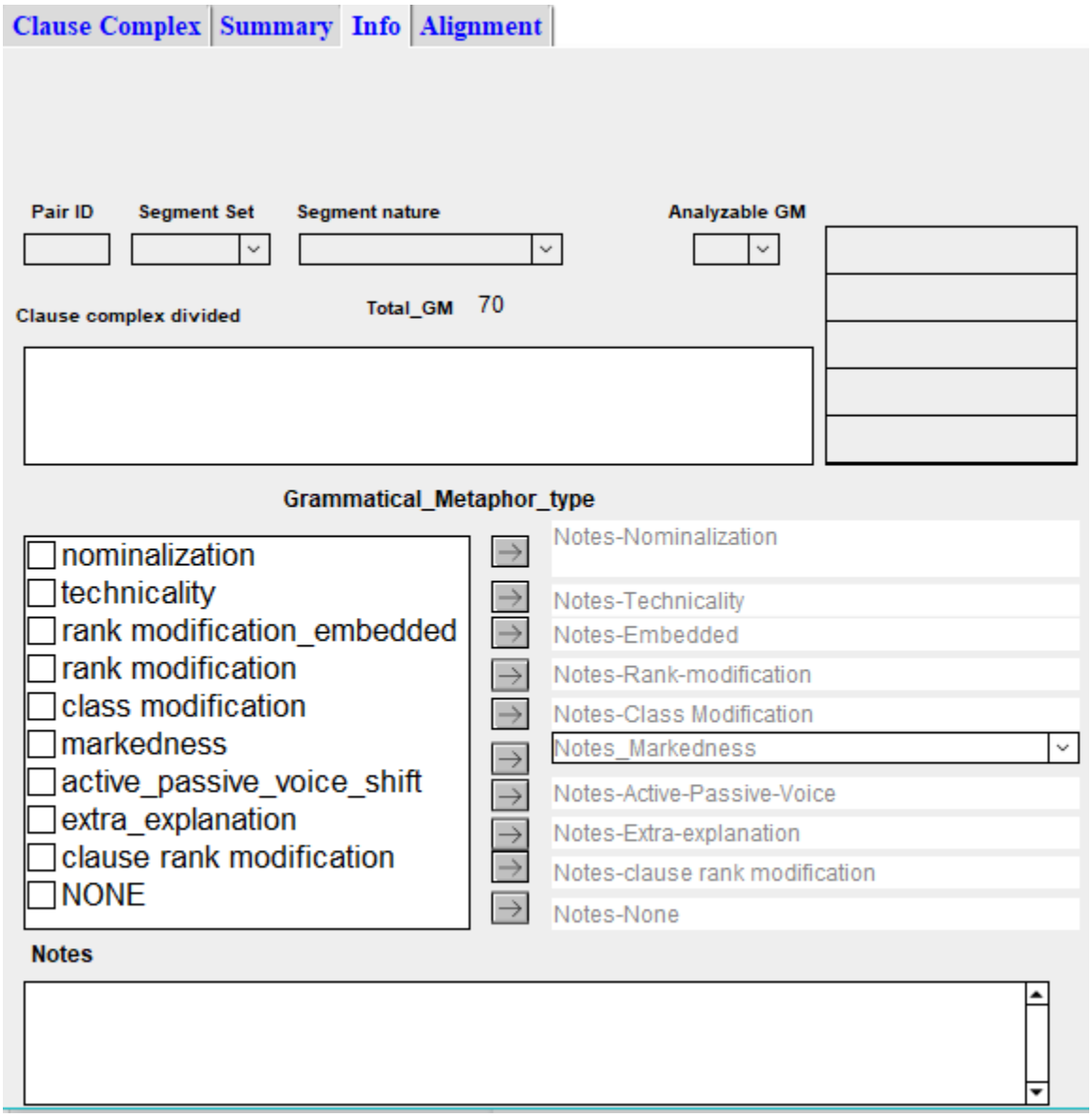
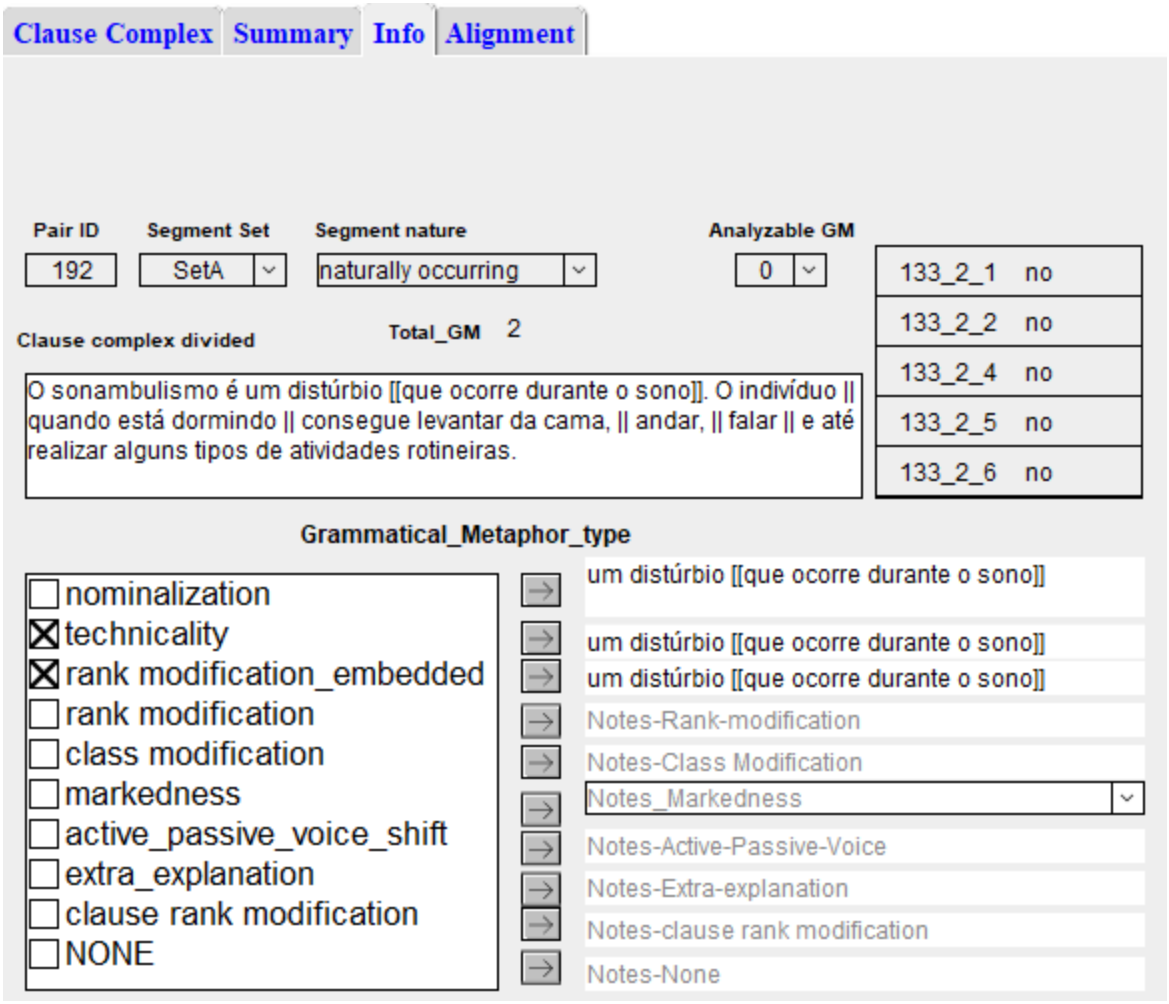


Figure 33 shows the template from Figure 32 filled with the information from segment pair 192.

Figure 33

Sysfan snapshot - analysis of the pair 192 in terms of text simplification criteria



In Figure 33, the segment in the box “Clause complex divided” was marked in terms of “technicality”, which was considered under the term “technical taxonomies” on Halliday and Martin (1993) (cf. section 2.3.2) as a feature of scientific texts, but it was disregarded later due to

lack of sufficient evidence of grammatical metaphor[42], and "rank modification_embedded" (the presence of an embedded clause).

Figure 34 shows a template to classify clauses according to linguistic categories, such as the CIRCUMSTANCE type.

Figure 34

Clause analysis Sysfan interface

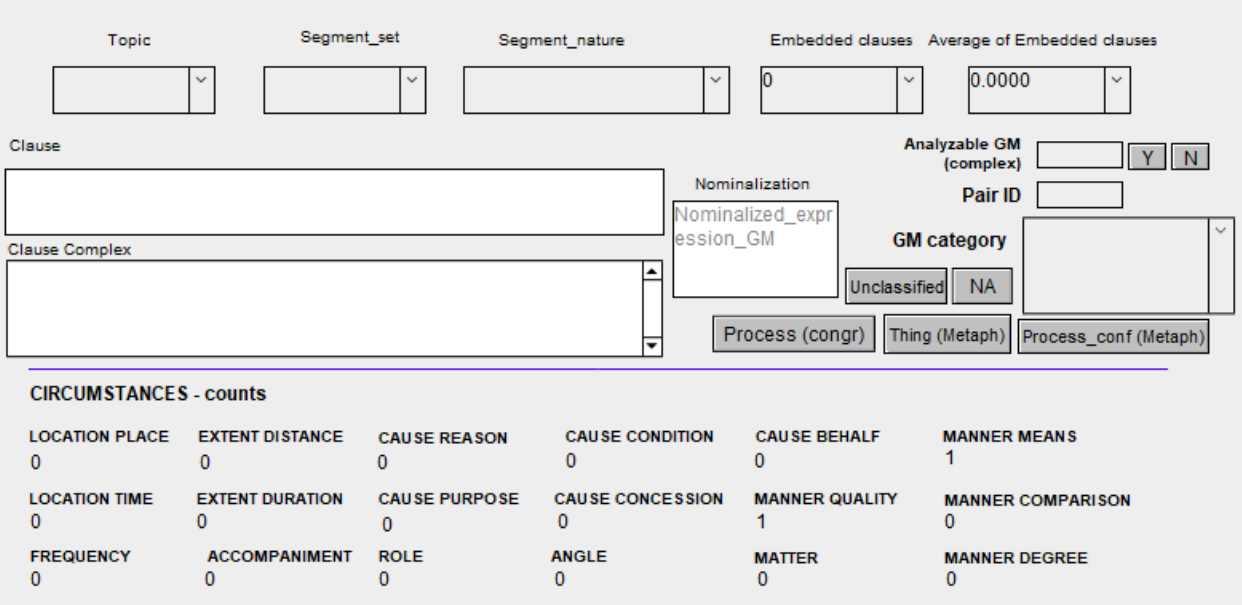


Figure 34 shows the categories in which the clauses were classified, for instance, the fields "embedded clauses", "Average of embedded clauses", "Analyzable GM (complex)" and "GM category". The field "embedded clauses" counts the number of embedded clauses in each

[42] The concept of "technicality" seemed promising at the beginning of this study, but compared to the evidence provided by other types of features, e.g., embedded as qualifiers in nominal groups, it was not enough to support this approach in association with experiential grammatical metaphor. The examples could not support the initial hypothesis of the relationship between technicality and this kind of grammatical metaphor. Therefore, the decision of focusing on more substantial evidence to draw relevant conclusions,

segment and the average number is shown in the field “Average of embedded clauses”. The field “Analyzable GM (complex)” counts the instances of grammatical metaphor in each clause. Finally, the field “GM category”, which stands for (ideational) Grammatical metaphor category, lists the categories used to classify the data and ascertain which categories would provide enough evidence. The categories are shown in Table 37.

Table 37

Analysis categories explained

Category	Meaning
UNCLASSIFIED	This clause was not classified yet.
NOT_APPLICABLE	This clause will receive no classification, as none of the categories fit.
free clause	This is a free clause.
bound clause	This is a bound clause.
thing	In this clause, the examined item is considered a “Thing”. This category matches one of the categories “simple: metaph: thing”, “simple: metaph: process”, “simple: metaph: quality” and “simple: metaph: circumstance” to indicate the less metaphorical agnated version
process	The examined item is considered a “Process” in this clause. This category is considered less metaphorical than “Simple: metaph: process”.
quality	In this clause, the examined item is considered a “Quality”. This category matches one of the categories “simple: metaph: thing”, “simple: metaph: process”, “simple: metaph: quality” and “simple: metaph: circumstance” to indicate the less metaphorical agnated version.
embedded	An embedded clause is realizing meanings in this clause.

simple: metaph: thing	In this simple clause, the examined item is considered a “Thing”. This category matches one of the categories “thing”, “process”, “quality” to indicate the more metaphorical agnated version.
simple: metaph: process	In this simple clause, the examined item is considered a “Process”. This category matches one of the categories “thing”, “process”, “quality” to indicate the more metaphorical agnated version.
simple: metaph: quality	In this simple clause, the examined item is considered a “Thing”. This category matches one of the categories “thing”, “process”, “quality” to indicate the more metaphorical agnated version.
simple: metaph: circumstance	In this simple clause, the examined item is considered a “Thing”. This category matches one of the categories “thing”, “process”, “quality” to indicate the more metaphorical agnated version.

Table 37 presents the description of each category used to obtain evidence of grammatical metaphor and the metaphoricity relation between some of them. For instance, a segment from Set A in which part of it was selected to be compared with another segment from Set B. The first selection was classified as “process” and the second as a “thing” in a more metaphorical simple clause (thus the classification “simple:metaph:thing”, as explained in Table 37), showing that the second segment is more metaphorical than the first.

Figure 35 shows the analysis of the first clause from segment pair 192.

Figure 35

Sysfan snapshot - the first clause of pair 192 analyzed

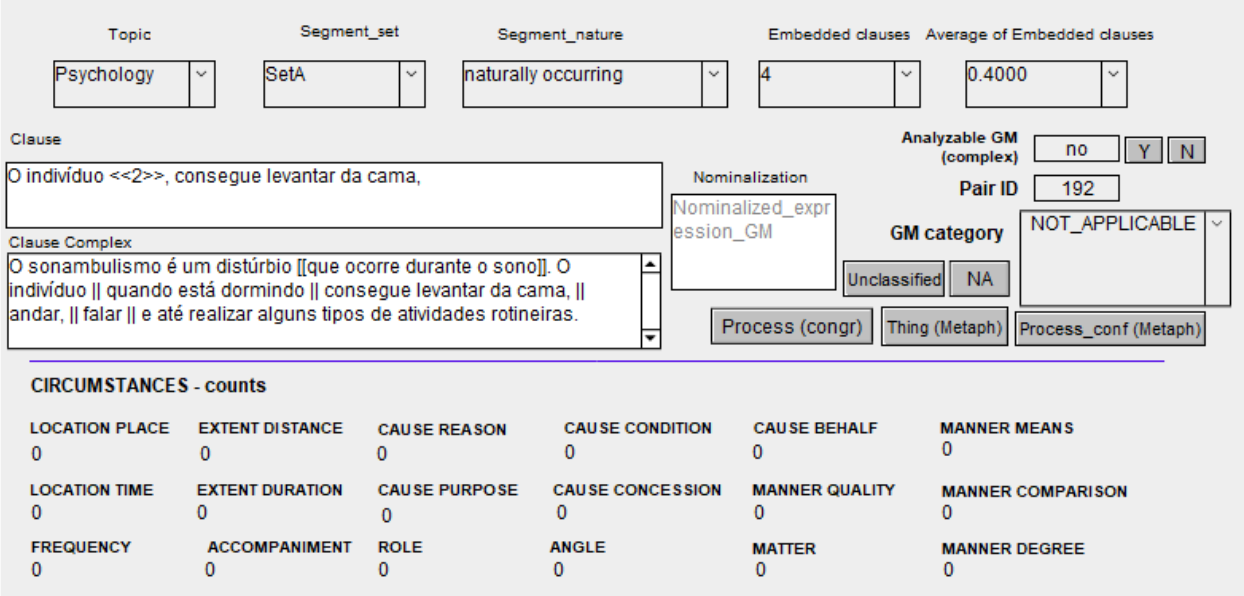


Figure 35 shows several fields that provide information about this clause, including the segment pair identifier (192), its domain, its segment set, its type (in the "Segment_nature" field), the number of embedded clauses in the segment, among other fields. The field "GM_category" is very important, because it indicates the evidence ("yes") or the lack of evidence ("no") of ideational grammatical metaphor in that specific clause.

Considering that all segments were classified similarly, the same for every clause complex and every clause, it was possible to obtain alignments containing the main classifications. Figure 36, which shows the same information as Figure 17, shows a full alignment of this sort, also with segment pair 192.

Figure 36

Sysfan snapshot - automatic alignment of segment pair

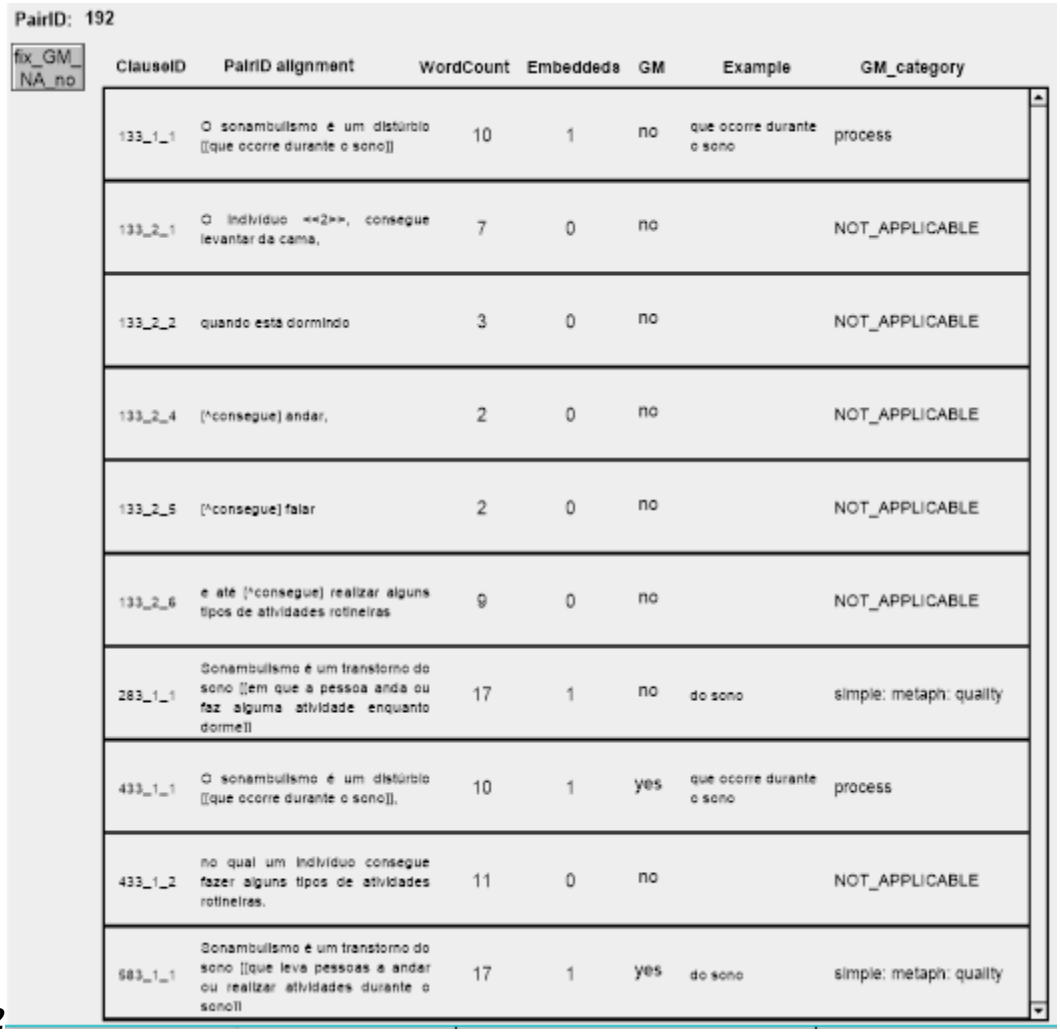


Figure 36 shows how the automatic alignment is presented in Sysfan, the information being presented in the columns “ClauseID”, “WordCount”, “Embeddeds”, “GM”, “Example” and “GM_category”. “ClauseID” informs automatic and unique indexing for each clause, “WordCount” the automatic measurement of the number of words, “Embedded” the number of embedded clauses, “GM” the verification if each segment contained evidence of grammatical

metaphor, and the last column, "GM_category" brings the classification of the clauses according to the categories shown in Table 37. This alignment, in contrast with most of the classification, is performed by Sysfan after the customizations described in section 3.2.1.1 and the segmentation and classification of the data.

The following steps in the next sections characterize the analysis of the lexicogrammar and the patterns indicating text simplification.

3.2.1.8 Clause metafunctional analysis

As Sysfan (Wu, 2000) was developed for the analysis of English texts drawing on SFL, the English SFL description was taken into account, mainly Halliday & Matthiessen (1994) and Halliday & Matthiessen (1999). However, as SFL is not yet fully described in Brazilian Portuguese and this thesis' corpus compiles texts in Brazilian Portuguese, it was necessary to take into account current descriptions in Brazilian Portuguese. As mentioned in Chapter 2, section 2.2.2., some of the SFL categories for Brazilian Portuguese were described by Araújo (2007), Figueredo (2007, 2011), Pagano, Ferreguetti e Figueredo (2011) Ferreguetti (2014, 2018), Figueredo, Pagano e Ferreguetti (2014), Sá (2016), Braga (2016), Monteiro (2016), Rosa (2017), A. Paula (2017), Alves (2017, 2018. These descriptions were taken into account and the systems of AGENCY, PROCESS, POLARITY, MOOD, THEME, and CONJUNCTION in English, already provided in Sysfan, were used for analysis.[43]

[43] Other relevant references on the account of Brazilian Portuguese can be found at SAL Project (*Projeto Sistêmica, Ambientes e Linguagens* – Project Systemic, Environments and Languages), from Federal University of Santa Maria (UFSM)/Brazil. Further information on https://www.ufsm.br/cursos/pos-graduacao/santa-maria/ppgletras/livro-e-books/.

Based on these descriptions and SIM-Pt, Figure 37, Figure 38, and Figure 39 show the procedures for the systemic analysis, which supplement the structural analysis presented in Figure 40 and Figure 41. These figures illustrate systems with more delicate options organized by the principle of realization.

Figure 37

Experiential metafunction categories classification

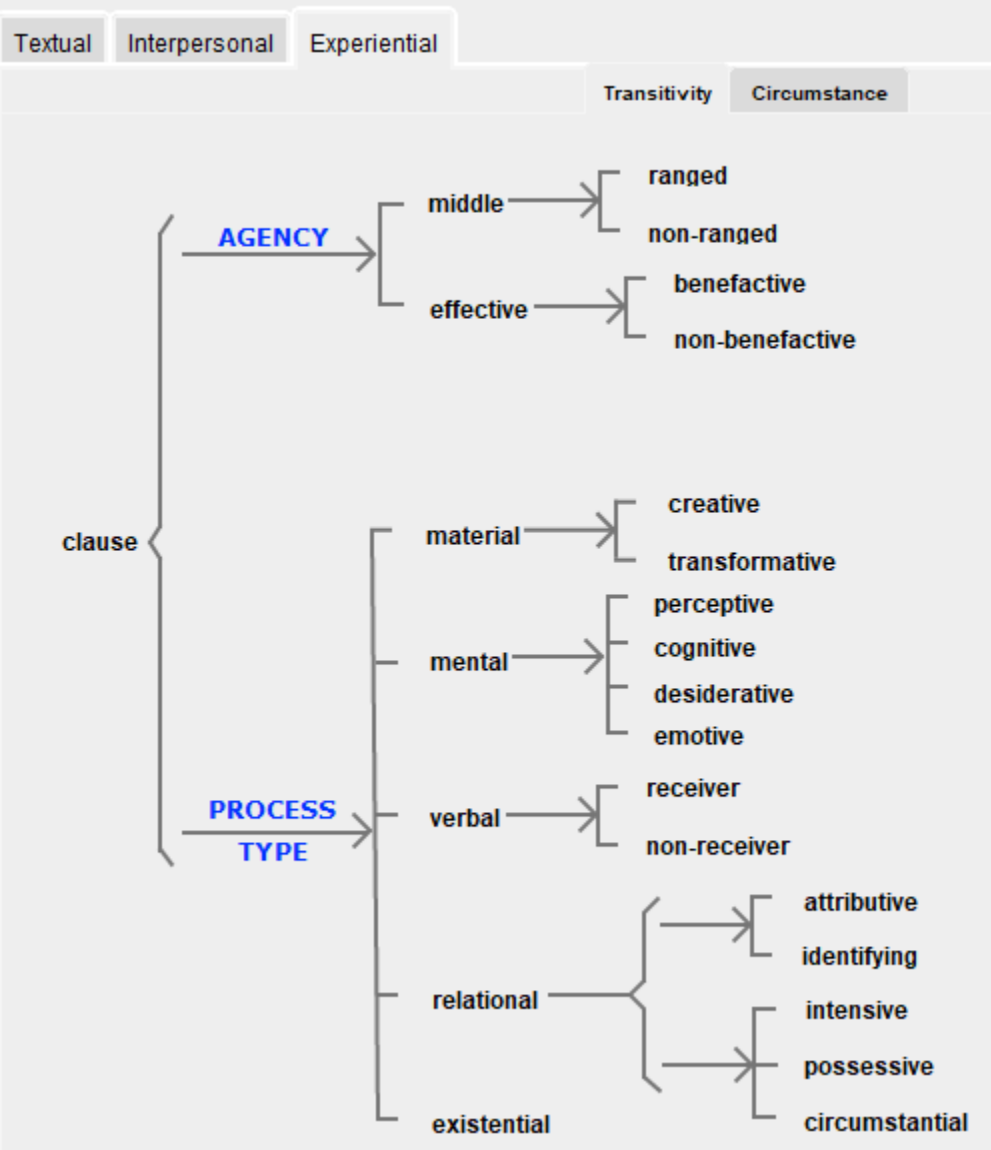


Figure 37 shows the systemic choices associated with the IDEATIONAL metafunction, especially the EXPERIENTIAL type, concerning the systems of AGENCY and PROCESS TYPE. The entry condition of these systems is a clause. The least delicate options of the system of the AGENCY are middle and effective, and, on a more delicate level, “ranged” and “non-ranged”, “benefactive” and “non-benefactive”. The more delicate options of the system of PROCESS TYPE options are Material, Mental, Verbal, Relational, and Existential. Each of these systems can be

further classified into more delicate options. The combination of these selections allows many combinations, which can co select with the options from the other systems, shown in Figure 38 and Figure 39, to ascertain the existence of linguistic patterns.

Figure 38

Interpersonal metafunction categories classification

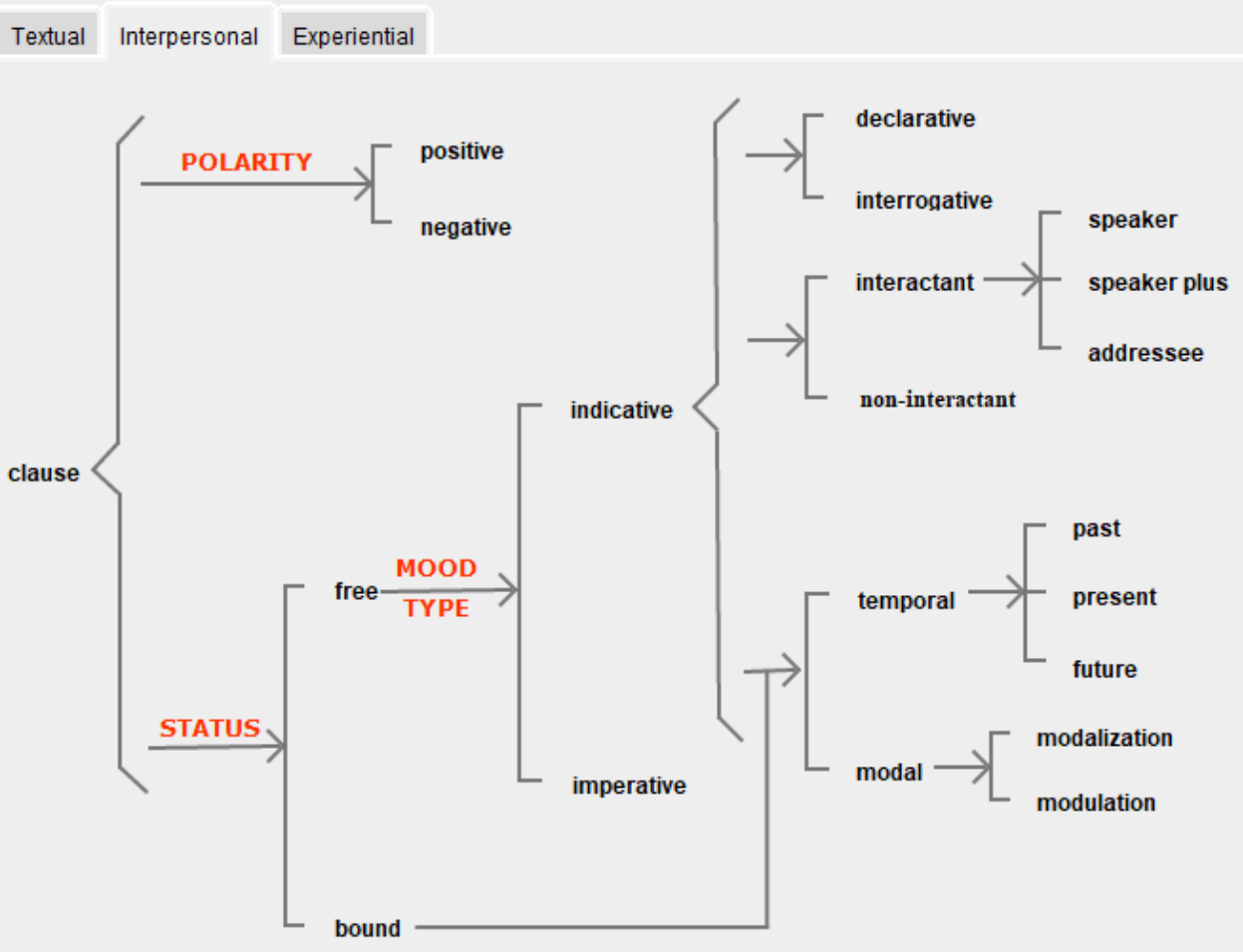


Figure 38 shows the systemic choices associated with the INTERPERSONAL metafunction, especially concerning the systems of POLARITY, STATUS, and MOOD. The options of the POLARITY system are “positive” and “negative”, while the STATUS system can be further classified into four levels of delicacy, associated with the system of MOOD TYPE.

