

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

ELISA MATTOS

**A CORPUS-BASED STUDY OF HYPHENATED PREMODIFIERS IN
COMPLEX NPs IN BIOLOGY RESEARCH ARTICLES**

BELO HORIZONTE
2020

ELISA MATTOS

**A CORPUS-BASED STUDY OF HYPHENATED PREMODIFIERS IN
COMPLEX NPs IN BIOLOGY RESEARCH ARTICLES**

Dissertação apresentada ao PosLin - 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 Mestra em Linguística Aplicada.

Área de concentração: Linguística Aplicada

Linha de pesquisa: Ensino e Aprendizagem de Línguas Estrangeiras

Orientadora: Deise Prina Dutra

Belo Horizonte
Faculdade de Letras da UFMG
2020

Ficha catalográfica elaborada pelos Bibliotecários da Biblioteca FALE/UFMG

M444c

Mattos, Elisa.
A corpus-based study of hyphenated premodifiers in complex NPs in biology research articles [manuscrito] / Elisa Mattos. – 2020.
160 f., enc. : il., tabs., p&b., color.

Orientadora: Deise Prina Dutra.

Área de concentração: Linguística Aplicada.

Linha de pesquisa: Ensino e Aprendizagem de Línguas Estrangeiras.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras.

Bibliografia: f. 129-140.

Anexos: f. 141-160.

1. Linguística de corpus – Teses. 2. Linguística – Processamento de dados – Teses. 3. Redação acadêmica – Teses. 4. Língua inglesa – Sintagma nominal – Teses. I. Dutra, Deise Prina. II. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD : 410



UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGÜÍSTICOS



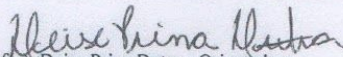
FOLHA DE APROVAÇÃO

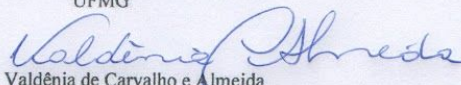
A CORPUS-BASED STUDY OF HYPHENATED PREMODIFIERS IN COMPLEX NPs IN BIOLOGY RESEARCH ARTICLES

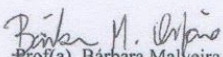
ELISA MATTOS DE SÁ

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGÜÍSTICOS, como requisito para obtenção do grau de Mestre em ESTUDOS LINGÜÍSTICOS, área de concentração LINGÜÍSTICA APLICADA, linha de pesquisa Ensino/Aprendizagem de Línguas Estrangeiras.

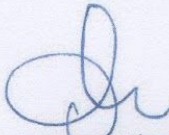
Aprovada em 14 de fevereiro de 2020, pela banca constituída pelos membros:


Prof(a). Deise Prina Dutra - Orientadora
UFMG


Prof(a). Valdênia de Carvalho e Almeida
UFV


Prof(a). Bárbara Malveira Orfanó
UFMG

Belo Horizonte, 14 de fevereiro de 2020.


Prof(a). Ana Larissa Adorno Marcocato Oliveira
Subcoord. Programa de Pós-Graduação
em Estudos Linguísticos
FALE/UFMG

To my mother, who I hope is watching over me.

We travel, some of us forever, to seek other states, other lives, other souls. (Anais Nin)

I am only a shell where the ocean is still sounding (Марина Цветаева)

AGRADECIMENTOS

Agradeço primeiramente a meu pai, Nazareno, (e minha mãe), por todo o esforço despendido na minha formação acadêmica e profissional, desde os livros e as mensalidades escolares às conversas sobre gramática no café da manhã. I have benefited immensely from your support.

Agradeço também às amizades acadêmicas feitas no mestrado: Vanessa e Gustavo (the other two musketeers), Nivia e Wellington, Camila Bunny. E aos amigos de sempre: Eric e Marcellus. Telminha. Sem vocês eu não teria chegado à conclusão desta etapa.

Agradeço à minha orientadora, Deise Prina Dutra, pelos comentários e pelas oportunidades de aprendizado e desenvolvimento nesses últimos dois anos. E pela paciência.

Ao GECEA, pelas perguntas sempre pertinentes. Ao NELFA, pelas trocas criadas.

Às minhas supervisoras de estágio docente, Bárbara Orfano e Climene Arruda, e aos alunos do IFA, que em muito me inspiram a continuar aprendendo. Ao Leo Nunes, também.

À Pró-Reitoria de Graduação (PROGRAD) e à Diretoria de Relações Internacionais (DRI), ambas da UFMG, pelas bolsas de estágio e oportunidades de crescimento.

Esta dissertação também é reflexo das várias disciplinas cursadas no PosLin de 2017 a 2019, sem as quais estas páginas não teriam alçado vôo. Agradeço especialmente às professoras Lúcia Ferrari e Heliana Mello, for always taking the time to answer my (many) questions.

Finally, agradeço à querida Albertina, por ser meu elo familiar, por tudo o que ela representa em minha vida. À Ângela, por me ajudar a me entender e, nesse processo, a entender o mundo.

And to other numerous trailblazing women who have bravely broken invisible glass ceilings in the professional world and in everyday life.

GRACIAS.

RESUMO

Esta dissertação objetiva investigar sintagmas nominais complexos em textos especializados produzidos em inglês. Especificamente, esta pesquisa visa examinar o uso de modificadores pré-nominais hifenizados em artigos acadêmicos de Biologia. Segundo Biber e Gray (2016), Gray (2015), Pirrelli, Guevara e Baroni (2010) e Biber et al (1999), a escrita científica tende a ser caracterizada por construções nominais complexas, compactadas, dado seu forte potencial de compactação (BIBER; GRAY, 2016; GRAY, 2015; HERRERO-ZORITA; SANDOVAL, 2016). Isso pode ser vantajoso para a escrita de textos restritos em número de palavras ou páginas. Conforme os princípios básicos da Linguística de Corpus (SINCLAIR, 2005; SARDINHA, 2004; LÜDELING; KYTÖ, 2008; GRIES, 2009; McENERY; HARDIE, 2012; DAVIES, 2015) e com base na concepção de *English as a Lingua Franca* (JENKINS, 2013; JENKINS; LEUNG, 2013; MAURANEN; HYNINEN; RANTA, 2016; SEIDLHOFER, 2013) este estudo utiliza textos autênticos cuidadosamente compilados para ser processados e tratados computacionalmente. Para tanto, um corpus de 250 artigos de Biologia foi compilado com base em cinco periódicos de alto impacto, totalizando 1.294.161 *tokens* distribuídos em textos de 3.500 e 7.500 palavras, publicados entre 2015 a 2019. Para a compilar os artigos automaticamente, uma extensão computacional foi desenvolvida. Softwares de Processamento da Linguagem Natural (PLN) foram empregados na extração e análise dos dados, conforme as diretrizes de *Constituency e Dependency Grammar* (JURAFSKY; MARTIN, 2019), em forte diálogo com a Linguística Computacional. A análise voltou-se para a frequência e distribuição dos sintagmas nominais complexos extraídos e para um total de 5.789 sintagmas complexos com pré-modificados hifenizados, todos etiquetados morfossintaticamente de forma manual. Os resultados confirmam preferência por estruturas compactas como substantivos compostos, hifenização e acrônimos, verificadas estatisticamente, evidenciando a escrita científica como mais compactada e menos explícita gramatical e semanticamente, em inglês. Em situações de co-ocorrência, pré-modificadores hifenizados são favorecidos.

Palavras-chave: sintagmas nominais complexos, pré-modificadores hifenizados, Linguística Computacional; Linguística de Corpus; escrita científica.

ABSTRACT

The purpose of this thesis is to investigate noun phrase (NP) complexity in specialized texts produced in English. Specifically, this research examines the use of hyphenated premodifiers in complex NPs in Biology research articles (RAs). As argued in Biber and Gray (2016), Gray (2015), Pirrelli, Guevara and Baroni (2010), and Biber et al (1999), science writing has as one of its defining features the use of compressed, complex nominal structures. Such preference is often associated with the strong compacting potential nominal compression of these structures (BIBER; GRAY, 2016; GRAY, 2015; HERRERO-ZORITA; SANDOVAL, 2016). This can be advantageous for word/page-restricted texts. Following the basic tenets of Corpus Linguistics (SINCLAIR, 2005; SARDINHA, 2004; LÜDELING; KYTÖ, 2008; McENERY; HARDIE, 2012; GRIES, 2009; DAVIES, 2015) and based on the notion of English as a Lingua Franca (JENKINS, 2013; JENKINS; LEUNG, 2013; MAURANEN; HYNINEN; RANTA, 2016; SEIDLHOFER, 2013), this thesis employs naturally-occurring texts carefully compiled in 250 Biology RAs from five high impact journals, leading to a total 1,294,161 tokens distributed in 3,500-7,500-word texts published from 2015 to 2019. A computational extension was devised to automatically retrieve the RAs. Natural Language Processing (NLP) software were used for data extraction and analysis, following the guidelines of Constituency and Dependency Grammar (JURAFSKY; MARTIN, 2019), in dialogue with Computational Linguistics. The extracted NPs were analyzed for frequency and distribution. 5,789 hyphenated premodifiers were then morpho-syntactically labeled. The statistically verified results confirm a preference for compact structures such as compound nouns, hyphenation and acronyms, showing scientific writing to be more compact and less explicit grammatically and semantically, in English. For co-occurrences, hyphenated premodifiers are favored.

Keywords: complex noun phrases; hyphenated premodifiers; Corpus Linguistics; scientific writing; Computational Linguistics.

LIST OF ABBREVIATIONS AND ACRONYMS

L2	Additional/Second Language
AmE	American English
AB	Animal Behaviour
ATG	Article Text Grabber
ATG1	Article Text Grabber version 1
ATG2	Article Text Grabber version 2
BC	Biological Conservation
BP	Brazilian Portuguese
BrE	British English
CSV	Comma-Separated Value
CEFR	Common European Framework of Reference for Languages
CG	Constituency Grammar
CGr	Constraint Grammar
COORD	Coordination
CorABio	Corpus of Articles in Biology
CorAChem	Corpus of Articles in Chemistry
CoBRA	Corpus of Biology Research Articles
CoCRA	Corpora of Cultural Studies Research Articles
CoERA	Corpora of Education Research Articles
CoLRA	Corpora of Linguistics Research Articles
CoRA	Corpora of Research Articles
CL	Corpus Linguistics
EL	Ecology Letters
EAP	English for Academic Purposes
ELF	English as a Lingua Franca
FALE	<i>Faculdade de Letras</i>
UFMG	Federal University of Minas Gerais
L1	First Language
GECEA	<i>Grupo de Estudos em Corpora Especializados e de Aprendizizes</i>
HTML	HyperText Markup Language
IFA	<i>Inglês para Fins Acadêmicos</i>
C-ORAL-ROM	Integrated Reference Corpora for Spoken Romance Languages

ISSN	International Standard Serial Number
IMRaD	Introduction, Methods, Results, and Discussion
IMRD(C)	Introduction, Methodology, Results, Discussion, and Conclusion
MDA	Multidimensional Analyses
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
NP	Noun Phrase
NPs	Noun Phrases
OE	Oecologia
POS	Part-of-Speech
PSG	Phrase Structure Grammar
PPs	Prepositional Phrases
C-ORAL BRASIL I	Reference Corpus for Informal Spoken Brazilian Portuguese
RAs	Research Articles
SJR	Scimago Journal and Country Rank
TSV	Tab-Separated Value
VP	Verb Phrase
VISL	Visual Interactive Syntax Learning
ZS	Zoologica Scripta

LIST OF FIGURES

Chapter 2 – Literature Review

Figure 2.1: a CorABio sentence example	27
Figure 2.2: tree diagram for the CorABio sample sentence	28
Figure 2.3: CorABio sample sentence dependency structure	29
Figure 2.4: decrease in predicative clauses in science RAs	31
Figure 2.5: increase in compressed <i>-ed</i> and <i>-ing</i> noun-initial NPs	31
Figure 2.6: Biber and Gray's cline of structural compression	32
Figure 2.7: Huddleston and Pullum's hyphenated compound list	40

Chapter 3 – Materials and Methods

Figure 3.1: RA metadata in plain text	44
Figure 3.2: CorABio's 25 most frequent words	46
Figure 3.3: CorABio's 25 most frequent nouns	47
Figure 3.4: CorABio's 25 most frequent verbs	48
Figure 3.5: foreign characters found post-cleanin	50
Figure 3.6: CoBRA word counts	53
Figure 3.7: Zoologica Scripta article webpage	54
Figure 3.8: Zoologica Scripta landing page source code	55
Figure 3.9: the ATG1 at play	56
Figure 3.10: the ATG2 at play	57
Figure 3.11: CoBRA's 25 most frequent words	60
Figure 3.12: data visualization with Unitex/GramLab IDE 3.1	61
Figure 3.13: compound visualization with Unitex/GramLab IDE 3.1	61
Figure 3.14: data output	64
Figure 3.15: PP output	64
Figure 3.16: tabulated NP categorization	65
Figure 3.17: data extraction output	68

Chapter 4 – Results and Discussion

Figure 4.1: dependency structure	60
Figure 4.2: dependency structure	64
Figure 4.3: POS-categorization	73
Figure 4.4: partial concordance plot for <i>density-dependent</i> .	83
Figure 4.5: partial concordance plot for <i>land-use</i>	84
Figure 4.6: left- and right-branching dependencies for <i>child language acquisition</i> .	104

LIST OF GRAPHS AND CHARTS

Chapter 3 – Materials and Methods

Graph 3.1 journal and corpus word count	59
---	----

Chapter 4 – Results and Discussion

Graph 4.1: distribution of the hyphenated premodifiers in the 2-item NP set	91
Graph 4.2: distribution of the hyphenated premodifiers in the 3-item NP set	101
Graph 4.3: distribution of the hyphenated premodifiers in the 4-item NP set	104
Graph 4.4: distribution of the hyphenated premodifiers in the 5-item NP set	106

Chapter 4 – Results and Discussion

Chart 4.1: breakdown of the NP sets extracted.	70
Chart 4.2: number of NPs extracted and cleaned, with trendlines	73
Chart 4.3: number of NPs deep-cleaned, initiated by functional and content words	81
Chart 4.4: number of hyphenated premodifiers	83
Chart 4.5: number of content NPs and hyphenated premodifiers.	84
Chart 4.6: past participle in hyphenated premodifiers in the 2-5 range of NP sets	108
Chart 4.7: ten most frequent hyphenated past participles in the 2-5 range of NP sets	109
Chart 4.8: ten most frequent prefixes in the 2-5 range of NP sets	113

LIST OF TABLES

Chapter 3 – Materials and Methods

Table 3.1: CoBRA journal selection	52
Table 3.2: journal and corpus word counts	59

Chapter 4 – Results and Discussion

Table 4.1: number of NPs extracted and cleaned	73
Table 4.2: number of NPs deep-cleaned, initiated by functional and content words	80
Table 4.3 A: number of NPs with hyphenated premodifiers and hyphenated heads	82
Table 4.3 B: number of NPs containing hyphenated premodifiers (<i>sans</i> hyphenated heads)	82
Table 4.4: number of content NPs and hyphenated premodifiers	84
Table 4.5: 20 most frequent hyphenated premodifiers in the 2-item content NP set	87
Table 4.6: number of items and tokens and their frequency range	89
Table 4.7: hyphenated vs. solid uses for <i>long-term</i>	92
Table 4.8: tri-constituent hyphenated premodifiers	97
Table 4.9: twenty most frequent hyphenated premodifiers in the 3-item content NP set	100
Table 4.10: twenty most frequent hyphenated premodifiers in the 4-item content NP set	103
Table 4.11: twenty most frequent hyphenated premodifiers in the 5-item content NP set	105
Table 4.12: past participle in hyphenated premodifiers in the 2-5 range of NP set	107
Table 4.13: ten most frequent hyphenated past participles in the 2-5 range of NP sets	108
Table 4.14: etymology for the ten most frequent past participles	110
Table 4.15: ten most frequent prefixes in the 2-5 range of NP sets	112
Table 4.16: hyphenated vs solid five most frequent past participles.	116
Table 4.17: hyphenated premodifiers and their corresponding segmentation in BP	122
Table 4.18: hyphenation preferences in English compounds	123

CONTENTS

Chapter 1 – Introduction	18
1.1 Justification	19
1.2 Objectives	21
1.3 Organization	21
Chapter 2 – Literature Review	23
2.1 Overview	23
2.2 The noun phrase in English	23
2.2.2 <i>Observations on Constituency Grammar</i>	24
2.2.1 <i>NP constituent order</i>	25
2.3 Noun Phrase complexity	26
2.3.1 <i>Why investigate NP complexity?</i>	26
2.3.2 <i>Measuring NP complexity</i>	28
2.4 NP complexity in science writing	30
2.4.1 <i>A shift towards structural compression</i>	30
2.4.2 <i>Hyphenation</i>	32
2.4.3 <i>Word Formation</i>	34
2.4.3.1 <i>Derivation</i>	35
2.4.3.2 <i>Compounding</i>	36
2.4.4 <i>Hyphenation in Compounding</i>	38
2.4.5 <i>A move towards discipline-specific NP complexity</i>	40
2.5 Corpus Linguistics	42
Chapter 3 – Materials and Methods	43
3.1 Overview	43
3.2 The pilot corpus	43
3.2.1 <i>Compilation</i>	44
3.2.2 <i>Understanding the corpus</i>	46
3.2.3 <i>Metadata</i>	49
3.2.4 <i>Issues</i>	49
3.2.5 <i>A possible solution</i>	50
3.3 The research corpus	50
3.3.1 <i>Design criteria</i>	51
3.3.1.1 <i>Journal selection</i>	51

3.3.1.2 <i>Representativeness and balance</i>	52
3.3.1.3 <i>Article format</i>	52
3.3.1.4 <i>Text length</i>	53
3.3.2 <i>Compilation</i>	54
3.3.2.1 <i>The text-grabbing browser extension</i>	54
3.3.2.2 <i>Version 1</i>	56
3.3.2.3 <i>Version 2</i>	57
3.3.2.4 <i>Cleaning</i>	58
3.3.3 <i>Corpus counts</i>	59
3.3.4 <i>Metadata</i>	62
3.4 <i>Data extraction</i>	63
3.4.1 <i>Data output</i>	63
3.4.2 <i>Initial NP categorization</i>	65
3.4.3 <i>Issues</i>	66
3.5 <i>NP chunking revisited</i>	67
Chapter 4 – Results and Discussion	69
4.1 <i>Overview</i>	69
4.2 <i>NP chunks</i>	69
4.2.1 <i>Overall counts</i>	70
4.2.2 <i>Extracted and cleaned NPs</i>	70
4.2.3 <i>Observations about +6-item NPs</i>	75
4.3.4 <i>NP sets initiated by content words</i>	79
4.3 <i>Hyphenated premodifiers</i>	82
4.3.1 <i>POS-categorization</i>	84
4.3.2 <i>Hyphenated premodifiers in 2-item content NPs</i>	86
4.3.2.1 <i>Frequency and distribution</i>	88
4.3.2.2 <i>Verifying dispersion</i>	93
4.3.2.3 <i>Tri-constituent premodifiers</i>	97
4.3.3 <i>Hyphenated premodifiers in 3-item content NPs</i>	100
4.3.4 <i>Hyphenated premodifiers in 4-item content NPs</i>	102
4.3.5 <i>Hyphenated premodifiers in 5-item content NPs</i>	104
4.4 <i>Exploratory observations</i>	107
4.4.1 <i>The past participle</i>	107
4.4.2.1 <i>Decomposition</i>	109
4.4.2.2 <i>Etymological remarks</i>	110
4.4.2 <i>Prefixation</i>	111

4.5 Discussion	114
4.5.1 <i>The role of semantic and grammatical transparency</i>	114
4.5.2 <i>Hyphenation vs. non-hyphenation</i>	116
4.5.3 <i>Dependency links</i>	117
4.5.4 <i>Hyphenation as a compression strategy</i>	119
4.5.4.1 <i>Implications for EAP teaching</i>	121
4.5.4.2 <i>Trends in hyphenation</i>	123
4.5.4.3 <i>Pedagogical suggestions</i>	123
4.5.5 <i>Limitations</i>	125
Chapter 5 – Conclusion	126
References	129
Attachments	141
<i>Attachment A: Corpus-driven EAP applications (MATTOS, 2019)</i>	141
<i>Attachment B: Corpus-driven EAP applications (SÁ, 2019)</i>	146
<i>Attachment C: Corpus-driven L1 applications (MATTOS, 2018)</i>	151
<i>Attachment D: Chunking issues (partial output)</i>	155
<i>Attachment E: Chunking script (partial)</i>	157

1. INTRODUCTION

Academic prose is viewed as a highly specialized register (BIBER, 1988) that presents specific lexico-grammatical features, which, albeit employed in other registers, are used quite distinctively in academic writing (cf. BIBER; GRAY, 2016). Academic prose, in general, and science research writing, in particular, makes use of complex noun phrases (NPs) substantially more than other registers and in a wide range of grammatical functions by means of nominal pre- and post-modification, with adjectives recognized as the preferred premodifier, closely followed by nouns (BIBER; GRAY, 2016; GRAY, 2015; BIBER et. al, 1999).

Corpus Linguistics (CL) studies have consistently shown that noun pre-modification patterns vary across disciplines, in regard to complexity, with multiple nouns and adjectives increasingly more employed as premodifiers in so-called hard sciences when contrasted to the science writing in humanities (BIBER; GRAY, 2016; GRAY, 2015; DUTRA et al, 2020).

Such a pattern may not be generally employed in non-native speaker academic writing¹, probably due to lower language proficiency, since it is believed that phrasal compression such as noun premodification is usually acquired last² in academic writing development (cf. BIBER; GRAY; POONPON, 2011).

Moreover, as observed in Biber and Gray (2016, 2011, 2010), variation in academic prose is more pervasive than initially believed. In this sense, a more comprehensive study of the lexico-grammatical patterns employed in a particular register/genre³ may offer insight into how writers linguistically organize their ideas, as per communication purposes and intended audience, particularly when such examination pertains to the specificity of a given discipline or area of academic expertise. As stated in Biber and Gray (2016), **research writing is made by specialists for specialists.** (my emphasis).

¹ As witnessed in my own professional experience teaching English for Academic Purposes (EAP) at the Federal University of Minas Gerais (UFMG), in which I found that intermediate language learner texts tend to rely more on noun post-modification where premodified compressed structures would be more natural (e.g.: *species of animal, the strong influence of baroque and rococo*). I have also observed such preferences when proofreading Brazilian graduate students' research articles (RAs) written in English. A very recent study on NP complexity in learner corpora can be found in Queiroz (2019), for Brazilian learners' written productions.

² The same could be said for novel compounding, a pervasive word formation process in the data analyzed.

³ Register studies focus on associations between common linguistic features and communicative purposes, while genre studies are concerned with a text's conventional structures (BIBER; CONRAD, 2009). This thesis pertains to the former; however, it would be remiss not to mention the genre perspective herein.

In this vein, research articles (RAs) traditionally aim at informing highly specialized readerships about investigations from remarkably distinct fields of scientific research. It is assumed that their target audience has sufficient background and domain knowledge about the publication content (BIBER, GRAY, 2016). This may be one of the reasons why it is typical for writers to employ compressed structures, such as noun premodification, in RAs, as these constructions pack large amounts of semantic-pragmatic content in concise ways.

Of relevance to the present research are the investigations of González-Díaz (2009), Martínez-Insua and Pérez-Guerra (2011), Davidse (2016), and Adamson and González-Díaz (2009), who observe that in spite of the abundance of more formality-driven studies about the English NP (e.g., BIBER et al., 1999; HUDDLESTON; PULLUM, 2002, among others), few of those probe the semantic-pragmatic aspects of this phenomenon.

As much as possible for the time frame of a master's degree research, this thesis has attempted to address some of these matters, as seen in the discussion section of chapter four. This venture into semantic-pragmatic aspects of noun premodification was also driven by the methodological specificities stemming from the computational data extraction tools employed in this research.

1.1 Justification

As a specialized register, academic writing is believed to be not as easily acquired⁴ as other registers (BIBER; GRAY; POONPON, 2011), which might be why universities around the world offer year-round undergraduate and graduate-level academic writing courses. Either in a first (L1) or additional language (L2)⁵, university students and academics are expected to have a solid command of the rhetorical and discourse conventions and the lexico-grammatical patterns of academic prose in their respective disciplines.

As Brazilian scientific publications in English grow (cf. FINARDI, 2014), so does the implementation of internationalization processes⁶, which has led to numerous course offerings of English for Academic Purposes (EAP) at the university level in this country. At the Federal University of Minas Gerais (UFMG), the *Inglês para Fins Acadêmicos* (IFA) program is one such example.

⁴ A distinction between acquisition and learning will not be made. See Ortega (2009) for a discussion.

⁵ An L2 is hereby understood as “any language learned after the L1 (or L1s)” (ORTEGA, 2009, p. 5).

⁶ A discussion on internationalization is found in Sarmiento et al. (2017) and Finardi and França (2016).

Hence, understanding how English is used in academic texts is of utmost importance, particularly when it comes to advanced linguistic phenomena such as noun premodification in complex NPs, or hyphenated noun compounding in academic writing. These uses are not only pervasive in science research writing, but they also might be indicative of specificities and/or preferences in discipline-focused science writing practices (cf. BIBER; GRAY, 2016; GRAY, 2015; DUTRA et al, 2020).

Inspecting discipline-specific language patterns in specialized academic writing may therefore open a window, however small, into a deeper understanding of how expert English language writers and/or experienced academics linguistically share specialized knowledge in a globally reaching language. This research may shed light on academic writing and serve as a basis of reflection and supplementary instruction to novice and experienced researchers in the scientific area of Biology.

In this sense, the research hereby reported does not hold normative implications. This thesis therefore does not view language use in an explicit, clear-cut distinction between native and non-native speaker academic writing performance. Rather, this thesis relies on the notion of *English as a Lingua Franca*⁷ (ELF), which, in agreement with Tribble (2017), privileges an expert-apprentice approach to EAP.

Any observation about L2 writing in English made in this thesis therefore pertains to the proficient command of the English language and is meant to highlight its pervasiveness in academia. As hinted at throughout this thesis, the final goal of language analysis should be an understanding of the usage patterns, with the aim of facilitating communication. Specialized language use in English does not entail native-like proficiency. It does require familiarity with academic and linguistic practices in a given discipline, however.

Finally, by keeping these studies openly accessible to the academic community, more specifically to the IFA teachers and students at UFMG, and by strategically sharing them with the academic community more generally, we may be contributing to a still ongoing wave of highly focused research aimed at improving the English language skills⁸ of Brazilian students and researchers. As stressed in Dutra et al. (2017), this is an important step for technological and academic growth in Brazil, especially in the current socio-political scenario.

⁷ ELF accounts are found in Jenkins and Leung (2013) and Jenkins (2013). For ELF in academia, see Seidlhofer (2013), and Mauranen, Hynninen and Ranta (2016). For ELF and CL, see Conrad and Mauranen (2003).

⁸ Examples can be seen in Dutra, Queiroz, and Alves (2017), among others.

1.2 Objectives

Initially, the purpose of this thesis was to investigate noun premodification in complex NPs in Biology RAs. The main research question pondered was: *how is noun premodification manifested in specialized science writing, specifically in the area of Biology?* As a very broad question, this idea was later refined towards examining noun premodification in complex NPs as a means of compacting and compressing information.

While this initial goal remained in sight, the focus of this thesis somewhat changed as I started delving into the data. Looking closely at the extracted NPs and reflecting upon texts I had to grade as part of my academic internship⁹, new ideas came about. Specifically, a deeper, careful examination of the extracted NPs revealed that several of such nominal constructions presented hyphenated premodifiers.

Hence, because hyphenation use in premodifiers in complex NPs stood out, I decided to slightly veer the focus of this research to **the hyphenated premodifiers identified in 2- to 5-item complex NPs initiated by content words**, as extracted from the corpus compiled for this research. To this end, this thesis aims to investigate the use of hyphenated premodifiers in complex NPs in Biology research articles written in English by:

- Exploring the overall semantic and morpho-syntactic traits of the research corpus;
- Analyzing the frequency and distribution of the extracted complex NP chunks;
- Identifying hyphenated premodifiers in the extracted complex NP chunks;
- Mapping hyphenated pre-modification patterns in 2- to 5-item NP sets;
- Verifying the quantitative data by statistical means whenever possible;
- Analyzing the degree of complexity in the extracted and mapped NPs.

1.3 Organization

This thesis is organized in five chapters. After this introduction, chapter two presents a literature review in which the operational definition of NP complexity is provided, as related to premodification. Its importance is outlined in this chapter, particularly in regard to its use in RAs. Observations are also made in regard to the underlying grammar perspectives taken in this thesis, as well as in relation to hyphenation and compounding and CL.

⁹ *Estágio de docência*, an academic internship at the Faculdade de Letras (FALE), UFMG, teaching EAP under the supervision of Dr. Barbara Orfano and Dr. Climene Arruda.

The third chapter regards the materials and methods utilized in this research. It offers a full description of the compilation processes of the two specialized corpora meant to be used in the analysis of premodified complex NPs. It touches upon the computational methods used for data extraction and processing. The two attempts made at automated NP chunking are also described.

Chapter four reports on and discusses the findings from the automated NP extraction, both quantitatively and qualitatively. In addition, it informs the reader of the rationale behind the selection of hyphenated premodifiers, indicating usage patterns in the extracted NP sets. The fourth chapter discusses hyphenation and compounding as compression strategies and the pedagogical implications of the research findings.

Finally, chapter five offers concluding remarks followed by the references cited herein and five attachments. These are: 1) a pedagogical activity on noun compounding, 2) an extract from Sá (2019), in which complex NPs have been addressed in a pedagogical way, as related to CL and Data-Driven Learning (DDL), 3) slides from Mattos (2018), in which vocabulary in Portuguese, verified by corpora, is used for the development of intercultural competence, 4) issues identified in the output from the first data extraction, and 5) a copy of the script devised to extract the NPs for analysis.

2. LITERATURE REVIEW

2.1 Overview

This chapter summarizes the main theoretical assumptions behind this research. It also presents the operational definitions of the phenomena under investigation and some remarks about CL as an area of linguistic research.

2.2 The noun phrase in English

Fundamentally, noun phrases (NPs) consist of a noun or a pronoun forming the head of the NP. NPs may be composed simply by a head noun or by the head noun and dependent constituents before and/or after it. For the sake of clarification, this thesis understands simple NPs as i) a determiner and a head noun (*the house, a house*), ii) just a noun (*houses*), and iii) just a pronoun (*I [write]*), with one-noun NPs called *bare nominals*. In contrast, complex NPs are understood as consisting of multiple words acting as premodifiers and/or post-modifiers connected to the noun head (e.g.: *the blue house, that beautiful blue house on the hill, the blue house from my childhood, etc.*).