Figure 39

Textual metafunction categories classification

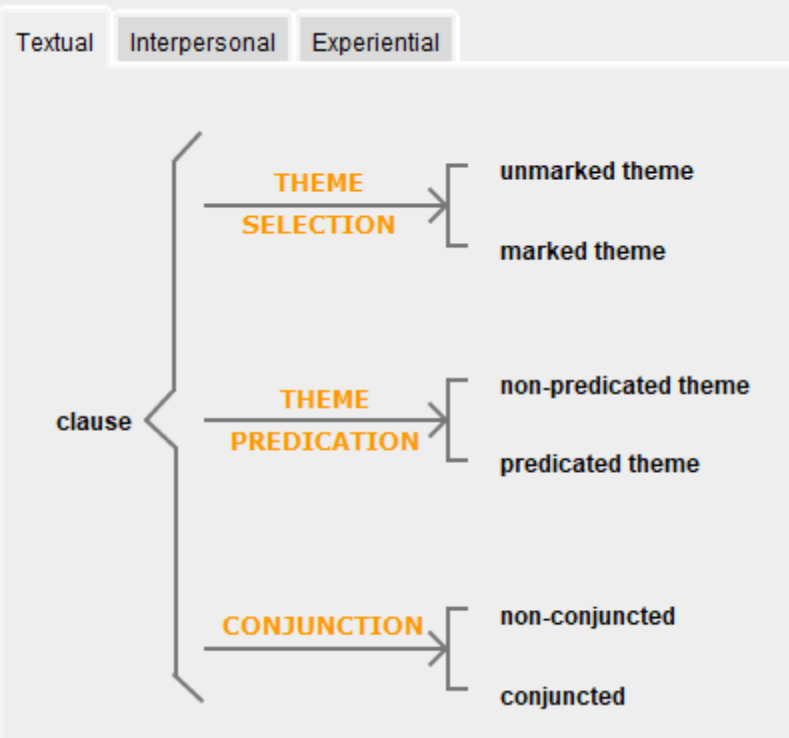


Figure 39 shows the systemic choices associated with the TEXTUAL metafunction, especially concerning the systems of THEME SELECTION, THEME PREDICATION, and CONJUNCTION. All of these systems have as an entry condition a clause. The least delicate options of the system of THEME, here referred to as THEME SELECTION, are unmarked and marked. The options of the system of THEME PREDICATION are non-predicated and predicated and the more delicate options of the system of CONJUNCTION are non-conjuncted and conjuncted. Combining these selections,

one per system, it is possible to reach many combinations, which can co select with the options from the other systems, shown in Figure 38 and Figure 39.

Figure 40 and Figure 41 show the procedures of the structural analysis in more detail, performed in parallel with the systemic analysis, to complement these results.

Figure 40

Sysfan snapshot of structural analysis

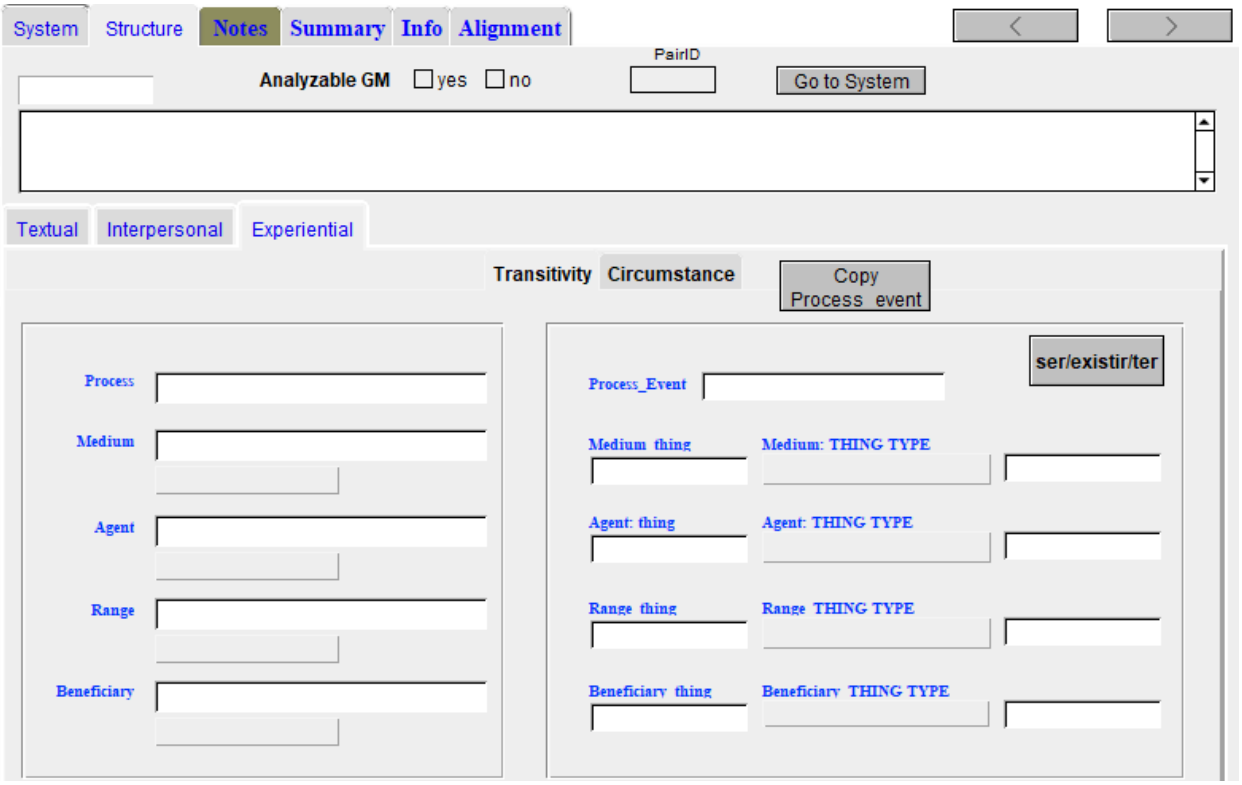


Figure 41 shows the structural (in this case, TRANSITIVITY) analysis of the clause, checking in the group rank which functions operate in the clause. For instance, the nominal groups can be analyzed on the left and the verbal group on the right.

Figure 41

Sysfan snapshot of structural analysis

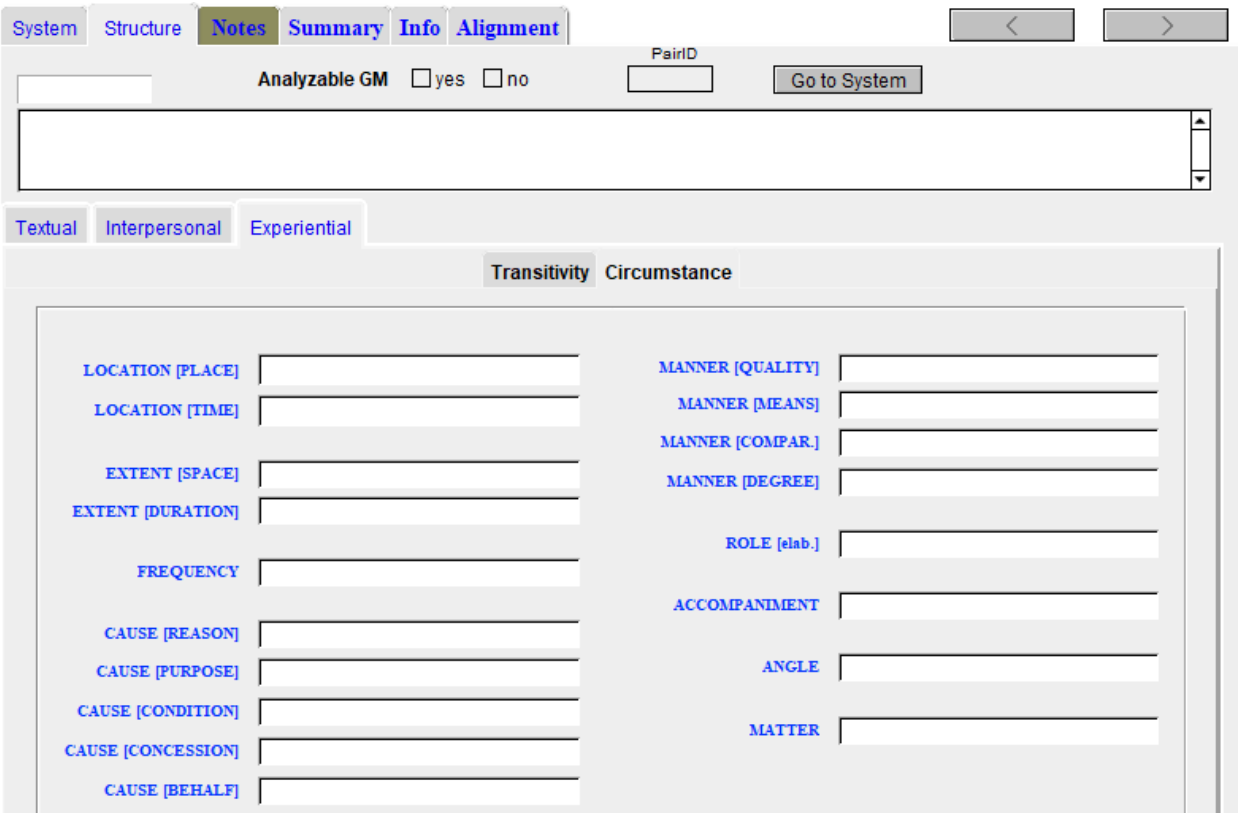
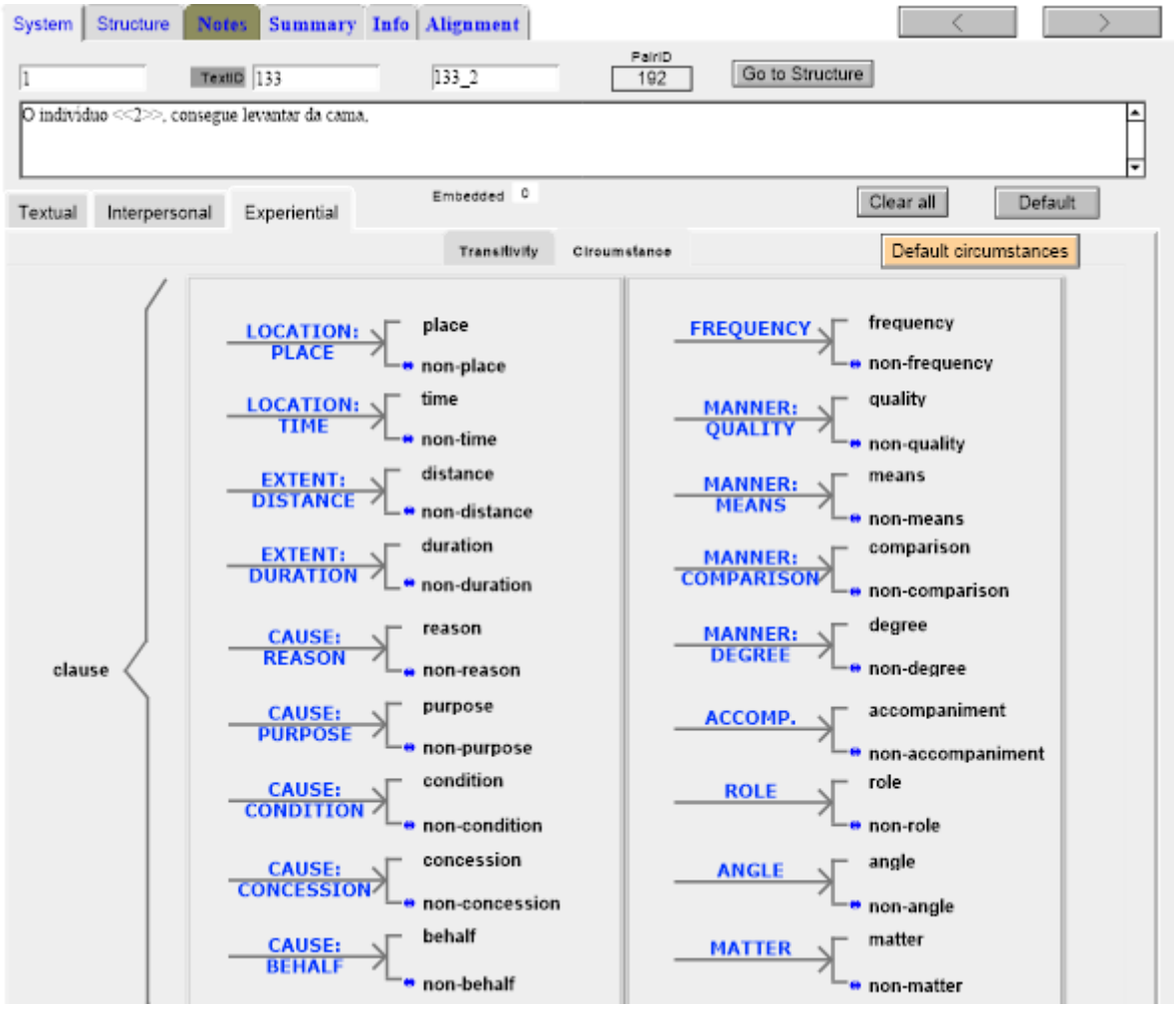


Figure 41 shows the circumstantial analysis in terms of a range of circumstantial types, in which the instances within certain clauses were annotated. The accounting of these instances is made on another screen, which requires the selection of the circumstance type. This selection is shown in Figure 42.

Figure 42

Circumstance types selected



The selection in Figure 42 consists of the systemic selection associated with the structural selection in Figure 41, complementing one another. Also, the systemic selection allows the user to count the frequency of each circumstance. This resource was used to assess if CIRCUMSTANCES would yield relevant results on the association between experiential grammatical metaphor and

text simplification, but, contrary to the positive indications from the preliminary study, it was not possible to obtain enough evidence. For this reason, this approach was not considered productive.

.

3.2.1.9 Extracting annotated categories frequencies

To obtain patterns from the annotation, it is necessary to prepare this great amount of information for the analysis. A summary is thus necessary, by counting the relative frequencies of each category from each set.

The following Figures show the way Sysfan automatically calculates the relative frequencies of each feature of the systems, after the manual classification drawing on SFL.

Figure 43

Sysfan snapshot of ideational metafunction classification

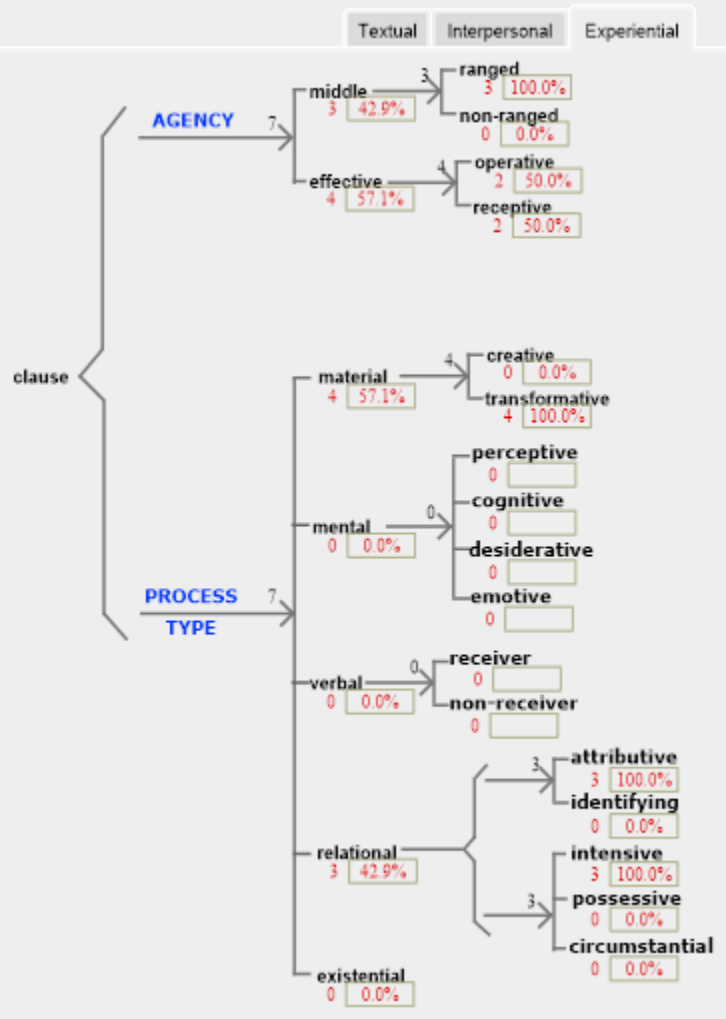


Figure 43 shows the classification according to the IDEATIONAL metafunction categories, in terms of the relative frequency of the systems of AGENCY and PROCESS. In the system of AGENCY, for instance, the options "middle" and "effective" sum up 100%, just like the least

delicate categories within these systems.

Figure 44

Sysfan snapshot of interpersonal metafunction classification

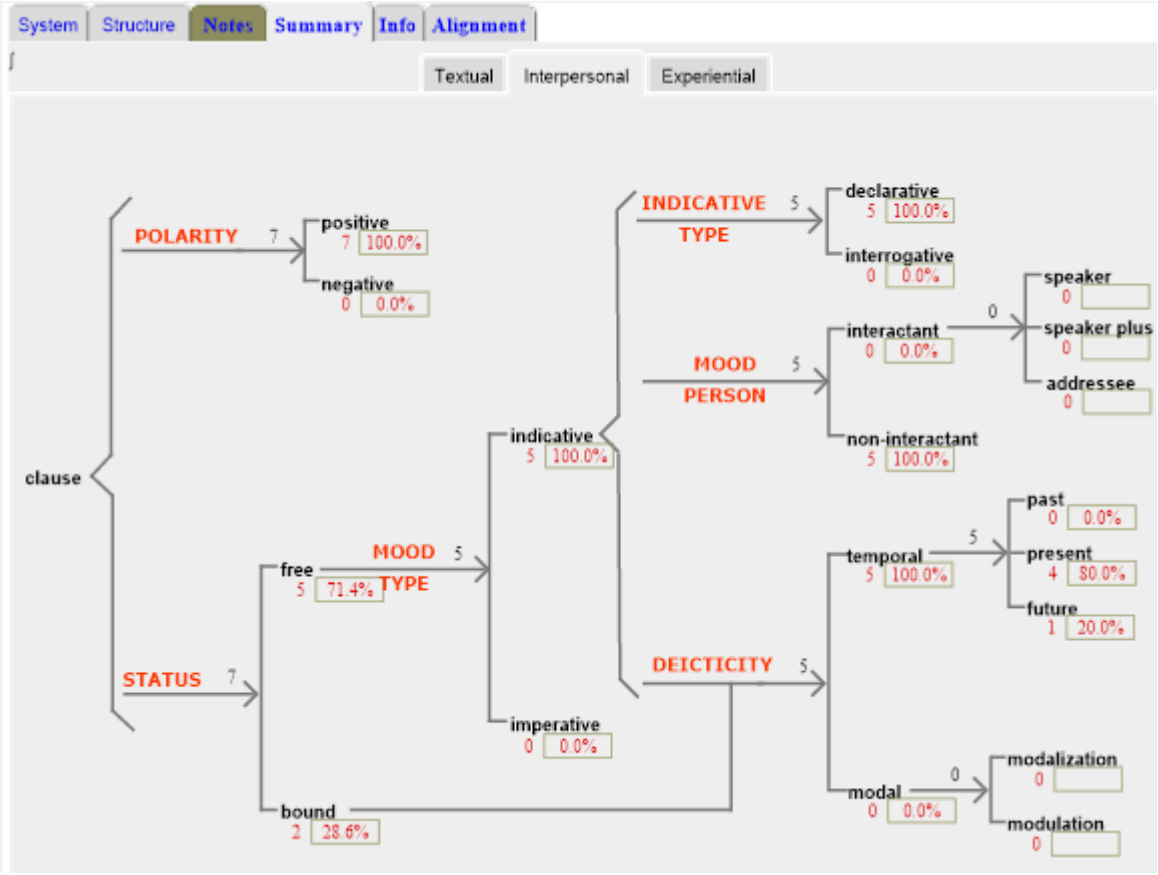


Figure 44 shows the classification based on the INTERPERSONAL metafunction categories, in terms of relative frequency. The systems illustrated are the systems of POLARITY, STATUS, MOOD, and their more delicate systems INDICATIVE TYPE, MOOD PERSON, and DEICTICITY.

Figure 45

Sysfan snapshot of textual metafunction classification

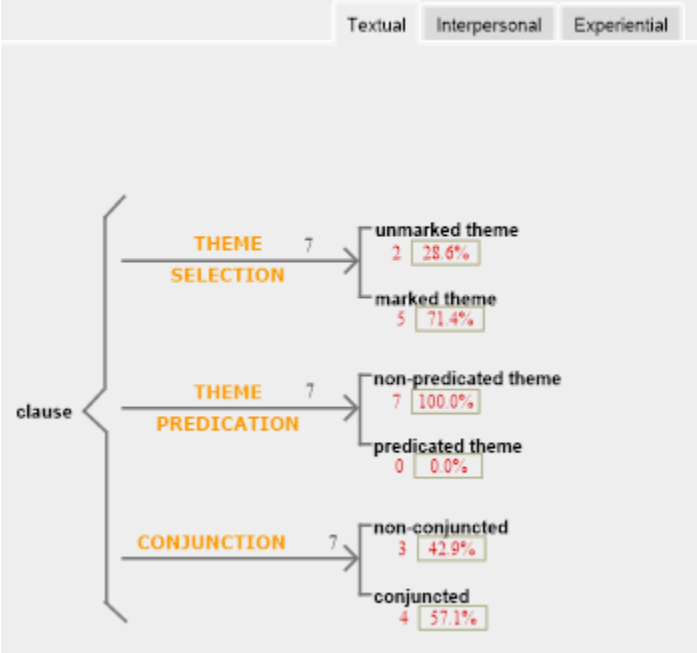


Figure 45 shows the classification regarding the TEXTUAL metafunction categories, in terms of relative frequency. These categories are associated with the systems of THEME, THEME PREDICATION, and CONJUNCTION, in which two options for each system are presented at a less delicate level.

The final steps of the analysis presented in the next sections describe the procedures to export data onto a spreadsheet (section 3.2.1.10), to carry out a manual analysis of the shifts (section 3.2.1.11), and to analyze the patterns in the data (section 3.2.1.12) to reach conclusions on the text simplification phenomena.

3.2.1.10 Exporting data onto a spreadsheet

One of the final steps in the analysis was obtaining the most frequent patterns to select the most relevant ones to obtain evidence of experiential grammatical metaphor. The following tables and figures show the procedures to find these patterns, which consist of three steps to export the data and a fourth one that is importing the spreadsheet and counting their patterns using an *R* script. These steps are listed below:

1) Choosing the file that will be exported
2) Choosing the worksheet details
3) Choosing the fields to be exported
4) Counting the patterns using an *R* script

From these steps, only the second is considered optional, though it is recommended if the user wishes to collect extra metadata for the analysis or organize the files more efficiently. These four steps are described in detail in the next topics.

1) Choosing the file that will be exported

The first step was to create a file that will be exported, as shown in Figure 46.

Figure 46

Sysfan snapshot - choosing a folder in the exporting menu

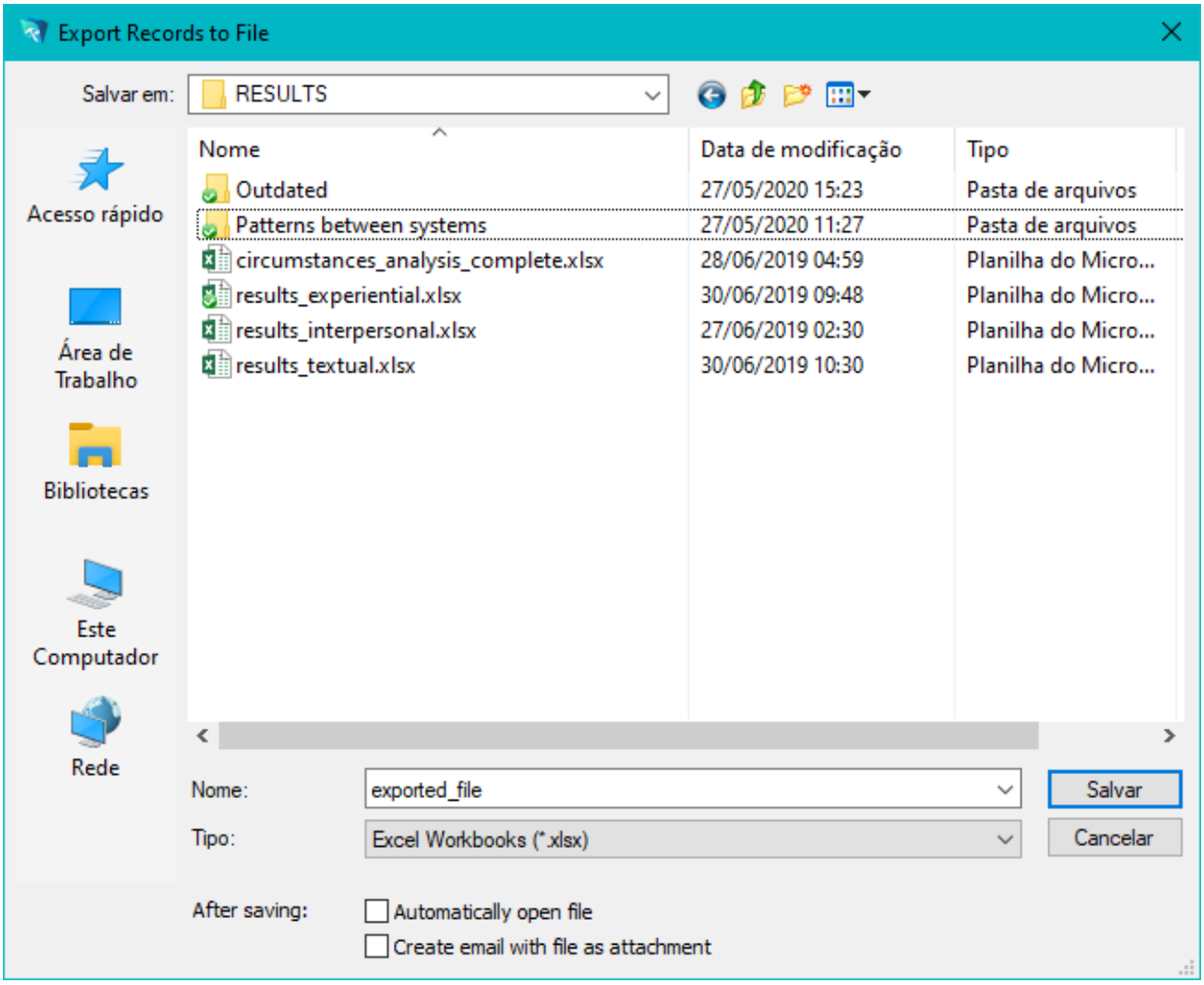


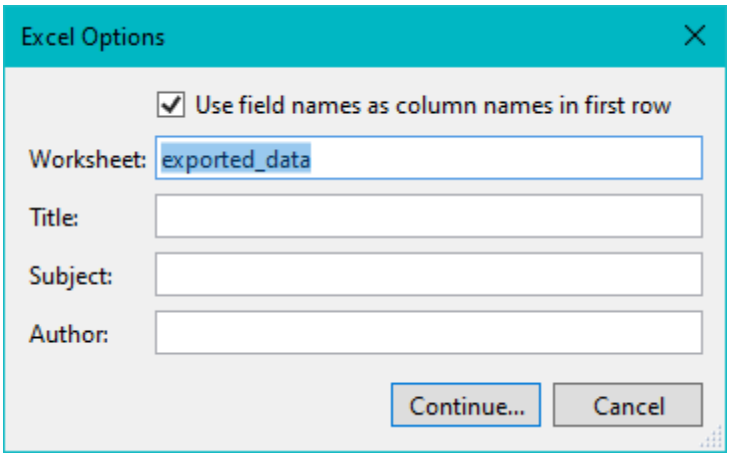
Figure 46 shows that the folder “RESULTS” was chosen to store the file “exported_file.xlsx” with the results that will be selected.

2) Choosing the worksheet details

The next step, in Figure 47, which is optional, was choosing the name of the worksheet that will store the data.

Figure 47

Sysfan snapshot - defining the worksheet details



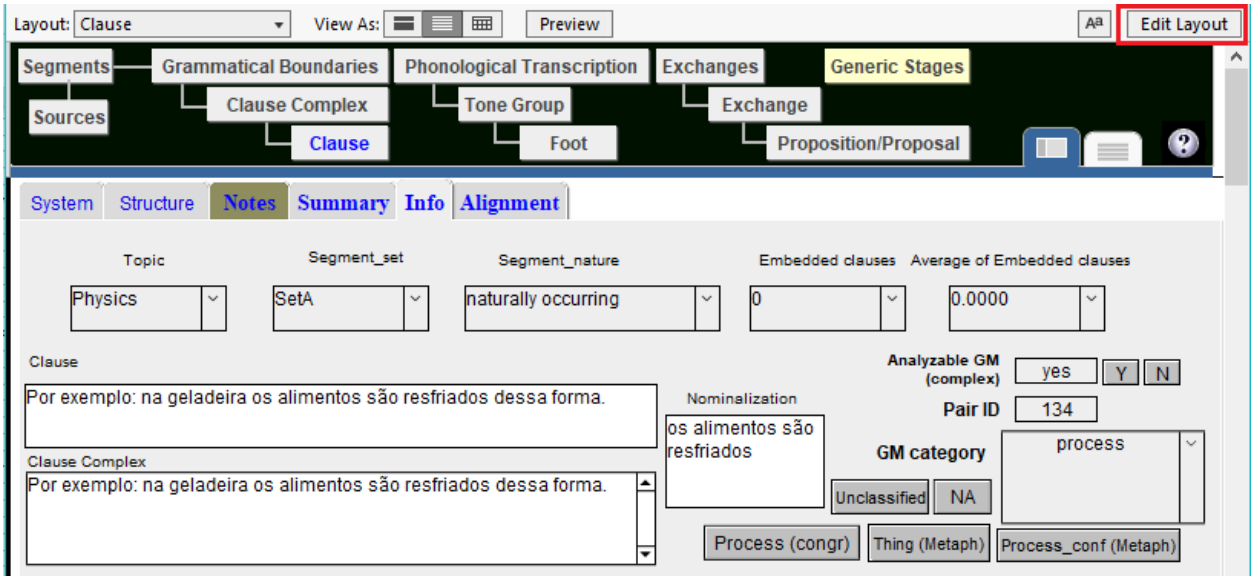
Although this step is optional, it might be relevant for certain users to make explicit the worksheet name in which the data will be stored. This is shown in Figure 47, in which the chosen name is “exported_data”.

3) Choosing the fields to be exported

The final and most crucial step was to choose the database fields that will be selected out of the fields of all tables. This process is usually guided by the interface, which lets you access the layout details and the whole database in the “Edit layout” button, highlighted in red in Figure 48.

Figure 48

Sysfan snapshot - Edit layout button



Note. The red square showing the "Edit layout" was added later in the image for emphasis.