Albeit with slight terminological variations, the NP definitions from the corpus-based reference grammars and guides consulted for this thesis are unanimously similar, in particular, Quirk et al (1985), Greenbaum (1996), Biber et al (1999), Huddleston and Pullum (2002)¹⁰, Parrott (2010), Carter and McCarthy (2006), Greenbaum and Nelson (2009), Gelderen (2010), Carter et al (2011), Downing (2015), Garner (2016), the Collins COBUILD English Grammar (2017). While these grammars and guides do not explicitly refer to the types of NP as simple or complex, such distinction is inherently stated therein.

The research proposed in this thesis is guided by the syntactic categorizations found in the above-cited grammar reference works, particularly Biber et al. (1999), in which extensive corpus-based examinations show an overall preference for head nouns to be pre-modified by adjectives and nouns, with determiners, possessive nouns, and numerals indicating reference. These authors are also conclusive about the ubiquitousness of noun premodification in written registers, most notably in news and academic prose, which serves the purpose of the present thesis.

¹⁰ More specifically Huddleston and Payne (2002).

Discussions on hyphenation and compounding have been grounded mostly in the work of Huddleston and Pullum (2002), specifically in Bauer and Huddleston (2002) and Nunberg, Briscoe, and Huddleston (2002). Likewise, the work of Pirrelli, Guevara, and Baroni (2010) and Sanchez-Stockhammer (2018) has been used to inform these discussions.

In the vein of Biber et al. (1999) and Biber and Gray (2016), the selected bibliography regarding the use of hyphenation and compounding is corpus-based. Given the close relation of CL and Computational Linguistics in this thesis, this selection oftentimes refers to NLP for theoretical and analytical support.

2.2.1 Observations on Constituency Grammar

It should be noted that, by looking at nominal groups in a structural, phrase-based way, we are accepting the main theoretical assumptions of Phrase Structure Grammar (PSG), also often referred to as Constituency Grammar (CG). The underlying notion is that sentences are formed by words and words are categorized into parts of speech (nouns, adjectives, adverbs, verbs, prepositions). Grammar is a set of rules by which these parts of speech are combined to form sentences, while phrases are a sequence of words surrounding a headword, as explained in Jurafsky and Martin (2019).

This syntax view is also one of the major driving forces behind the Natural Language Processing (NLP) tools employed in the present thesis, particularly the Stanford Core NLP (MANNING et al, 2014). Hence, although this research may often refer to CG/PSG, this is a methodological choice bound by the NLP information extraction tools and methods adopted, and the Chomskyan formalisms to which data processing is often bound.

As clarified in Sarkar (2019), context-free grammars are seen as the building blocks of text processing and information retrieval in NLP. But when it comes to understanding and analyzing language in use, CG/PSG tends to fall short by not taking into account contextual information and otherwise ill-fitting sentences that language users actually produce in social interaction. A context-free approach may be ideal for NLP and machine learning tasks, but in natural language in use, context is essential.

In this sense, a more inclusive grammar use outlook should consider the amalgamation between form and meaning (cf. LANGACKER, 1987) as well as the interdependence between grammar and lexicon (cf. HALLYDAY, 1985), a perspective often favored in CL research, as argued in Sardinha (2019).

2.2.2 NP constituent order

The industrially advanced countries¹¹
Os países industrialmente desenvolvidos¹²
Los países industrialmente avanzados¹³
Экономически развитые страны¹⁴
(промышленно развитые страны)

As can be seen in the above examples, the order of the constituents in an NP is subject to the syntactic possibilities and pragmatic preferences of each linguistic system, as well as to the semantic content of the items. The four NPs listed above display somewhat different word orders, with English and Russian favoring premodification and Portuguese and Spanish opting for post-modification.

While the English and the Russian modifiers could be rearranged as post-modification, such alteration would entail the addition of other items in order for these NPs to be considered syntactically and pragmatically natural. Hence, *the industrially advanced countries* would be rendered *the countries that are industrially advanced*, in which a verb phrase (VP) in the form of a relative clause is added, and *Экономически развитые страны* would be re-arranged as *страны, которые экономически развиты*, which can be roughly translated as *the countries which are economically developed*.

As for the Brazilian Portuguese and Spanish NPs, the following could be rendered: *os industrialmente desenvolvidos países* and *los industrialmente avanzados países*, respectively. Although these NPs are indeed grammatical, they do not sound natural on their own; it is as if a complement needed to be added in the form of a VP.

Interestingly, this also seems to apply to the changes in the English NP and may be justified by the presence of a definite article acting as a determiner in pre-modification, a feature not applicable to the NP in Russian, since articles do not exist in this Slavic language. These are a very elementary way of arguing that any changes made to NP constituents trigger a ripple effect whereby other constituents are also altered, both syntactically and semantically.

¹¹ Example extracted from Biber et al (1999).

¹² My translation.

¹³ My translation.

¹⁴ My translation, verified by two Russian native speakers, one of whom is a linguist.

2.3 Noun phrase complexity

Noun phrase (NP) complexity may be defined as “the complexity directly arising from the number of linguistic elements and their interrelationships” (PALLOTTI, 2015, p. 18), or it may also be taken as “(1) the number and the nature of the discrete components that the entity consists of, and (2) the number and the nature of the relationships between the constituent components”, as explained by Boulté and Housen (2012, p. 22).

Boulté and Housen (2012, p. 22) are not explicit as to what they understand by “the **nature** of the relationships between the constituent components” (my emphasis). Specifically, I will be looking mostly at the morphosyntactic properties of complex NPs. However, as will be seen in chapter four, considerations are made in regard to the role of semantic transparency in the complex NPs extracted and subsequently analyzed.

Throughout this thesis, I will be referring to the use of complex NPs as NP complexity with an understanding that NP complexity regards the syntactic arrangement of the NP items in relation to the head noun, as also governed by the semantic-pragmatic interplay established therein. In spite of sometimes focusing more on the grammatical side of complex NPs, their semantic-pragmatic properties are not meant to be downplayed.

2.3.1 *Why investigate NP complexity?*

An important reason for investigating NP complexity is the prominent role nouns have in communication. Because nouns are understood as “the main conveyors of information in a sentence” (HERRERO-ZORITA; SANDOVAL, 2016, p. 2), they pack a substantial amount of semantic-pragmatic information, particularly in complex NPs. As extensively demonstrated in chapter four of the present thesis, complex NPs with hyphenated premodifiers may condense even more information.

Information packing is a trait quite clearly illustrated in the sample sentence seen in Figure 2.1, in which nouns are highlighted in neon green, while adjectival premodifiers are marked in orange. Extracted from the Corpus of Articles in Biology (CorABio), this sample sentence has nouns accounting for 28% of all words, with a much higher percentage obtained once the head noun and its “orbiting” elements are observed.

We can also see that premodifiers may be composed of different grammatical classes. In this example, we see a noun (*microplastics*), an adjective (*coastal*), and two adjectival *-ed*

past participle (*urbanized, impacted*). Such variation is also confirmed in the NPs extracted for analysis in this thesis.

Figure 2.1: CorABio sentence example; nouns in green, adjectives in orange.



In this sample sentence, the NPs carry the bulk of the informational content. It could also be said that the semantic-pragmatic load of written sentences may be correlated to the syntactic sophistication of their corresponding NPs. In the sentence examples containing hyphenated premodifiers illustrated in chapter three, this correlation is readily visible.

Accordingly, it is nominal, and not verbal, groups that allow “the counting, specifying, describing, classifying and qualifying of Things”, as emphasized in Cullip (2000, p. 86). This high propensity for lexical extension may result in “an increase in the density of nouns and other lexical members of the nominal group” (*ibid.*), which can make writing more lexically dense and thus hamper understanding.

As argued in Bartning, Arvidsson, and Lundell (2015), complexity may be viewed as an indicator of L2 proficiency and is relevant for the study of advanced L2 learners. Because NP complexity may be consistent with more syntactically (therefore semantically) complex language users can (effortlessly) produce, it may accordingly be understood as “an interesting measure for the debate on native-likeness”, as these authors contend (p. 197).

While in this thesis native-likeness is not considered a goal for Brazilian speakers of English, clarity and compliance with discipline-specific uses are. In this sense, it is possible to agree with Bartning, Arvidsson, and Lundell (2015, p. 197) when they mention the relevance of NP complexity as a measure “for high level proficiency in an L2”, explaining that since complex NPs may have a “higher number of words, thus a higher degree of complexity, [they] might be more difficult for NNS”.

From the aforementioned definitions and observations, it might be quite reasonable to say that longer NPs may yield more complex information, since NP complexity seems to be closely related to length. In reality, however, this view might be misleading, as will be seen.

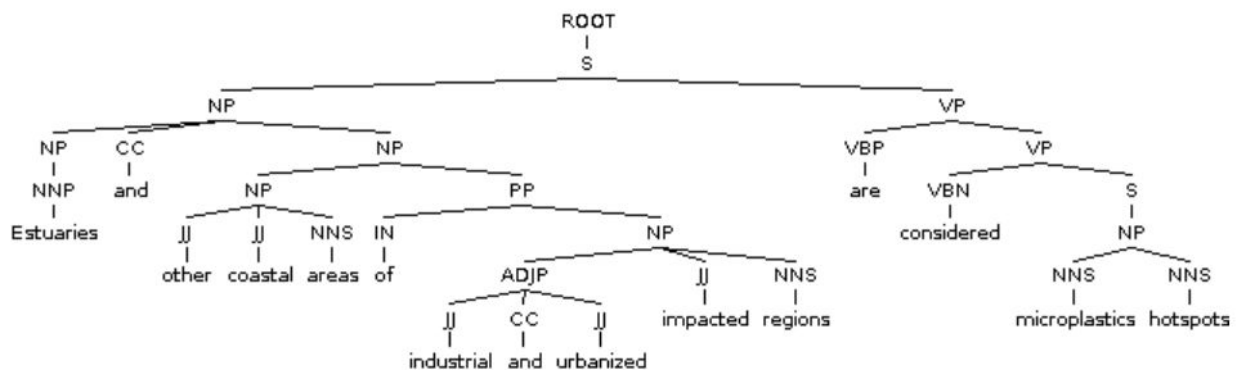
2.3.2 Measuring NP complexity

NP complexity can be measured in different ways. In her corpus-based study, Berlage (2014) lists length, structure, and hierarchy¹⁵. Length refers to the size of the NPs in terms of number of words: the more words an NP contains, the more complex it tends to be. Berlage points out that this is the simplest method for measuring complexity because it only requires counting the modifying words in the NP, in pre and post-modification capacity, a task that can be done quickly through automatization.

On the other hand, not much information can be gathered by word count alone beyond the ‘size’ of the NPs, since no morphosyntactic or pragmatic correlation is made between the head noun and the modifying elements and/or between the modifiers. The implication is that knowing the number of words in an NP does not necessarily yield an understanding of any type of relationship between the elements – semantic, syntactic, or otherwise, thus providing only a partial picture of nominal complexity.

Measuring NP complexity in terms of structure might solve this problem. Structure is related to the number of phrasal nodes in an NP. Since the guiding element, in this case, is the node, structure-based measuring implicitly requires understanding the relationships between the NP constituents. Such relationships can be visually displayed in tree diagrams or bracket format. Figure 2.2. below is a tree diagram for the CorABio sample sentence presented, with six NPs rendered, four of which branch out and dominate other elements. In structure-guided complexity measuring, structure does not necessarily correlate to length (BERLAGE, 2014).

Figure 2.2: tree diagram for the CorABio sample sentence, made with [Apache Tomcat Jesper/Catalina](#) (2018)



¹⁵ Berlage’s methodology for measuring NP complexity is anchored in noun post-modification. Nevertheless, as the author herself states, the same reasoning may apply to nominal premodification.

As mentioned, the head noun and its relationship to the NP constituents (or among the constituents) can be displayed in other formats. An alternative to tree diagrams is dependency, based on dependency grammar (TESNIÈRE, 1959), which represents a different grammatical paradigm also highly employed in NLP (cf. JURAFSKY; MARTIN, 2019). Figure 2.3 shows the dependency structure for the CorABio sample sentence, in which the NP heads and their dependent relations are highlighted in red and semantic information have been annotated.

Figure 2.3: CorABio sample sentence dependencies, made with [Visual Interactive Syntax Learning](#) (2018)

```

Estuaries [estuary] <*> <Lwater> <idf> <nhead> N P NOM @SUBJ §STP §§TOPIC #1->13
and [and] <co-subj> KC @CO #2->5
other [other] DET P @>N #3->5
coastal [coastal] <jpert> ADJ POS @>N #4->5
areas [area] <f-q> <L> <comp2> <idf> <nhead> N P NOM @SUBJ §TH #5->11
of [of] PRP @N< #6->5
industrial [industrial] <jpert> ADJ POS @>N @P< #7->9
and [and] <co-prenom> KC @CO #8->9
urbanized [urbanize] V PCP2 STA @>N @P< #9->11
impacted [impact] <v.contact> V PCP2 STA @>N #10->11
regions [region] <Lregion> <nhead> N P NOM @<ACC §PAT #11->0
are [be] <aux> V PR -1/3S @FS-STA #12->0
considered [consider] <v-cog> <mv> <fn:think> V PCP2 PAS @ICL-AUX< #13->12
microplastics [microplastic] <heur> <idf> <ncomp> N P NOM @>N #14->15
hotspots [hotspot] <Labs> <idf> <nhead> N P NOM @<SC §ATR #15->13

```

The third and final method for measuring NP complexity is hierarchy. As related to the number of clauses in an NP, this method follows Ross’s (2004 [1973]) scale of *nouniness*, in which NPs range from being very noun-like to being more clause-like, with the sentential NPs seen as more complex¹⁶. Drawing from this work, Berlage (2014, p. 15) argues that “the more sentence-like an NP is, the more complex it is”.

Based on a methodological choice by which only post-modification was investigated, she equates “the property of being nominal [...] with the quality of being less complex and the property of being sentential with the quality of being more complex”. (p. 41). Despite being in apparent opposition, such rationale works for premodification, particularly if we consider the need for segmentation of past participles in hyphenated premodifiers (as seen in chapter four).

¹⁶ Originally published in 1973 in response to issues from Chomsky’s transformational grammar, Ross’s lengthy and very abstract account places verbs as higher than nouns (in a syntax tree), in a continuum in which more noun-like NPs are less complex, and more clause-like NPs are more complex. Ross also maintained that “the traditional view of the categories verb, adjective, and noun, under which these three are distinct and unrelated, is incorrect” (p. 351), advocating for the fuzziness of linguistic categories, a hallmark in non-discrete grammar.

2.4 NP complexity in science writing

As contented in Biber, Grieve, and Ibbi-Shea (2009), NP complexity may correspond to “a fuller exploitation of the production possibilities of the written mode” (p. 182). This is a rather accurate observation about more specialized registers, seeing that these can be “slowly produced and carefully revised and edited” (BIBER; GRAY, 2011, p. 225). Hence, the reader also has the possibility of complementing his/her comprehension by re-reading sentences and paragraphs, which may be required for interpreting highly compressed NPs (BIBER, GRAY, 2016).

In this sense, the relevance of NPs in academic prose¹⁷ may rely on their pervasiveness and function. NPs are more common in academic writing possibly because, when complex, they can compress large quantities of semantic-pragmatic information by means of pre- and post-modification, which is of particular convenience in word/page limit-bound texts such as RAs. So do hyphens, abbreviations, and acronyms, which may be used to further compress information in RAs.

2.4.1 A shift towards structural compression

As explained in Biber and Gray (2016, p. 168), informational written registers package information in NPs, presenting it “in phrasal modifiers rather than clausal modifiers”, which is consistent with Biber et al’s (1999) finding that 60% of nouns in academic writing display some modification. This was not always the case. As Biber and Gray (2016) explain, science writing has shifted from a more clausal-like style to more phrasal-bound patterns.

The historical shift towards more phrasal-like compressed patterns in scientific writing is shown in Figure 2.4, reproduced from Biber and Gray (2016), where a gradual decrease in the use of predicative relative clauses in science RAs is illustrated. As can be seen, the past 50 years have experienced an overall significant decline in the use of such clauses.

Relatedly, the increase of compressed *-ed* and *-ing* participles in noun premodification is showcased in Figure 2.5, as also reproduced from Biber and Gray (2016), with my added emphasis (in purple) to the growing *-ing* trend and the spike in *-ed* use. The correlation made between these trends is that the decrease in clausal structure seems to propel the increase in compressed NPs.

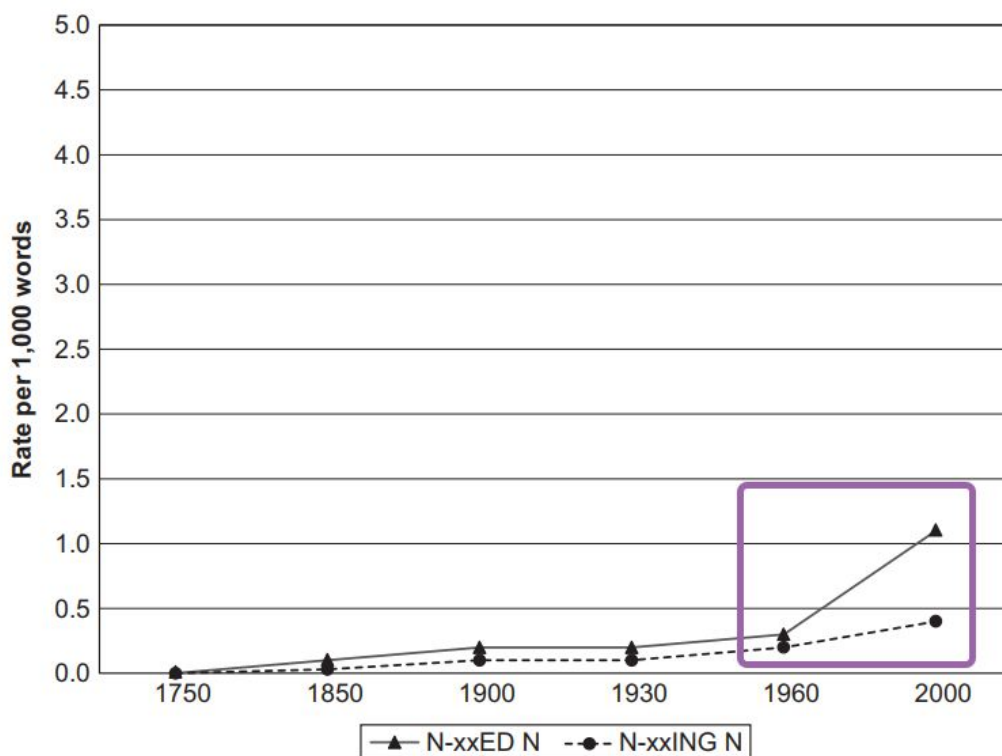
¹⁷ Scientific/science writing and academic prose/writing are often used interchangeably in this thesis.

Figure 2.4: historical view of the decrease in predicative clauses in science RAs.



Source: Biber and Gray (2016, p. 209).

Figure 2.5: historical view of the increase in compressed *-ed* and *-ing* noun-initial NPs.

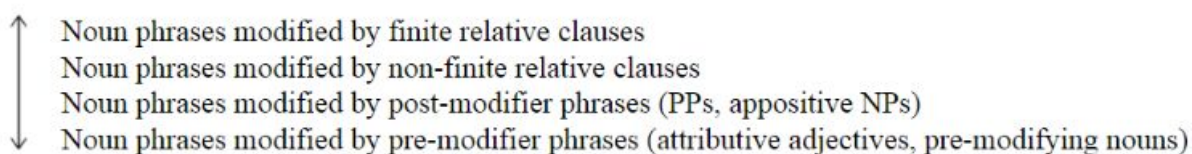


Source: Biber and Gray (2016, p. 209).

Biber and Gray (2016) have proposed a *cline of structural compression*, as reproduced in Figure 2.6, in which more structurally-condensed NPs are seen as more compressed. This is a historically, usage-based rationale signalling a longitudinal shift towards NP compression in scientific writing, particularly in RAs. As stated in Carter and McCarthy (2006), Biber et al. (2002), and Biber and Gray (2016), complex, compressed NPs provide the research writer the possibility of conveying more information in (much) fewer words.

Figure 2.6: Biber and Gray's (2016, p. 277) cline of structural compression.

Least compressed



Most compressed

2.4.2 Hyphenation

As mentioned, for a genre usually bound by word/page-limit restrictions such as RA, NP compression can be an appealing alternative. As discussed in chapter four, this strategy can be deployed by means of hyphenated, abbreviated, and acronymized forms. *chl-a-related variables*, *a separate correlation-based PCA*, *329 intra-specific environmental gradients*, and *This newly-implemented Kalman-filtering algorithm* are examples from the corpus compiled for this thesis, in which complex NPs with a high degree of premodification can yield obscure and less explicit meanings – possibly not in regard to form, morpho-syntactically speaking¹⁸.

Example 2.2

→ **This newly-implemented Kalman-filtering algorithm** provides more estimated positions and significantly improves position accuracy.

(CoBRA file BC.2016.02.txt)

Example 2.1

→ We consider **329 intra-specific environmental gradients** in adult body size across latitude, altitude and with seasonal temperature variation (...)

(CoBRA file OE.2019.10.txt)

¹⁸ Some of the constituents may have their grammatical function easily recognized via derivational morphology, which can also facilitate processing.

Example 2.3

→ Next, we performed **a separate correlation-based PCA** for all (1) pH-related variables, (2) **chl-a-related** and (3) temperature-related variables (...)

(CoBRA file EL.2016.06.txt)

What could make these NPs more explicit in meaning, as Biber and Gray (2016) have suggested for similar examples, is the background knowledge about specialized terminology, such as *SPAs* or *chl-a*. Additionally, Pirrelli, Guevara and Baroni (2010) and Biber and Gray (2016) might also recommend understanding the semantic-pragmatic correlation between the constituents in these NPs, a point also made in Dutra et al. (2020).

Regardless, it could be generally agreed that these NPs house considerable amounts of information in premodification. It could also be agreed that they take up less space than their verbal segmentations would. For instance, *chl-a-related variables*, with two items and three or four constituents¹⁹, could be segmented as *variables which are related to chl-a*.

Rendered in six/seven²⁰, this segmentation may indeed be of easier understanding for less proficient readers and English language users. On the downside, however, it takes double the space when compared to its compressed counterpart *chl-a-related variables*. Moreover, it yields more tokens than *chl-a-related variables*, especially seeing that *chl-a-related* is counted as one item in any word processor. Hyphenation may be employed as a compression device in this sense.

Another argument in support of hyphens viewed as a compression device is the role this punctuation mark has in forming compounds in the English language. More often than not, as noted in Nunberg, Briscoe and Ruddleston (2002), hyphens may be used to “join the bases of a compound (...) or the affix and base of a derivative” (p. 1760). Compounds are discussed in subsection 2.4.4, in association with hyphenation.

Furthermore, the use of hyphens in concatenated nominal premodifiers, as seen in the NPs exemplified, signals the semantic links between the NP constituents and in relation to the NP head, which may help comprehension. As explained in Nunberg, Briscoe, and Ruddleston (2002), Kösling and Plag (2009), and Sanchez-Stockhammer (2018), hyphens assist readers in selecting the correct interpretation for less explicit and sometimes ambiguous NPs.

¹⁹ *Chl-a* stands for *Chlorophyll a*. As a name, it would be considered a single unit. For those unfamiliar with this term, it might be understood as two separate units connected by a hyphen, namely *chl* and *a*.

²⁰ See previous note.

As stated in Nunberg, Briscoe, and Ruddleston (2002), Sanchez-Stockhammer (2018), hyphenation use may vary on a more individual level (at the level of the language user), or on a more collective level (in the form of academic writing conventions). Nunberg, Briscoe, and Ruddleston (2002) explain that hyphenation preferences may also be more closely associated with language variety, citing specificities for American (AmE) and British English (BrE).

According to Nunberg, Briscoe, and Ruddleston (2002), BrE tends to favor hyphens if compared to AmE. Hyphenation is also preferred when expressing kinship, numbers between twenty-one and ninety-nine, and fractions (NUNBERG; BRISCOE; RUDDLESTON, 2002). Examples are seen in *great-grandfather*, *twenty-five*, and *two-thirds*, respectively. These uses are also confirmed in Sanchez-Stockhammer (2018).

This is an important observation. Despite the ELF perspective adopted in this study, a closer look at the journals selected for compilation show a predilection for BrE spelling and punctuation conventions. This, however, does not negatively impact the research done in this thesis. Rather, it adds support to a more variational view based on the notion of *competing variants*, as will be seen in chapter four.

2.4.3 Word formation

In regard to the production of the complex NPs exemplified throughout this thesis, as will be presented in chapter four, the present research would be lacking by not addressing an important aspect inevitably involved in NP complexity, hyphenated or not: word-formation²¹. This mention is also justified by the challenges word-formation processes often present to the Brazilian speakers of English with whom I work at UFMG.

More broadly understood not as a single theory, but as a diversified group of processes (BAUER, 1983; HUDSON, 2000; BAUER, HUDDLESTON, 2002; JURIDA, 2018), among them compounding, affixation, and back-formation, these processes are conventionally taken to be the primary foundation for complex forms²² (BAUER, 1983; BAUER, HUDDLESTON, 2002), hence their importance for the present research. This subsection deals with derivation and compounding as prime examples of word formation in English.

²¹ Word-formation processes in English owe much to the intense language contact periods, by numerous words were borrowed into this Germanic language, especially during the Middle English period, as discussed in Baugh and Cable (2013) and Crystal (2002, 2005), among others.

²² Complex in the sense of 'produced by derivation' (BAUER, 1983).

2.4.3.1 Derivation

In terms of the operational definitions hereby employed, derivation is understood as a morphological process by which new words are created via derivational affixation (BAUER, 1983; HUDSON, 2000; BAUER, HUDDLESTON, 2002). For Hudson (2000), derivation is a major word-formation procedure favored in STEM-based²³ disciplinary fields such as natural sciences, computer science, and physics, a tendency readily verified in this research.

The main processes of derivation in English are prefixation and suffixation, which aid in forming new word arrangements on the basis of adding prefixes and/or suffixes to the word root. *Newly*, for instance, has taken the prototypical suffix *-ly*, which turns the adjective *new* into an adverb. Likewise, the suffix *-al* is added to *environmental*, making an adjective from the noun form *environment*.

In *variables*, suffixation brings about changes to the base *vary* by eliminating the *y* to receive the suffix *able*. Interestingly in this case is the addition of the inflectional *s* to mark the plural of *variable*, illustrating the interplay between morphological processes. We also see that the same suffix can indicate different word classes: *-able* may be more often employed in adjectives. Examples 2.4 to 2.7 below illustrate these uses.

Example 2.4

→ This species reproduces all year round, with recorded observations of gravid females, eggs and **newly hatched juveniles every month** (...)

(CoBRA file AB.2019.08.txt)

Example 2.5

→ As an alternative to the niche overlap tests we also performed a principal component analysis (PCA) of **the nine environmental variables** (...)

(CoBRA file ZS.2017.04.txt)

Example 2.6

→ The inclusion of **this variable** also accounts for whether individuals had experienced these tasks as chicks (...)

(CoBRA file AB.2019.08.txt)

Example 2.7

→ The result implies that (...) the impacts on the underlying OTUs were **relatively more variable** and therefore more difficult to detect using the sample sizes available.

(CoBRA file EL.2018.04.txt)

²³ Science, Technology, Engineering, and Mathematics (STEM) is conventionally understood as “a coordinated strategy for pre-college education” (BYBEE, 2010, p. 996) currently in place in research universities around the world (cf. APPLEBY, 2012; REYNOLDS, 2012), in close association with in-house academic writing programs (HARRISON; BROOKS, 2017).

By the same token, prefixation also assists in word formation, in this case by attaching an affix to the very beginning of the word. As mentioned in Bauer and Huddleston (2002) and Bauer (1983), prefixes may be class-maintaining or not. This means that by adding a prefix to a word base, the grammatical class of that word may or may not change. In *regeneration*, for instance (Example 2.8), the prefix *re-* does not produce any grammatical changes to the noun.

Example 2.8

→ Understanding the factors that promote or inhibit **seismic line forest regeneration** is therefore critical for determining future thresholds (federal targets) for caribou habitat
(CoBRA file BC.2015.04.txt)

Prefixation is also visible in Example 2.9, in which *non* presents a hyphen that visually signals its connection to *target*. Together, *non-* and *target* modify the acronymized (and quite possibly lexicalized) noun *DNA*. Prefixes generally do not occur in stand-alone mode, that is, they are largely not used on their own. Example 2.10 shows an NP with a solid use of *non*.

Example 2.9

→ Environmental DNA assays reduce missed detections resulting from samples dominated by **non-target DNA** (...)
(CoBRA file BC.2019.01.txt)

Example 2.10

→ We have employed longitudinal studies since 1990, (...) **applying nontoxic acrylic paint** for identification of individuals from a distance (...)
(CoBRA file AB.2019.03.txt)

2.4.3.2 Compounding

As stated in Pullum and Huddleston (2002), compounding is part of word-formation processes and refers to “forming a new base” by combining two or more smaller bases (p. 28). According to Bauer and Huddleston (2002, p. 1646), most compounds are “subordinative, in that one base can be regarded as head, the other as dependent”. This would mean that *gender equality* (Example 2.11) has *equality* as its head, and *gender* as a dependent much in the same way as the head-modifier relationship in an NP.

Example 2.11

→ How the observed effects might change as a function of sociocultural factors (e.g. typical mating strategies or **gender equality**) remains to be investigated.
(CoBRA file OE.2018.02.txt)

Additionally, Bauer and Huddleston (2002) explain that compound bases may display equal status, meaning that “neither component is dependent on, subordinate to, the other.” (p. 1646) and that coordination-based compounds are normally segmented with the conjunction *and*. For example, *secretary-treasurer* refers to “someone who is both secretary and treasurer, not (or not just) a kind of treasurer” (ibid). This is a compound anchored in coordination.

From the research corpus compiled for this thesis, an example would be found in the hyphenated *stimulus-response* in the complex NP *a stimulus-response association* illustrated as Example 2.12 below. *Association* clearly establishes a subordination connection with the NP premodifier, where the modifier ‘responds’ to the head (*association*). But *stimulus-response* seems to be on the same level, in a trading relationship.

Example 2.12

→ The first follows the Thorndikian law of effect, which proposes that the strength of a **stimulus-response association** is directly related to reward magnitude and probability.