This button highlighted in Figure 48 allows the user to access the details of the layout and the database to make all the necessary adaptations. With this, it becomes more feasible to obtain the name of the fields of interest. These fields are explained in detail in Table 38.

Table 38

Descriptions of the exported fields

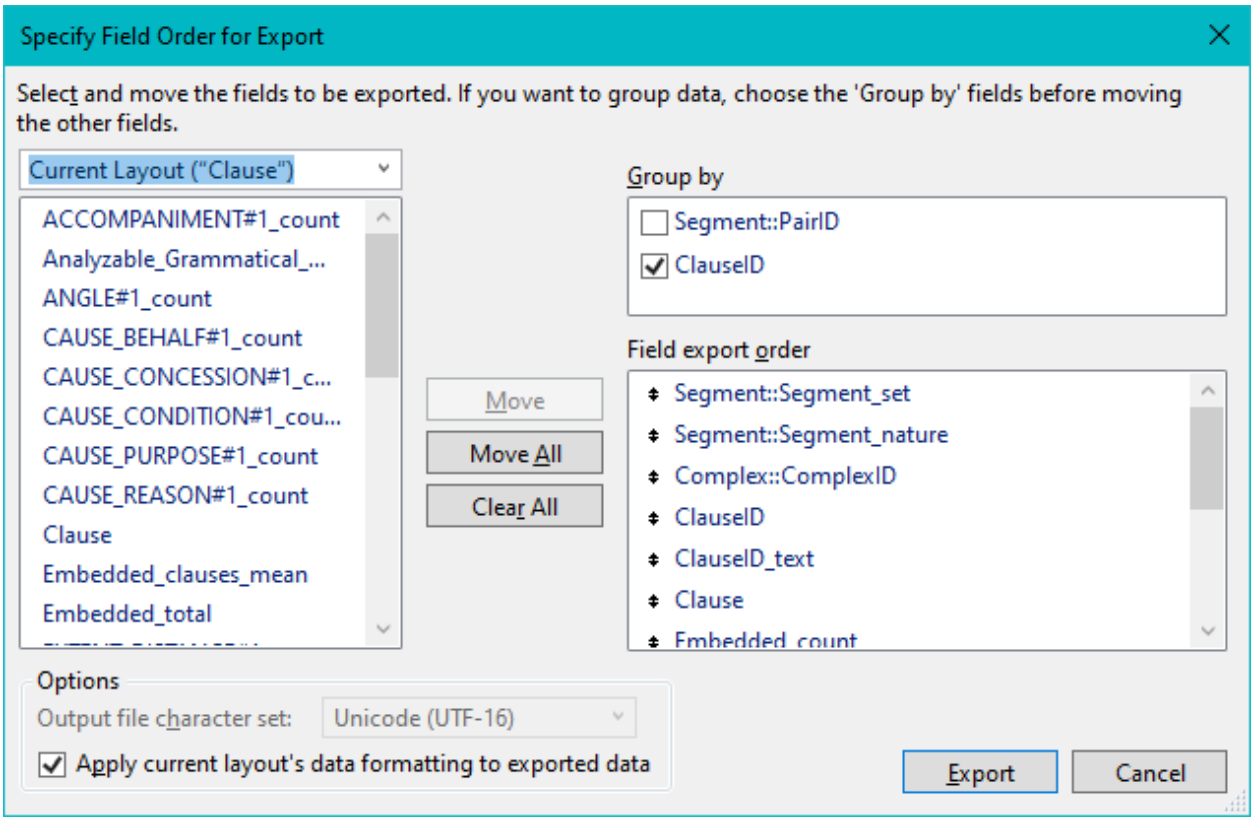
Column	Description
Text::Segment_Set	Text set (Set A, Set B, Set C, or Set D)
Text::Segment_nature	Text type (naturally occurring or manually constructed)
ComplexID	Unique identifier (ID) of the clause complex

ClauseID	Unique identifier (ID) of the clause
ClauseID_text	Unique identifier (ID) of the clause in text format (underline splits the ID into parts)
ClauseNo	Number of the clause within the complex (1 equals first, 2 equals second, etc)
Clause	The clause, if it is a finite clause
Embedded_count	Number of embedded clauses in the clause
Process_event	The Event of the verbal group
Agent	The Agent of the finite clause
Medium	The Medium of the finite clause
Range	The Range of the finite clause
Beneficiary	The Beneficiary of the finite clause

Based on the description of these fields in Table 38, it is possible to predict what the result will be after choosing the adequate fields based on the information which is of interest. In this case, the interest was on the classification of the clauses according to the systemic and structural categories to find evidence of experiential grammatical metaphor. This way, the exporting interface from Figure 49 can be filled with the necessary details.

Figure 49

Sysfan snapshot - grouping and exporting fields



In this interface, the fields can be selected from one of the tables (for instance, Segment) and grouped, for example, by ClauseID (id of the clauses) simply by selecting the field. Then, the fields can be chosen and ordered. The results are exported as an electronic spreadsheet with all the fields selected, as shown in Figure 50.

Figure 50

Example of the spreadsheet with the exported data

Text::Segment_set	Text::Segment_nature	ComplexID	ClauseID	ClauseID_text	ClauseNo	Clause	Embedded_count	Process_Event	Agent	Process	Medium	Range	Beneficiary
SetA	naturally occurring	1_1	111	1_1_1	1	Forças de contato: [ocorren	0	ocorrer		ocorrem	Forças de contato		
SetA	naturally occurring	1_1	112	1_1_2	2	Quando há contato direto e	0	haver		há	contato direto entre dois corpos		
SetA	naturally occurring	2_1	211	2_1_1	1	Forças de campo: [ocorrem	0	ocorrer		ocorrem	Forças de campo		
SetA	naturally occurring	2_1	212	2_1_2	2	Quando a atuação da força	0	ocorrer		ocorre	a atuação da força		

Note. These are the first four lines of data from the spreadsheet, which contains 973 observations, one per clause.

Figure 50 shows that the fields described in Table 38 and selected in Figure 49 were exported and the information is now available in this spreadsheet. This spreadsheet, in the format *.xlsx*, can then be used as input for a further analysis using R software, aiming at discovering the patterns.

3.2.1.11 Counting the frequency of the categories from the spreadsheet with the use of an R script

With the use of an R script[44], the spreadsheet from Figure 50 was imported into the R environment and the lexicogrammatical configuration of each clause was determined by combining the observations from the fields “Agent”, “Process”, “Medium”, “Range” and “Beneficiary”. One example of such configuration would be “Participant_Process_Range”, in relational clauses with the process “to be”:

Table 39 shows an example of semantic configuration from Set A, segment pair 192.

[44] This script is available at <https://github.com/rodrigoacastro/Castro2020>, more specifically in <https://github.com/rodrigoacastro/Castro2020/tree/master/counting_word_order_seqs>.

Table 39

Example of semantic configuration

O sonambulismo	é	um distúrbio [[que ocorre durante o sono]].
participant	process	range
Somnambulism	is	a disorder that takes place during someone's sleep

Note. This is the segment from segment pair 192, Set A.

Table 39 brings an example with a systemic gloss of the segment retrieved from pair 192, Set A, and its translation. This is one example of the configuration "Participant_Process_Range", one of many that were accounted for to obtain the results.

Table 40

Sample of the configuration sequences for each set

Set A and Set C		Set B and Set D	
Sequence	Frequency	Sequence	Frequency
Process_Medium_Range	111	Process_Medium_Range	149
Agent_Process_Medium	54	Process_Medium	37
Process_Medium	42	Agent_Process_Medium	25
Agent_Process	13	Agent_Process	4
Agent_Process_Range	4	Agent_Process_Medium_Beneficiary	1
Agent_Process_Medium_Range	2	Agent_Process_Range	1
Agent_Process_Beneficiary	1		
Process_Medium_Range_Beneficiary	1		

Table 40 shows the frequency of different sequences of semantic functions, which were manually analyzed based on the procedures described in the next section.

3.2.1.12 Manual analysis of the shifts

Firstly, this step also followed the first three steps proposed in the earlier section, section 3.2.2.9:

1) Choosing the file that will be exported

2) Choosing the worksheet details

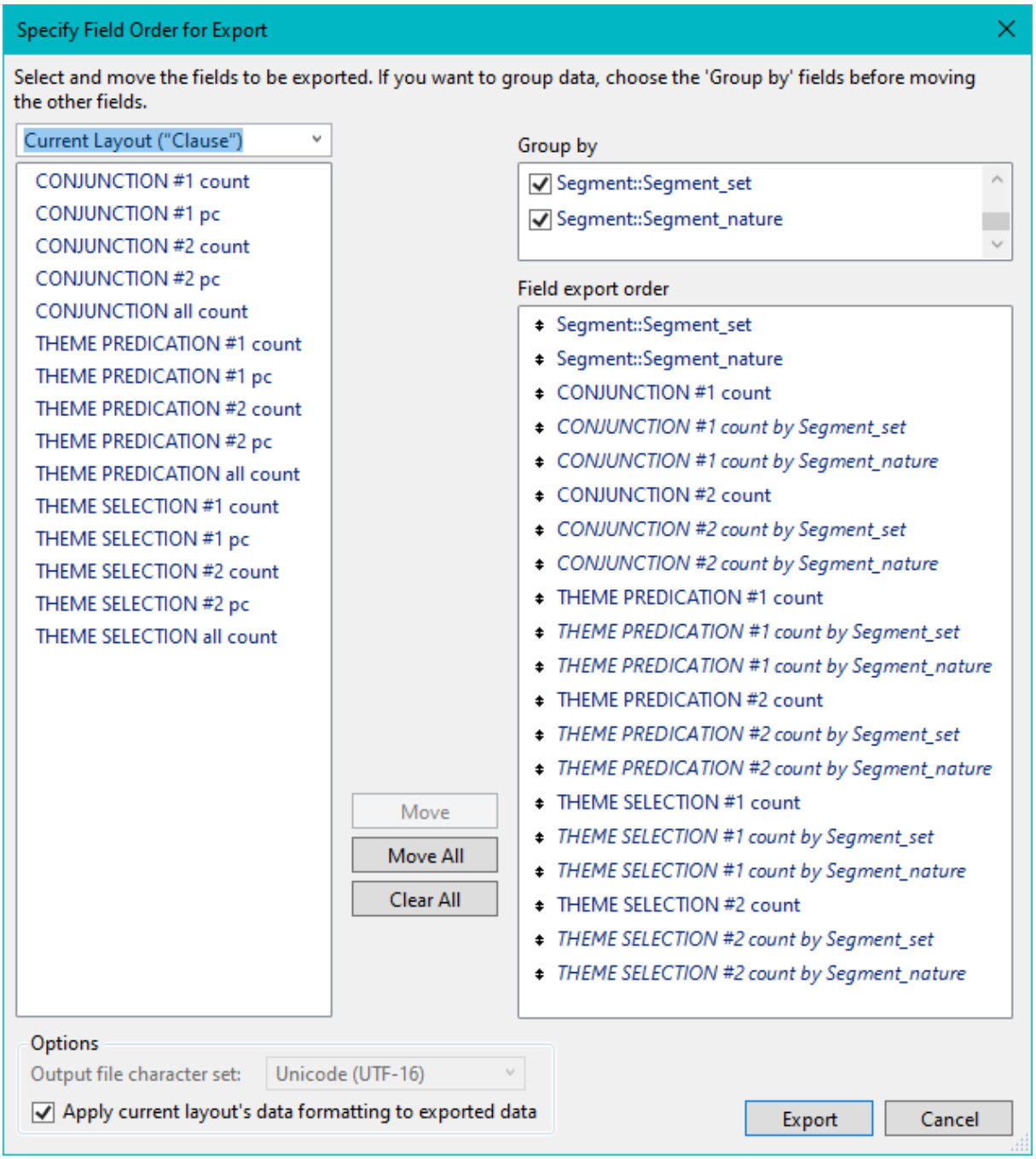
3) Choosing the fields to be exported

The difference between the previous exporting process and the current one was that the field selections in step 3 were not the same, as the systemic patterns were retrieved from the fields associated with the classification shown in section 3.2.1.9.

The field selection for exporting summarization results in Figure 51 indicates that the data was grouped by set (Set A, Set B, Set C, Set D) and type (“nature” in Sysfan).

Figure 51

Sysfan snapshot - exporting summarization results



In Figure 51, the Sysfan fields on the left were selected on the right to be exported, grouping the data by set and type ("nature", in Sysfan).

After the exporting process, the tables in *.xlsx* format were organized and processed in R[45]. One example of a table ascertaining the trends or patterns in the data is in Table 41, showing the absolute and relative frequency of systemic categories.

[45] This script is available at https://github.com/rodrigoacastro/Castro2020.

Table 41

Example of systemic patterns associated with AGENCY system and Relational PROCESS

SELECTION	**Assessed system**	**SET A and SET C**				**SET B and SET D**			
		Middle		**Effective**		**Middle**		Effective	
Marked theme	Agency	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		35	39.33	54	**60.67**	45	**75.00**	15	25.00
		Attributive		Identifying		**Attributive**		Identifying	
Marked theme	Relational Process	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		14	**51.85**	13	48.15	23	**58.97**	16	41.03

Note. The highest values are highlighted in bold to summarize the information. The complete results are available in Appendix D for a more detailed view.

Taking into account the sampled data systemic patterns associated with the AGENCY system and Relational Process, Table 41 shows the tendencies associated with the AGENCY system and Relational Process. The highlighted categories show that clauses from Set A and Set C tend to be classified mostly as “effective”, rather than “middle”. Also, the clauses that are

relational and belong to Set A or Set C tend to be attributive rather than identifying. The opposite, in both bases, tends to take place in the clauses from Set B and Set D.

The next and final step is interpreting the patterns to associate the structural and systemic patterns with the pieces of evidence of experiential grammatical metaphor.

3.2.1.13 Analyzing the structural and systemic patterns and experiential metaphoricity

This last step associated the patterns discovered in the data with the criteria for identifying experiential grammatical metaphor as indicators of text simplification. By analyzing the extracted patterns and contrasting them in light of the trends found (e.g., effective clauses tended to take place more often in simpler instances), these pieces of evidence formed the picture of how experiential grammatical metaphor could determine the degree of simplification in a segment or a text; in other words, how selection associates higher metaphorization with higher text complexity. An example of this manual analysis is shown in Table 42 and Table 43, based on the electronic spreadsheets used to perform the analysis, which will be analyzed further in the results chapter, section 4.5.2.2.

Table 42

Alignment of naturally occurring segments from segment pair 4

Pattern 1 - Segment pair 4 - Set A - Naturally occurring segment		
Forças de campo:	**[^são]**	aquelas [[que agem sobre os corpos]]
Field forces	*are*	*forces [[that act between two bodies]].*
participant	relational process	participant
Pattern 2 - Segment pair 4 - Set C - Manually constructed segment		
Forças de campo:	**[ocorrem]**	Quando a atuação da força ocorre a distância.
Field forces	*act*	*when the action of the force takes place*
participant	existential process	circumstance

Note. The explicit lexicogrammatical functions from the first clause were conserved in both clauses to allow the comparison in this example, even if the slot is empty in the other segment.

Table 43

Alignment of naturally occurring segments from segment pair 59

Pattern 1 - Segment pair 59 - Set A - Naturally occurring segment					
ou seja	uma mesma molécula de DNA [[submetida à ação de uma enzima]]	**é**	sempre	**cortada**	nos mesmos pontos
that is,	*The same DNA molecule submitted to the action of an enzyme*	*is*	*always*	*cut*	*in the same point*
	participant	relational process 1	circumstance 1	relational process 1	circumstance 2
Pattern 2 -Segment pair 59 - Set C - Manually constructed segment					
	os cortes	**ocorrem**			nos mesmos pontos
	the cuts	*take place*			*in the same point*
	participant	existential process			circumstance 2

Note. The explicit lexicogrammatical functions from the first clause were conserved in both clauses to allow the comparison in this example, even if the slot is empty in the other segment.

As Table 42 and Table 43 reveal, even though the structure of the segments seems to be similar taking into account the ordering of some elements, there are some differences between them. For instance, in Table 42 the relational process matches the existential process in the structure, the same that takes place in Table 43, in which there is also a conversion of passive into active voice. Other patterns such as this one could be found in the data as well (cf. section 4.2.2 from the Findings chapter).

The next section describes the measures taken to handle the exceptions found in the classification, given that unexpected behaviors that were not compatible with the rest of the results were not accounted for in the results, though they were initially taken into account. These exceptions were not analyzable cases and controversial cases.

3.2.2 Exceptions

Taking into consideration that distinct patterns of grammatical metaphor realization were observed, the probabilistic nature of language must be pointed out. In other words, not all instances presented such evidence, and the examples that could not be fully explored by this analysis were considered “special cases”, is divided into unclassified cases (for the analysis of experiential grammatical metaphor) and controversial cases (a few cases in which Set A and Set C proved to be, after the analysis, more metaphorical than Set B and Set D).

3.2.3.1 Cases that were not analyzable

The corpus initially consisted of 200 segments from three distinct fields (Physics, Biology, and Psychology), from which 150 segments were selected to be analyzed. Yet, not all pairs presented evidence of ideational metaphor. Some of them suggested that another type of grammatical metaphor, interpersonal metaphors, was in use instead. In other cases, even if there was evidence of experiential grammatical metaphor, it was still not possible to analyze some of them because the criteria used in this thesis were not able to explain the differences between the segment pairs in these cases. Due to these reasons, those two special cases are explained below.

- Dissimilarities between the main criteria for performing text simplification and metaphoricity degree, as shown in Table 44.

Table 44

Example of segment pairs showing dissimilarities between text simplification strategies and metaphoricity criteria

Pair	Type	Set	Example	English Translation
42	Naturally occurring	Set A	O androceu é constituído pelos estames, estruturas onde se formam os grãos de pólen, o gameta masculino da flor;	The androceutical is composed of stamens, structures in which the pollen grain, the male gamete of the flower, is formed.
42	Manually constructed	Set C	ANDROCEU: [é o] conjunto dos órgãos masculinos [[formados pelos estames]].	Androceutical: [it is] the set of male organs [[formed by stames]].
42	Naturally occurring	Set B	O androceu é o aparelho genital masculino e é formado pelos estames.	The androceutical is the male genital tract formed by stamens.
42	Manually constructed	Set D	Androceu é o conjunto de órgãos masculinos e é constituído pelos estames, nos quais ficam armazenados os gametas masculinos da flor.	Androceutical is the set of male organs formed by stamens, in which the male gametes of the flower are stored.

Taking into consideration that most examples illustrate evidence of different degrees of metaphoricity in which Set A and Set C were less metaphorical than Set B and Set D, it was expected that the same would take place in Table 44. Yet, it was not the case.

Taking into account the criteria proposed in this thesis (the occurrence of class shifts, rank shifts, and embedded clauses) the segments from Set A and Set C are more complex than segments from Set B and Set D. The first reason for that is the occurrence of the qualifier "estruturas onde se formam os grãos de pólen, o gameta masculino da flor;" (*structures in which the pollen grain, the male gamete of the flower*) in the complement realized by the prepositional phrase "pelos estames, estruturas onde se formam os grãos de pólen, o gameta masculino da flor;" (*of stamens, structures in which the pollen grain, the male gamete of the flower*) in Set A, which is not realized in Set B. This means that, according to this criterium, Set A is more metaphorical than Set B. The second reason is the occurrence of an embedded clause "formados pelos estames" (*formed by stamens*) in Set C instead of the free clause "e é constituído pelos estames" (*and is formed by the stamens*) in Set D.

Considering another set of criteria, for instance, from text simplification studies (see section 2.2.1), especially the criteria "addition of explanation or examples" (Hwang et al, 2015), Set A is simpler than Set C and Set B is simpler than set D. In both cases, the reason is the addition of more information in the segment. In Set A, compared to Set B, it was added "estruturas onde se formam os grãos de pólen, o gameta masculino da flor;" (*structures in which the pollen grain, the male gamete of the flower*)" and in Set D, compared to Set C, it was added "nos quais ficam armazenados os gametas masculinos da flor." (*in which the male gametes of the flower are stored*).

In sum, Table 44 presents segment pair 42 as an example in which, according to text simplification strategies, Set A and Set D are less metaphorical than Set B and Set C; yet, based on metaphoricity analysis (considering mainly nominalization through class shifts, rank shifts and use of embedded clauses), Set A and Set C are more metaphorical. In general, though, Set C and Set D were more metaphorical than Set A and Set B.

This lack of consistency between classifications based on distinct criteria leads to the conclusion that the analysis of the same occurrences can lead to different conclusions.

- The occurrence of other types of grammatical metaphor

  The analysis shows at least one occurrence of another type of grammatical metaphor: the interpersonal grammatical metaphor of modality. Table 45 shows an example of this type.

Table 45

Examples with interpersonal grammatical metaphor

Pair	Type	Set	Example	English Translation
13	Naturally occurring	Set A	A teoria mais aceita é que seja uma síndrome autoimune.	The most accepted theory is that it is an autoimmune disease.
13	Naturally occurring	Set B	**Teoria Imunológica: Admite** que o vitiligo é uma doença auto-imune pela formação de anticorpos antimelanócitos.	**Immunological theory: It admits** that vitiligo is an autoimmune disease due to the formation of anti melanocyte antibodies.
13	Manually constructed	Set C	O vitiligo é uma doença autoimune, ou seja, o corpo ataca a si próprio.	Vitiligo is an autoimmune disease, that is, the body attacks itself.

13	Manually constructed	Set D	Segundo a teoria imunológica, o vitiligo caracteriza-se como um doença auto-imune devido à formação de anticorpos antimelanócitos.	According to the immunological theory, vitiligo is characterized as an autoimmune disease due to the formation of anti melanocyte antibodies.

Table 45 indicates that in segment pair 13, the reason for the interpersonal grammatical metaphor is the usage of "Teoria Imunológica: Admite" **(***Immunological theory: admits"*) followed by a projected clause in Set B. This structure indicates that the speaker is dissimulating that this is his/her own opinion, as there are many ways to express or, as in this case, dissimulate one's opinion (Halliday & Matthiessen, 2014, p. 689). In contrast, in the segment from Set D, the PARTICIPANT "Teoria imunológica" (*immunological theory*) is part of a Circumstance "Segundo a teoria imunológica" (*according to the immunological theory*). In Set A and Set C, there is no realization of this PARTICIPANT. The main contrast is between Set B and Set D; in the latter, this PARTICIPANT is a realization of an interpersonal metaphor of modality, which, in this case, is more relevant to contrast these examples than other kinds of metaphor. For this reason, this instance was not accounted for in the thesis, though it could be explored further not only in this research but also in similar studies.

This chapter presented the compilation and the analysis methodologies used in SIM-Pt to achieve the results described in the next chapter.

Chapter 4 Findings

This chapter outlines the findings of this thesis regarding features indicating text complexity that may potentially inform text simplification. Evidence provided by the corpus analysis points to patterns of semantic configuration, class, and rank shifts analyzed to establish a relationship between text complexity and experiential grammatical metaphor. All examples that illustrate the findings were retrieved from the corpus and are illustrated following segment-to-segment correspondence.

4.1 Text complexity and experiential grammatical metaphor

As discussed in section 2.3, according to Steiner (2004), there is a relationship between text complexity and experiential grammatical metaphor. Furthermore, as discussed in section 2.7, Ravelli (1999) describes two types of complexity – lexical and grammatical complexity – which can indicate how compact or intricate ideational meanings are realized. These are measurable through lexical density and grammatical intricacy and can be associated with the metaphoricity level of texts and segments.

Results from average lexical density and average grammatical intricacy measurements from each set of the sample retrieved from the SIM-Pt corpus, as provided in Table 46, show that

these measures are associated with the degree of metaphoricity, according to Ravelli's (1999) hypothesis.

As explained in section 3.2.1.6, Set A and Set B consist of naturally occurring segments, collected from websites providing science texts, and Set C and D consist of manually constructed segments, which are rewritings of the naturally occurring segments. The pairing between Set A and Set C is because they present a lower metaphoricity degree than Set B and Set D. This is shown in Table 46.

Table 46

Lexical density and grammatical intricacy measurements for all segment sets

Nature	Set	Average lexical density (ALD)	Average grammatical intricacy (AGI)	Metaphoricity degree
Naturally occurring	Set A	9.92	**1.67**	Lower
	Set B	**11.42**	1.54	Higher

Manually constructed	Set C	8.44	**1.64**	Lower
	Set D	**10.74**	1.41	Higher

Note. The values in bold were selected because they are higher than their counterparts (from Set A and Set C or Set B and Set D). All values were obtained using Sysfan.

Table 46 shows that naturally occurring and manually constructed segments belonging to Set A and Set C present lower metaphoricity than Set B and Set D. Table 46 also suggests that higher metaphoricity is associated with higher lexical density (LD) and lower average grammatical intricacy (AGI).[46]

Table 47 shows that segment sets presenting the highest values are highlighted in bold.

[46] Average and mean are defined in section 3.2.1.1.2, footnote 27. This thesis takes into consideration the definition of average as "the sum of all the values divided by the total number of values in a given set" and the definition of sum as "the result of the addition of the largest and smallest numbers in the set and dividing them by 2". Thus, in Table 47, each mean is calculated with the results for the Average lexical density (ALD) and Average grammatical intricacy (AGI), both technical terms in this thesis.

Table 47

Average lexical density and Average grammatical intricacy measurements for all segment sets

Nature	Set	Average lexical density (ALD)[47]	Mean ALD[48]	Average grammatical intricacy (AGI)	Mean AGI	Metaphoricity degree
Set A	Naturally occurring	9.92		**1.67**		
			9.18		**1.655**	Lower
Set C	Manually constructed	8.44		**1.64**		
Set B	Naturally	**11.42**	**11.08**	1.54	1.475	Higher

[47] Average lexical density is the sum of all lexical density values for each set divided by the number of segments of such a set. The same applies for the average grammatical accuracy. These values are calculated automatically by Sysfan.

[48] Mean ALD (average lexical density) is the mean of the average lexical density values for two sets (Set A and Set C or Set B and Set D). The same calculation applies for the mean AGI (average grammatical intricacy).

	occurring		
Set D			
	Manually constructed	**10.74**	1.41
	Mean	10.13	1.565

Note. The values emphasized in bold were selected because they are higher than their counterpart (from Set A or Set B). All these values were obtained using Sysfan.

Table 47 shows that the segments from Set B and Set D have a higher average lexical density (ALD) and lower average grammatical intricacy (AGI) than the segments from Set A and Set C. Considering that, during corpus compilation, an assumption was the fact that segments from Set A and C are considered less complex than Set B and Set D and the fact that Table 46 indicates a relationship between higher metaphoricity with higher ALD and lower AGI. Thus, these measures, average lexical density, and average grammatical intricacy are most likely related.

These findings for Brazilian Portuguese segments are following Ravelli (1999), who investigated English text and suggested that higher ALD and lower AGI indicate a higher degree of grammatical metaphor. Table 47, in turn, shows that Set B's and Set D's ALD measurements are higher compared to the mean of all ALD and AGI values lower than the mean for all AGI

values. Thus, segments from Set B and Set D present a higher degree of metaphoricity than segments from Set A and Set C.

These findings are in accord with Ravelli (1999), who affirms that texts with varying degrees of grammatical metaphoricity can be produced - potentially leading to distinct levels of text complexity. The main mechanism for this to happen seems to be nominalization, "a special type of class shift related to two types of complexities" – lexical and grammatical complexity – which describes how compact or intricate "ideational information" can be (Ravelli, 1999, p. 99). Therefore, this relationship between text complexity and experiential grammatical metaphor is associated with class shifts and rank shifts, as well as systemic and structural patterns. Evidence on this is detailed in the next section.

4.2 Systemic and structural patterns

This section presents the main findings on experiential grammatical metaphor as evidence of varying degrees of text complexity leading to evidence to explain text complexity. This evidence can be investigated through systemic and structural patterns drawing on Systemic Functional Linguistics.

4.2.1 Systemic Categories

In terms of systemic categories, grammatical metaphor can be realized in the grammar through lexicogrammatical patterns and embedded clauses. These are presented in detail in the next subsections.

4.2.1.1 Lexicogrammatical patterns

After the analysis of the systemic classification of the segments, systemic patterns revealed key differences between Set A and Set B and between Set C and Set D. These differences are presented in Table 48, Table 49, and Table 50, as well as Table 51 and Table 52, which combine categories pertaining to different systemic choices.

Table 48

Comparison of the agency system between Set A and Set C and Set B and D

Set	Process type	Frequency	%	Total	%
Set A	Middle	303	58.49	518	100%

and Set C	Effective	215	41.51		
Set B	Middle	324	72.00		
and Set D	Effective	126	28.00	450	100%

In Table 48, Middle agency is more frequent in Set B and Set D, which is more metaphorical; in other words, Set B and Set D should present more relational clauses in combination with nominalizations. Other differences can be observed if the Process type is fixed and the sets compared. This comparison is presented in Table 49.

Table 49

Comparison of frequency of material and relational Processes between Set A, Set C, Set B, and Set D

Set	Process type	Frequency	%
Set A	relational	134	49.63

	material	90	33.33
	existential	16	5.93
	mental	25	9.26
	verbal	5	1.85
Subtotal		270	100%
Set C	relational	129	51.39
	material	93	37.05
	existential	15	5.98
	mental	9	3.59
	verbal	5	1.99

Subtotal		251	100%
Set B	relational	135	60.81
	material	66	29.73
	existential	8	3.60
	mental	3	1.35
	verbal	10	4.50
Subtotal		222	100%
Set D	relational	143	66.82
	material	43	20.09
	existential	21	9.81

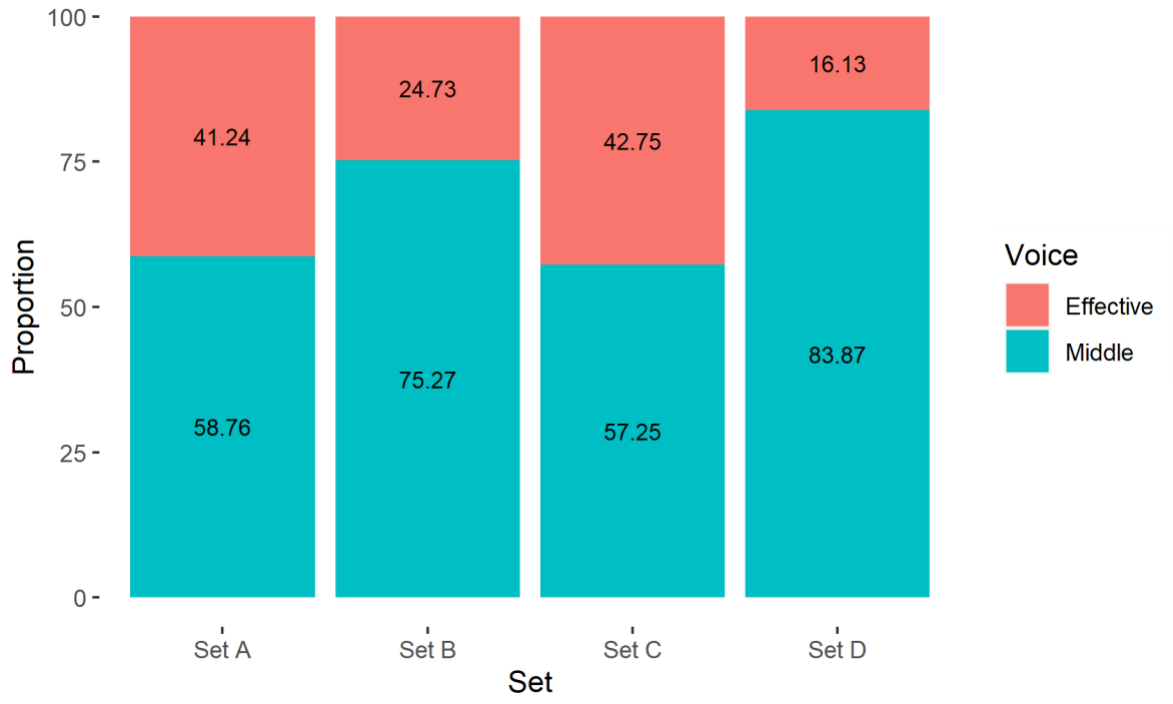
	mental	6	2.80
	verbal	1	0.47
Subtotal		214	100%

In Table 49, considering only the two most frequent Process types, the frequency of material Process is higher in Set A and Set C and lower in Set B and Set D, in which relational and existential Processes are more prevalent. In other words, relational and existential Processes favor the occurrence of nominalizations, which indicate a higher degree of text complexity of sets C and D. This can be observed by the use of nouns instead of verbs or other classes in the text flow (Halliday & Martin, 1993). This result can be associated with the results shown in Table 49, as material Processes most often occur on effective clauses, while relational clauses more frequently take place in middle clauses.