(CoBRA file AB.2015.10.txt)

Compounding is therefore the productive process of combining existing words/bases into new expressions. It may be represented in three distinct ways: 1) solid/juxtaposed, also called concatenated, such as in *whiteboard* and *dressmaker*; 2) hyphenated, as in *small-scale activity* and *word-play*; and 3) open/separated/spaced, as in *parking space* and *shopping mall*. It should be noted, however, that many of these compound spellings may be interchangeable (KUPERMAN; BERTRAM, 2015).

The present thesis could not possibly cover all the types and attributes of compounds as employed in the English language. For this, Bauer and Huddleston (2002) are an excellent reference, because they offer a comprehensive look at compounding in English, resorting to a number of examples from corpora and reference work. Still, some other important remarks in regard to English compounds may be made.

The first is that, as these authors demonstrate, nouns are the main recipient for English compounds. This in itself is highly relevant for any study in NP complexity. Compound nouns may be verb- or noun-centered. In general terms, the former are compound nouns “where the central element is ‘verbal’, its form being identical with that of the lexical base of a verb (...) or derived from it by suffixation” (BAUER, HUDDLESTON, 2002, p. 1652). As will be seen in chapter four, such compound nouns are very pervasive in the research corpus.

The latter type, termed noun-centered, refers to compounds that have a noun as their final base. As Bauer and Huddleston (2002, p. 1647) assert, in the majority of cases, “the first element is a dependent, the final one the head”. Such compounds are also largely found in the research corpus. A preview can be more readily identified in *oxygen consumption* in Example 2.13 below. In this noun-noun compound, *oxygen* is subordinated to *consumption*.

Example 2.13

→ *Maximum metabolic rate was then determined using an exhaustive chase protocol followed immediately by measurement of oxygen consumption during the recovery period using closed-system respirometry (...)*

(CoBRA file EL.2018.01.txt)

Finally, as stated in Bauer and Huddleston (2002, p. 1647), noun-noun compounds are “by far the most productive kind of compounding in English, and indeed the most productive kind of word-formation”. The authors highlight the wide range of semantic relations between the compound bases, arguing that the meaning of the compound is often not derived from its parts. They view noun-noun compounds as “lexical structures designed to act as mnemonics” (BAUER, HUDDLESTON, 2002, p. 1647, 1648).

It should be noted that, in this particular case, the authors seem to base their arguments on the relative ease with which an L1 speaker would experience noun-noun compounds, both in terms of production as well as in regard to interpretation. This might not be the case for L2 speakers of English, especially if their L1 is grounded on opposite patterns, which is often the situation Brazilian speakers face.

2.4.4 Hyphenation in compounding

Nunberg, Briscoe, and Ruddleston (2002) notice a correlation between hyphenation in compounding. Fixed expressions, as the authors explain, tend to be solid/juxtaposed. On the other hand, more recent compounds may welcome hyphens, possibly leading to higher uses of hyphenation in new, creative nominals such as in business names (e.g., *Bristol-Myers Squibb*, *Coca-Cola*).

In terms of compound interpreting, eye-tracking experiments have shown that hyphens and spaces play an important role in meaning-extraction (cf. INHOFF; RADACH; HELLER, 2000; JUHASZ; INHOFF; RAYNER, 2005) and that hyphenation may offer better visual

cues for compound processing. In addition, these studies confirm that frequency greatly determines processing speed: the more frequent the word, the faster the processing, since the reader may already be familiarized with the compound, both in form and meaning.

Spelling variation for English compounds is by no means a straightforward matter. As investigated by Sepp (2006) and Kuperman and Bertram (2015), the choice of hyphenated, solid, or open spelling can be explained by several factors, from phonological preferences to frequency of occurrence and semantic transparency.

In Nunberg, Briscoe, and Ruddlestone (2002), hyphens are understood as indicative of the semantic connection between two or more constituents. Similarly, Kuperman and Bertram (2015) found that in noun-noun compounds, writers alternate between hyphenated and spaced compounds to convey differences in grammatical/semantic function, selecting hyphenation to signal adjectival uses in the form of premodification to the head noun.

This alternation between hyphenated and spaced compounds was not extensively seen in the research corpus. As will be discussed in chapter four, alternation does occur. However, it does so increasingly more with adjectival uses, with hyphenation largely preferred. What is more, this alternation is often not bound to differences in grammatical/semantic function and might be rather justified as an idiosyncrasy.

Examples 2.14 to 2.16 illustrate this alternation for *power-law*, with the spaced variant using a premodifier and a noun in the same file/text. The only other use of the spaced variant is reproduced in Example 2.17. All other occurrences for *power-law* are hyphenated: 23 for hyphenated uses vs. 3 tokens for spaced occurrences, one of which is not a premodifier, as already stated.

Example 2.14

→ Further problems with **the power-law SAR** are that it is phenomenological, (...) and that it ignores scale-dependent variation in **the power-law exponent** observed in empirical data.
(CoBRA file EL.2018.08.txt)

Example 2.15

→ By 'slope', we refer to the slopes in logarithmic plots, corresponding to **the power law exponents** in the limiting cases of eqn 4.
(CoBRA file EL.2017.07.txt)

Example 2.16

→ If no inorganic nutrients are available or in the case of animals that cannot take up inorganic nutrients, the decline follows **a power law** with exponent -1 .
(CoBRA file EL.2017.07.txt)

Example 2.17

→ Handling times often exhibit **a negative power law** with increasing consumer mass.

(CoBRA file EL.2016.05.txt)

These findings and observations likely influenced my decision to slightly change the research focus of the present study. To finalize this section, a list of hyphenated preferences in compounding is reproduced below from Nunberg, Briscoe and Ruddleston (2002, p. 1761).

Figure 2.7: hyphenated compounds exemplified in Nunberg, Briscoe and Ruddleston (2002, p. 1761)

i	compound adjective	<i>bone-dry, oil-rich, red-hot, snow-white</i>
ii	contains transitive prep	<i>free-for-all, sergeant-at-arms, sister-in-law</i>
iii	intransitive prep as 2nd base	<i>break-in, build-up, drop-out, phone-in, stand-off</i>
iv	coordinative compound	<i>Alsace-Lorraine, freeze-dry, murder-suicide</i>
v	nominal compound + <i>·ed</i>	<i>one-eyed, red-faced, three-bedroomed</i>
vi	numerals and fractions	<i>twenty-one, ninety-nine, five-eighths</i>
vii	dephrasal compounds	<i>cold-shoulder (V), has-been (N), old-maidish</i>
viii	verb with noun as 1st base	<i>baby-sit, gift-wrap, hand-wash, tape-record</i>
ix	1st base is letter-name	<i>H-bomb, t-shirt, U-turn, V-sign</i>
x	rhyiming-base compounds	<i>clap-trap, hoity-toity, teeny-weeny, walkie-talkie</i>

2.4.5 A move towards discipline-specific NP complexity

As indicated in Biber and Gray (2016), most research on scientific writing in English has focused on rarely-occurring phenomena or on salient features more commonly found in other registers. By taking a different stance, and relying on CL principles, these authors argue that understanding academic prose entails examining its more frequent (therefore less salient) traits. This can easily be done automatically by generating wordlists from existing corpora, provided the corpora are morpho-syntactically annotated.

This approach involves a more fine-grained inspection of linguistic phenomena within the situational properties of the larger context where the linguistic feature is at use, which, in turn, may require looking at register/genre variation. As stated in Hyland (2016), genres are "socially recognized ways of using language and represent how writers typically respond to recurring situations" (p. 120), implying that disciplines see the world differently and make use of different linguistic and communicative conventions (HYLAND; HAMPS-LYON, 2002).

Cullip (2000, p. 85) states that because NPs “can be stretched syntactically and packed semantically”, academic prose tends to be lexically dense and grammatically complex, a point also made in Biber and Gray (2016) about register that convey large quantities of information in a limited number of pages, such as RAs. In this sense, a productive venue for investigating complex premodified NPs can be found in the writing of RAs in specific fields of scientific knowledge.

Moreover, heavily packed NPs can lead to ambiguity, since syntactic relationships and meanings are not explicit in compressed nominal groups and may not be readily understood. A resolution for potentially ambiguous constructions lies not only in familiarity with a given issue or conceptual content but also in the efficient processing of new and given information. Lexical density, syntactic complexity, and ambiguity may therefore present challenges for L1 and L2 language users, especially in scientific writing (HALLIDAY; MARTIN, 1993).

For Brazilian expert and novice academics who need to write and publish in English, densely pre-modified NPs may be even more complicated, seeing that Brazilian Portuguese tends to rely more on noun post-modification, which is more explicit in meaning and syntax. Consequently, understanding and producing complex pre-modified NPs may prove to be an arduous task, especially in research writing.

Noun premodification in RAs can vary from area to area. As thoroughly discussed in Biber and Gray (2016, 2010), CL research has revealed that lexico-grammatical patterns in STEM-based disciplines may be considerably different from those of more socially-oriented sciences, with the former employing more compression than the latter.

This tendency is also found in Dutra et al. (2020) in their cross-disciplinary analysis of RAs in Chemistry and Applied Linguistics. Not surprisingly, more discursive elements, such as the organizational layout of RAs in terms of sections also vary across disciplines, as noted in Dahl (2004).

In addition to Biber and Gray (2011) and Biber and Conrad (2015), Hutter (2015) investigated noun pre- and post-modification in different sections of Applied Linguistics RAs. Similarly, Hong, Hua, and Mengyu (2017) have examined premodification in International Business Management RAs, while Parkinson and Musgrave (2014), Yang (2015), Lu and Ai (2015), Lu (2011), and Queiroz (2019) are among those who have probed NP complexity in advanced/intermediate L2 academic writing.

2.5 Corpus Linguistics

As a highly versatile sub-area of Linguistics, CL principles and methodology can be employed in a plethora of linguistic investigations. Such multidisciplinaryity can also be seen in the ongoing dialogue CL has established with the greater areas of Statistics and Computer Science. In regard to the former, for instance, it has become more standard for corpus-based investigation results and annotation to be tested and/or validated statistically.

In addition to stressing the importance of Statistics in CL, Baroni and Evert (2006) and Gries (2008) advocate for the employment of statistical treatment of corpus data, which the authors believe may better complement corpus findings. Adopted in this thesis are notions such as dispersion and adjusted frequency, employed in chapter four. Simple statistical tests to verify significance are also utilized.

It is with Computational Linguistics and Computer Science in general, though, that CL may need to establish more permanent and fruitful exchanges, through the use of corpus data to inform machine learning processes or via data treatment techniques in NLP²⁴, as well as in corpus compilation and data analysis, on the CL side. This thesis, for instance, has attempted to do that by finding time-efficient solutions for corpus compilation and/or data processing, as shown in chapter three.

The usefulness of CL as a methodological tool in EAP is discussed in Hyland (2016), Coxhead (2010), Yoon (2008), and Lee and Swales (2006), who have observed that CL has been increasingly employed in teaching practices, particularly in academic writing instruction in English. In this regard, Tribble and Wingate (2013) maintain that CL-based studies can be used in discipline-specific supplementary material to inform language teachers about current patterns of language use (as based on large scale data).

²⁴ https://www.lancaster.ac.uk/fass/projects/spatialhum.wordpress/?page_id=40.

3. MATERIALS AND METHODS

3.1 Overview

This chapter describes the different methodological procedures undertaken for corpus compilation and NP extraction. Firstly, it first presents the pilot corpus, outlining some issues identified in this material. Secondly, it describes a proposed solution to those issues, including a description of the research corpus. Thirdly, it offers an account of the computational steps taken to extract the NPs for subsequent analysis. Figures accompany all steps described in this chapter²⁵.

3.2 The pilot corpus

The initial corpus (henceforth pilot corpus) was compiled within the *Grupo de Estudos em Corpora Especializados e de Aprendizizes* (GECEA) in the first half of 2018. The decisions on corpus design were made following the same criteria adopted in other specialized corpora previously compiled in the research group: 150 research articles (RAs) published between 2014 and 2018, that is, 50 RAs collected from three high impact international journals²⁶, ten RAs per year.

Just as it had been done with CorAChem – the corpus of Chemistry articles, the choice of journals was made based on a list provided by senior researchers at the Federal University of Minas Gerais (UFMG), from which I curated 62 journals by organizing them according to sub-area, scope, and frequency of publication.

Similarly, and in order to be a comparable corpus with other specialized corpora from the research group, the RAs collected had to follow the **IMRD(C)** model (cf. SWALES, 1990) by clearly displaying the following sections: **I**ntroduction, **M**ethodology, **R**esults, **D**iscussion, and **C**onclusion (and abstracts). Token count was not recorded. This means that no restrictions were made regarding the minimum or maximum number of words per text. The corpus was primarily anchored in the sub-area of zoology, with the following journals selected: *Frontiers in Plant Science*, *Frontiers in Zoology*, and *Zoologica Scripta*.

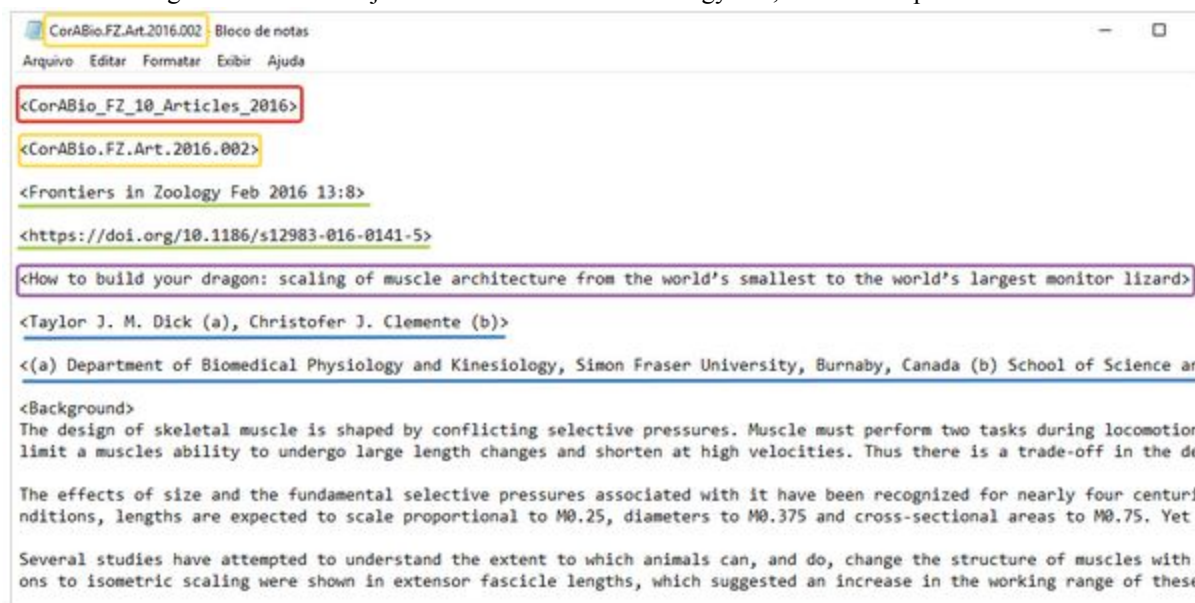
²⁵ Unless otherwise stated, all screenshots, tables, and figures have been produced by the author.

²⁶ The list provided contained only Qualis A1 journals. For more information on the CAPES metric, please see: <http://www.capes.gov.br/pt/acessoainformacao/7422-qualis> (in Portuguese).