Also, taking into account the same categories from Table 49, Figure 52 shows a similar profile.

Figure 52

Comparison of the agency system between Set A and Set C and Set B and Set C



Both Table 49 and Figure 52 confirm that Set B and Set D present a higher rate of middle clauses. They also highlight that the tendencies from the complete corpus, with 200 segments, agree with the balanced sample size, with 150 segments - 50 for each domain -, as discussed in section 3.2.1.3. The reason is the fact that the prevalence of middle clauses was higher in Set C and Set D not only in the complete corpus but also in the balanced sample, which shows that similar results could be found analyzing the complete corpus as well.

Besides, based on accounts for all the segments, Table 50 and Figure 53 present similar results.

Table 50

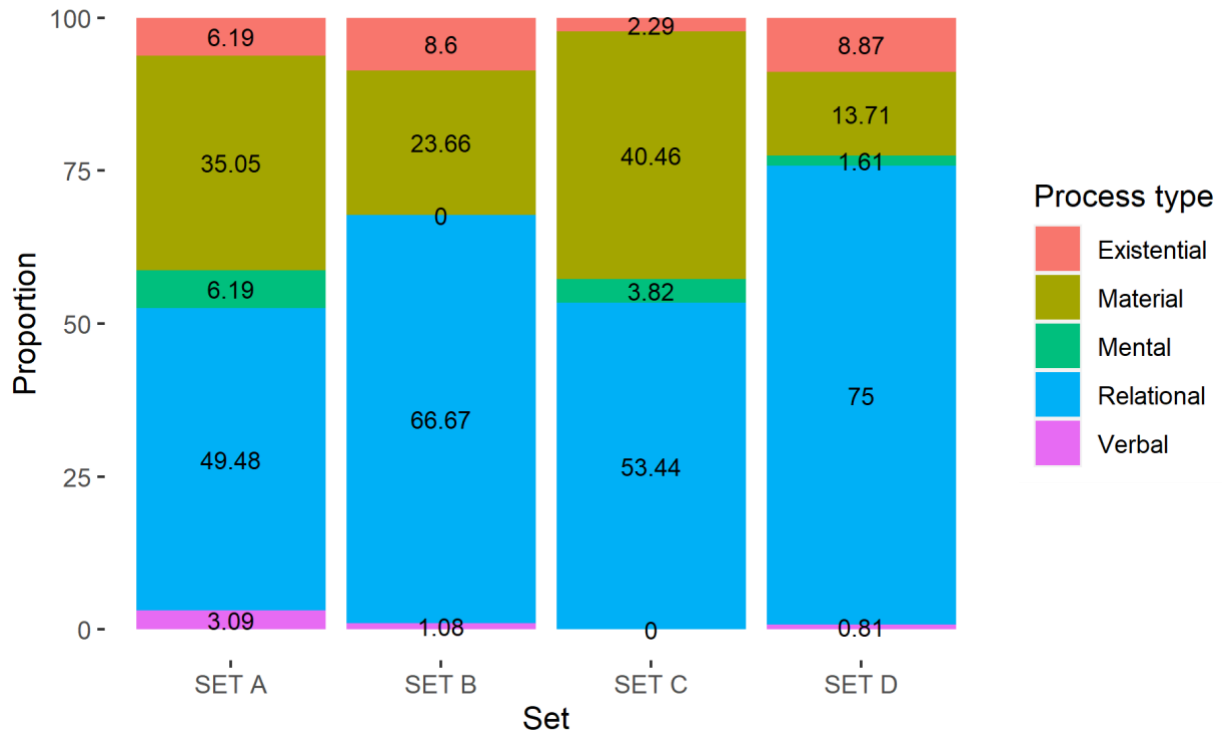
Comparison of the frequency of middle and effective Voice between Set A and Set C and Set B and Set D

Process type	Set	Frequency	%	Total	%
Set A	Middle	57	58.76	518	100%
	Effective	40	41.24		
Set C	Middle	70	75.27	450	100%
	Effective	23	24.73		
Set B	Middle	75	57.25	131	100%
	Effective	56	42.75		
Set D	Middle	104	83.87	124	100%

	Effective	20	16.13		

Figure 53

Comparison of the frequency of material and relational Processes between Set A and Set C and Set B and Set D



As expected in the light of the results shown in Table 49, Table 50 and Figure 53 also suggest a higher occurrence of material Process in Set A and C and mental, relational, and existential Processes in Set B and D. In contrast, Verbal Processes occurred similarly, with a smaller difference.

Finally, comparing categories from distinct metafunctions in Table 51 and Table 52, in this case, TEXTUAL and INTERPERSONAL, the conclusions mentioned before are reinforced.

Table 51

Key systemic differences between Set A and Set C and Set B and Set D with Marked theme fixed, varying Agency type

SELECTION	Related variable	Most frequent category in Set A and C	Most frequent category in Set B and D
Marked theme	Agency	Effective (54.66%)	Middle (63.90%)

Note. The options in the Agency system are Effective and Middle.

Table 52

Key systemic differences between Set A and C and Set B and D with Marked theme fixed, varying Process type

SELECTION	Related variable	Most frequent category in Set A and C	Most frequent category in Set B and D
Marked theme	Process type	material (46.61%)	relational (49.11%)

Through Table 51 and in Table 52, it can be observed that considering only Marked themes, Set A and Set C present effective, free clauses more often than Set B and Set D, as well as material clauses.

The next subsection analyzes the occurrence of embedded clauses occurring in the same environment as experiential grammatical metaphors.

4.2.1.2 Embedded clauses

The embedded clauses were marked for all the segments and the experiential grammatical metaphors in these segments were categorized and accounted for in each part of the corpus, as shown in Table 53. Table 54 shows the same data for the segments in the balanced corpora.

Table 53

Count of grammatical metaphor per set/nature

Nature	Set	Absolute frequency	Relative frequency (%)
Naturally occurring	Set A	154	24.92
	Set B	126	20.39
Manually constructed	Set C	190	30.74
	Set D	148	23.95
Total		618	100

Table 54

Count of complexes and clauses per set/nature

Nature	Set	Number of complexes	Number of clauses	Relative frequency (%)
Set A	Naturally occurring	90	97	51.05
Set C	Manually constructed	90	93	48.95
Set B	Naturally occurring	119	131	51.37
Set D	Manually constructed	119	124	48.63
	Total	418	445	100

In Table 53 and Table 54, contrasted frequency of complexes and clauses for both naturally occurring (Sets A and C) and manually constructed segments Sets B and D) shows that

Set A and Set C present more grammatical metaphors. This is not the expected results according to Ravelli (1999) and Steiner (2004), yet, a more detailed analysis is necessary to account for more variables, such as the number of embedded clauses, especially if taken into account considered their average in terms of the number of clauses and complexes in each set. These are shown in Table 55.

Table 55

Count of embedded clauses per set/nature

Nature	Set	Embedded frequency	Relative frequency (%)	Complex count	Avg. of embedded / complex	Clause count	Avg. of embedded/clause
Naturally occurring	Set A	92	28.66%	161	0.5714	272	0.3382
	Set B	85	26.48%	153	0.5556	236	**0.3602**
Manually constructed	Set C	72	22.43%	153	0.4706	251	0.2869
	Set D	72	22.43%	152	0.4737	214	**0.3364**

Total	321	100	619	Mean: 0.5178	973	Mean: 0.3304

Notes: No significance tests were taken here as well, as all results were used to indicate a tendency. Any non-significant results are likely to be due to the low size of the sample.

In Table 55, even though the average number of embedded clauses per complexes does not show a consistent result (pointing either to Set A and Set C or Set B and Set D), some results can be extracted from the average number of embedded clauses per clause. A higher result in Set B and Set D for both natures (in bold) indicates that this rate can be associated with metaphoricity.

The next section describes more patterns regarding structural categories, namely the order of elements (semantic functions), class, and rank shifts.

4.2.2 Structural Categories

Some of the mechanisms through which grammatical metaphor can be realized in the grammar are word order shifts, class shifts, and rank shifts. These are presented in detail in the next subsections.

4.2.2.1 Configuration sequences

The main sequences of semantic configurations are presented in Table 56 comparing Set A and Set B.

Table 56

Detailed table comparing configuration sequences in all sets

Set A and Set C			Set B and Set D		
Sequence	Freq	%	Sequence	Freq	%
Process + Medium + Range	111	48.68	Process + Medium + Range	149	68.66
Agent + Process + Medium	54	23.68	Process + Medium	37	17.05
Process + Medium	42	18.42	Agent + Process + Medium	25	11.52
Agent + Process	13	5.70	Agent + Process	4	1.84
Agent + Process + Range	4	1.75	Agent + Process + Range	1	0.46
Agent + Process + Medium + Range	2	0.88	Agent + Process + Medium + Beneficiary	1	0.46
Agent + Process + Beneficiary	1	0.44			

Process + Medium + Range + Beneficiary	1	0.44			
Total	228	100	Total	217	100

Table 56 shows that each sequence, such as Agent + Process + Medium, in general, can be verified in Set A and Set C and in Set B and Set D. Yet, some differences can be observed. The sequence Process + Medium is more frequent in Set B and Set D than in Set A and Set C, as it occupies the second position in the rank. In contrast, the sequence Agent + Process + Medium is more frequent in Set A and Set C. Also, the sequence Process + Medium + Range + Beneficiary only operates in Set A and Set C and Agent + Process + Medium + Beneficiary only in Set B and Set D. Table 57 was developed based on Table 56, summarizing the results according to the categories “Process + Medium” and “Agent + Process”, not taking consideration of more delicate structures.

Table 57

Summarized table comparing configuration sequences

Set A and Set C	Set B and Set D

Sequence	Freq	%	Sequence	Freq	%
Process + Medium	154	67.54	Process + Medium	186	85.71
Agent + Process	74	32.45	Agent + Process	31	14.28

Table 57 shows that the word order Process + Medium is more frequent in Set B and Set D, whereas the word order Agent + Process is more frequent in Set A and Set C, confirming the hypothesis that the segments in Set B and Set D are more metaphorical than those in Set A and Set C. This claim corroborates the assumption that nominalization is the main mechanism for Grammatical metaphor, represented by structures realizing, in the grammar, Medium, and relational Process.

According to Ravelli's (1999) claim that the two main ways of expressing grammatical metaphor are through class shifts or rank shifts, each of these is explained in the following sections.

4.2.2.2 Class shifts

Many instances that show evidence of class shifts were retrieved from the corpus for the reason that they can show patterns indicating grammatical metaphor. Some of these instances also present evidence of a rank shift, an aspect that will be discussed in detail in the next section.

Table 58 and Table 59 show some instances of grammatical metaphor through class shift, in which some patterns are observed – the shift from Pattern 1 to Pattern 2 as a way to increase the degree of metaphoricity.

Table 58

Segment 4 as class shift instance with nominalization

Pattern 1 - Segment pair 4 - Set A - Naturally occurring segment		
Forças de campo:	**[^são]**	aquelas [[que agem sobre os corpos]]
Field forces	*are*	*forces [[that act on two bodies]].*
Participant	relational Process	Participant
Pattern 2 - Segment pair 4 - Set C - Manually constructed segment		
Forças de campo:	**[ocorrem]**	Quando a atuação da força ocorre a distância.
Field forces	*take place*	*when a non-contact force acts*

Participant	existential Process	circumstance

Table 59

Segment 59 as class shift instance with nominalization

Pattern 1 - Segment pair 59 - Set A - Naturally occurring segment					
ou seja	uma mesma molécula de DNA [[submetida à ação de uma enzima]]	**é**	sempre	**cortada**	nos mesmos pontos
that is,	*The same DNA molecule submitted to the action of an enzyme*	*is*	*always*	*cut*	*in the same point*
	Participant	relational Process 1	Circumstance 1	relational Process 1	Circumstance 2

Pattern 2 -Segment pair 59 - Set C - Manually constructed segment					
	os cortes	**ocorre m**			nos mesmos pontos
	the cuts	*occur*			*in the same point*
	Participant	existential Process			Circumstance 2

Note: The explicit lexicogrammatical functions from the first clause were conserved in both clauses to allow comparison in this example, even though the slot is empty in the other segment.

In Table 58 and Table 59, the shifts shown are in terms of Process type, from relative Process to existential Process in the first example and from material Process to existential Process in the second. In both examples, nominalizations can be found, namely from “agem” (act) to “atuação” (action) in the first example and from “é cortada” (is cut) to “os cortes” (the cuts) in the second.

Table 60

Instances of class shifts producing nominalization

Pair	Pattern 1		Pattern 2	
	Category	Example	Category	Example
203	Process	A trajetória é o espaço das posições [[que um corpo que se **move** ocupa]] *Trajectory is the space taken by positions [[that a body that **moves** occupies]]*	Thing	O conjunto de posições sucessivas [[ocupadas por um **móvel** no decorrer do tempo]] pode ser chamado de trajetória *The set of successive positions [[occupied by a **particle** over time]] can be called trajectory*
202	Quality	A onda é um pulso [[que é **capaz** de se propagar]] *A wave is a pulse [[that is **capable** of propagating itself]]*	Thing	De acordo com a física, a onda é uma perturbação com a **capacidade** [[de se propagar no espaço ou em qualquer outro meio]] *According to physics, the wave is*

				a perturbation with the capacity to propagate in space or in any other environment.

Table 60 presents evidence of nominalization, from the lexical item of the verbal class realizing the Process “move” (*move*) to the lexical item of the nominal class realizing the Thing “móvel” (*mobile*) in the first example and from the lexical item of the nominal class realizing the Quality “capaz” (*capable*) to the lexical item of the nominal class realizing the Thing “capacidade” (*capacity*). More examples of class shifts are presented in Table 61.

Table 61

Instances of class shifts producing grammatical metaphor

Pair	Pattern 1		Pattern 2	
	Category	Example	Category	Example
104	Circumstance	[^Podem ocorrer] Alterações no comportamento **devido aos ataques de pânico** ***[**^there may be] changes in the behavior **due to the panic attacks.***	Quality	[^Pode ocorrer] Resposta comportamental **desadaptativa aos ataques de pânico** ***[**^there may be] an **unadaptive** behavioral response **to panic attacks**.***

99	Circumstance	Tratamento: [^é] [[Criar um bom relacionamento **entre o médico e o paciente**]] *[^the] treatment: is creating a good **doctor-patient** relationship*	Participant	Tratamento: Visa estabelecer **relação médico-paciente** efetiva *[^the] treatment: aims to establish an effective **doctor-patient** relationship*
	Thing	**As proteínas conjugadas** são proteínas [[que por hidrólise liberam aminoácidos mais um radical não peptídico, denominado grupo prostético]] *Conjugated proteins are proteins that, through hydrolysis, release amino acids and a non-peptide radical called the*	Circumstance	**Nas proteínas conjugadas**, além de aminoácidos, existe um radical de origem não peptídica, *In conjugated proteins, besides amino-acids, there is a radical with a non-peptide origin,*

		prosthetic group.		

In all the examples in Table 61, the degree of grammatical metaphor increases from Set A and Set C to Set B and Set D, by different class shifts. These shifts are from CIRCUMSTANCE to QUALITY and PARTICIPANT and from Thing to CIRCUMSTANCE.

Finally, all the results confirm Ravelli's (1999) and Halliday & Matthiessen's (2014) hypothesis concerning nominalization as the main resource for increasing the degree of grammatical metaphor.

In the next section, to illustrate these general tendencies, systemic and morphosyntactic glosses following Leipzig glossing rules (Max Plank, 2015) and the systemic gloss (Max Plank, 2020) were used to facilitate the understanding of the examples.

4.2.2.2.1 Class shifts Instances

The instances in this section focus on class shifts, which present nominalizing tendencies, referring to "nominalization", which, according to Ravelli (1999), is a special case of a class shift in which some semantic category (like a PROCESS) shows correspondence with a THING.

These examples consist of instances from segment pairs 61 and 139 and illustrate the main shifts related to metaphoricity found in the corpus: Process to Quality and Process to Thing.

The morphological gloss clarifies the morphemic structure of each word and the systemic gloss provides the structure of the clauses. The contrast between the segments in Brazilian Portuguese and English shows differences related to gender, number, verb tense, among others. The systemic gloss elucidates the figures in each clause, presenting pieces of evidence of different degrees of metaphoricity.

4.2.2.2.2.1 Process to Quality

The findings showed a shift from Process to Quality based on the glosses from Example 1, Example 2, Example 3, and Example 4. These examples are related to the distribution of the semantic categories, namely PARTICIPANT, PROCESS, and CIRCUMSTANCE, and QUALITY. The examples glossed are summarized in Table 61, with the semantic category of the main evidence of experiential grammatical metaphor (in bold). The way the shifts in the segments bring this kind of evidence is described in detail in the examples glossed.

Table 62 and Table 63 compare, respectively, the segments in Example 5 to Example 8 and Example 9 to Example 12, providing evidence of experiential grammatical metaphor.

Table 62

An overview of class shift instances (PROCESS to QUALITY) showing experiential grammatical metaphor

Set	Nature	Example number	Example	Translation	Semantic category + realization
Set A	Naturally occurring	1	[^Na] Irradiação: [^,] ao contrário dos processos de condução e convecção [[que **necessitam** de um meio material para a transferência de calor,]] a irradiação é o processo [[que	Irradiation: in contrast to conduction and convection processes, [[which **require** a physical medium for heat transfer to take place]], irradiation is a process [[that can take place without a physical medium]].	**Process - verb**

			pode acontecer sem que exista meio material]].		
Set B	Naturally occurring	2	A condução e a convecção são formas de propagação de calor [[que para ocorrer é **necessário** que haja meio material]], contudo, existe uma forma de propagação de calor [[que não **necessita** de	Conduction and convection are ways of heat propagation [[so that a physical medium is **necessary**]]; yet, there is a means of transferring heat [[that does not **need** a physical medium (vacuum) to propagate]]; this is thermal radiation.	**Quality - noun**

			um meio material (vácuo) para se propagar]], esta é a irradiação térmica.		
Set C	Manually constructed	3	A condução e a convecção **precisam** de um meio material para ocorrer mas a irradiação pode ocorrer sem um meio material, ou seja, no vácuo.	Conduction and convection **need** a physical medium to take place, but irradiation can take place without a physical medium, that is, in the vacuum.	**Process - verb**

Set D	Manually constructed	4	Para ocorrer condução e convecção, é **necessário** [[que exista um meio material]]; no entanto, a irradiação não possui essa restrição e pode ocorrer também no vácuo	For conduction and convection to take place, a physical medium is **necessary**; yet, irradiation does not have that restriction and and can similarly take place in vacuum.	**Quality - noun**

Examples 1 to 4 are shown in detail in Table 62. The columns “Set”, “Nature”, “Example”, “Translation” and “Semantic category + realization” represent, respectively, the segment set (Set A, Set B, Set C or Set D), the segment type (naturally occurring or manually constructed), the segment, the segment translation into English and the classification of the term in bold with its realization. In each segment, the embedded clauses are enclosed in double brackets (“[[“ and “]]”) and the term in bold, which will be analyzed further, is classified in the last column. Table 63 and Table 64 follow the same organization.

Set A and Set C proved to be less complex than Set B and Set D. This shows an overview of how experiential grammatical metaphor can be associated with text simplification, with respect to evidence of rank shift and class shift, focusing on nominalization.

The following examples, namely Example 1, Example 2, Example 3, and Example 4, present systemic glosses as a detailed analysis of the segments that Table 62 shows.

Each example has the following structure: the segment in Brazilian Portuguese, the gloss, and the segment translated into English. Each gloss has four levels: i) a classification in terms of the categories PARTICIPANT, PROCESS, CIRCUMSTANCE, and QUALITY; ii) a group rank classification (nominal group, verbal group, prepositional phrase); iii) a morphological gloss following Leipzig rules; iv) a word-by-word translation based on the gloss.

Example 1 – A naturally occurring segment from Set A (Pair 139) - Process realized by a verb

[^Na] Irradiação: [^,] ao contrário dos processos de condução e convecção [[que **necessitam** de um meio material para a transferência de calor,]]
a irradiação é o processo [[que pode acontecer sem que exista meio material]]

n-a	irradiação	ao contrário d-os	processos	de	condução	e	convecção
Circumstance		Circumstance					
Prep. phrase		Prep. phrase					
in-PREP-ART.DEF.F.SG	irradiation-F.SG	in_contrast_to-CONJ-ART.DEF.M.PL	process-M.PL	of-PREP	conduction-F.SG	and-CONJ	convection-F.SG
in-the	irradiation	in_contrast_to-the	processes	of	conduction	and	convection

que	**necessitam**	de	um	meio	material	para	a	transferência	de	calor
	Process		Participant				Participant			
	Verb. group		Nom. group				Nom. group			
that-REL	need-PRS.3.PL.IND	of-PREP	ART.INDF.M.SG	mean-M.SG	material-M.ADJ.SG	for-PREP	ART.DEF.F.SG	transference-F.SG	of-PREP	heat-M.SG
that	need		a	medium	material	for	the	transference	of	heat

a	irradiação	é	o	processo	que	pode	acontecer	sem	que	exista	meio	material
Participant		Process	Participant									
Nom. group		Verb. Group	Nom. group									
ART.DEF.F.SG	irradiation-F.SG	be-PRS.3.SG.IND	ART.DEF.M.SG	process-M.SG	that-REL	can-PRS.3.SG.IND	happen-INF	without-PREP	PRON	happen-PRS.3.SG.SBJV	medium.M.SG	material-ADJ.SG
the	irradiation	is	the	process	that	can	happen	without		happen	medium	material

Irradiation: in contrast to conduction and convection processes, [[which require a physical medium for heat transfer to take place]], irradiation is a process [[that can take place without a physical medium]].

In Example 1, the categories at the first level are realized by groups, shown at the second level, which are formed by words segmented in morphemes at the third level. Below the third level, the word-by-word translation is the first step to translate the segment into English. The fourth level is the segment translation into English.

From the first level, we can obtain the sequence followed by the segment in terms of PARTICIPANT, PROCESS, CIRCUMSTANCE. In Example 1 the sequence is "Circumstance (location-place) - Circumstance (manner - comparison) - Participant - Process (relational) – Participant". In this case, this sequence describes a clause simplex, not a clause complex.

The findings show that PARTICIPANTS are realized by nominal groups, PROCESSES by verbal groups, and CIRCUMSTANCES by prepositional phrases. This characterizes the default behavior of these realization patterns, which is shown in detail with further examples provided. When the configuration of the elements in the clauses is not the default option, some kind of grammatical metaphor is involved, either interpersonal or experiential. Analyzing the instances emphasized in bold, evidence of experiential grammatical metaphor can be found, concerning mostly rank and class shifts, according to Ravelli (1999) and Halliday (1985). These shifts are shown in Examples 1 to 4.

In Example 1, the term in bold, namely 'necessitam' (*need*) is an item of verbal class operating in a PROCESS realized by a verbal group. In the following examples, even though they are similar in terms of experiential meaning, the equivalent to the focused term is often not from the same grammatical class or sometimes does not function at the same rank. This means that the comparison between Examples 1 to 4 illustrates the experiential grammatical metaphor analysis, which presents a strong piece of evidence of experiential grammatical metaphor, as for the fact

that this kind of class shift suggests that the nominalization phenomenon increases text complexity in Examples 2 and 4. As text complexity varies comparing one segment to the other, we can relate experiential grammatical metaphor shifts with text simplification.

Example 2 – A naturally occurring segment from Set B - Quality realized by a noun

A condução e a convecção são formas de propagação de calor [[que para ocorrer é **necessário** que haja meio material]], contudo, existe uma forma de propagação de calor [[que não **necessita** de um meio material (vácuo) para se propagar]], esta é a irradiação térmica.

a	condução	e	a	convecção	são	formas	de	propagação	de	calor														
Participant					Process	Participant																		
Nom. Group					Verb. Group	Nom. group																		
ART.DEF.F.SG	conduction-F.SG	and-CONJ	ART.DEF.F.SG	convection-F.SG	be-PRS.3.PL.IND	way-F.PL	of-PREP	propagation-F.SG	of-PREP	heat-M.SG														
the	conduction	and	the	convection	are	ways	of	propagation	of	heat														
que	para	ocorrer	é	necessário	que	haja	meio	material																
		Process	Process	Quality		Process	Participant	Process																
		Verb. group	Verb. group	Nom. group		Verb. Group	Nom. group	Verb. group																
that-REL	in_order_to-CONJ	happen-INF	be-PRS.3.SG.IND	necessary-ADJ	that-CONJ	have-PRS.3.SG.SBJV	mean-M.SG	material-M.ADJ																
that	in_order_to	happen	it_is	necessary	that	it_may_have	medium	material																
contudo	existe	uma	forma	de	propagação	de	calor	que	não	necessita	de	um	meio	material	para	se	propagar							
	Process	Participant																						
	Verb. group	Nom. group																						
however-CONJ	exist-PRS.3.SG.IND	ART.INDF.F.SG	way-F.SG	of-PREP	propagation-F.SG	of-PREP	heat-M.SG	that-REL	NEG.ADV	need-PRS.3.SG.IND	of-PREP	ART.INDF.M.SG	medium-M.SG	material-M.ADJ	to-PREP	PRON.REFL	propagate-INF							
however	exists	a	way	of	propagation	of	heat	that	does not	need	of	a	medium	material	to	itself	propagate							

Conduction and convection are ways of heat propagation [[so that a physical medium is **necessary**]]; yet, there is a means of transferring heat [[that does not **need** a physical medium (vacuum) to propagate]]; this is thermal radiation.

The systemic gloss in Example 2 describes the segment as the complex equivalent to the simplex from Example 1. This complex is formed by a relational clause followed by an existential clause and another relational clause. The sequence in Example 2 is “Participant - Process (relational) - Participant // Process (existential) – Participant”. Compared to Example 1, the findings are that meanings construed in the clause rank from a clause simplex (the segment in Example 1) are construed by embedded clauses in nominal groups in a clause complex. These nominal groups are: “formas de propagação de calor [[que para ocorrer é **necessário** que haja meio material]]” and “uma forma de propagação de calor [[que não **necessita** de um meio

material (vácuo) para se propagar]]". This way, new information can be provided in addition to what was mentioned before. This leads to an increase in complexity.

The term "necessário" (necessary) is a lexical item of nominal class operating in a nominal group realizing Quality, which contrasts with the Process realized by a verbal group in Example 1. This is the first evidence of a class shift, which indicates experiential grammatical metaphor when both segments are compared because the class shift describes how nominalization is increasing the complexity of segments.

Example 3 – A manually constructed segment from Set C (Pair 139) - Process realized by a verb

A condução e a convecção **precisam** de um meio material para ocorrer mas a irradiação pode ocorrer sem um meio material, ou seja, no vácuo.

a	condução	e	a	convecção	**precisam**		de	um		meio	material	para		ocorrer		
ART.DEF.F.SG	conduction-F.SG	and-CONJ	ART.DEF.F.SG	convection-F.SG	need-PRS.3.PL.IND		of-CONJ	ART.INDF.M.SG		medium-M.SG	material-ADJ.SG	to-PREP		happen-INF		
Participant					Process		Participant									
Nom. Group					Verb. group		Prep. phrase									
the	conduction	and	the	convection	need		of	the		medium	material	to		happen		
mas		a		irradiação	pode	ocorrer	sem	um	meio	natural	ou seja	n-o		vácuo	material	
		Participant			Process		Circumstance					Circumstance				
		Nom. group			Verb. Group		Prep. phrase									
but-CONJ		ART.DEF.F.SG		irradiation-F.SG	can-PRS.3.SG.IND	happen-INF	without-PREP	ART.INDF.M.SG	mean-M.SG	natural-M.ADJ	that_is-CONJ	at-PREP	ART.DEF.M.SG	vacuum-M.SG	material.M.ADJ	
but		the		irradiation	can	happen	without	the	mean	natural	that is	at	the	vacuum	material	

Conduction and convection **need** a physical medium to take place, but irradiation can take place without a physical medium, that is, in the vacuum.

The systemic gloss in Example 3 illustrates a complex of two relational clauses. The sequence in this complex is: Participant - Process (relational) - Participant // Participant - Process (existential) - Participant - Circumstance (accompaniment) - Circumstance (location-place). The term "precisam" (need), just like in Example 1, is a lexical item of verbal class operating in a verbal group realizing Process. This term can be compared with the bold term in Example 4.

Also, the findings elucidate that the circumstance "sem um meio material" (*without a physical medium*) is simpler than the embedded clause "que exista um meio material" (*that a physical medium exists*) from Example 4.