3.2.1 Compilation

The corpus was compiled almost completely manually²⁷ through the manual collection of the following information: title, author(s) name(s), institution(s) and contact information, abstract, and the body of the article. Such information was then copied and pasted at Notepad and saved as plain text with UTF-8²⁸ encoding, a standard procedure in CL research. In order to optimize collection time, Notepad++ (HO, 2019) was chosen to replace Notepad halfway through compilation. Information such as title, author(s) name(s), institution(s), and contact information was then manually annotated between chevrons (<>), so that such data would not be read by concordancers or other software during data processing. This means the metadata and the body of the article were kept together in the same file, as shown in Figure 3.1²⁹.

Figure 3.1: metadata just above a Frontiers of Zoology RA, in the same plain text file.



The next step was saving the abstracts, which were stored separately from the body of the text, each in an individual plain text file. Abstract files also included the article metadata to which they refer. The choice of keeping the abstract separate from the body of the RA was made according to previous genre studies in which abstracts are oftentimes examined on their

²⁷ An attempt had been made at collecting the journal texts automatically on Sketch Engine (KILGARRIFF et al, 2014, 2018). Unfortunately, however, such attempts did not yield positive, productive results. One such problem was the amount of unnecessary information included in the automated compilation, which would have required various degrees of cleaning, thus becoming counterproductive.

²⁸ Unicode Transformation Format with 8-bit blocks for character recognition. For more information, please see: <http://www.utf-8.com/> and <https://www.ime.usp.br/~pf/algoritmos/apend/unicode.html> (in Portuguese).

²⁹ As with all other chapters, Tables and Figures follow a sequence first signalled by the number of the chapter.

own terms (VENTOLA, 1994; MELANDER; SWALES; FREDRICKSON, 1997; LORÉS, 2004; SALAGER-MEYER; SEGURA; RAMOS, 2016). The process of abstract separation was done manually.

Likewise, RA sections may also be inspected on their own, perhaps not as full-fledged genres, but as registers³⁰, in which lexico-grammatical patterns and specific discourse features can be examined (cf. GLEDHILL, 2000; MARTINEZ, 2005; PARKINSON, 2011; CORTES, 2013; MACEDO, 2018). For this reason, individual plain text files for each RA section were created. Sections were sorted manually and saved in individual files. The same had been done for other specialized corpora compiled in the group, thus allowing cross-disciplinary analyses as well as cross-section examinations.

After the collection, unnecessary information had to be removed from the texts, a task done manually by various research group members. This clean-up aimed at removing figures and graphs, formulas, and equations. By selecting, copying, and pasting this information at Notepad++, most of this data (e.g.: figures and graphs) were automatically not included, since this software does not read picture-format information. This helped save time. Nevertheless, alphanumeric information (H₂O) and section headings had to be kept between chevrons (< >), since they were part of the text. This mark-up process was carried out manually.

This corpus was also initially compiled to provide data for multidimensional analyses³¹ (MDA) (cf. BIBER, 1988, 1992a,b), which requires the Biber tagger (BIBER, 1988) or the Multidimensional Analysis Tagger (NINI, 2018). In order for these software to work properly, more information had to be removed from the corpus (numerals, foreign characters, and other number-letter associations such as H₂O). A second cleaning was done semi-automatically in Notepad++ and then checked manually.

Following other specialized corpora compiled in the research group, the corpus was annotated with CLAWS 7³², a part-of-speech (POS) annotation tool. Separate folders housed the POS-tagged files. Both the abstracts and the RA sections previously stored independently underwent POS tagging. The corpus creation, from text collection to cleaning and annotation, was concluded in the first semester of 2018.

³⁰ See Biber and Conrad (2009) for a discussion on the distinction between genre and register.

³¹ See Dutra and Sardinha (2018) for an MDA-based study using this corpus.

³² The Constituent Likelihood Automatic Word-tagging System (cf. RAYSON; GARSIDE, 1988). For tagset and further details, see: <http://ucrel.lancs.ac.uk/claws7tags.html> and <http://ucrel.lancs.ac.uk/claws/>.

3.2.2 Understanding the corpus

In order to gain insight into the corpus content, two corpus-related software were used: Sketch Engine (KILGARRIFF et al, 2014, 2018) and AntConc (ANTHONY, 2019). The pilot corpus contains 825,863 word tokens and 30,763 word types, with function words such as *the* and *of* at the top of the list, as expected.

Interestingly, as shown in figure 3.2 below, is the use of *et al* and *we* among the most frequent words, which may be indicative of the collective authorship involved in the writing practices of the biological sciences. Equally interesting is the verb *be* in five forms in this list (*were*, *was*, *is*, *are*, and *be*), which might be an indicator of passive constructions. Also worthy of mention is the *species*, a content word in the midst of 25 mostly functional words.

Figure 3.2: the twenty-five most frequent words in the pilot corpus.

Rank	Freq	Word
1	47026	the
2	29066	of
3	28367	and
4	22220	in
5	15269	to
6	12770	a
7	8041	for
8	8031	were
9	7620	with
10	6629	et
11	6625	al
12	6389	was
13	5693	as
14	5649	that
15	5386	from
16	5382	is
17	4948	by
18	4290	species
19	4148	on
20	4118	at
21	3952	are
22	3529	this
23	2976	be
24	2934	we
25	2897	s

Sketch Engine presents a built-in POS-tagger and automatically runs morphosyntactic annotation in any corpora. Similarly to the NLTK (Natural Language Toolkit) project (BIRD, LOPER, KLEIN, 2009), Sketch Engine’s standard POS-tagset³³ follows the Penn Treebank Project³⁴, which means its tagset is different from CLAWS7. The corpus files were then added to the platform for automatic wordlist generation aimed at providing snapshots of word count, word/token ratio, and amount of nouns, adjectives, and verbs. A list of total counts for nouns and verbs is shown in the following tables.

Figure 3.3 exhibits the twenty-five most frequent nouns in the pilot corpus, indicating the overall semantic content of the RAs. The most frequent semantic fields mentioned refer to two categories: a) biology-related lexicon (*species, plant*) and b) academic-scientific practices (*study, analysis, sample*). In the former, units and proper names specific to the natural world (*C.* for *Celsius*) can be found. In the latter, visual resources used in scientific communication (*figure, table*) can be seen.

Figure 3.3: the twenty-five most frequent nouns in the pilot corpus

lemma	tag	
specie	NNS	2669
datum	NNS	1636
figure	NP	1546
species	NN	1501
fig.	NN	1481
p	NN	1412
table	NP	1342
c	NP	1324
study	NN	1312
s	NP	1208
gene	NNS	1208
analysis	NN	1108
plant	NNS	1082
group	NN	1002
gene	NN	996
analysis	NNS	926
result	NNS	911
study	NNS	862
plant	NN	844
sample	NNS	836
time	NN	829
group	NNS	805
number	NN	795
e	NP	765
condition	NNS	759

³³ <https://www.sketchengine.eu/english-treetagger-pipeline-2/>.

³⁴ For overviews, see Marcus, Marcinkiewicz, and Beatrice (1993); and Taylor, Marcus, and Santorini (2003).

Figure 3.4 displays the twenty-five most frequent verbs in the pilot corpus. There are 2,315 types of verbal occurrences, in which *be* is the most widely used verb in the past (*were* and *was*), in the present (*is* and *are*), and in the infinitive (*be*), followed by *have* (3,955), *use* (3,807), *show* (1,633), and *include* (1,319).

Be is so productive that it functions almost as if it were a function word, as opposed to a content word. In this lemmatized list, we see that it occupies the five most frequent verb slots in the corpus, with *were* displaying nearly four times as many occurrences as the next verb in line (*be* – *were* 8,031 vs. *use* 2,103).

Figure 3.4: the twenty-five most frequent verbs in the pilot corpus

lemma	tag	word	
be	VBD	were	8031
be	VBD	was	6387
be	VBZ	is	5382
be	VBP	are	3951
be	VB	be	2974
use	VVG	using	2103
be	VBN	been	1365
have	VHP	have	1358
have	VHZ	has	1242
use	VVN	used	1227
base	VVN	based	869
find	VVN	found	843
show	VVD	showed	677
have	VHD	had	615
have	VH	have	588
include	VVG	including	584
observe	VVN	observed	544
compare	VVN	compared	522
identify	VVN	identified	488
show	VVN	shown	457
increase	VVN	increased	425
accord	VVG	according	425
do	VVD	did	423
perform	VVN	performed	417
see	VV	see	407

Such productiveness may be related to *be* being used as an auxiliary verb, most likely in passivized constructions. It could also be an indicator of descriptions, in which *be* (as *were*, *was*, *is*, and *are*) is used as copular verb, i.e., employed to connect adjectives and/or nouns to the subject of the sentence. I tried to verify this possibility by means of a regular expression: `[tag = "VV"] - v & [lemma = "be"]` and the results partly confirmed my suspicion.

3.2.3 Metadata

As stated, the 150 RAs in the pilot corpus displayed the metadata, the section headings and the figure/chart captions between chevrons to prevent their reading by concordancers and other software. Even though this is entirely appropriate, I believe my research would be better served by having the metadata kept separately from the body of the RAs. This measure would avoid eventual processing issues. It would also keep the actual corpus more visually clean and the corpus files more content-focused. Because I had been considering adding information of other nature to the metadata, keeping them in a separate file would give me more freedom and space to store this new information. Finally, storing the corpus text independently from other types of information is a recommendation made in Sinclair (2005).

3.2.4 Issues

When checking for *hapax legomena* (i.e., words that only once in the corpus), which are helpful to determine the lexical density of a corpus, special characters and foreign letters were found. These characters and letters should have been removed in the process of cleaning but were kept in the corpus. While their presence would not be an issue for the kind of study I intended to conduct, the fact that they remained in the corpus after the cleanings carried out in the research group is an indication that a clean-up verification may be needed, to minimally confirm the accuracy of such tasks. Figure 3.5 below shows some of these special characters, as verified with AntConc.

Furthermore, the deletion of alphanumeric information initially kept between chevrons (< >) meant items of this type, e.g., *5-mg tissue samples*, (ZS.2018.02), *the complete ATP6 sequence* (ZS.2018.05), *16S, ITS1-5.8S and ITS2 sequences* (ZS.2018.08), and *the partial 28S gene dataset* (ZS.2018.10), would have been deleted³⁵, thus altering the composition of these complex NPs.

In these premodified NPs, the letter-number items are just as important as the other NP constituents, in some cases perhaps even more so. In *16S, ITS1-5.8S and ITS2 sequences*, for instance, deleting these combinations would have transformed the premodified NP into a bare nominal, that is, a simple, non-modified NP (*sequences*), since *16S, ITS1-5.8S and ITS2* would have been removed.

³⁵ Examples extracted from the research corpus, in which items made-up of alphanumeric associations were kept as they were. See the next section for further details.

Figure 3.5: foreign characters found in the pilot corpus after cleaning.

Rank	Freq	Word
30739	1	ögren
30740	1	öresund
30741	1	čandek
30742	1	đureje
30743	1	šlósarski
30744	1	şekerociođlu
30745	1	šivickis
30746	1	đod
30747	1	đr
30748	1	đrn
30749	1	θg
30750	1	θxz
30751	1	θyz
30752	1	κmo
30753	1	λ
30754	1	μ
30755	1	μem
30756	1	μmole
30757	1	υ
30758	1	ψl
30759	1	ψt
30760	1	waragonite
30761	1	wcalcite
30762	1	wvib
30763	1	cell

3.2.5 A possible solution

A straightforward solution for the issues previously mentioned would be simply to fix the problems identified by re-checking the corpus data. However, the very existence of these issues meant that the corpus and its compilation process could be improved, as it is often the case with most scientific endeavors. This seemed to me like an opportunity in disguise: by not fixing the issues and instead by looking for ways to improve corpus design and compilation I would be building on my (however small) experience in CL. More specifically, by working on the compilation tasks by myself, from beginning to end, I would be honing in on my skills as a corpus linguist. These improvements are presented in the next section of this chapter.

3.3 The research corpus

This section deals with the compilation of the research corpus, which was undertaken following the best practices in CL as recommended in Wynne (2005), particularly in Sinclair (2005) for compilation, Burnard (2005) for metadata, Leech (2004) for annotation, McEnery and Xiao (2005) for encoding, and Wynne (2004) for distribution. Such recommendations are supplemented by Bonelli (2010).

As is traditional in CL, the corpus required a name. I decided to name it the **Corpus of Biology Research Articles (CoBRA)**, which is part of the CoRA³⁶ – **Corpora of Research Articles**. The CoRA currently covers RAs in Education (CoERA), Linguistics (CoLRA), and Cultural Studies (CoCRA). It is still in development. Its main objective is to provide carefully selected linguistic data for cross-disciplinary analyses following the design criteria and compilation methodology described in this section.

3.3.1 Design criteria

3.3.1.1 Journal selection

The first step in re-imagining the corpus involved gaining more insight into the journal list and the greater area of Biological Sciences. This was done through a more rigorous look at the Biology journals and subareas, by learning more about each listed publication in terms of high impact indexes.

While the CAPES Qualis index may be a reliable measure for Brazilian publications, internationally-based indicators such as the Scimago Journal and Country Rank³⁷ (SJR) may be better suited to determine the influence of internationally produced journals.

It was by consulting the SJR page³⁸ for each of the 62 journals previously selected that I found more precise sub-area categorizations and learned about other relevant metrics such as *international collaboration* and *quartiles*. These metrics and indicators seemed particularly relevant for the research at hand and therefore guided a second curation of the journal list.

Also helpful were the clearly defined SJR areas/categories. The final and improved list thus contained only Q1 journals belonging to the following categories: *Ecology, Evolution, Behavior and Systematics, Aquatic Science, Animal Science and Zoology, and Plant Science*. Journal selection also had to display a minimum **overall** 40% of international collaboration³⁹. Including such indicators and metrics meant reducing the number of selected journals from 62 to 20.

³⁶ This corpus ‘family’ was largely inspired by Prof. Deise Dutra’s cross-disciplinary research work.

³⁷ <https://www.scimagojr.com/>.

³⁸ In addition to basic information (e.g., *country, subject area and category, and publisher*), the SJR informs the *International Collaboration* indicator for each publication from 1999 to the last available year. This means that it is possible to understand how much material was produced by researchers from more than one country. *Quartiles* ranks the quality of each publication according to its SJR index divided into four previously set groups, from 1999 onward, with the highest values placed at Q1. For each journal SJR index and additional details, please see: <https://www.scimagojr.com/SCImagoJournalRank.pdf>.

³⁹ Percentage reached by calculating the international collaboration mean of each of the 62 journals from 1999 to 2018, then by adding each journal mean and dividing the total sum by 62, rounded up to a decimal number.

3.3.1.2 Representativeness and balance

Always a thorny issue, representativeness, in this case, would ideally mean collecting RAs from more academic journals. I initially considered compiling about half the curated list, which would have totaled 10 journals or 500 articles. While ideal, such a plan would not have been very realistic for quality analysis in the timeframe of a master's thesis. A choice of 5 journals or 250 articles seemed sensible enough. Hence, 5 journals were elected to inform the research corpus, with RAs ranging from 2015 to 2019 (10 articles per publishing year).

Ecology, Evolution, Behavior and Systematics presented a larger number of journals (9 in total), which is why 3 of the 5 journals were picked from this category, namely: *Biological Conservation*, *Ecology Letters*, and *Oecologia*. *Animal Science and Zoology* had 6 journals and produced the following choices: *Animal Behaviour* and *Zoologica Scripta*. Due to a much lower amount of journals listed in *Aquatic Science* and *Plant Science*, 3 and 1 publications, respectively, these Biology sub-areas were not contemplated in the CoBRA⁴⁰. Table 3.1 below lists the journals selected, offering basic information about each journal.