Example 4 – A manually constructed segment from Set D (Pair 139) - Quality realized by a noun

Para ocorrer condução e convecção, é **necessário** [[que exista um meio material]]; no entanto, a irradiação não possui essa restrição e pode ocorrer também no vácuo.

para	ocorrer	condução	e	convecção	é	**necessário**	que	exista	um	meio	material
	Process	Participant			Process	Quality		Process	Participant		
	Verb. group	Nom. group			Verb. group	Nom. group		Verb. group	Nom. group		
to-PREP	occur-INF	conduction-F.SG	and-CONJ	convection-F.SG	be-PRS.3.SG.IND	necessary-ADJ.M.SG	that-CONJ	exist-PRS.3.SG.SBJV	ART.INDF.M.SG	medium-M.SG	material-M.SG
to	occur	conduction	and	convection	is	necessary	that	exists	a	means	material

no entanto	a	irradiação	não	possui	essa	restrição	e	pode	ocorrer	n-o	vácuo
	Participant		Proc	Parti		Process	Circumstance		Participant		

			ess	cipant							
	Nom. group		Verb. group	Nom. group			Verb. group	Prep. phrase		Nom. group	
yet-CONJ	ART.DEF.FSG	irradiation-MSG	NEG.ADV	possess-PRS.3SG.IND	PRON.DEM.FSG	restriction-FSG	and-CONJ	can-PRS.3SG.IND	take place-INF	at-PREP-ART.DEF.MSG	vacuum-MSG
yet	the	irradiation	does_not	possess	this	restriction	and	can	take place	in	vacuum

For conduction and convection to take place, a physical medium is **necessary**; yet, irradiation does not have that restriction and can take place in vacuum.

In Example 4, the complex described by the systemic gloss follows the sequence "minor_clause - Participant - Process (relational) - Quality - Participant_clause". The term "necessário" (necessary), just like in Example 2, is a lexical item of nominal class operating in a nominal group realizing Quality. As mentioned in Example 3, the analysis showed that the embedded clause "que exista um meio material" in "é necessário que exista um meio material (a

physical medium is necessary) (Example 4) is more complex than the circumstance “sem um meio material” (*without a physical medium*) (Example 3), because there is a rank shift from the clause, in which the prepositional phrase functions, to the group, in which the embedded clause is functioning.

Comparing these examples, the evidence of experiential grammatical metaphor showed that segments from Set B are more metaphorical than Set A. This evidence is that in Example 1 and Example 3, the bold terms are lexical items of the verbal class operating in verbal groups realizing a Process and, in Example 2 and Example 4, the terms were lexical items of the nominal class functioning in a nominal group realizing qualities. This indicates that a nominalization phenomenon is taking place. The first two examples were associated with Set A and Set C, which proved to have a lower degree of metaphoricity than the other examples, associated with Set B and Set D.

The next subsection shows more examples of class shifts, specifically through nominalization, which proved to be essential to allow remapping between different metaphoricity degrees

4.2.2.2.2.2 Process to Thing

The results showed a shift from Process to Thing, indicating nominalization tendencies, referred to by Ravelli (1999) as a special case of class shift, based on the analysis of glossed examples organized according to the segment-to-segment correspondence. These examples consisted of instances from segment 139, summarized in Table 63.

Table 63

An overview of class shift instances (PROCESS to THING) showing experiential grammatical metaphor

Set	Nature	Example number	Example	Translation	Semantic category + realization
Set A	Naturally occurring	5	A sinalização celular é a forma [[como uma célula **comunica**-se com outra a partir de sinais por elas emitidos]].	Cell signaling is the manner [[cells **communicate** with each other through signals they send]].	Process - verb

Set B	Naturally occurring	6	A sinalização celular faz parte de **um complexo sistema de comunicação** [[que governa e coordena as atividades e funções celulares]]	Cell signaling is part of **a complex communication system** [[that governs and coordinates cellular activities and functions]].	**Thing - noun**
Set A	Manually constructed	7	A sinalização celular é a maneira [[pela qual as células se **comunicam** entre si para determinar as atividades celulares]].	Cell signaling is the manner [[cells **communicate** with each other to assign cell activities]].	**Process - verb**

Set B	Manually constructed	8	A sinalização celular é **um sistema de comunicação** responsável pelo controle das atividades e funções celulares.	Cell signaling is **a communication system** for controlling cellular activities and functions.	**Thing - noun**

Examples 5 to 8 are shown in detail in Table 63, organized the same way as Table 62. As was the case in the latter, the columns “Set”, “Nature”, “Example”, “Translation” and “Semantic category + realization” represent, respectively, the segment set (Set A, Set B, Set C or Set D), the segment type (naturally occurring or manually constructed), the segment, the segment translation into English and the classification of the term in bold with its realization. In each segment, the embedded clauses are enclosed in double brackets (“[[“ and “]]”) and the term in bold, which will be analyzed further, is classified in the last column.

On the basis of Set A and Set C being less complex than Set B and Set D, Table 63 shows that these sets can be compared to obtain evidence of how experiential grammatical metaphor can be associated with text simplification, based on evidence of rank shift and class shift, focusing on nominalization.

The subsequent examples, namely Example 5, Example 6, Example 7, and Example 8, are clarified by systemic glosses and are followed by a detailed analysis of the segments that Table 63 shows.

Example 5 – A naturally occurring segment from Set A (pair 61) - Process realized by a verb

A sinalização celular é a forma [[como uma célula **comunica**-se com outra a partir de sinais por elas emitidos]].

a	sinalização	celular	é	a	forma	como	uma	célula
Participant			Process	Partic-				
Nom. group			Verb. group	Nom.				
ART.DEF.F.PL	sinalization-F.PL	cell-ADJ.F.SG	be-PRS.3.SG.IND	ART.DEF.F.PL	manner-F.SG	how-CONJ	ART.INDF.SG	cell-F.SG
the	signalling	cellular	is	the	manner	how	a	cell

comunica	se	com	outra	a partir de	sinais	por	elas	emitidos
-ipant								
group								
communic	REFL	with-	another-	by_means	signal	by-PREP	PRON.REF	emit-

ate-PRS.3.SG.IND		PREP	PRON.F.SG	_of- PREP	-M.PL		L.F.PL	PST.PTCP.M.PL
communicates	itself	with	another	by_means_of	signal	by	themselves	emitted

Cell signaling is the manner [[in which cells **communicate** with each other through signals they send]].

The systemic gloss in Example 5 describes the segment as a relational clause formed by the sequence "PARTICIPANT-PROCESS(relational)-PARTICIPANT". The latter is a lexical item of nominal class operating in a nominal group with an embedded clause as a QUALIFIER. The term "comunica" (*communicate*) is a lexical item of the verbal class operating in a verbal group from an embedded clause realizing a PARTICIPANT, which is contrasted with "comunicação" (*communication*) from a nominal group, also realizing a PARTICIPANT in Example 6.

Example 6 – A naturally occurring segment from Set B (pair 61) - Thing realized by noun

A sinalização celular faz parte de **um complexo sistema de comunicação** [[que governa e coordena as atividades e funções celulares]]

a	sinalização	celular	faz	parte	de	um	complexo	sistema	de	**comunicação**
ART.DEF.F.SG	sinalization-F.SG	cell-ADJ.SG	do-PRS.3.SG.IND	part-F.SG	of-PREP	ART.INDF.M.SG	complex-M.ADJ	system-M.SG	of-PREP	communication-F.SG
Participant			Process			Partic-				
Nom. group			Verb. group		Prep.					
the	signaling	cellular	does	part	of	a	complex	system	of	communication
que	governa	e	coordena	as	atividades	e	funções	celulares		
-ipant										
phrase										
that-REL	govern-PRS.3.SG.IND	and-PREP	coordinate-PRS.3.SG.IN	ART.DEF.F.PL	activity-F.PL	and-PREP	function-F.PL	cell-ADJ.F.PL		

			D					
that	governs	and	coordinates	the	activity	and	function	cell

Cell signaling is part of **a complex communication system** [[that governs and coordinates cellular activities and functions]].

Based on example 6, the systemic gloss illustrates a relational clause realized as a material clause. In the relational clause, the lexical item of the verbal class realizing the relational PROCESS "é" (is) functions as a part of the term "é parte de" (is part of). In the material clause, the lexical item of the verbal class realizing the Process is "faz" (does) in "faz parte de" (is part of). The structure follows the sequence "PARTICIPANT-PROCESS(relational)-PARTICIPANT". The latter is a nominal group with an embedded clause as a QUALIFIER. The term "um complexo sistema de comunicação [[que governa e coordena as atividades e funções celulares]]" includes a Qualifier realized by an embedded clause. Furthermore, the analysis also indicates that this Qualifier leads to a higher degree of complexity in Example 6 compared to Example 5 due to a rank shift, which can be observed by analyzing the contrast between the groups containing the terms in bold ("comunicam" in Example 5 and "communication" in Example 6).

Example 7 – A manually constructed segment from Set C (pair 61) - Process realized by a verb

A sinalização celular é a maneira [[pela qual as células se **comunicam** entre si para determinar as atividades celulares]].

a	sinalização	celular	é	a	maneira
Participant			Process	Part-	
Nom. group			Verb. group	Nom.	
ART.DEF.F.SG	sinalization-F.PL	cell-ADJ.F.SG	be-PRS.3.SG.IND	ART.DEF.F.PL	manner-F.SG
the	signaling	cellular	is	the	manner

pela	qual	as	células	se	**comunicam**
icip-					
gr-					
in-PREP-ART.DEF.F.SG	which-PRON	ART;DEF.F.PL	cell-F.PL	REFL	communicate-PRS.3.PL.IND
in-the	which	the	cells	themselves	communicate

entre	si	para	determinar	as	atividades	celulares
ant						
-oup						

betwee n.PREP	PRON.RE FL.F.PL	to-PREP	determin e-INF	ART.DE F.F.PL	activity-F.PL	cell-ADJ.F.P L
ant						
betwee n	themselve s	to	determin e	the	activity	cell

Cell signaling is the way [[cells **communicate** with each other to assign cell activities]].

The systemic gloss in Example 7 illustrates a relational clause formed by the sequence "Participant-Process(relational)-Participant". The latter PARTICIPANT is realized by a nominal group with an embedded clause as a QUALIFIER. The term "comunicam" (*communicate*) is an item of the verbal class that realizes a PROCESS and can be compared to "comunicação" (*communication*) from a nominal group shown in detail in Example 8. This comparison is also related to class shift, namely nominalization, which increases the degree of complexity in texts. In this case, one more time the segment from Set C is less complex than the one from Set D.

Example 8 – A manually constructed segment from Set D (pair 61) - Thing realized by noun

A sinalização celular é **um sistema de comunicação** responsável pelo controle das atividades e funções celulares.

a	sinalizaçã	celular	é	um	sistem	de	**comunicação**

	o				a		
Participant			Proce ss	Partic-			
Nom. group			Verb. group	Nom.			
ART.DEF .F.SG	sinalizatio n-F.SG	cell- ADJ.F.S G	be- PRS.3 .SG	ART.INDF. F.SG	syste m- M.SG	of- PREP	communication- F.SG
the	signaling	cellular	is	a	syste m	of	communication
responsáv el	pelo	controle		das	atividade s	e	funções celulare s
ipant							
group							
responsibl e-M.ADJ	for-PREP- ART.DEF .M.SG	control.M.SG		of-PREP- the- ART.DEF. F.PL	activity- F.PL	and- PRE P	functio n-F.PL cell- ADJ.F. PL
responsibl e	for-the	control		of-the	activity	and	functio n cellular

Cell signaling is **a communication system** for controlling cellular activities and functions.

In Example 8, a relational clause is organized by the sequence “Participant-Process(relational)-Participant”. The latter PARTICIPANT is a lexical item of nominal class operating in a nominal group with an embedded clause as a QUALIFIER. The term “é um sistema de comunicação [[responsável pelo controle das atividades e funções celulares]]” contains a QUALIFIER realized by an embedded clause. In this group, “um sistema de comunicação” is within Thing, realized by a nominal group, in a relation of experiential grammatical metaphor with “comunicam” (communicate) from Example 7.

More examples are presented, this time related to pair 134, also with the gloss following Leipzig glossing rules (Max Plank, 2015) and the systemic gloss (Max Plank, 2020). They are summarized in Table 64, with a comparison between the semantic category of the term “são resfriados” (*are cooled*) and their correspondent examples in the other segments.

Table 64

Additional class shift instances (PROCESS to THING) showing grammatical metaphor

Set	Nature	Example number	Example	English Translation	Semantic category + realization
Set A	Naturally	9	Por exemplo, os alimentos da	For example, food in the fridge	**Process -**

	occurring		geladeira **são resfriados** dessa maneira	is cooled down in such manner	**verb**
Set B	Naturally occurring	10	Um exemplo disso ocorre, por exemplo, no **resfriamento** dos alimentos dentro da geladeira.	An example of this [phenomenon] takes place, for instance, in cooling down food in the fridge.	**Thing - noun**
Set A	Manually constructed	11	Os alimentos da geladeira **são resfriados** assim.	The food in the fridge is cooled down this way.	**Process - verb**
Set B	Manually constructed	12	Isso ocorre dessa forma no **resfriamento** dos alimentos na	This takes place this way when cooling down food in the fridge.	**Thing - noun**

			geladeira.		

Examples 8 to 12 are shown in detail in Table 64, which is organized the same way as Table 62 and Table 63. The columns “Set”, “Nature”, “Example”, “Translation” and “Semantic category + realization” represent, respectively, the segment set (Set A, Set B, Set C or Set D), the segment type (naturally occurring or manually constructed), the segment, the segment translation into English and the classification of the term in bold with its realization. In each segment, the embedded clauses are enclosed in double brackets (“[[“ and “]]”) and the term emphasized in bold, which will be analyzed further, is classified in the last column.

As in Table 63, as Set A and Set C were less complex than Set B and Set D, these sets can be compared to obtain evidence of how experiential grammatical metaphor can be associated with text simplification, based on evidence of rank shift and class shift, focusing on nominalization.

Further evidence is provided in the detailed analysis of the following examples, namely Example 9, Example 10, Example 11, and Example 12, which present morphological and systemic glosses that illustrate a detailed analysis of the segments that Table 64 shows.

Example 9 - A naturally occurring segment from Set A (pair 134) - Process realized by a verb

Por exemplo: na geladeira os alimentos são **resfriados** dessa forma.

por exemplo	n-a	geladeira	os	alimentos	são	**resfriados**	d-essa	forma
Circumstance	Circumstance		Participant		Process		Circumstance	
Prep. phrase	Prep. phrase		Nom. group		Verb. group		Prep. phrase	
for_example-CONJ	in-PREP-ART.DEF.F.SG	fridge-F.SG	ART.M.PL	food-M.PL	be-PRS.3.PL.IND	cool_down-PST.PTCP.M.PL	of-PREP-PRON.DEM.F.SG	manner-F.SG
for_example	in-the	fridge	the	food	are	cooled_down	of-this	manner

For example, food in the fridge is cooled down in such manner

The systemic gloss in Example 9 describes the simplex with the sequence "Circumstance(manner-comparison)-Circumstance(location-place)-Participant-Process(material)-Circumstance (manner-means)". The term "são resfriados" (*are cooled*) is an item of verbal class operating in a PROCESS that has a lower degree of metaphoricity than the noun "resfriamento" (*cooling*) from the nominal group within the prepositional group "no resfriamento dos alimentos dentro da geladeira" (*when cooling down the food in the fridge*) in Example 10. The analysis indicates that this also takes place due to the nominalization, which in general increases the complexity of texts.

.Example 10 - A naturally occurring segment from Set B (pair 134) - Thing realized by noun

Um exemplo disso ocorre, por exemplo, no **resfriamento** dos alimentos dentro da geladeira.

um	exemplo	d-isso	ocorre	por exemplo	n-o	resfriamento	d-os	alimentos	dentro	d-a	geladeira
Participant			Process	Circumstance	Circumstance						
Nom. group			Verb. group	Prep. phrase	Prep. phrase						
ART.INDF.M.SG	example-M.SG	of-PREP-PRON.DEM.M.SG	take_place-PRS.3.SG.IND	for_example-CONJ	in-PREP-ART.DEF.M.SG	cooling-M.SG	of-PREP-ART.DEF.M.PL	food-M.PL	inside-PREP	of-PREP-ART.DEF.F.SG	fridge-F.SG
an	example	of_this	takes place	for_example	in-the	cooling	of-the	food	inside	of-the	fridge

An example of this [phenomenon] takes place, for instance, in cooling down the food in the fridge.

The systemic gloss in Example 10 illustrates a detailed analysis of the sequence "Participant-Process(existential)-Circumstance (manner-comparison)-Circumstance(location-place)". The term "resfriamento" (*cooling*) is a lexical item of a nominal class operating in a nominal group inside a prepositional phrase realizing the CIRCUMSTANCE "no resfriamento dos alimentos dentro da geladeira". The analysis proves that this term has a higher degree of metaphoricity than "são resfriados" (*are cooled*) from Example 9.

Example 11 - A manually constructed segment from Set C (pair 134) - Process realized by a verb

Os alimentos da geladeira **são resfriados** assim.

os	alimentos	d-a	geladeira	são	**resfriados**	assim	
Participant				Process		Circumstance	
Nom. group				Verb. group		Verb. group	
ART.DEF.M.PL	food-M.PL	of-PREP-ART.DEF.F.SG	fridge-F.SG	be-PRS.3.PL.IND	cool_down-PTCP.M.PL	like_this-ADV	
the	food	of-the	fridge	are	cooled_down	of-this	manner

The food in the fridge is cooled down like this.

The systemic gloss in Example 11 describes the sequence" Participant-Process(material)-Circumstance(manner-means)". The term emphasized in bold, "é resfriado" (*is cooled down*), is an item of nominal class operating in a PROCESS. This group is compared with "resfriamento", from the circumstance "no resfriamento dos alimentos na geladeira" (*when cooling down the food in the fridge*) in Example 12.

Example 12 - A manually constructed segment from Set D (pair 134) - Thing realized by noun

Isso ocorre dessa forma no **resfriamento** dos alimentos na geladeira.

isso	ocorre	d-essa	forma	n-o	resfriamento	dos	alimentos	na	geladeira
Participant	Process	Circumstance		Circumstance					
Nom. group	Verbal group	Prepositional phrase		Prepositional phrase					
PRON.DEM.M.SG	take_place-PRS.3.SG.IND	of-PREP-ART.DEM.F.SG	manner-F.SG	in-PREP-ART.DEF.M.SG	cooling-M.SG	of-PREP-ART.DEF.M.PL	food-M.PL	in-PREP-ART.DEF.F.SG	fridge-F.SG
this	takes_place	in_this_manner		in_the	colling	of_the	food	in_the	fridge

This takes place this way when cooling down food in the fridge.

The systemic gloss in Example 12 shows the sequence “Participant-Process(existential)-Circumstance(cause-reason)”. The term “resfriamento” (*cooling down*) is part of the Circumstance “no resfriamento dos alimentos na geladeira”. This nominal group within the CIRCUMSTANCE has a higher degree of metaphoricity than the verbal group “são resfriados” (*are cooled down*) from Example 11.

By analyzing the segments from Table 64 in detail with the examples, the contrast between the semantic category from the examples emphasized in bold suggests evidence of experiential grammatical metaphor. In other words, correspondent segments from Set B and Set D show an increasing degree of metaphoricity compared to Set A and Set C, indicating a nominalizing tendency associated with an increase in experiential grammatical metaphor.

4.2.2.3 Rank shifts

On the basis of the class shift results, Table 65 presents some of those instances in which there were also rank shifts, especially from the clause rank to the group/phrase rank. This is the main direction since rank shifts increase the information density by compressing the information into a lower rank and allows more information to be expressed.

Table 65

Instances of rank shift

Pair	Pattern 1		Pattern 2	
	Category	Example	Category	Example
75	Free cause	Na maioria dos casos, o uso contínuo de medicamentos **reduz os sintomas e essa porcentagem para cerca de 30%.** I*n most of the cases, the continuous use of drugs* ***alleviates the symptoms and this rate falls to approximately 30%.***	Embedded clause	O uso persistente de drogas antipsicóticas tem a capacidade [[**de reduzir essa taxa para aproximadamente 30%**]]. *The persistent use of antipsychotic drugs has the capacity [[****for reducing this rate to approximately 30%****]].*

104	Circumstance	[^Podem ocorrer] Alterações no comportamento **devido aos ataques de pânico** *There may be behavior changes **due to panic attacks.***	Quality	^Pode ocorrer] Resposta comportamental **desadaptativa aos ataques de pânico** *There might be **non-adaptive behavior responses to panic attacks.***
74	Bound clause	**Se não for feito o tratamento com medicamentos**, entre 70% e 80% dos pacientes apresentam outro episódio nos doze meses seguintes. *If the treatment using drugs is not performed, 70% to 80% of the patients suffer another episode in the next twelve months.*	Circumstance	**Na ausência do tratamento com drogas antipsicóticas** após o primeiro episódio, entre 70% e 80% dos pacientes apresentam outro episódio nos próximos 12 meses. *In the absence of the treatment using antipsychotic drugs after the first episode, 70% to 80% of the patients suffer another episode in the next 12 months.*

38	Participant	**As proteínas conjugadas** são proteínas [[que por hidrólise liberam aminoácidos mais um radical não peptídico, denominado grupo prostético]] *Conjugated proteins are proteins that, through hydrolysis, release amino acids and a non-peptide radical called prosthetic group.*	Circumstance	**Nas proteínas conjugadas**, além de aminoácidos, existe um radical de origem não peptídica, *In conjugated proteins, there is a radical with a non-peptide origin,*

In contrast to the previous tables, Table 65 shows rank shifts, preserving approximately the same meanings from Pattern 1, less metaphorical, to Pattern 2, more metaphorical.

In the first example, a free clause in Pattern 1 is realized as an embedded clause in the nominal group “a capacidade de reduzir essa taxa para aproximadamente 30%”. (*the capacity for reducing this rate to approximately 30%*). In the second case, a Circumstance is realized as a

Quality, which specifies a nominal group. In the third case, a bound clause is realized as a Circumstance. In the fourth instance, a Participant (Thing) is realized as a Circumstance. In the first three examples, the meaning that was expressed in Pattern 1 in the rank of the clause was realized in Pattern 2 working in the rank of the group. In the fourth example, however, the meaning that was previously realized in the clause rank was also realized in Pattern 2 in the group rank, with a relevant shift, which was in the Theme – in this case, the Theme that was unmarked in Pattern 1 became marked in Pattern 2, increasing the complexity of the information.

The next section will present examples of rank shifts, which indicate experiential grammatical metaphor leading to different levels of text complexity.

4.2.2.3.1 Rank shifts instances

The analysis of the examples 13, 14, 15, and 17 (from segment pair 199) was performed in the same manner as the ones already presented, and show evidence of a rank shift, as well as nominalization.

Example 13 - A naturally occurring segment from Set A (pair 199) - Qualifier realized by nominal group

A Mecânica é o ramo da Física [[responsável pelo **estudo** do movimento]]

a	Mecânica	é	o	ramo	d-a	Física
Participant		Process	Partic-			
Nom. group		Verb. group	Nom.			
ART.DEF.F.SG	mechanics-F.SG	be-PRS.3.SG.IND	ART.DEF.M.SG	field-M.SG	of-PREP-ART.DEF.F.SG	physics-F.SG
the	mechanics	is	the	field	of-the	physics

responsável	pel-o	estudo	d-o	movimento
-ipant				
group				
responsible-M.SG	for-PREP-ART.DEF.M.SG	study-M.SG	of-PREP-ART.DEF.M.SG	movement-M.SG
responsible	for-the	study	of-the	movement

Mechanics is the field of physics concerned with the study of motion.

The systemic gloss in Example 13 illustrates the relational clause formed by the sequence "Participant-Process(relational)-Participant". The latter PARTICIPANT is realized by a nominal group with an embedded clause as a QUALIFIER. The QUALIFIER "responsável pelo estudo do movimento" (responsible for the study of motion) is realized by an embedded clause that characterizes "o ramo da Física" (*the field of physics*), forming the nominal group "o ramo da Física [[responsável pelo estudo do movimento]]" (*the field of physics responsible for the study of motion*).

Also, the term in bold "estudo" (*study*) is the THING in the nominal group inside the embedded clause "responsável pelo estudo do movimento" (*responsible for the study of motion*). This term can be compared with terms in bold in the next examples to obtain evidence of nominalization.

Example 14 - A naturally occurring segment from Set B (pair 199) - QUALIFIER realized by embedded clause

Mecânica é o ramo da física [[que compreende o **estudo** e análise do movimento e repouso dos corpos e sua evolução no tempo]]

Mecânica	é	o	ramo	d-a	Física		
Participant	Process	Part-					
Nom. group	Verb. group	Nom.					
mechanics-F.SG	be-PRS.3.SG.IND	ART.DEF.M.SG	field-M.SG	of-PREP-ART.DEF.F.SG	physics-F.SG		
mechanics	is	the	field	of-the	physics		
que	compreende	o	estudo	e	análise	d-o	movimento
-ici-							
gr-							
that-PRON.REL	comprehend-PRS.3.	ART.DEF.M.S	study-M.SG	and-CONJ	analysis- F.SG	of-PREP-ART.DEF.M.SG	movement-M.SG

	SG.IND	G						
that	comprehends	the	study	and	analysis	of-the	movement	
e	repouso	d-os		corpos	e	sua	evolução	no	tempo
-pant									
oup									
and-CONJ	rest-M.SG	of-PREP-ART.DEF.F.PL		body-M.PL	and-CONJ	PRON.POSS.M.PL	evolution- F.SG	over-PREP-ART.DEF.M.SG	time-M.SG
and	rest	of-the		body	and	their	evolution	over-the	time

Mechanics is the field of physics that concerns the study and analysis of the bodies and their evolution in time.

Example 14 describes a relational clause that follows the sequence “Participant-Process(relational)-Participant”, like in Example 13. The second PARTICIPANT is realized by a nominal group with an embedded clause as a QUALIFIER. The QUALIFIER “que compreende o estudo e análise do movimento e repouso dos corpos e sua evolução no tempo” (that

comprehends the study and analysis of the bodies and their evolution over time) is realized by an embedded clause, specifying the type of domain of physics.

Unlike in Example 13, in which the QUALIFIER is realized by an embedded non-finite clause, in Example 14 a finite embedded clause functions as a QUALIFIER. This shows a rank shift from the group to clause that indicates experiential grammatical metaphor.

Nominalization can be observed as well by comparing the terms emphasized in bold "estuda" (*studies*)", a PROCESS realized by a verbal group, and "estudo" (*estudo*), a PARTICIPANT realized by a nominal group. The same phenomenon can be observed by comparing Examples 15 and 16.

Example 15 - A naturally occurring segment from Set B (pair 199) - Qualifier realized by embedded clause

A mecânica é a parte da física [[que **estuda** o movimento]]

a	Mecânica	é	o	ramo	d-a	Física
Participant		Process	Partic-			
Nom. group		Verb. group	Nom.			
ART.DEF.F.SG	mechanics-F.SG	be-PRS.3.SG.IND	ART.DEF.M.SG	field-M.SG	of-PREP-ART.DEF.F.SG	physics-F.SG
the	mechanics	is	the	field	of-the	physics

que	estuda	o	movimento
-ipant			
group			

that-PRON.REL	study-PRS.3.SG.IND	ART.DEF.M.SG	movement-M.SG
that	studies	the	movement

Mechanics is the field of physics concerned with the study of motion.

Example 15 clarifies a relational clause that follows the sequence “Participant-Process(relational)-Participant”, just like Example 14. The second PARTICIPANT is realized by a nominal group with an embedded clause as a QUALIFIER, namely “a parte da física [[que estuda o movimento]]” (*that studies motion*). This QUALIFIER is realized by an embedded clause, which decodes the meaning of “mechanics”, a subfield of physics.

In contrast with Example 16, in which the structure of the QUALIFIER is realized by a non-finite embedded clause, in Example 15 an embedded clause functions as a QUALIFIER. This shows a rank shift that indicates experiential grammatical metaphor.

Example 16 - A manually constructed segment from Set D (pair 199) - Qualifier realized by non-finite embedded clause

A mecânica é a parte da física [[responsável pelo **estudo** e pela análise do movimento e do repouso dos corpos em sua evolução no tempo]].

Mecânica	é	o	ramo	d-a	Física
Participant	Process	Partic-			
Nom. group	Verb. group	Nom.			
mechanics-F.SG	be-PRS.3.SG.IND	ART.DEF.M.SG	field-M.SG	of-PREP-ART.DEF.F.SG	physics-F.SG
mechanics	is	the	field	of-the	physics

que	compreende	o	estudo	e	análise	d-o	movimento
-ip-							
gr-							
that-PRON.REL	comprehend-PRS.3.SG.IND	ART.DEF.M.SG	study-M.SG	and-CONJ	analysis-F.SG	of-PREP-ART.DEF.M.SG	movement-M.SG
that	comprehends	the	study	and	analysis	of-the	movement

e	repouso	d-os	corpos	em	sua	evolução	no	tempo
-ant								
-oup								
and-CONJ	rest-	of-PREP-	bod	in-	PRON.POSS.	evolutio	in-PREP-	time-

	M.SG	ART.DEF. M.PL	y- M.PL	CONJ	M.PL	n- F.SG	ART.DEF. M.SG	M.SG
and	rest	of-the	body	in	their	evolution	in-the	time

Mechanics is the field part of physics concerned with the study of motion and rest state of the bodies in their evolution over time.

The relational clause in Example 16 also follows the sequence "Participant-Process(relational)-Participant". The second PARTICIPANT is realized by a nominal group with an embedded clause as a QUALIFIER, namely: "que compreende o estudo e análise do movimento e repouso dos corpos e sua evolução no tempo" (*responsible for the study and the analysis of the movement and rest state of the bodies in their evolution over time*). This QUALIFIER is realized by a nominal group decoding the meaning of "mechanics", one of the subfields in physics.

In contrast with Example 15, in which the structure of the QUALIFIER is realized by a finite embedded clause, in Example 16, an embedded clause functions as a CLASSIFIER. This shows one more time a rank shift, from group rank to clause rank, which suggests experiential grammatical metaphor.

Finally, similar to Examples 13 and 14, the term in bold "estuda" (*studies*) in Example 15 is an item of verbal class operating in a Process. In Example 16, the term "estudo" (*study*) is the

THING in a nominal group that realizes a PARTICIPANT. This is another evidence of nominalization increasing text complexity due to rank shift.

The next section summarizes the results presented in the earlier sections, associating the categories rank shift and class shift as evidence of distinct metaphoricity degrees.

4.4.3 Overview

All the instances shown in this subsection are summarized in Table 66, which shows the association between each category, illustrated by the examples.

Table 66

Overview of the main experiential grammatical metaphor evidence with examples

Category	Example	Pair ID	Nature	Set	Category degree
Class shifts	1		Naturally occurring	Set A	lower
	2	139		Set B	higher
	3		Manually	Set C	lower

	4		constructed	Set D	higher
Class shifts - Nominalizations	5	61	Naturally occurring	Set A	lower
	6			Set B	higher
	7		Manually constructed	Set C	lower
	8			Set D	higher
	9	134	Naturally occurring	Set A	lower
	10			Set B	higher
	11		Manually constructed	Set C	lower
	12			Set D	higher
Rank shifts	13	199	Naturally	Set A	lower

14	occurring	Set B	higher
15	Manually constructed	Set D	lower
16		Set D	higher

In Table 66 the columns “Category”, “Examples”, “Pair ID”, “Nature”, “Set” and “Category degree” represent, respectively, the phenomenon investigated (in this case class shifts or rank shifts), the number of each example, the segment pair ID, the segment type (naturally occurring or manually constructed), the segment set (Set A, Set B, Set C or Set D) and the metaphorization level of each segment compared to another segment. For instance, Example 1 is contrasted with Example 2, Example 3 with Example 4, and so on.

Finally, taking into account Table 66, in general, Set B and Set D present a higher degree of metaphoricity than Set A and Set C, confirming what had been hypothesized initially, as well as the fact that the compilation criteria were successful.

The next chapter presents the discussion of each research question, bringing evidence and theoretical framework together to consolidate the findings of this thesis.

Chapter 5 Discussion

This chapter discusses the findings from Chapter 4 in accordance with the studies on text simplification from NLP and linguistics and their implications for text complexity, seeking to contribute to text simplification.

In doing so, drawing on Systemic Functional Linguistics (SFL), this thesis investigates text complexity in-depth to find out which variables are associated with different levels of text complexity that lead to the text simplification of science texts in Brazilian Portuguese.

For this purpose, this thesis sought to address the following research questions in the next section.

I. How can Systemic Functional Linguistics map the association between textual complexity and experiential grammatical metaphor in science texts in Brazilian Portuguese?

II. According to Systemic Functional Linguistics (SFL), what evidence of varying metaphoricity degrees in terms of structure and system can be found in science texts in Brazilian Portuguese?

III. Which linguistic patterns discriminate between congruent and non-congruent clauses; in other words, different degrees of experiential grammatical metaphor in Brazilian Portuguese?

IV. Which linguistic patterns indicating experiential grammatical metaphor lead to varying degrees in text complexity in Brazilian Portuguese?

The next section clarifies in detail how this thesis contributes to the main theoretical framework on text complexity and text simplification, from a linguistic and a computational perspective, by addressing the four research questions this thesis proposed.

5.1 Contributions to the theoretical framework

The aim of this section is to set out, through evidence of experiential metaphoricity in the structure and system of Brazilian Portuguese text, the strong association between text complexity and experiential grammatical metaphor. To this aim, linguistic patterns that discriminate between congruent and non-congruent clauses (that is, linguistic patterns that indicate experiential grammatical metaphor leading to different levels of text complexity) were mapped in the light of the literature on linguistic and computational studies.

5.1.1 Research question 1

How can Systemic Functional Linguistics (SFL) establish an association between textual complexity and experiential grammatical metaphor in science texts in Brazilian Portuguese?

This study has found that generally the higher the experiential grammatical metaphor, the higher the text complexity. This finding is consistent with that of Steiner (2004, 2005), who proposed hypotheses on the relationship between information density and grammatical metaphor and information explicitness.

Further analysis of the data revealed that metaphoricity allows information to accumulate while the text unfolds, mainly through nominalization, a special type of class shift (Ravelli, 1999). In other words, the use of nominal groups was critical in realizing (IDEATIONAL) meanings of clauses that were previously mentioned. These results corroborate the findings of Ravelli (1999), who highlights the relevance of nominalization in English for experiential grammatical metaphor. Furthermore, for the text complexity to vary, other linguistic strategies were found to play an essential role.

One contributing strategy observed in this thesis was the use of free or bound clauses in a relationship of experiential grammatical metaphor, with embedded clauses which operate as a qualifier in a nominal group. This can be related to rank shifts, allowing the realization of (most of) the clause meanings as part of the meaning from a nominal group. In other words, the use of these clauses is the same as using the clausal variant rather than the nominal variant. These results are consistent with those of Halliday (2004, p. 36). Halliday further suggested that a nominal variant tends to be more metaphorical than the clausal variant due to three reasons, each standing for one of these time perspectives: the three histories of text - phylogenetic; ontogenetic; and logogenetic.

This study, then, supports previous research into text complexity, particularly the issues which link Ravelli (1999) and Steiner (2004, 2005), extending such research to Brazilian

Portuguese and confirming that similar patterns are in play. This thesis investigated common text complexity measures, namely lexical density and grammatical intricacy, as well as other possible indicators.

Drawing on SFL, lexicogrammatical patterns, both systemic and structural perspectives were investigated thoroughly. The systemic patterns were regularities in the systemic choices observed, mainly in the following systems: AGENCY, PROCESS, POLARITY, MOOD, THEME, and CONJUNCTION. The structural patterns related to the distinct configurations of elements, focusing specifically on rank shifts (nominalization included), in fact mainly embedded clauses functioning as qualifiers in the nominal group.

Different types of configurations were analyzed in search of regular patterns. These patterns were grouped as PROCESS + MEDIUM and AGENT + PROCESS, which showed that the sequence PROCESS + MEDIUM is more frequent in Set B and Set D and sequence AGENT + PROCESS in Set A and Set C. The first sequence, PROCESS + MEDIUM, is prototypical of middle clauses, and the second, AGENT + PROCESS, prototypical of effective clauses, which leads to the conclusion that Set A and Set C are more congruent, or less metaphorical than Set B and Set D.

The findings regarding distinct configurations of semantic elements (PARTICIPANT, PROCESS, CIRCUMSTANCE) support the notion that rank shifts are related to embedded clauses functioning as qualifiers in the nominal group. This evidence again supported Ravelli's (1999) statements concerning agnation which, together with rank and class shifts, are another way of viewing experiential grammatical metaphor. Even though the concept of agnation was not the focus of this thesis, this phenomenon emerged when comparing multiple texts with similar meanings.

In sum, by investigating the differences between simpler and more complex segments, this thesis found that the most important factor for varying text complexity levels was nominalization - a special case of class shift. Nominalization was investigated by analyzing systemic and structural patterns in the sequences of semantic functions (PARTICIPANT, PROCESS, CIRCUMSTANCE).

5.1.2 Research question 2

What evidence of varying metaphoricity degrees in terms of structure and system can be found in science texts in Brazilian Portuguese?

The structural patterns most concerned in this issue were the distribution of the semantic functions in the clauses with respect to the categories PARTICIPANT, PROCESS, CIRCUMSTANCE, as well as the types of circumstances that occurred. The systemic patterns consisted of systemic categories assigned to the clauses, drawing on SFL, according to all three metafunctions (IDEATIONAL, INTERPERSONAL, and TEXTUAL).

These results concerning the distribution of the semantic functions in the clauses based on the categories PARTICIPANT, PROCESS, CIRCUMSTANCE showed that some patterns prevailed in more metaphorical segments and others in less metaphorical ones. These were associated with the complexity of the examples analyzed, confirming that a higher degree of metaphoricity implies a higher information density. This finding was also reported by Steiner (2005, pp. 21-23), who hypothesized this possibility.

In accordance with Halliday and Martin (1993), the systemic choices indicating a higher degree of nominalization were more frequent in the more metaphorical segments, and vice

versa. These patterns were mainly the high frequency of relational processes (middle voice) in comparison with material processes (effective voice). A possible explanation for this might be that the occurrence of relational PROCESSES is associated with a higher level of metaphoricity, while material processes indicate lower metaphoricity.

The choices that were most prominent in the system were in accord with the findings obtained by analyzing the structure of the clauses, confirming the role played by nominalization as a device to gradually accumulate meanings thereby increasing the degree of text complexity. These results provide further support for the hypothesis that both structural and systemic patterns contribute to varying degrees in text complexity, which would mean that some patterns may lead to simpler versions of science text.

5.1.3 Research question 3

Which linguistic patterns discriminate between congruent and non-congruent clauses; in other words, different degrees of experiential grammatical metaphor in Brazilian Portuguese?

Aiming at addressing the issue of discriminating between congruent and non-congruent clauses, this study compiled SIM-Pt, a corpus of 200 text segments that were later on sampled into 150 segments encompassing the domains of psychology, biology, and physics to spread the sample. These segments were grouped into two groups: the first was text segments collected from science websites; the second was rewritings of the latter performed by researchers from LETRA.

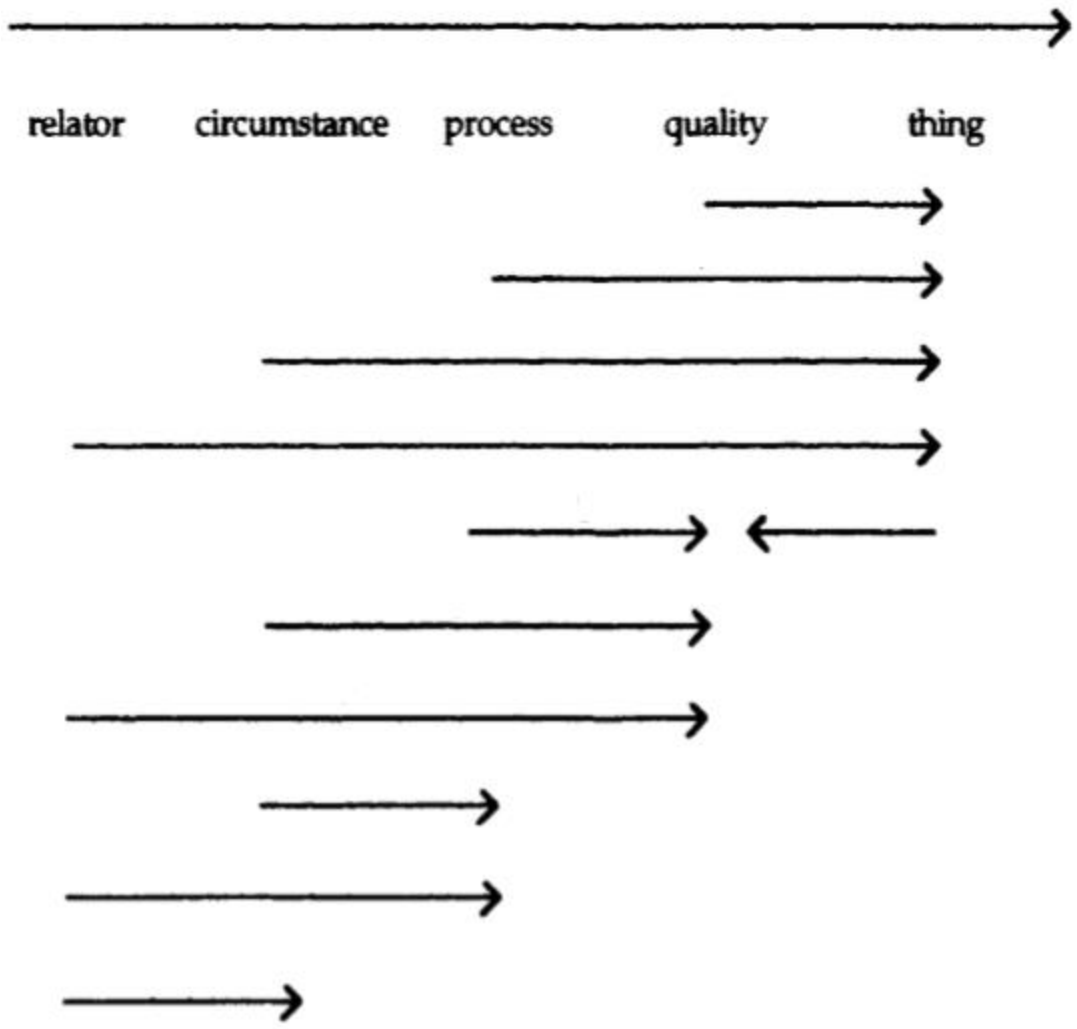
The segments were organized in pairs and labeled as belonging to different sets. Set A and Set B were the segments retrieved from science websites and Set C and Set D were their rewritings. This thesis assumed that Set A and Set C were simpler than Set B and Set D, which would suggest that Set A and Set C were more congruent (or less metaphorical). This assumption is based on a process of rewriting, which necessarily requires metaphorization; as a consequence, this would change the level of complexity of the text. This assumption was tested with the use of Translation studies' undergraduate students, which led to a reclassification of all segments based on their complexity level. The analysis of the segments from the perspective of the system and the structure followed mostly Halliday & Martin (1993), Halliday & Matthiessen (1999), Ravelli (1999). Other authors contributed as well, such as the authors from the computational studies who provided the basis for the definition of text simplification in the light of text complexity, mainly Siddharthan (2004, 2006), Specia, Aluísio & Pardo (2008), Štajner (2015), and Hwang et al (2015), who provided the main text simplification strategies interpreted in the light of the SFL. Some practical strategies of simplification were shared across the authorities, e.g., dis-embedding of relative clauses and separation of subordinate clauses (Siddharthan, 2004, 2006), and sentence splitting (Štajner, 2015). But the operations were understood in quite different textual contexts. Computational authorities were not allocating the changes in clause structure to a global registerial shift, but rather as isolates. In SFL, the shifts towards complexity are viewed as part of a cultural 'syndrome' – a wave or 'drift' of features which are a response to the evolving demands of manipulating abstractions, that is, symbolic formulations that rely on mutual values.

The findings confirmed that nominalization was the most relevant factor for determining different levels of text complexity, based on class and rank shifts, including the use of embedded

clauses as nominal group qualifiers. Furthermore, most segments suggested that the distinction between Set A to Set B and from Set B to Set D follows the direction of nominalization, as shown in Figure 54.

Figure 54

Direction of nominalization



Note: Adapted from Halliday & Matthiessen (1999, p. 262).

Figure 54 shows at the top, from left to right, an arrow pointing out to the right, e.g., the direction in which nominalization increases depending on the semantic category used to realize meanings. Further arrows below show possible modifications that lead to an increase or decrease (if you consider the arrow pointing left) in metaphoricity, which results in an increase in text complexity.

Regarding the issue of the discrimination between congruent and non-congruent clauses, Set A and Set C were found to be simpler and less complex than Set B and Set D, which suggested that Set A and Set C were congruent and Set B and Set D non-congruent (or more metaphorical). A possible explanation for this might be that text congruence is a continuum measured by the level of text complexity, which can be quantified by analyzing experiential grammatical metaphor. This relationship implies that varying degrees in experiential grammatical metaphor - those that determine a lower metaphoricity level - should simplify texts.

5.1.4 Research question 4

Which linguistic patterns indicating experiential grammatical metaphor lead to varying degrees in text complexity in Brazilian Portuguese?

To obtain a reliable measure of text complexity, this thesis investigated several approaches of text simplification from both the computational and linguistic studies, aiming at describing criteria for producing varying text complexity levels. The first step was investigating the viability of using lexical density to measure text complexity since even though lexical density

has been used extensively in the literature, there was no consensus on the way it is measured, that is, when studies from computational and linguistic perspectives were compared.

According to Sokolova and Bobicev (2018, p. 4), lexical density is the measure of the frequency of lexical terms ("content words") compared to the sum of all words from a text, both lexical and grammatical terms. This implies that the higher the lexical density, the more meanings, and consequently the higher is the cognitive effort needed to interpret the text. In contrast, according to Ravelli (2006, p. 55) and Halliday (1993, 76), lexical density is the "measure of the proportion of lexical items in a clause".

Surprisingly, the comparison of both ways of measuring lexical density to assess the viability of one metric over another, as discussed in the exploratory study, proved that lexical density is not such a reliable measure for text complexity. This inconsistency may be due to Ravelli's and Halliday's affiliation to Systemic Functional Linguistics, which often regards the clause rank as the main unit of analysis to be able to explain linguistic phenomena; therefore, the need to calculate text complexity in relation to the clause. Due to this inconsistency, a new method drawing on Systemic Functional Linguistics was proposed to obtain linguistic patterns relevant for explaining differences in text complexity. This method was the count of semantic functions (PARTICIPANT, PROCESS, CIRCUMSTANCE) in the clauses from SIM-Pt corpus, which allowed for comparison, i.e. between the first group as Set A and Set C and the second group as Set B and Set D.

Again, systemic and a structural perspective complement each other. The main systemic patterns that emerged from the analysis were the higher frequency of effective clauses of the material type in the simpler segments and more congruent segments (Set A and Set C), in contrast with

the higher occurrence of middle clauses of the relative and existential types in the more complex and more metaphorical segments (Set B and Set D). However, these were not unexpected findings, as they were in accordance with SFL. Other structural patterns that emerged were more relevant in agreement with these systemic results, such as the class and the rank shifts (including the cases of embedded clauses as qualifiers) indicating experiential grammatical metaphor, corroborating Ravelli (1999) and Steiner (2004, 2005).

Further structural results were the counts of the semantic functions PARTICIPANT, PROCESS, CIRCUMSTANCE, and MEDIUM. These were further classified into categories from the group rank, namely AGENT, MEDIUM, RANGE, and BENEFICIARY. These findings were grouped by segment metaphoricity level (Set A with Set C and Set B with Set D) and showed that while the most frequent pattern at the clause rank (PARTICIPANT + PROCESS + PARTICIPANT) was consistent for all segments; at the group rank, different patterns prevailed for each group. For Set A and Set C, the most frequent pattern was "AGENT + PROCESS", in comparison with "PROCESS + MEDIUM", more frequent in Set B and Set D. This also accords with the earlier results from the systemic perspective, clarifying that "AGENT + PROCESS", typical of material clauses, were less complex than relational clauses, typically with the pattern "PROCESS + MEDIUM". Thus, these results proved that Set A and Set C were less complex and less metaphorical because their patterns reveal mostly clausal variants rather than nominal variants, more frequent in Set B and Set D.

5.2 Conclusion

This thesis investigated manipulations in the degree of complexity of the clauses retrieved from the science text segments on the domains of psychology, biology and physics through different levels of experiential grammatical metaphor. Some simplification strategies used by

Computer Science studies were consistent with these findings, specifically those associated with class or rank shifts. Although some strategies were quite similar to another, the interpretation of these strategies in the light of SFL allowed this study to be performed without splitting the simplification strategies into further groups. The findings also showed that systemic patterns related to the IDEATIONAL, INTERPERSONAL and TEXTUAL metafunctions indicated different text complexity levels.[49] These manipulations were mapped on the basis of structural patterns that were counted and allowed a descriptive statistical description of varying degrees in metaphoricity and text complexity. Yet, for the analysis to be performed as mentioned, some linguistic requirements are necessary, namely the adequate linguistic classification according to SFL.

Fulfilling these conditions, the findings provide some support for the conceptual premise that it is feasible and effective to apply a hybrid proposal drawing on the linguistic background of SFL and the computational methods available nowadays, potentially improving their efficiency and/or reducing their processing time if properly given. Yet, it is necessary to account for text complexity using methods compatible with the theory, as there was an inconsistency between lexical density measurements typically used in linguistic and computational studies, possibly explained by the affiliation of some authors to a linguistic theory, such as SFL. Nevertheless, whereas lexical density measurement is not an efficient method by itself to

[49] Some authors argue that the systemic description of Brazilian Portuguese is particularly different compared to other languages, namely Gouveia (2010), who suggests that Portuguese is a Finite-less language, and Ventura and de Lima-Lopes (2020) and Gouveia and Barbara (2004), who propose contrasting approaches on how a concept of Theme might be applied to a "pro-drop" language like Portuguese. Yet, this issue was beyond the scope of this thesis, which focused on text complexity. Due to this scope, for Brazilian Portuguese description, this thesis follows Figueredo's (2007, 2011), as well as other works on systemic description of Brazilian Portuguese, namely Araújo (2007), Pagano, Ferreguetti e Figueredo (2011) Ferreguetti (2014, 2018), Figueredo, Pagano e Ferreguetti (2014), Sá (2016, 2020), Braga (2016), Monteiro (2016), Rosa (2017), A. Paula (2017), Alves (2017, 2018).

discriminate between less and more metaphorical text segments, combined with grammatical intricacy, lexical density can still be used to investigate text complexity.

As a conclusion, text congruence seems to configure a continuum measured by the level of text complexity, quantifiable by analyzing experiential grammatical metaphor. Therefore, varying degrees in experiential grammatical metaphor that decrease the metaphoricity should reduce text complexity, improving text comprehension. This way, texts could become more accessible to audiences by manipulating experiential grammatical metaphor to reduce their complexity.

Further studies will need to ascertain these assumptions and improve on the most specific points. The next chapter presents the conclusions of this thesis, on the basis of the development of this thesis' objectives with a sound methodology and drawing on the relevant literature.

Chapter 6 Conclusions

The main goal of the current study was to investigate, from a systemic functional perspective, underlying lexicogrammatical patterns from science text segments in Brazilian Portuguese to produce a linguistic representation that may characterize different degrees of text complexity.

This study has identified some text simplification strategies from the NLP literature (Siddharthan, 2004, 2006; Specia, Aluísio & Pardo, 2008; Štajner, 2015; and Hwang et al, 2015)

mostly related to syntactic and lexical substitution strategies that can be associated with some concepts from Systemic Functional Linguistics (SFL), drawing on mostly on Halliday & Martin (1993), Halliday & Matthiessen (1999), and Ravelli (1999). Other authors contributed as well, such as the authors from the computational studies who provided the basis for the definition of text simplification in the light of text complexity, for example: Siddharthan (2004, 2006), Specia, Aluísio & Pardo (2008), Štajner (2015), and Hwang et al (2015), who provided the main text simplification strategies interpreted in this thesis in the light of the SFL.

The second major finding was that the most frequent systemic and structural patterns are indicative of experiential grammatical metaphor as a phenomenon associated with varying degrees of text complexity, either to increase or decrease text comprehension levels; in the latter case, these modifications can make texts simpler. Evidence associating varying degrees in text complexity and experiential grammatical metaphor was also unraveled, aligned with Steiner's (2015) hypotheses and Ravelli's (1999) studies on experiential grammatical metaphor. Pieces of evidence for this include instances in which class shifts, rank shifts, and embedded clauses were used to increase the complexity and the metaphoricity of texts.

This research supports the idea that Systemic Functional Linguistics is a broad enough linguistic perspective to provide a linguistic background for computational implementations aiming at making texts simpler or more complex. That is to say, the grammatical patterns can be used to train systems to perform text simplification tasks and to develop guidelines for simplifying science texts in Brazilian Portuguese. As a consequence, these representations may be adapted to the study of text simplification in other languages.

The simplification strategies mapped from NLP literature allowed the compilation of SIM-Pt, a corpus of text segments from the domain of psychology, biology, and physics with 150 segment pairs organized in pairs. In each pair, one segment is simpler and less metaphorical and the other more complex and more metaphorical. SIM-Pt comprises four sets of segments: Set A and Set C, which are less metaphorical than Set B and Set C. Set A and Set B were naturally occurring segments retrieved from science texts available at websites, and Set C and Set D were rewritings of the latter; Set A being rewritten as Set C and Set B being rewritten as Set D. The compilation methodology validated the hypothesis that rewriting is a valid method to produce further segments for analysis. The corpus compilation suggests that similar corpora can be obtained for different domains or even different registers to contribute to studies in both NLP and applied linguistics.

This is the first study of text simplification of science texts in Brazilian Portuguese in the light of Systemic Functional Linguistics, providing a new understanding of which linguistic features are associated with distinct degrees of text complexity. This approach can assist not only people learning in a range of scientific domains but also people with learning disabilities and professionals who rewrite texts aiming at audiences with different levels of literacy.

This work also contributes greatly for translation studies technology, which can include simplification systems, according to Holmes (1988/2000) and Quah (2006)., by providing a text simplification corpus and guidelines. The guidelines provided by this thesis can assist both translators and researchers interested in text simplification, not only from Linguistics but also from NLP.The limitations of this work were, firstly, the scarcity of literature relating text simplification to experiential grammatical metaphor, though Steiner (2005, pp. 21-23) raises this

relationship. Secondly, no software was available to classify texts automatically according to the categories from SFL; so, it was not possible to work with a great number of texts. Thirdly, no corpus was readily available for analyzing experiential grammatical metaphor in Brazilian Portuguese, which required the compilation of the SIM-Pt corpus. Fourthly, due to time constraints, it was not possible to use a second annotator or reviser to assess the quality of the annotation; however; to reduce the possibility of bias the segments' classification in terms of complexity was revised on the basis of student feedback. Lastly, due to the specificities of the experiential grammatical metaphor analysis, the size of the corpus and the scope of the analysis were limited, making it necessary to go further into the ideational patterns associating experiential grammatical metaphor and text simplification.

Despite its limitations, the study adds to the understanding of text complexity in science texts in Brazilian Portuguese. The study also reviews the text simplification strategies from the NLP literature and an interpretation of these strategies in the light of Systemic Functional Theory. Also, although the current study analyzed a small sample of texts, the findings suggest that patterns associated with either a lower or higher degree of experiential grammatical metaphor may guide writers to produce more comprehensible science texts in Brazilian Portuguese. In sum, notwithstanding these limitations, the study suggests the analysis of ideational grammatical metaphor as input to text simplification strategies which would draw on both NLP and applied linguistics to obtain more effective results. Considerably more work will need to be done to determine text complexity features regarding INTERPERSONAL and TEXTUAL metafunctions, and for in-depth analysis drawing on pieces of evidence of all three metafunctions.

A natural progression of this work is to implement text simplification systems. This can be performed by rendering these patterns in the linguistics representation as a computational representation (which usually uses mathematical logic) to implement the patterns or rules as software. This progression is shown in Figure 55, with the focus of this study emphasized in bold.

Figure 55

A natural progression of this study

Note. Elaborated by the author for this research.

Figure 55 shows a way to create a direct interface between this linguistics study and computational linguistics that concerns the application of linguistic knowledge in real-world problems. To perform this task, it is necessary, first, to investigate the problem and the possible solutions to it, organizing which patterns can be used to describe and solve it. Then these patterns would be rendered as computational logic to become the root of a specific computational system designed to solve this problem.

The findings of this study have some practical implications, namely the possibility of training text simplification, text summarization, or readability assessment systems to perform these tasks automatically. This information can be used to develop text simplification representations and systems for other domains or even other languages.

References

Aluísio, S. M., Specia, L., Pardo, T. A. S., Maziero, E. G., Caseli, H., & Fortes, R. P. (2008). A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems. In *Proceedings of the 26th Annual International Conference on Design of Communication* (pp. 15–22). Lisbon, Portugal. Retrieved from https://www.researchgate.net/publication/220961980_A_corpus_analysis_of_simple_account_texts_and_the_proposal_of_simplification_strategies_First_steps_towards_text_simplification_systems

Aluísio, S. M.; Specia, L.; Gasperim, C.; Scarton, C.; (2010). Readability Assessment for Text Simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1-9). São Paulo, Brasil. Retrieved from http://www.aclweb.org/anthology/W10-1001

Alves, L. 2017. *Uma proposta de descrição sistêmico-funcional das orações materiais do português brasileiro orientada para os estudos multilíngues*. [Master ‘s Thesis, Universidade Federal de Ouro Preto]. Retrieved from https://www.repositorio.ufop.br/handle/123456789/10170

Alves, R. J. 2018. *Para além da oração: uma descrição sistêmico-funcional do sistema de conjunção do português brasileiro*. [Master ‘s Thesis, Universidade Federal de Ouro

Preto]. Retrieved from https://www.repositorio.ufop.br/bitstream/123456789/10619/1/DISSERTA%C3%87%C3%83O_Al%C3%A9mOra%C3%A7%C3%A3oDescri%C3%A7%C3%A3o.pdf

Aranzabe, M. J., Ilarraza, A. D. de, & Gonzalez-Dios, I. (2013). Transforming Complex Sentences using Dependency Trees for Automatic Text Simplification in Basque. *Procesamiento Del Lenguaje Natural*, 50(0), 61–68. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4660

Araújo, C. 2007. *O sistema semântico de PROJEÇÃO e sua dispersão gramatical em português brasileiro: uma descrição sistêmico-funcional orientada para os estudos linguísticos da tradução*. [Master ‘s thesis, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/VCSA-77SQAQ

Barlacchi, G., & Tonelli, S. (2013). ERNESTA: A Sentence Simplification Tool for Children’s Stories in Italian. *Computational Linguistics and Intelligent Text Processing*, 476–487. https://doi.org/10.1007/978-3-642-37256-8_39. Retrieved from https://gbarlacchi.github.io/papers/2013/Cicling-Ernesta.pdf

Bateman, J. A. (1990). Finding translation equivalents: an application of grammatical metaphor. Retrieved from https://www.aclweb.org/anthology/C90-2003.pdf

Bhagat , R., & Hovy, E. (2013). What Is a Paraphrase? *The MIT Press Journals, 39*(3), 463-472. Doi:https://doi.org/10.1162/COLI_a_00166

Barlacchi, G., & Tonelli, S. (2013). ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. *Computational Linguistics and Intelligent Text Processing*, 476–487. https://doi.org/10.1007/978-3-642-37256-8_39

Byrnes, H. (2009). Emergent L2 German writing ability in a curricular context: A longitudinal study of grammatical metaphor. *Elsevier*. *20*(1), 50-66. Doi: 10.1016/j.linged.2009.01.005. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0898589809000084

Braga, A. B. C. 2016. *O sistema de Transitividade no inglês e no português brasileiro: caracterização da função Circunstância com base em textos originais e traduzidos*. [Master 's thesis, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/MGSS-ADHMCX

Burstein, J. (2009). Opportunities for Natural Language Processing Research in Education. *Computational Linguistics and Intelligent Text Processing*, 6–27. https://doi.org/10.1007/978-3-642-00382-0_2

Candido Jr. A., Maziero E., Gasperin, C., Pardo, T., Specia, L. and Aluísio, S. (2009, June). Supporting the Adaptation of Texts for Poor Literacy Readers: a Text Simplification Editor for Brazilian Portuguese. In *Proceedings of the NAACL HLT Workshop on Innovative Use of NLP for Building Educational Applications*, (pp. 34–42). Boulder, Colorado. Retrieved from www.aclweb.org/anthology/W09-2105

Carnie, A. (2013). *Syntax: A Generative Introduction*. New Jersey: John Wiley & Sons. Retrieved from https://archive.org/stream/GenerativeGrammar/Generative%20grammar_djvu.txt

Carroll, J., Minnen, G., Pearce, D., Canning, Y., Devlin, S., & Tait, J. (1999). *Simplifying Text for Language-Impaired Readers*. Retrieved April 24, 2021, from https://www.aclweb.org/anthology/E99-1042.pdf

Caseli, H. M.; Pereira, T.F., Specia, L.; Pardo, T.A.S.; Gasperin, C.; Aluísio, S.M.; (2009). Building a Brazilian Portuguese parallel corpus of original and simplified texts. In Alexander Gelbukh (ed), *Advances in Computational Linguistics, Research in Computer Science*, vol 41, pp. 59-70. 10th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2009), March 01–07, Mexico City.

Chen, H., Huang, H., Chen, H., & Tan, C. (2012). A simplification translation-restoration framework for cross-domain SMT applications. In *Proceedings of COLING 2012* (pp. 545-560). Mumbai, India: The COLING 2012 Organizing Committee. Retrieved from https://www.aclweb.org/anthology/C12-1034/

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Massachusetts: MIT Press.