Table 3.1: research corpus journal selection

Journal	Country	Intl. Collab.*	Publisher
<i>Animal Behaviour (AB)</i>	United States	43.80	Elsevier Inc.
<i>Biological Conservation (BC)</i>	Netherlands	52.76	Elsevier BV
<i>Ecology Letters (EL)</i>	United Kingdom	58.70	Blackwell Publishing Inc.
<i>Oecologia (OE)</i>	Germany	38.98	Springer Verlag
<i>Zoologica Scripta (ZS)</i>	United Kingdom	68.42	Blackwell Publishing Inc.

* 2018 stats.

3.3.1.3 Article format

Article collection was carried out in the first half of 2019. It involved examining each journal more closely, to better understand their main features, from author guidelines to RA sections. This examination proved fruitful since it was possible to verify that the typical RA pattern in biology-driven academic publications does not include the *Conclusions* section. The RAs procured for the research corpus, therefore, display the **(IMRaD)** pattern: **I**ntroduction,

⁴⁰ These decisions were made to achieve more representativeness and balance in regards to the list of academic journals provided by senior researchers from the Institute of Biological Sciences at UFMG. This thesis never set out to cover all the sub-areas contemplated in the journal list. The CoBRA, however, may be complemented in the future, given the ease of compilation brought about by the compilation device developed for this research, as shown in sub-section 2.3 of the present chapter.

Methods, Results, and Discussion, which has been traditionally understood as the dominant, prototypical structure of RAs⁴¹ (cf. SOLLACI; PEREIRA, 2004), particularly in STEM-based publications (cf. THOMPSON, 2011; WOLFE; BRITT; ALEXANDER, 2011; NAIR; NAIR, 2014; MOGULL, 2018).

3.3.1.4 Text length

After selecting RAs in the desired format, I checked the number of words in each text. This verification⁴² revealed that most IMRaD RAs in the 20 selected journals contain around 5,000-6,000 words, with tail ends ranging from \approx 2,200 words to \approx 12,000 words. From these numbers, I procured texts with a minimum of 3,500 words and a maximum of 7,500 words.

A Google Sheets document was created to record each article word count per journal and year. These records were supplemented by statistics collected for each journal, as seen in Figure 3.6. Mean, median, and range, with the largest and smallest values, were calculated automatically with an online mathematical calculator.

Figure 3.6: word count and other statistics for one of the journals included in the CoBRA.

Animal Behaviour						all RAs are open access.	
TOTAL		268,465					
	2015	2016	2017	2018	2019		
1	4781	4919	4257	6312	4686		
2	6943	5665	3843	6657	5229		
3	5310	3981	5610	4975	4727		
4	4992	4263	7062	5470	5463		
5	6309	5890	7275	4613	6055		
6	4387	6596	5683	5496	6002		
7	4964	5841	5373	4646	4705		
8	6137	5712	5295	5008	5631		
9	4405	4729	4658	5606	5706		
10	5768	6480	4603	4007	5740		
TOTAL	53996	54076	53659	52790	53944		
						Mean (Average)	5369.3
						Median	5418
						Range	3432
						Mode	All values appeared just once.
						Geometric Mean	5307.1405467041
						Largest	7275
						Smallest	3843
						Sum	268465
						Count	50

⁴¹ While the importance and rhetorical function of *Conclusions* is indeed not disputed in the literature (cf. NAIR; NAIR, 2014; THOMPSON, 2016), this section is generally viewed as an appendix of *Discussion*. Its purpose is to “help readers move *out* of the article”, which is why it tends to be brief, consisting of “one or two paragraphs focusing on specific aspects of the Discussion” (GLASMAN-DEAL, 2010, p. 154)

⁴² Text length varied substantially, with few RAs on the extremes of word count. Most texts rested in the middle, so I opted for a middle ground range calculated by the mode value of the RAs under consideration for the corpus (those exhibiting the IMRaD format), that is, by the value that occurs more often.

3.3.2 Compilation

3.3.2.1 The text-grabbing browser extension

A possible solution for the time-consuming process of manual corpus compilation was designing a text-grabbing browser extension to automate and streamline this task. Written by Russian software engineer Maksim Ustiantsev in Javascript⁴³, the *Article Text Grabber* (ATG) yielded two versions – ATG1, created and deployed as a pilot extension to retrieve data from *Zoologica Scripta*, the first journal selected for this task; and ATG2 – designed for the other journals.

Figure 3.7: Zoologica Scripta article webpage (table 1 highlighted in blue)

Material suitable for genetic analyses was obtained for 13 of the 40 *Labrundinia* species included in the morphological character matrix. The aligned CAD sequences were 899 bp long with 524 variable sites (58.3%), of which 424 (80.9%) were potentially parsimony informative. Most variable sites occurred in the third codon position (Table 1). The sequences were somewhat AT-biased, especially in third position, which exhibited a combined average AT composition of 64.0% (Table 1). No introns were recognized in CAD sequences of *Labrundinia*.

Table 1. Variable and informative sites and average nucleotide composition in the analysed

`table.table.article-section_table` 706 × 223

Nucleotide position	% Variable sites	% Informative sites	% Adenine	% Cytosine	% Guanine	% Thymine
1st	26.5	12.5	33.6	17.3	28.1	21.0
2nd	17.9	0.5	35.2	19.2	18.1	27.5
3rd	55.5	87.0	27.8	17.5	18.5	36.2
All	58.2	95.6	32.2	18.0	21.6	28.2

Highlighted in light blue is Table 1 of the ZS article and its source code. The extension reads this source code, which is written in Javascript (inserted and displayed on the landing page – HTML structure), and then overrides the element corresponding to Table 1, as based on the aforementioned criteria (i.e., the element is not supposed to be retrieved). Both versions run on Google Chrome in *developer mode* by retrieving information directly from each RA webpage, based on the following criteria: i) retrieve article sections only, ii) do not retrieve tables, figures, graphs, charts, acknowledgments, and reference list, and iii) maintain the article format (i.e., keep headings, spaces, etc. intact).

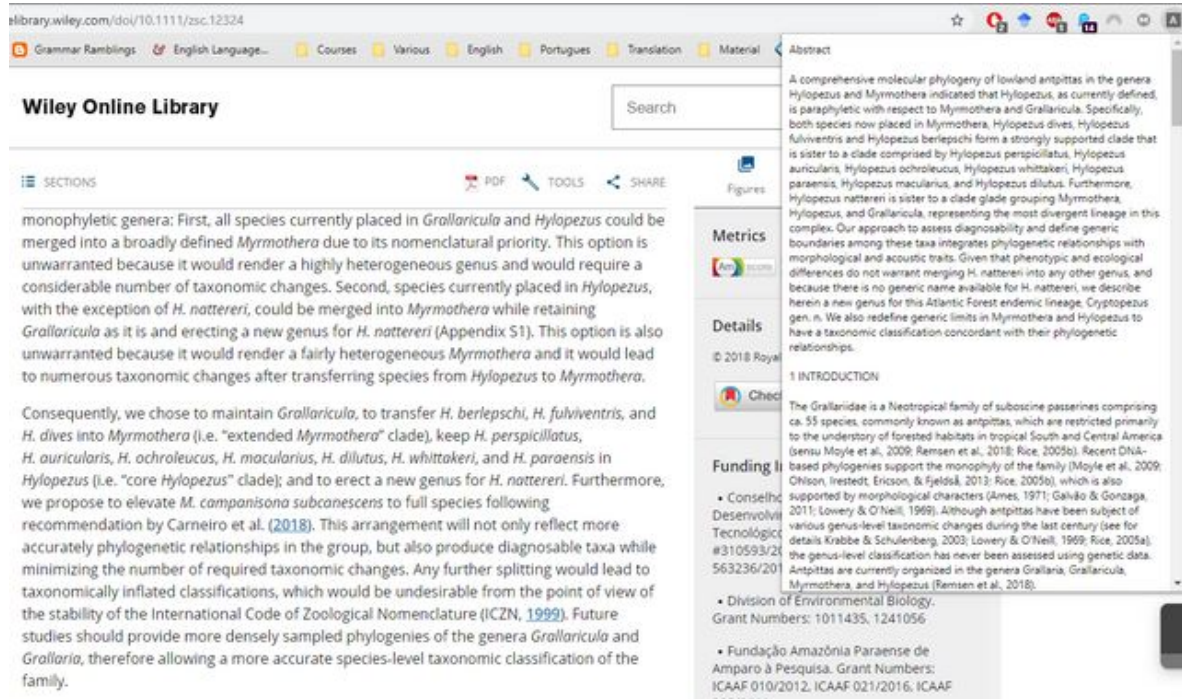
⁴³ <https://www.javascript.com/>.

3.3.2.2 Version 1

As seen in figures 3.7 and 3.8, the ATG1 reads the webpage source code and retrieves the desired text in its original format, retaining subtitles, paragraphs, and line spaces, in an extremely time-efficient fashion, since it “cleans” the text as it grabs it (e.g. it does not grab unnecessary information such as figures and reference lists, as mentioned previously).

The extension could have included a built-in save-to-text function, therefore saving more time. However, I opted for keeping it out⁴⁶, so that the retrieved information could be saved as plain text by *selecting all* (Ctrl-A), *copying* (Ctrl-C), *pasting* (Ctrl-V), which gave me the chance to check each retrieval individually for any inconsistencies between the original texts and the retrieved texts⁴⁷.

Figure 3.9: the ATG1 at play as a pop-up box.



⁴⁶ The development of this extension was a joint effort between Mr Ustiantsev and myself, a process made easier by his familiarity with CL and mine with (very) basic programming. The ongoing dialogue between programmer and linguist made it possible to find a simple and time-saving solution for an ultimately computation-based issue.

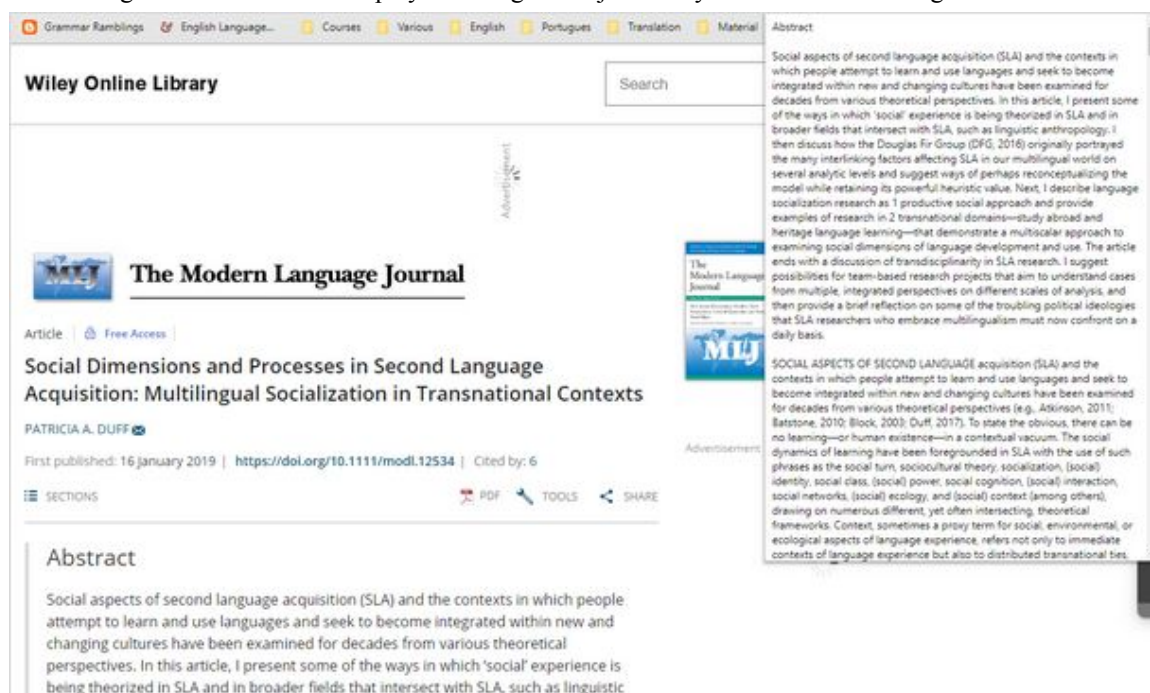
⁴⁷ Out of the 50 RAs extracted from *Zoologica Scripta*, only 1 retrieval presented problems. For some reason, the RA corresponding to file *ZS.04.2019* had not been fully retrieved and the three last paragraphs were missing. The issue was identified and immediately corrected during the *select all* step. The missing paragraphs were manually selected, copied, and pasted to the corresponding RA file on Notepad++.

In addition to saving an insurmountable amount of time⁴⁸, the ATG1 is very clean and user-friendly. It works similarly to the Google Scholar button. After installation, all it takes is the click of a button for the extension to work. The ATG1 displays the text grabbed in a clean, organized way, as a pop-up window (Figure 3.9), and does not require previous specialized technical skills. While the extension may not be economical to compile large corpora, for this master's thesis, it felt like an elegant, simple solution that required minimum manual efforts and produced practical and mostly error-free results.

3.3.2.3 Version 2

Based on the ATG1, a second version of the extension was developed to run on the remaining journals. Because the extension reads the landing page source code, open access RAs randomly selected from *Animal Behaviour*, *Biological Conservation*, *Ecology Letters*, and *Oecologia* were shared with Mr. Ustiantsev. As with the ATG1, the ATG2 was developed from these samples, which were used as models for the JavaScript parameters, following the same criteria established for the first extension.

Figure 3.10: the ATG2 at play in a Linguistics journal by Blackwell Publishing Inc.



⁴⁸ In an [Intel Pentium N3700](#), with a 10 giga-byte broadband connection, compiling one journal batch (10 texts) took approximately 60 minutes, from article selection to text retrieval with the extension, cleaning, total word count, and text storage. This means that total corpus compilation time took an estimated 25-30 hours.

Although on the surface it may seem otherwise, the aforementioned academic journals display very similar HTML layouts, which facilitated the process of developing the extension and lead us to create a single extension to be used on all journals, as opposed to devising one extension for each journal.

This was the most time-efficient decision for both parties involved. For me, the young linguist compiling the corpus, it is practical because with a single device any open access RA can be retrieved from its landing page, with minimum cleaning required. For Mr. Ustianstvev, the software engineer, it is efficient because no iterations needed to be made⁴⁹.

Another advantage of the ATG2 is that it retrieves information from **any open access** RA hosted on **any Elsevier BV, Elsevier Inc., Blackwell Publishing Inc., and Springer Verlag** journal. This means that the ATG2 works not only on the aforementioned journals, but on **any journal whose webpage is housed at these publishers' online platforms**: *ScienceDirect* for Elsevier, under <https://www.sciencedirect.com>, the *Wiley Online Library* for Blackwell, under <https://onlinelibrary.wiley.com>, and *SpringerLink* under <https://link.springer.com>, for Springer Verlag. Figure 3.10 illustrates a free access RA captured from the Modern Language Journal.

3.3.2.4 Cleaning

Because the extensions grabbed only previously defined information, leaving out the unnecessary elements (tables, acknowledgments, reference list, etc.), very little cleaning was required. The major cleaning undertaken refers to the bulleted section *highlights* displayed in the Elsevier journals (*Animal Behaviour* and *Biological Conservation*). This was carried out manually immediately after pasting the grabbed text onto a Notepad++ file.

The cleaning process could have included the removal of the references used in in-text citations (e.g., name + year of publication). However, these references are a rather important part of the RAs, as they provide credibility to the research reported in the texts, which would strongly justify keeping⁵⁰ them in the corpus.