Chung, J., H. J., M., & Kim, J. (2013). Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics* (pp. 30:1–30:10). New York, NY, USA: Association for Computing Machinery. Retrieved from https://dl.acm.org/doi/abs/10.1145/2479787.2479808

Daelemans, W., Hothker, A., & Sang, E. (2004). Automatic sentence simplification for subtitling in Dutch. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, (pp. 1045-1048). Retrieved from https://www.clips.uantwerpen.be/~walter/papers/2004/dhs04.pdf

Dagan, I., Roth, D., Sammons, M., & Fabio Massimo, Z. (2013). Recognizing Textual Entailment Models and Applications. Em I. Dagan, D. Roth, M. Sammons, F. Zanzotto, & M. Claypool, *Recognizing Textual Entailment: Models and Applications* (pp. 1-220). San Rafael, California: Morgan and Claypool.

Damay, J. J. S., Lojico, G. J. D., Lu, K. A. U., Tarantan, D. E., Ong, E. C. (2006, February). SIMTEXT: Text Simplification of Medical Literature. In *3rd National Natural Language Processing Symposium – Building Language Tools and Resources* (pp. 34-38). Manila, Philippines, 3. Retrieved from https://www.researchgate.net/publication/237251066_SIMTEXT_Text_Simplification_of_Medical_Literature

Derewianka, B. (2003). Grammatical metaphor in the transition to adolescence. In A.-M. Simon-Vandenbergen, M. Tavernier, & L. Ravelli (Eds.), Grammatical metaphor: Views from systemic functional linguistics (pp. 185–219). Amsterdam/Philadelphia: Benjamins.

Drndarevic, B., Štajner, S., Bott, S., Bautista, S., & Saggion, H. (2013). Automatic text simplification in spanish: A comparative evaluation of complementing modules. Em A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing ser. Lecture Notes in Computer Science. 7817*, p. 7817. Berlin, Heidelberg: Ed. Springer.

De Belder, J. & Moens, M. (2012, March). A dataset for the evaluation of lexical simplification. In *Proceedings of the 13th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing-2012)* (pp. 426–437). New Delhi, India. Retrieved from https://liir.cs.kuleuven.be/publication_files/1044lseval-1.pdf

Fajardo, I., Tavares, G., Ávila, V., & Ferrer, A. (2013). Towards text simplification for poor readers with intellectual disability: When do connectives enhance text cohesion? *Research in Developmental Disabilities*, *34*(4), 1267–1279. https://doi.org/10.1016/j.ridd.2013.01.006

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. *Applied Psycholinguistics*. 22(1), 131-134. Retrieved from https://www.cambridge.org/core/journals/applied-psycholinguistics/article/wordnet-an-electronic-lexical-database-christiane-fellbaum-ed-cambridge-ma-mit-press-1998-pp-423-/8A9F540FB453B327C1AF0AC74E2F7D4D

Ferreguetti, K. 2014. *As orações existenciais em inglês e português brasileiro: um estudo baseado em corpus*. [Master ‘s thesis, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/MGSS-9PMPQA

Ferreguetti, K. 2018. *A frase preposicional com função de Qualificador no grupo nominal: um estudo de equivalentes textuais no par linguístico inglês e português brasileiro*. [Doctoral dissertation, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/LETR-B2JH2G

Figueredo, G. P. 2007. *Uma descrição sistêmico-funcional da estrutura do grupo nominal em português orientada para os estudos lingüísticos da tradução*. [Master ‘s thesis, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/MGSS-77ZJ7W

Figueredo, G. P. 2011. *Introdução ao perfil metafuncional do português brasileiro: contribuições para os estudos multilíngues*. [Doctoral dissertation, Federal University of

Minas Gerais]. Retrieved from http://www.bibliotecadigital.ufmg.br/dspace/handle/1843/DAJR-8GLS6E

Figueredo, G. P.; Pagano, A. S.; Ferreguetti, K (2014). Os sistemas textuais de focalização na organização funcional da gramática do Português Brasileiro. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 30(2), 88–115.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *John Benjamins Publishing Company, 14*(3), pp. 393–417. Doi:10.1075/ijcl.14.3.05flo

Gasperin, C.; Maziero, E.; Aluísio, S. M.; (2010, April). Challenging Choices for Text Simplification. In *PROPOR 2010: Computational Processing of the Portuguese Language*, (pp. 40-50). São Carlos, Brazil. Retrieved from https://link.springer.com/book/10.1007/978-3-642-12320-7

Gonzalez-Dios, I. 2016. *Readability Assessment and Automatic Text Simplification: The Analysis of Basque Complex Structures*. [Doctoral dissertation, University of the Basque Country]. Retrieved from http://www.ixa.eus/sites/default/files/dokumentuak/4102/TESIS_GONZALEZ_DIOS_ITZIAR(eng).pdf

Gries, S. T. (2009). What is Corpus Linguistics?. *Language and Linguistics Compass,* 3(5), 1225–1241. https://doi.org/10.1111/j.1749-818x.2009.00149.x

Halliday, M.A.K. 1956. *Grammatical categories in Modern Chinese*. Transactions of the Philosophical Society 1: 177–224. https://doi.org/10.1111/j.1467-968X.1956.tb00567.x

Halliday, M.A.K. (1985). *Introduction to functional grammar*. London: Arnold.

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. *In: Computational and quantitative studies* (pp. 74–86). New York: Continuum.

Halliday, M. A. K. (1993). Some Grammatical Problems in Scientific English. In: Halliday, M. A. K., Martin, J. *Writing Science: Literacy and Discursive Power* (pp. 76–94). London: The Falmer Press.

Halliday. M. A. K (2002). On grammar. London: Continuum.

Halliday, M. A. K. Language and Knowledge: the 'Unpacking' of Text. In Webster, J. (Ed.) *The Language of Science*. London/New York: Continuum, 2004. P. 24-48.

Halliday. M. A. K., & Matthiessen, M. I. M. (1994). *Introduction to Functional Grammar* (2nd ed.). London: Edward Arnold.

Halliday. M. A. K., & Matthiessen, M. I. M. (1999). *Construing Experience Through Meaning: A Language-Based Approach to Cognition*. London: Continuum.

Halliday. M. A. K., & Matthiessen, M. I. M. (2014). *Halliday's Introduction to Functional Grammar* (4nd ed.) . New York: Routledge.

Halliday, M. A. K., Martin, J. (1993). *Writing Science: Literacy and Discursive Power*. London: The Falmer Press.

Halliday, M. A. K. (2007). *Language and Education*. London: Continuum.

Halliday, M. A. K. (2008). *Complementarities in Language*. Beijing: The Commercial Press.

Halliday, M. A. K., & Hasan, R. (2013). *Cohesion in English*. London: Routledge. https://www.taylorfrancis.com/books/cohesion-english-halliday-ruqaiya-hasan/10.4324/9781315836010 (Original work published 1976)

Harris, Z. (1957). *Co-occurrence and Transformations in Linguistic Structure*. *Language,* 33(3), 283-340.

Harris, Z.S. (1965). *Transformation theory*. Reprinted in Harris (1970), pp. 533-577.

Hjørland, B. (2014). Are relations in thesauri "context-free, definitional, and true in all possible worlds"?. *Journal of the Association for Information Science and Technology*, *66*(7), 1367–1373. https://doi.org/10.1002/asi.23253

Holmes, J. (1988/2000). The name and nature of translation studies. In *Translated! Papers on Literary Translation and Translation Studies*, (pp. 81–92). Amsterdam: Rodopi.

Horn, C., Manduca, C., & Kauchak, D. (2014). Learning a lexical simplifier using wikipedia. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, (pp. 458–463). Baltimore, USA, 52. Retrieved from http://www.aclweb.org/anthology/P14-2075

Hung, B., Minh, N., & A., S. (2012). Sentence splitting for Vietnamese-English machine translation. *Knowledge and Systems Engineering (KSE) 2012 Fourth International Conference*, (pp. 156-160). Retrieved from https://ieeexplore.ieee.org/document/6299413

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT,* (pp. 211–217). Denver, Colorado. Retrieve from https://www.aclweb.org/anthology/N15-1022.pdf

Indig, B, Knap, A., Sárköozi-Lindner, Z., Timári, M. and Palkó, G. (2020, May). The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata. In *Proceedings of the 12th Web as Corpus Workshop (LREC 2020)*, (pp. 33–41). Marseille. Retrieved from https://www.aclweb.org/anthology/2020.wac-1.5.pdf

Ji, Y. & Eisenstein, J. 2013. Discriminative Improvements to Distributional Sentence Similarity. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 891-896, Seattle, Washington.

Kauchak, D. (2013) Improving Text Simplification Language Modeling Using Unsimplified Text Data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, (pp. 1537-1546). Middlebury, USA, 51. Retrieved from http://www.aclweb.org/anthology/P13-1151.pdf

Kajiwara, T, & Komachi, M. (2016, December). Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, (pp. 1147–1158). Osaka, Japan, 26. Retrieved from http://aclweb.org/anthology/C16-1109

Klaper, D., Ebling, S., & Volk, M. (2013). Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 11-19). Sofia, Bulgaria: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W13-2902.pdf

Lee, J. S.Y.; Cheung, L. M. E.; Saberi, Dariush; Webster, J. J. (2019) Expanding students' registerial repertoire with a writing assistance tool. *Science Direct*, *42*. doi: https://doi.org/10.1016/j.jeap.2019.100777. Available at https://www.sciencedirect.com/science/article/abs/pii/S1475158519300797

Lester, J. N., O'Reilly, M. (2018). *Applied conversational analysis: social interaction in institutional settings*. New York: SAGE Publications.

Liddy, E. D. Natural Language Processing. (2001). In Miriam A Drake (Ed), *Encyclopedia of Library and Information Science*. (2nd ed.). New York: Marcel Dekker Inc

Lin, Dekang and Lin Pantel. 2001. DIRT - Discovery of Inference Rules from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 323–328, San Francisco, CA

Madnani, N., & Dorr, B. (2010). Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *36*, pp. 341-387. Association for Computational Linguistics.

Mahboob, A., & Knight, N. K (Eds). (2010). *Appliable Linguistics. Sydney:Continuum.*

Mair, C. (2019). Corpus linguistics meets sociolinguistics: The role of corpus evidence in the study of sociolinguistic variation and change. *Language and Computers*, 69(1), pp. 7-32. doi:10.1163/9789042025981_003

Martin, J. R. (1992). *English Text: System and Structure*. Philadelphia/Amsterdam: John Benjamins.

Martin, J. R. (2013). *Systemic Functional Grammar: a next step into the theory - Axial relations*. Beijing: Higher Education Press.

Matthiessen (2015). Register in the round: registerial cartography. *Functional Linguistics*, 2 (9), pp. 1-48. doi:10.1186/s40554-015-0015-8

Matthiessen, C. M. I. M. (2012). Systemic Functional Linguistics as appliable linguistics: social accountability and critical approaches. *DELTA: Documentação de Estudos Em Lingüística Teórica e Aplicada*, 28(spe), 435–471. https://doi.org/10.1590/s0102-44502012000300002

Max, A. (2005). Simplification interactive pour la production de textes adapts aux personnes souffrant de troubles de la compréhension. *Proceedings of Traitement Automatique Des Langues Naturelles (TALN)*.

Max Planck Institute. (2015). *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses.* Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology - Department of Linguistics. Retrieved from https://www.eva.mpg.de/lingua/resources/glossing-rules.php

Max Planck Institute. (2020). *Systemic Functional glossing conventions.* Leipzig, Germany: Max Planck Institute for Evolutionary Anthropology - Department of Linguistics.

Monteiro, G. F. 2016. *Da organização oracional ao fluxo do discurso: o Adjunto e a vírgula sob perspectiva sistêmico-funcional*. [Master 's thesis, Universidade Federal de Ouro Preto].

Navigli, R, & Ponzetto, S. P. (2010, July). BabelNet: Building a Very Large Multilingual Semantic Network. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (pp. 216–225), Uppsala, Sweden, 48. Retrieved from http://www.mt-archive.info/ACL-2010-Navigli.pdf

Nevěřilová, Z. (2014). Paraphrase and Textual Entailment Generation. *Text, Speech and Dialogue*, 293–300. https://doi.org/10.1007/978-3-319-10816-2_36

NILC – Interinstitutional Center for Computational Linguistics – Núcleo Institucional de Linguística Computacional. Retrieved from http://www.nilc.icmc.usp.br/nilc/index.php/projetos?layout=edit&id=27.

Neumann, S., & Hansen-Schirra, S. (2005). *The CroCo Project Cross-linguistic corpora for the investigation of explicitation in translations*. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/ContrastiveCorpusLinguistics/thecrocoproject.pdf

Neumann, Stella. (2013). *Contrastive register variation. A quantitative approach to the comparison of English and German*. De Gruyer: Berlin. https://doi.org/10.1515/9783110238594.

Paetzold, G. H., & Specia, L. (2017). A Survey on Lexical Simplification. *Journal of Artificial Intelligence Research*, 60, 549–593.

Pagano, A. S. (2015). A linguagem na Construção das práticas educativas nas ciências da saúde. In *Empoderamento do pesquisador nas Ciências da Saúde*. (pp. 19-36). Belo Horizonte: Tribo da Ilha.

Pagano, A. S.; Ferreguetti, K.; Figueredo, G. P. (2011). Significados relacionais em tradução: uma abordagem da equivalência baseada em corpus. *Caderno de Letras*, 17, 88–115.

Palumbo, G. 2008. *'Translating Science': an empirical investigation of grammatical metaphor as a source of difficulty for a group of translation trainees in English-Italian translation*. [Doctoral dissertation, University of Surrey]. Retrieved from https://core.ac.uk/download/pdf/40060582.pdf

Pasqualini, B. *CORPOP: um corpus de referência do português popular escrito do brasil*. [Doctoral dissertation, Federal University of Minas Gerais]. Retrieved from https://www.lume.ufrgs.br/bitstream/handle/10183/177566/001065569.pdf?sequence=1

Paula, A. 2017. *Orações verbais – uma descrição sistêmico-funcional dos processos de representação do dizer do português brasileiro*. [Master 's thesis, Universidade Federal de Ouro Preto].

Petersen, S., & Ostendorf, M. (n.d.). *Text Simplification for Language Learners: A Corpus Analysis*. Retrieved April 14, 2021, from https://www.isca-speech.org/archive_open/archive_papers/slate_2007/sle7_069.pdf

Quah C.K. (2006) Translation Studies and Translation Technology. In *Translation and Technology. Palgrave Textbooks in Translating and Interpreting*. Palgrave Macmillan, London. https://doi.org/10.1057/9780230287105_3a

R: A language and environment for statistical computing. Version 4.0.3. Vienna: R. Foundation for Statistical Computing, 2021. Retrieved from http://R-project.org/

Ravelli, L. J. (1996). *Making Language Accessible: Successful Text Writing for Museum Visitors*. Linguistics and Education, 8, 367-387. Retrieved from https://www.sciencedirect.com/science/article/pii/S0898589896900170

Ravelli, L. J. (1999). *Metaphor, Mode and Complexity: An exploration of co-varying patterns*. Nottingham: Nottingham Trent University.

Ravelli, L. J. (2006). *Museum Texts - Communication Frameworks*. New York: Routledge.

Rigat, M. V. 2013. *Paraphrase Scope and Typology: A Data-Driven Approach from Computational Linguistics*. [Doctoral dissertation, Universitat de Barcelona]. Retrieved from https://www.semanticscholar.org/paper/Paraphrase-scope-and-typology.-A-data%E2%80%91driven-from-Rigat/2434535043795aeaa0657e6d5ea98e36429efec6

Rocha, G., & Cardoso, H. (2018). Recognizing Textual Entailment: Challenges in the Portuguese Language. *Information, 9*(4), 1-19. doi:https://doi.org/10.3390/info9040076

Rosa, A. L. 2017. *Descrição sistêmico-funcional do verbo no português brasileiro orientada para os estudos da tradução: o sistema de modificação da experiência*. [Master 's thesis, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/LETR-ANRH7T

Sá, A. de M. 2016. *Uma descrição sistêmico-funcional do grupo verbal do português brasileiro orientada para os estudos da tradução*. [Master 's thesis, Federal University of Minas Gerais]. Retrieved from https://repositorio.ufmg.br/handle/1843/MGSS-ACXPQ8

Sá, A. de M. (2020) *Parâmetros temporais em português brasileiro: investigação das estruturas lexicogramaticais orientada para os estudos linguísticos da tradução*. [Doctoral dissertation, Federal University of Minas Gerais]. Retrieved from http://www.poslin.letras.ufmg.br/tese_defesas_detalhes.php?aluno=1837

Seretan, V. (2012). Acquisition of syntactic simplification rules for french. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12).* Istanbul, Turkey: European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L12-1138/

Shardlow, M. A. (2014). Survey of Automated Text Simplification. (IJACSA) International Journal of Advanced Computer Science and Applications - Special Issue on Natural Language Processing 2014, pp. 58-70. Manchester: United Kingdom. Retrieved from http://thesai.org/Downloads/SpecialIssueNo9/Paper_9-A_Survey_of_Automated_Text_Simplification.pdf

Siddharthan, A. 2003. *Syntactic simplification and Text Cohesion*. [Doctoral dissertation, University of Cambridge]. Retrieved from https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-597.pdf.

Siddharthan, A. (2004). *Syntactic simplification and text cohesion.* Cambridge, United Kingdom: University of Cambridge. Retrieved from https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-597.pdf

Siddharthan, A. (2006). Syntactic Simplification and Text Cohesion. *Res Lang Comput,* 4, pp. 77-109. https://doi.org/10.1007/s11168-006-9011-1

Siddharthan, A. (2014). A survey of research on text simplification. John Benjamins, 2, pp. 259 - 298. Retrieved from https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-597.pdf.

Silva, I. A. L. 2007. *Conhecimento experto em tradução: aferição da durabilidade de tarefas tradutórias realizadas por sujeitos não-tradutores em condições empírico-experimentais*. [Master 's thesis, Federal University of Minas Gerais]. Retrieved from http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/ALDR-797K7C/igor_silva_diss.pdf?sequence=1

Silva, I. A. L. 2012. *(Des)compactação de significados e esforço cognitivo: um estudo da metáfora gramatical na construção do texto traduzido*. [Doctoral dissertation, Federal University of Minas Gerais]. Retrieved from http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/LETR-96NP8J/silvatese27nov2012.pdf?sequence=1

Sinclair, J.; Carter, R. (Eds). (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Sokolova, M., Bobicev, V. (2018). Corpus Statistics in Text Classification of Online Data. *Computing Research Repository (CoRR)*. Retrieved from https://www.researchgate.net/publication/323867523_Corpus_Statistics_in_Text_Classification_of_Online_Data

Specia, L. (2010). Translating from Complex to Simplified Sentences. In *Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language*, (pp.

30–39). Berlin, Heidelberg, 9. Retrieved from https://pdfs.semanticscholar.org/828e/7cff761d3203a39d848934ec74f6cedc1453.pdf.

Specia, L., Aluísio, M. S., & Pardo, T. A. S. (2008). *Manual de simplificação sintática para o português*. São Carlos: NILC - ICMC-USP. Retrieved from https://sites.icmc.usp.br/taspardo/NILCTR0806.pdf

Štajner, S. 2015. *New Data-Driven Approaches to Text Simplification*. [Doctoral dissertation, University of Wolverhampton]. Retrieved from https://wlv.openrepository.com/bitstream/handle/2436/554413/thesis-afterViva.pdf;jsessionid=F59E0A7061CFEA3FCE0EEC8DAAA2BAC6?sequence=1

Štajner, S., Drndarević, B., & Saggion, H. (2013). Corpus-based Sentence Deletion and Split Decisions for Spanish Text Simplification. http://www.scielo.org.mx/pdf/cys/v17n2/v17n2a15.pdf

Štajner, S., Mitkov, R., & Corpas Pastor, G. (2014). Simple or Not Simple? A Readability Question. *Language Production, Cognition, and the Lexicon*, 379–398. https://doi.org/10.1007/978-3-319-08043-7_22. Retrieved from https://www.researchgate.net/profile/Gloria-Corpas-Pastor/publication/269828008_Stajner_S_R_Mitkov_and_G_Corpas_Pastor_Simple_or_not_Simple_A_Readability_Question/links/5541f6140cf232222731794f/Stajner-S-R-Mitkov-and-G-Corpas-Pastor-Simple-or-not-Simple-A-Readability-Question.pdf

Popović, M., & Štajner, S. (2016). Machine Translation Evaluation Metrics for Quality Assessment of Automatically Simplified Sentences. *Qats2016: LREC 2016 Workshop &*

Shared Task on Quality Assessment for Text Simplification (QATS), 28th May 2016, Portorož, Slovenia ; Proceedings, 32–37. Retrieved from https://www.researchgate.net/publication/301229033_Machine_Translation_Evaluation_Metrics_for_Quality_Assessment_of_Automatically_Simplified_Sentences

Steiner, E. (2002). Ideational grammatical metaphor: Exploring some implications for the overall model. Languages in Contrast, 4, 137–164.

Steiner, E. (2004). Ideational grammatical metaphor. Languages in Contrast, 4(1), 137–164. https://doi.org/10.1075/lic.4.1.07ste

Steiner, E. (2005). Explicitation, its lexicogrammatical realization, and its determining (independent) variables -towards an empirical and corpus-based methodology Explicitation, its lexicogrammatical realization, and its determining (independent) variables -towards an empirical and corpus-based methodology 1. https://www.hf.uio.no/ilos/forskning/prosjekter/sprik/pdf/Report_36_ESteiner.pdf

Steiner, E. (2019). Theorizing and Modelling Translation. In G. Thompson, W. Bowcher, L. Fontaine, & D. Schönthal (Eds.), *The Cambridge Handbook of Systemic Functional Linguistics* (Cambridge Handbooks in Language and Linguistics, pp. 739-766). Cambridge: Cambridge University Press. doi:10.1017/9781316337936.030

Stymne, S., Tiedemann, J., & Hardmeier, C. (2013). Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, (pp. 375-386). Retrieved from https://www.aclweb.org/anthology/W13-5634.pdf

Teruya, K., Lam, M. & Matthiessen, C. M. I. M (2010). Key terms in Systemic Functional Linguistics. London/New York: Continuum.

Vila, M., Rodríguez, H., & Martí, M. (2010). WRPA: A System for Relational Paraphrase Acquisition from. *Procesamiento del Lenguaje, 45*, 11-19.

Webster, J. (Ed.) (2004). *The Language of Science*. London/New York: Continuum, pp. 24-48.

Wu, C. 2000. *Modelling Linguistic Resources: A Systemic Functional Approach*. [Doctoral thesis, Macquarie University]. Retrieved from http://www.api.adm.br/GRS/referencias/thesis_liguistica_sistemica.pdf

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3, 283–297. Retrieved from https://cocoxu.github.io/publications/tacl2015-text-simplification-opinion.pdf

Zaenen, A., Karttunen, L., & Crouch, R. (2005). Local Textual Inference: Can it be Defined or Circumscribed? *ACL Anthology*, 31–36. https://www.aclweb.org/anthology/W05-1206/

Zhu, Z.; Bernhard, D. and Gurevych, I. (2010, August). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING)*, (pp. 1353-1361) Beijing, China, 23. Retrieved from http://www.aclweb.org/anthology/C10-115

Zhu, X and Wu, X. (2004). Class Noise vs. Attribute Noise: A Quantitative Study of Their Impacts. *Artificial Intelligence Review*. 22. pp. 177–210. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.886.3083&rep=rep1&type=pdf

Appendices

Appendix A

Complete data on the websites used to extract the naturally occurring text segments analyzed ordered by the number of segments extracted from each

Websites	Count
https://www.msdmanuals.com	43
https://mundoeducacao.bol.uol.com.br	25
https://www.todamateria.com.br	12
https://pt.wikipedia.org	11
http://escolakids.uol.com.br	5
http://alunosonline.uol.com.br	4
http://brasilescola.uol.com.br	4

http://www.dicionarioinformal.com.br	3
http://www.ufrgs.br	3
https://brasilescola.uol.com.br	3
https://conceito.de	3
https://mundoestranho.abril.com.br	3
https://www.coladaweb.com	3
https://www.estudopratico.com.br	3
https://www.infoescola.com	3
https://www.manualdaquimica.com	3
https://www.minhavida.com.br	3
http://www.if.ufrgs.br	2
http://www.scielo.br	2

https://conceitos.com	2
https://www.significados.com.br	2
https://www.todabiologia.com	2
escola.tecnico.ulisboa.pt	1
http://bioprofessor.webnode.com.br	1
http://draanabeatriz.com.br	1
http://engenhariacotidiana.com	1
http://labs.icb.ufmg.br	1
http://leg.ufpi.br	1
http://medicopsiquiatra.pt	1
http://meuartigo.brasilescola.uol.com.br	1
http://pataniscassatanicas.blogspot.com.br	1

http://pepsic.bvsalud.org	1
http://psiquiatriabh.com.br	1
http://qnesc.sbq.org.br	1
http://quimicanova.sbq.org.br	1
http://radiologia.blog.br	1
http://reinpec.srvroot.com:8686	1
http://w3.ufsm.br	1
http://wikiciencias.casadasciencias.org	1
http://www.abc.med.br	1
http://www.cienciasecognicao.org	1
http://www.digimed.ufc.br	1
http://www.educacaoprofissional.seduc.ce.gov.br	1

http://www.fis.uc.pt	1
http://www.inicepg.univap.br	1
http://www.itad.pt	1
http://www.revista-fi.com	1
http://www.saudemental.net	1
http://www.uff.br	1
https://accbarroso60.wordpress.com	1
https://books.google.com.au	1
https://books.google.com.br	1
https://dicasdeciencias.com	1
https://irradiacaoecateg.blogspot.com	1
https://meradius.com.br	1

https://meuartigo.brasilescola.uol.com.br	1
https://prezi.com	1
https://pt.khanacademy.org	1
https://sigarra.up.pt	1
https://www.abcdasaude.com.br	1
https://www.actamedicaportuguesa.com	1
https://www.colegioweb.com.br	1
https://www.estudaetal.com	1
https://www.facebook.com	1
https://www.google.com.br	1
https://www.ic.unicamp.br	1
https://www.inpaonline.com.br	1

https://www.portaleducacao.com.br	1
https://www.psiquiatriageral.com.br	1
www.cnen.gov.br	1
www.scielo.mec.pt	1

Appendix B

List of all websites used to extract the text segments analyzed

The following table shows the 196 websites from which texts were retrieved. Due to technical issues on some websites, during the collection of texts, 3 of the 196 complete texts could not be stored. Yet, it was possible to retrieved all the segments to be analyzed. The full list of 196 websites is shown in the table below. Websites with ids 87, 98 and 100 remain, up to this date, inaccessible, due to technical issues with their websites.

Source_id	Source
1	https://mundoeducacao.bol.uol.com.br/quimica/fissao-nuclear.htm
2	https://www.todamateria.com.br/efeito-joule/
3	https://mundoeducacao.bol.uol.com.br/fisica/forca-centripeta.htm
4	https://accbarroso60.wordpress.com/2011/02/27/biologia-celulas-tecidossistemas-etc/
5	https://conceitos.com/fecundacao/
6	https://mundoestranho.abril.com.br/ciencia/como-acontece-uma-mutacao-no-dna/
7	http://www.abc.med.br/p/sinais.-sintomas-e-doencas/345159/catalepsia+o+que+e+isso.htm
8	http://www.digimed.ufc.br/wiki/index.php/Psicomotricidade_e_suas_Altera%C3%A7%C3%B5es
9	https://mundoestranho.abril.com.br/saude/o-que-e-o-vitiligo-como-ele-surge/

10	http://brasilescola.uol.com.br/doencas/vitiligo.htm
11	https://www.significados.com.br/pareidolia/
12	https://mundoeducacao.bol.uol.com.br/fisica/asteroides.htm
13	http://www.dicionarioinformal.com.br/anag%C3%AAnese/
14	http://pataniscassatanicas.blogspot.com.br/2014/10/flora-saprofita.html
15	http://www.dicionarioinformal.com.br/heter%C3%B3trofo/
16	https://www.todabiologia.com/ecologia/autotrofos_heterotrofos.htm
17	https://www.estudaetal.com/thebox/theboxficheiros/2041925989074a9f38e3414005d5a27c860d4
18	https://www.psiquiatriageral.com.br/glossario/a.htm
19	https://pt.wikipedia.org/wiki/Acatisia
20	http://escolakids.uol.com.br/divisao-celular.htm
21	https://www.estudopratico.com.br/aplicacao-medicinal-da-radioterapia/
22	http://bioprofessor.webnode.com.br/products/conceito%20e%20defini%C3%A7%C3%B5es%20das%20celulas/
23	http://radiologia.blog.br/medicina-nuclear/o-que-sao-radiofarmacos-e-suas-aplicacoes
24	http://escolakids.uol.com.br/o-que-sao-proteinas.htm
25	https://mundoeducacao.bol.uol.com.br/biologia/proteinas.htm
26	https://www.estudopratico.com.br/dna-cromossomos-genes-genoma-e-rna/
27	http://escolakids.uol.com.br/composicao-da-flor.htm

28	http://brasilescola.uol.com.br/o-que-e/biologia/o-que-e-anticorpo.htm
29	http://escolakids.uol.com.br/anestesia.htm
30	https://mundoeducacao.bol.uol.com.br/doencas/anemia.htm
31	http://escolakids.uol.com.br/arterias-veias-capilares.htm
32	https://www.todamateria.com.br/daltonismo/
33	https://pt.wikipedia.org/wiki/Daltonismo
34	https://mundoeducacao.bol.uol.com.br/biologia/daltonismo.htm
35	http://brasilescola.uol.com.br/biologia/pleiotropia-interacoes-genicas.htm
36	https://www.todamateria.com.br/pleiotropia/
37	http://wikiciencias.casadasciencias.org/wiki/index.php/Enzima_de_Restri%C3%A7%C3%A3o
38	http://labs.icb.ufmg.br/lbcd/grupo2/estrutura/enzimas.html
39	https://www.infoescola.com/biologia/enzimas-de-restricao/
40	https://pt.khanacademy.org/science/physics/centripetal-force-and-gravitation/centripetal-forces/a/what-is-centripetal-force
41	https://mundoeducacao.bol.uol.com.br/biologia/sinalizacao-celular.htm
42	https://mundoestranho.abril.com.br/saude/o-que-e-disturbio-de-deficit-de-atencao/
43	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/s%C3%ADndrome-de-digeorge
44	http://www.saudemental.net/transtorno_obsessivo_compulsivo.htm

45	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/esquizofrenia-e-transtorno-delirante/esquizofrenia
46	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/defici%C3%AAncia-seletiva-de-anticorpos-com-imunoglobulinas-normais
47	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-do-cora%C3%A7%C3%A3o-e-dos-vasos-sangu%C3%ADneos/endocardite/defini%C3%A7%C3%A3o-de-endocardite
48	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-do-cora%C3%A7%C3%A3o-e-dos-vasos-sangu%C3%ADneos/endocardite/endocardite-infecciosa
49	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/transtornos-de-personalidade/transtornos-de-personalidade#v6580311_pt
50	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/fobia-social
51	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/fobia-social
52	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/agorafobia
53	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/transtorno-de-estresse-agudo-tea
54	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/transtorno-do-estresse-p%C3%B3s-traum%C3%A1tico-tept
55	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/esquizofrenia-e-transtorno-delirante/transtorno-delirante
56	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/transtorno-de-ansiedade-generalizada-tag
57	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/ataques-de-p%C3%A2nico-e-s%C3%ADndrome-do-p%C3%A2nico

58	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/ansiedade-e-transtornos-relacionados-ao-estresse/transtornos-f%C3%B3bicos-espec%C3%ADficos
59	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/biologia-do-sistema-imunol%C3%B3gico/considera%C3%A7%C3%B5es-gerais-sobre-o-sistema-imunol%C3%B3gico
60	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/vis%C3%A3o-geral-de-imunodefici%C3%AAncias
61	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/ataxia-telangiectasia
62	https://mundoeducacao.bol.uol.com.br/fisica/reflexao-luz.htm
63	https://brasilescola.uol.com.br/fisica/dilatacao-termica-calorimetria.htm
64	https://mundoeducacao.bol.uol.com.br/fisica/radiacao-conducao-conveccao.htm
65	https://www.todamateria.com.br/conducao-termica/
66	http://meuartigo.brasilescola.uol.com.br/fisica/conducao-conveccao-irradiacao.htm
67	https://mundoeducacao.bol.uol.com.br/fisica/transmissao-energia-termica.htm
68	http://www.if.ufrgs.br/mpef/mef008/mef008_02/Beatriz/irradiacao.htm
69	https://mundoeducacao.bol.uol.com.br/fisica/corrente-eletrica.htm
70	https://mundoeducacao.bol.uol.com.br/fisica/eletricidade.htm
71	http://alunosonline.uol.com.br/fisica/eletricidade-.html
72	http://alunosonline.uol.com.br/fisica/eletricidade.html
73	https://www.todamateria.com.br/eletricidade/
74	https://pt.wikipedia.org/wiki/Corrente_el%C3%A9trica

75	https://www.infoescola.com/fisica/corrente-eletrica/
76	http://brasilescola.uol.com.br/o-que-e/fisica/o-que-e-corrente-eletrica.htm
77	http://www.if.ufrgs.br/tex/fis01043/20021/Gusmao/
78	https://pt.wikipedia.org/wiki/Tens%C3%A3o_el%C3%A9trica
79	http://alunosonline.uol.com.br/fisica/fontes-campo-magnetico.html
80	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/s%C3%ADndrome-de-ch%C3%A9diak-higashi
81	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/candid%C3%ADase-mucocut%C3%A2nea-cr%C3%B4nica
82	https://www.msdmanuals.com/pt/casa/doen%C3%A7as-imunol%C3%B3gicas/doen%C3%A7as-decorrentes-de-imunodefici%C3%AAncia/defici%C3%AAncia-de-ades%C3%A3o-leucocit%C3%A1ria
83	https://www.todamateria.com.br/fissao-nuclear/
84	https://www.todamateria.com.br/forca/
85	apostila radiacoes ionizantes www.cnen.gov.br
86	https://www.todabiologia.com/dicionario/fecundacao.htm
87	http://www.uff.br/genetica_animal/mutacao
88	http://medicopsiquiatra.pt/glossario/flexibilidade-cerea/
89	http://reinpec.srvroot.com:8686/reinpec/index.php/reinpec/article/view/108
90	https://www.abcdasaude.com.br/dermatologia/vitiligo
91	http://www.scielo.br/pdf/%0D/rbp/v27n2/a17v27n2.pdf

92	https://pt.wikipedia.org/wiki/Asteroide#cite_note-vest-2
93	https://www.google.com.br/search?q=Dicion%C3%A1rio#dobs=anag%C3%AAnese
94	https://pt.wikipedia.org/wiki/Saprotrofia
95	https://pt.wikipedia.org/wiki/Heterotrofismo
96	http://www.dicionarioinformal.com.br/aut%C3%B3trofo/
97	https://sigarra.up.pt/icbas/en/pub_geral.pub_view?pi_pub_base_id=28777
98	http://www.ufrgs.br/bibicbs/livros-novos/tortora-corpo-humano-10.-ed (p.64)
99	https://meradius.com.br/radioterapia
100	https://books.google.com.au/books?id=Rx56DwAAQBAJ&pg=PA79&lpg=PA79&dq=%22As+radia%C3%A7%C3%B5es+ionizantes+s%C3%A3o+eletromagn%C3%A9ticas+ou+corpusculares+e+carregam+energia.+Ao+interagirem+com+os+tecidos,+d%C3%A3o+origem+a+el%C3%A9trons+r%C3%A1pidos+que+ionizam+o+meio+e+criam+efeitos+qu%C3%ADmicos+como+a+hidr%C3%B3lise+da+%C3%A1gua+e+a+ruptura+das+cadeias+de+ADN.%22&source=bl&ots=c-vd21Xxvg&sig=ACfU3U0_EumTijX9KpvIAFd0ccNQJouB_w&hl=en&sa=X&ved=2ahUKEwiW6uKei6niAhWb8XMBHez2BQYQ6AEwAXoECAgQAQ#v=onepage&q=%22As%20radia%C3%A7%C3%B5es%20ionizantes%20s%C3%A3o%20eletromagn%C3%A9ticas%20ou%20corpusculares%20e%20carregam%20energia.%20Ao%20interagirem%20com%20os%20tecidos%2C%20d%C3%A3o%20origem%20a%20el%C3%A9trons%20r%C3%A1pidos%20que%20ionizam%20o%20meio%20e%20criam%20efeitos%20qu%C3%ADmicos%20como%20a%20hidr%C3%B3lise%20da%20%C3%A1gua%20e%20a%20ruptura%20das%20cadeias%20de%20ADN.%22&f=false
101	http://www.ufrgs.br/livrodehisto/pdfs/1Celula.pdf
102	http://qnesc.sbq.org.br/online/cadernos/06/a08.pdf
103	http://www.revista-fi.com/edicoes/38/files/assets/downloads/publication.pdf
104	http://www.educacaoprofissional.seduc.ce.gov.br/images/material_didatico/quimica/quimica_bioquimica_industrial.pdf
105	https://pt.wikipedia.org/wiki/%C3%81cido_desoxirribonucleico

106	http://w3.ufsm.br/herb/glossario.pdf
107	http://quimicanova.sbq.org.br/imagebank/pdf/Vol25No2_316_19.pdf
108	http://leg.ufpi.br/subsiteFiles/lapnex/arquivos/files/Farmacologia%20dos%20anestesicos%20gerais.pdf
109	https://books.google.com.br/books?id=HWU9DwAAQBAJ&pg=PA1159&dq=anemia+%C3%A9&hl=pt-BR&sa=X&ved=0ahUKEwjr2baU-o7YAhXBHpAKHcCgCggQ6AEISDAF#v=onepage&q=anemia%20%C3%A9&f=false
110	http://www.ufrgs.br/livrodehisto/pdfs/6Circulat.pdf
111	http://www.inicepg.univap.br/cd/INIC_2010/anais/arquivos/0342_0216_02.pdf
112	https://www.infoescola.com/genetica/pleiotropia/
113	https://dicasdeciencias.com/2013/04/27/pleiotropia-e-interacao-genica/
114	http://e-escola.tecnico.ulisboa.pt/pt/topico.asp?id=279
115	https://mundoeducacao.bol.uol.com.br/biologia/enzimas-restricao.htm
116	https://prezi.com/g4ft9sm_m-9w/copy-of-biotecnologia/
117	http://www.fis.uc.pt/data/20052006/apontamentos/apnt_067_7.pdf
118	https://pt.wikipedia.org/wiki/Sinaliza%C3%A7%C3%A3o_celular
119	https://www.actamedicaportuguesa.com/revista/index.php/amp/article/viewFile/686/364
120	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/s%C3%ADndrome-de-digeorge
121	http://draanabeatriz.com.br/portfolio/mentes-e-manias-toc-transtorno-obsessivo-compulsivo-intro/
122	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/esquizofrenia-e-transtornos-relacionados/esquizofrenia

123	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/defici%C3%AAncia-seletiva-de-anticorpos-com-imunoglobulinas-normais-sadni
124	https://www.msdmanuals.com/pt/profissional/dist%C3%BArbios-cardiovasculares/endocardite/defini%C3%A7%C3%A3o-de-endocardite
125	https://www.msdmanuals.com/pt/profissional/dist%C3%BArbios-cardiovasculares/endocardite/endocardite-infecciosa
126	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/transtornos-de-personalidade/transtorno-de-personalidade-narcisista-tpn
127	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/agorafobia
128	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/transtorno-de-estresse-agudo-tea
129	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/transtorno-de-estresse-p%C3%B3s-traum%C3%A1tico-tept
130	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/esquizofrenia-e-transtornos-relacionados/transtorno-delirante
131	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/transtorno-de-ansiedade-generalizado-tag
132	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/ataques-e-transtorno-de-p%C3%A2nico#v1025529_pt
133	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/transtornos-f%C3%B3bicos-espec%C3%ADficos
134	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/biologia-do-sistema-imune/vis%C3%A3o-geral-do-sistema-imune
135	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/vis%C3%A3o-geral-das-imunodefici%C3%AAncias

136	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/ataxia-telangiectasia
137	https://www.todamateria.com.br/reflexao-da-luz/
138	https://pt.wikipedia.org/wiki/Irradia%C3%A7%C3%A3o_t%C3%A9rmica
139	https://www.todamateria.com.br/irradiacao-termica/
140	https://www.todamateria.com.br/corrente-eletrica/
141	https://www.todamateria.com.br/circuito-eletrico/
142	https://www.todamateria.com.br/campo-magnetico/
143	http://alunosonline.uol.com.br/fisica/efeitos-corrente-eletrica.html
144	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/s%C3%ADndrome-de-ch%C3%A9diak-higashi
145	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/candid%C3%ADase-mucocut%C3%A2nea-cr%C3%B4nica
146	https://www.msdmanuals.com/pt/profissional/imunologia-dist%C3%BArbios-al%C3%A9rgicos/imunodefici%C3%AAncias/defici%C3%AAncia-de-ades%C3%A3o-leucocit%C3%A1ria
147	https://www.msdmanuals.com/pt/casa/dist%C3%BArbios-de-sa%C3%BAde-mental/transtornos-do-humor/transtorno-bipolar
148	https://mundoeducacao.bol.uol.com.br/fisica/dilatacao-superficial-dos-solidos.htm
149	https://www.manualdaquimica.com/quimica-geral/alotropia.htm
150	https://www.coladaweb.com/quimica/fisico-quimica/alotropia
151	https://www.coladaweb.com/quimica/fisico-quimica/catalise-e-catalisadores

152	https://www.manualdaquimica.com/fisico-quimica/termoquimica.htm
153	https://www.coladaweb.com/quimica/fisico-quimica/termoquimica
154	https://brasilescola.uol.com.br/psicologia/motivacao-psicologica.htm
155	https://conceito.de/psiquiatria
156	https://mundoeducacao.bol.uol.com.br/psicologia/criatividade.htm
157	https://mundoeducacao.bol.uol.com.br/psicologia/hipnose.htm
158	https://mundoeducacao.bol.uol.com.br/psicologia/agorafobia.htm
159	https://mundoeducacao.bol.uol.com.br/psicologia/inteligencia-emocional.htm
160	https://mundoeducacao.bol.uol.com.br/psicologia/sonambulismo.htm
161	https://mundoeducacao.bol.uol.com.br/psicologia/deficiencia-mental.htm
162	https://mundoeducacao.bol.uol.com.br/psicologia/autoestima.htm
163	https://www.minhavida.com.br/saude/temas/psicose
164	https://conceito.de/depressao
165	https://www.significados.com.br/ansiedade/
166	https://www.msdmanuals.com/pt/profissional/transtornos-psiqui%C3%A1tricos/transtornos-do-humor/transtornos-bipolares
167	https://www.manualdaquimica.com/fisico-quimica/catalisador.htm
168	http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1517-55452005000100012
169	http://psiquiatriabh.com.br/faq-items/o-que-e-psiquiatria/

170	http://www.cienciasecognicao.org/pdf/v14_3/m96.pdf
171	https://www.ic.unicamp.br/~wainer/cursos/906/trabalhos/hipnose.pdf
172	www.scielo.mec.pt/pdf/aps/v25n4/v25n4a12.pdf
173	https://meuartigo.brasilescola.uol.com.br/psicologia/inteligencia-emocional-equilibrio-comportamental.htm
174	https://www.minhavida.com.br/saude/temas/sonambulismo
175	http://www.itad.pt/deficiencia-mental/
176	https://www.inpaonline.com.br/auto-estima/
177	https://conceitos.com/autoestima/
178	https://conceito.de/psicose
179	https://www.minhavida.com.br/saude/temas/depressao
180	http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-44462000000600006
181	https://www.msdmanuals.com/pt-pt/profissional/transtornos-psiqui%C3%A1tricos/ansiedade-e-transtornos-relacionados-a-estressores/fobia-social
182	https://irradiacaoecateg.blogspot.com/
183	https://pt.wikipedia.org/wiki/Regress%C3%A3o_(psicologia)
184	https://www.portaleducacao.com.br/conteudo/artigos/psicologia/psicanalise-para-leigos-entendendo-os-principais-conceitos-da-area/51504
185	https://www.facebook.com/permalink.php?id=582165688916747&story_fbid=610862512713731
186	https://www.estudopratico.com.br/conceito-de-trajetoria-cinematica/

187	https://www.colegioweb.com.br/fundamentos-da-cinematica-escalar/trajetoria.html
188	https://mundoeducacao.bol.uol.com.br/fisica/ondas-2.htm
189	https://brasilescola.uol.com.br/fisica/ondas.htm
190	https://mundoeducacao.bol.uol.com.br/fisica/o-que-sao-ondas-eletromagneticas.htm
191	http://engenhariacotidiana.com/conceitos-da-fisica-conheca-os-principais/
192	https://mundoeducacao.bol.uol.com.br/fisica/mecanica.htm
193	https://mundoeducacao.bol.uol.com.br/fisica/eletromagnetismo.htm
194	https://www.estudopratico.com.br/referencial-movimento-espaco-e-repouso/
195	https://br.answers.yahoo.com/question/index?qid=20100131153426AAKoFtq
196	https://fisicapassoapassojp.blogspot.com/2012/06/conceitos-basicos.html

Appendix C

Tutorial for analyzing discourse using Sysfan

Importing data or entering it manually

In Filemaker Pro, after the layout is complete, data can be either imported from files, e.g., spreadsheets, or manually entered, as presented in this section.

Figure 1 presents how data can be automatically imported to the software in order to be analyzed.

Figure 1

Snapshot of Sysfan importing menu

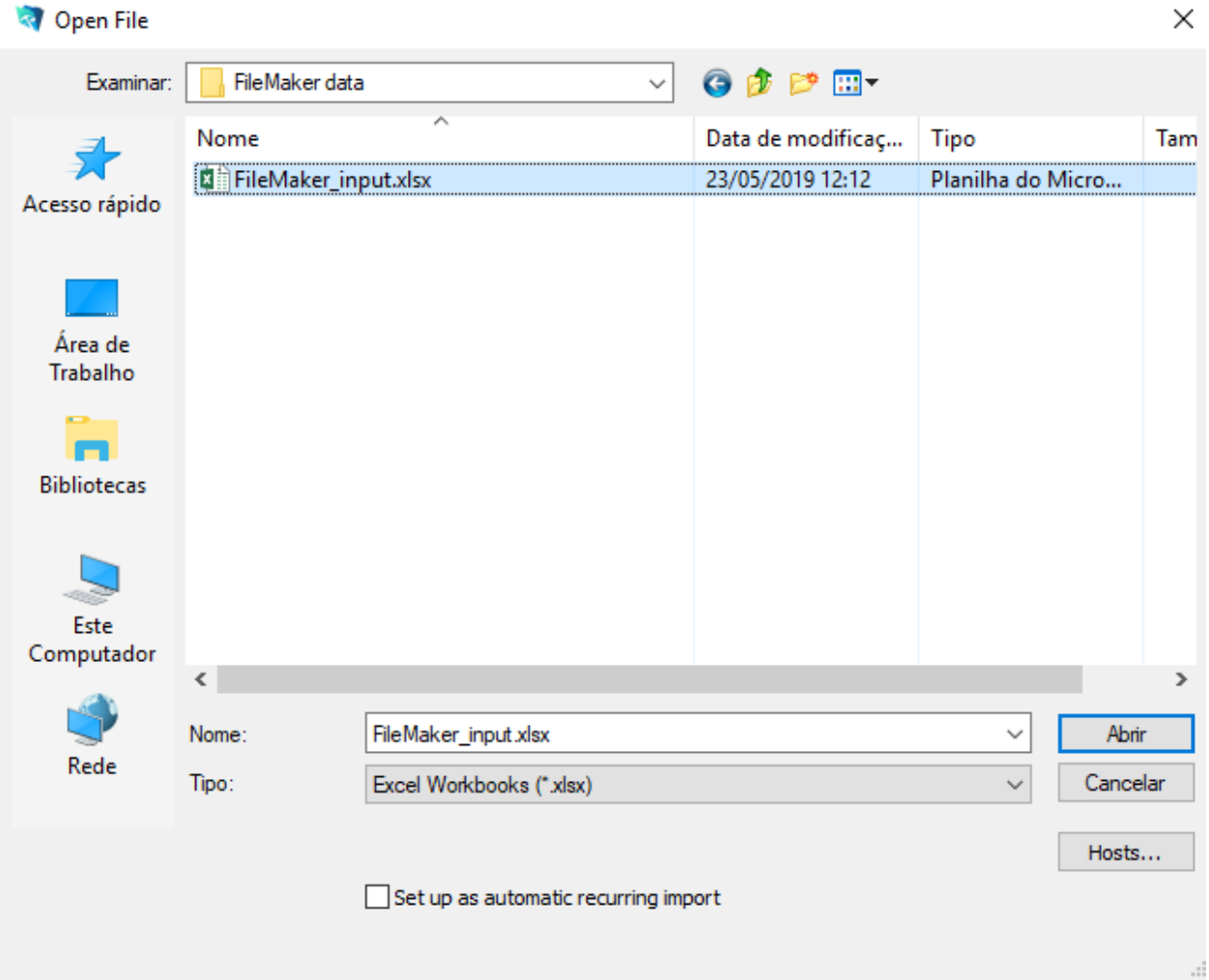


Figure 1 presents a snapshot of the importing menu that allows a user to select a spreadsheet file, such as “FileMaker_input.xlsx”, and use the file as input to add data to the database that is being created. It can also be done manually, by typing, as Figure 1 shows.

Figure 2

Snapshot of Sysfan layout

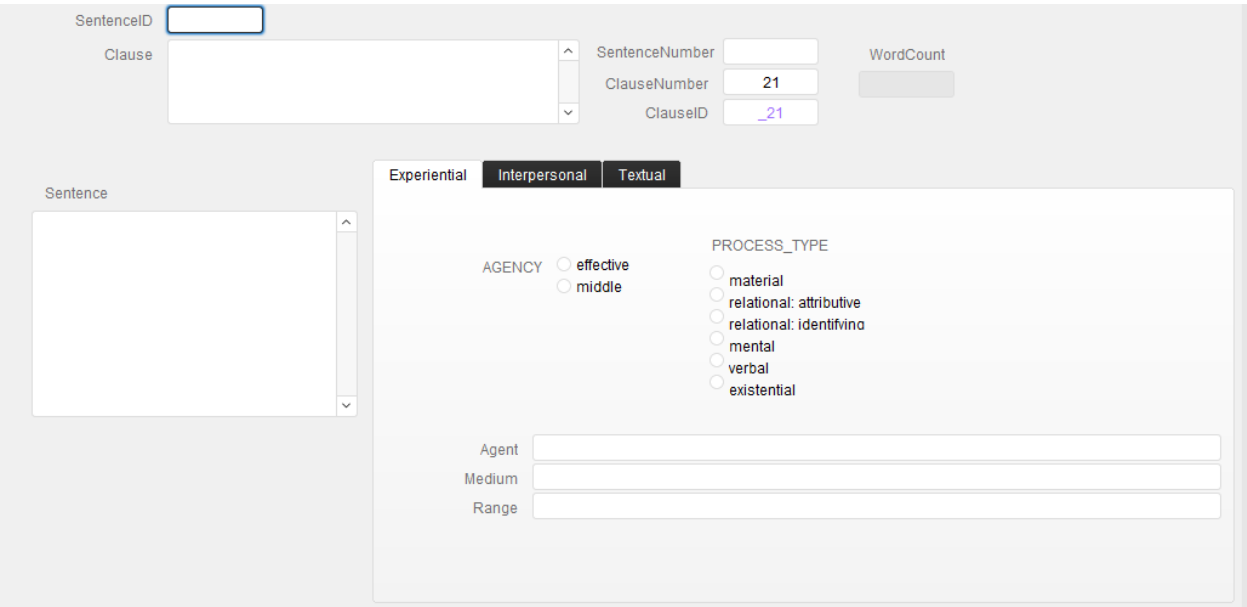


Figure 2 is the snapshot of the layout used to extract clauses from sentences retrieved from the text sets. It is done filling the “Clause” field with the relevant information.

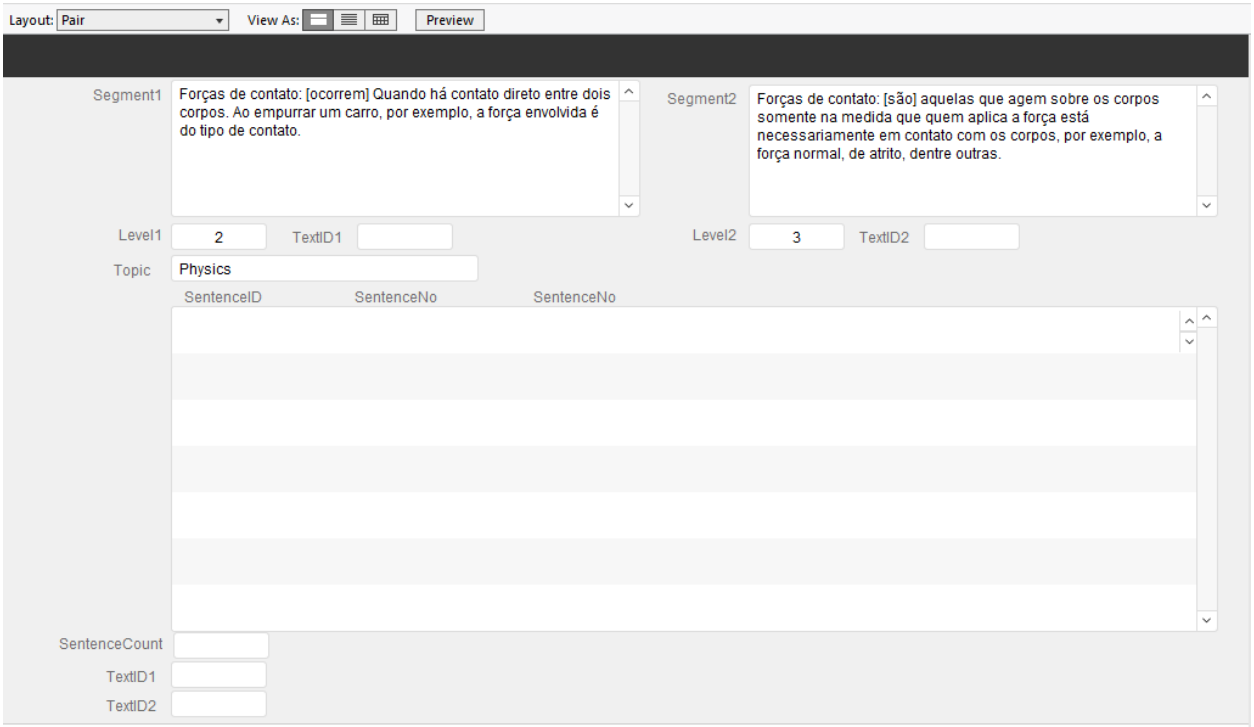
The next step, after adding data to the database, is to annotate the data with the GUI. This process will be shown in the next section.

Annotating data using GUI

As presented in Figure 2, the GUI allows the user to classify the data within a range of categories of interest for the research being conducted. In Figure 3, for instance, it is possible to classify the clauses into different systemic categories, as the type of agency and the process type, and select which examples represent the Agent, the Medium and the Range in a clause.

Figure 3

Snapshot of Sysfan classification layout



Appendix D

Frequency of the main systemic patterns analyzed with the highest values between the sets highlighted in bold

SELECTION	**Assessed system**	**SET A and SET C**						**SET B and SET D**					
Marked theme	Agency	Middle		**Effective**				**Middle**				Effective	
		Abs	Rel	Abs	Rel			Abs	Rel			Abs	Rel
		99	41.95	137	**58.05**			99	**59.28**			68	40.72
		Abs	Rel	Abs	Rel			Abs	Rel			Abs	Rel
		114	**59.07**	79	40.93			57	42.54			77	**57.46**
Marked theme	Relational Process	**Attributive**		Identifying				**Attributive**				Identifying	
		Abs	Rel	Abs	Rel			Abs	Rel			Abs	Rel
		40	**50.63**	39	49.37			44	**57.14**			33	42.86
Marked theme	Relational Process	**Intensive**		Possessive		Circumstantial		**Intensive**		Possessive		Circumstantial	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel

		46	**58.23**	7	8.86	26	32.91	61	**79.22**	8	10.39	8	10.39
Unmarked theme	Relational Process	**Intensive**		Possessive		Circumstantial		**Intensive**		Possessive		Circumstantial	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		140	**79.55**	14	7.95	22	12.50	155	**89.60**	18	10.40	16	9.25
Unmarked theme	Relational Process	**Attributive**		Identifying				Attributive				**Identifying**	
		Abs	Rel	Abs		Rel		Abs		Rel		Abs	Rel
		89	**50.63**	87		49.37		94		49.74		95	**50.26**
Material Process	Theme Selection	Unmarked theme		**Marked theme**				Unmarked theme		Marked theme			
		Abs	Rel	Abs		Rel		Abs	Rel	Abs		Rel	
		75	39.68	114		**60.32**		56	49.56	57		50.44	
Material Process	Conjunction	**non-conjuncted**		conjuncted				**non-conjuncted**		conjuncted			
		Abs	Rel	Abs		Rel		Abs	Rel	Abs		Rel	
		117	**61.90**	72		38.10		83	**73.45**	30		26.55	
Bound	Agency	Middle		**Effective**				**Middle**		Effective			
		Abs	Rel	Abs		Rel		Abs	Rel	Abs		Rel	

		53	46.09	62	**53.91**	47	**53.41**	41	46.59
Free	Agency	**Middle**		Effective		**Middle**		Effective	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		239	**59.16**	165	40.84	261	**73.11**	96	26.89

Free	Relational Process	**Intensive**		Possessive		Circumstantial		**Intensive**		Possessive		Circumstantial	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		165	**77.10**	16	7.48	33	15.42	186	**81.22**	22	9.61	21	9.17

Frequency of co-selections between different systemic choices from different systems

SELECTION	**Assessed system**	**SET A and SET C**				**SET B and SET D**			
Marked theme and Material process	Conjunction	non-conjuncted		**conjuncted**		non-conjuncted		**conjuncted**	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		42	36.84	72	**63.16**	28	49.12	29	**50.88**
Marked theme	Relational	**Attributive**		Identifying		**Attributive**		Identifying	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel

and Material process	Process	40	**50.63**	39		49.37		44	**57.14**	33		42.86	
Marked theme and Relational process	Relational Process	**Intensive**		Possessive		Circumstantial		**Intensive**		Possessive		Circumstantial	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		46	**58.23**	7	8.86	26	32.91	61	**79.22**	8	10.39	8	10.39
Unmarked theme and Material process	Deicticity	**Temporal**		Modal				**Temporal**		Modal			
		Abs	Rel	Abs		Rel		Abs	Rel	Abs		Rel	
		55	**77.46**	16		22.54		46	**85.19**	8		14.81	
Unmarked theme and Relational process	Relational Process	Attributive		Identifying				Attributive		Identifying			
		Abs	Rel	Abs		Rel		Abs	Rel	Abs		Rel	
		89	50.57	87		49.43		94	49.74	95		50.26	
Marked	Agency	Middle		**Effective**				**Middle**		Effective			

theme and Free		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		47	37.90	77	**62.10**	52	**65.00**	28	35.00
Marked theme and Free	Conjunction	Non-conjuncted		**conjuncted**		**Non-conjuncted**		conjuncted	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		54	43.55	70	**56.45**	41	**51.25**	39	48.75
Marked theme and Bound	Agency	Middle		**Effective**		**Middle**		Effective	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		52	46.85	59	**53.15**	47	**54.02**	40	45.98
Marked theme and Bound	Relational Process	Attributive		**Identifying**		**Attributive**		Identifying	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		18	45.00	22	**55.00**	20	**54.05**	17	45.95

SELEC	**Asses**	**SET A and SET C**	**SET B and SET D**

TION	**sed syste m**																				
Marked theme and Bound	Proce ss type	**Materia l**		Mental		Verbal		Relation al		existenti al		Material		Mental		Verbal		**Relation al**		existent ial	
		A bs	Re l	A bs	Rel	A bs	Re l	A bs	Rel	A bs	Rel	Abs	Re l	A bs	R el	A bs	Rel	A bs	Rel	A bs	Rel
		4 9	**44. 14**	1 3	11. 71	2	1.8 0	40	36. 04	7	6.3 1	35	40. 23	5	5. 7 5	0	0.0	37	**42. 53**	1 0	11. 49
Marked theme and Free	Proce ss type	**Materia l**		Mental		Verbal		Relation al		existenti al		Material		Mental		Verbal		**Relation al**		existent ial	
		A bs	Re l	A bs	Rel	A bs	Re l	A bs	Rel	A bs	Rel	Abs	Re l	A bs	R el	A bs	Rel	A bs	Rel	A bs	Rel
		6 4	**51. 16**	7	5.6 5	6	4.8 4	39	31. 45	8	6.4 5	22	27. 50	2	2. 5 0	5	6.25	40	**50. 0**	1 1	13. 75
Bound and middle	Proce ss type	Material		Mental		Verbal		**Relation al**		existenti al		Material		Mental		Verbal		**Relation al**		existent ial	
		A bs	Re l	A bs	Rel	A bs	Re l	A bs	Rel	A bs	Rel	Abs	Re l	A bs	R el	A bs	Rel	A bs	Rel	A bs	Rel
		1	1.8 9	4	7.5 5	1	1.8 9	40	**75. 47**	7	13. 21	0	0.0 0	1	2. 1 3	0	0.0	36	**76. 60**	1 0	21. 28

Bound and middle	Relational process	**Intensive**		Possessive		Circumstantial		**Intensive**		Possessive		Circumstantial	
		Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel	Abs	Rel
		21	**52.50**	5	12.50	14	35.00	29	**80.56**	4	11.11	3	8.33