⁴⁹ While no iteration was needed, iterations can be made to improve the ATG and/or to adjust it to specific needs, e.g., retrieving RAs from journals housed by other publishers.

⁵⁰ Differently, figures, tables, graphs, and charts either incorporate or are anchored on other semiotic modes (cf. HALLIDAY, 1978, KRESS; van LEEUWEN, 2001; KRESS, 2010), not just the (written) verbal mode, which is the case of references in in-text citations. This decision was also directed by the fact that in-text citations often refer to people, organizations, processes, and places, and are denoted by proper names, which means that NPs containing proper names can be parsed via Named Entity Recognition (NER). NER extraction can indicate the number of NPs referring to real world-world entities, as well as the categories of these entities

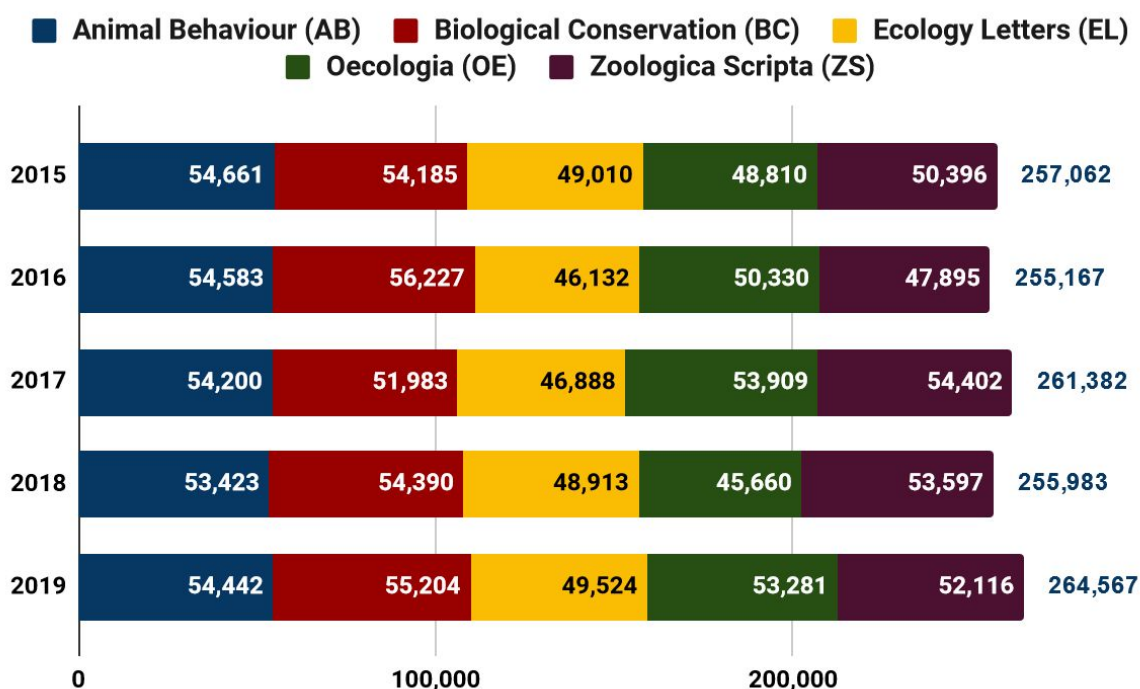
3.3.3 Corpus counts

Table 3.2 displays the token counts for each journal in the CoBRA, according to the year of publication. It indicates the total number of tokens in the corpus: 1,294,161 tokens. In spite of being a small corpus, if compared to the large corpora currently available for research at Sketch Engine, for instance, the CoBRA is large enough and appropriate for the purposes of the research intended in this thesis.

Table 3.2: journal word counts and corpus total count

<i>Journal</i>	2015	2016	2017	2018	2019	Total
<i>Animal Behaviour (AB)</i>	54,661	54,583	54,200	53,423	54,442	271,309
<i>Biological Conservation (BC)</i>	54,185	56,227	51,983	54,390	55,204	271,989
<i>Ecology Letters (EL)</i>	49,010	46,132	46,888	48,913	49,524	240,467
<i>Oecologia (OE)</i>	48,810	50,330	53,909	45,660	53,281	251,990
<i>Zoologica Scripta (ZS)</i> ⁵¹	50,396	47,895	54,402	53,597	52,116	258,406
Total						1,294,161

Graph 3.1: journal word counts and corpus total count



⁵¹ The RAs selected from this journal do not overlap with those from the CorABio. This means that both corpora may be used in cross-reference studies.

Corpus data visualization was performed with three different software: Sketch Engine (KILGARRIFF et al, 2014, 2018), AntConc (ANTHONY, 2019), and Unitex/GramLab IDE 3.1 (cf. MARTINEZ, 2020). Retrieved from AntConc, Figure 3.11 shows a list of the most frequent tokens in the CoBRA. Expectedly, functional words make up the most frequent words in the corpus, in absolute frequencies, a sign of Zipf’s Law (2016 [1949]), as discussed in the following chapter. With a total of 1,229,918 tokens⁵² distributed over 36,588 items, the CoBRA displays a type-token ratio of c. 33.7%, indicating high lexical density.

Figure 3.11: the twenty-five most frequent words in the CoBRA.

Word Types: 36588		Word Tokens: 1229918
Rank	Freq	Word
1	63256	the
2	42565	of
3	39549	and
4	29649	in
5	24618	to
6	19812	a
7	13223	al
8	13216	et
9	12801	for
10	10842	that
11	10012	with
12	9105	were
13	8390	species
14	8141	as
15	8006	is
16	7937	we
17	7507	on
18	7283	from
19	7226	by
20	7092	was
21	5996	are
22	5446	this
23	4826	at
24	4589	be
25	4553	or

⁵² Notepad++ lists the CoBRA with 1,294,161. Slight differences in the number of tokens occur due to software heuristics. Total token count varies at .05 between AntConc and Notepad++.

As a grammar-based NLP tool, the Unitex/GramLab IDE 3.1 provides insightful input to corpus data. It automatically reads each file and generates outputs on number of sentences, tokens, and items per corpus file, simple and compound words as well as unknown words (i.e., words the program does not recognize as being part of its input data, such as foreign words). Figures 3.12 and 3.13 are examples.

Figure 3.12: data visualization with Unitex/GramLab IDE 3.1.

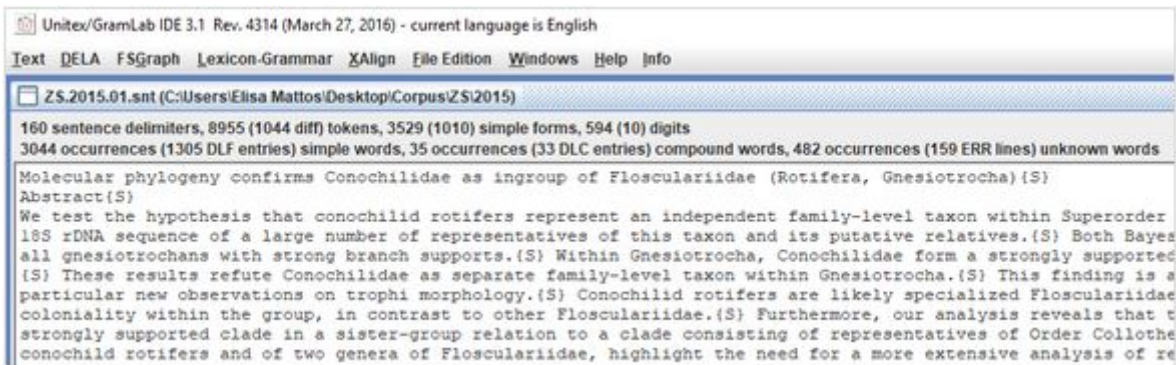
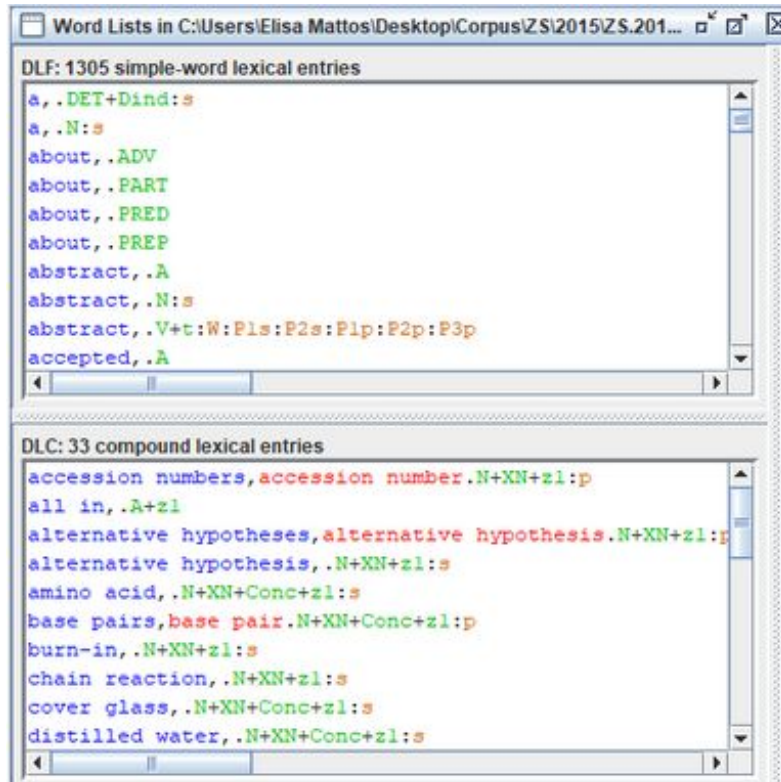


Figure 3.13: compound visualization with Unitex/GramLab IDE 3.1.



3.3.4 Metadata

The metadata was organized in four different levels, saved in individual spreadsheets created with Google Sheets and in plain text to facilitate portability and annotation. This task was carried out in the first half of 2019, but it was conceptualized in the second half of 2018. The first level of metadata concerns the corpus main features, including type, sociolinguistic information (e.g. diatopic variation), total word count, and count per main word classes⁵³ (nouns, verbs, adjectives, adverbs), as well as the number of sentences and paragraphs. The counts are usually recorded during *data processing* and were retrieved with Sketch Engine.

The second level of metadata is focused on the selected journals. It displays thorough, comprehensive sets of information organized in five categories: a) **journal**: ISSN, number of publications to date (volumes and/or issues), and publishing company (name and country); b) **publications**: types (genres contemplated in the journal), identification (name carried)⁵⁴, and minimum/maximum number of words (for RAs only); c) **language**: variety, editing services, d) **metrics**: SJR impact factor, indexation; and e) **legal**: copyright, including information about open access. Such information was collected during *discovery*⁵⁵.

The third level regards the RAs and includes linguistic and non-linguistic information on each article. Linguistic information refers to total word count, count per main word classes (nouns, verbs, adjectives, adverbs, etc.), and the number of sentences and paragraphs. On the other hand, non-linguistic information pertains to the number of authors who penned the RA, as well as their respective academic institutions.

The fourth level of metadata relates to any treatment the corpus may have been subject to, from POS tagging to other types of annotation. This level includes information on who did what to the corpus, in the case of manual annotation and/or cleaning. This decision was made following standards seen in corpora such as the Nordic Dialect Corpus and Syntax Database (LINDSTAD et al., 2009) and the C-ORAL-ROM family (CRESTI; MONEGLIA, 2005), in particular the C-ORAL BRASIL (RASO; MELLO, 2012).

⁵³ As seen on reference corpora such as the C-ORAL BRASIL I (RASO; MELLO, 2012).

⁵⁴ This refers to the identification of the RAs in each journal. For example, in *Ecology Letters*, RAs are listed as *Letters*, in what must have been a stylistic choice. The texts are indeed original research articles, as mentioned in the *About* section of the journal: “The journal publishes concise **papers** that merit urgent publication by virtue of their originality [...] Three types of **article** are published [...]: ***Letters: exciting findings in fast-moving areas**; *Ideas and Perspectives: novel essays for a general audience; *Reviews and Syntheses: syntheses of important subjects meriting urgent coverage.” (bolds not present in the original).

⁵⁵ While examining the publications to confirm their suitability for the corpus.

3.4 Data extraction

This section deals with the methodology behind the automated NP extraction. The NPs extracted served as the main data for the analysis developed in chapter four. The focus of the NP extraction was on premodifiers, that is, words subordinated to a head noun. Two attempts at extracting the NPs from the CoBRA were made. This section describes the first one and the final section of this chapter addresses the second.

3.4.1 NP extraction

NLP is mostly guided by two formal grammatical theories: constituency grammar and dependency grammar. Stanford Core NLP (MANNING et al, 2014) provides data extraction modules shaped by these grammar models. The Stanford Core NLP (MANNING et al, 2014) module used for NP chunking in this thesis is based on constituency grammar, that is, the idea that languages are governed by sentences composed of different constituents, as stated in chapter two.

A similar method based on Queiroz (2019) was devised by the same programmer, Mr. Euller Borges, who worked on this authors' thesis and noun-headed phrases were extracted in Python, with NLTK and the Stanford Core NLP, which, as an NLP tool, condenses many tasks (tokenizing, POS-tagging) often undertaken separately in Python. This points to a more mixed programming-language approach: since Stanford Core NLP is written/designed for Javascript, its use with Python, a high-level programming language different from Java, required using a wrapper⁵⁶.

3.4.2 Data output

The extracted data were saved in comma-separated value (CSV) files that can be better visualized in spreadsheet format. Figure 3.14 showcases a simple Google Sheets spreadsheet, in which the first NP extracted from file *ZS.2015.08.txt* is shown. NP chunks are accompanied by the following information: *corpus file*, *sentence number*, *NP phrase tag sequence*, *NP POS tag sequence*, *NP word sequence*, and *number of words per NP*. As illustrated in figure 3.14 as follows, the NP extracted is displayed in constituency terms, that is, driven by constituency grammar, with constituents nested between parenthesis and arranged hierarchically.

⁵⁶ For more information, please see: <https://stanfordnlp.github.io/CoreNLP/other-languages.html>.

Figure 3.14: data output

A sisters ' story comparative phylogeography and taxonomy of Hierophis viridiflavus and H. gemonensis (Serpentes Colubridae)							
A	B	C	D	E	F	G	H
Corpus File	Sentence #	NP extracted from parser	NP (phrase tag sequer	NP (POS tag sequen	NP (word sequen	Number of words per NP	
ZS.2015.08.txt	1	(NP (NP (NP (DT A) (NNS sisters) (POS ')) (NN story)) (:) (NP (NP (NP (NP (JJ comparative) (NN phylogeography) (CC and) (NN taxonomy)) (PP (IN of) (NP (NP (NNP Hierophis) (NN viridiflavus) (CC and) (NN H.) (NN gemonensis)) (PRN (-LRB- -LRB-) (NP (NNP Serpentes)) (,) (NP (NNP Colubridae) (-RRB- -RRB-))))))	NP NP NP NP NP NP DT NNS NN JJ NN C	A sisters ' story co	15		

Because NPs can often be found as complements of prepositions, that is as part of PPs – Prepositional Phrases, a second script was run with the first script, so that NPs found within PPs could be extracted separately, generating a second output of extracted data. The second output was also saved in CVS, as done with all other data extraction files.

Figure 3.15 below shows three PPs from the CoBRA file *ZS.2015.08.txt*, accompanied by the same type of information found in the NP spreadsheet (*corpus file, sentence number, phrase tag sequence, POS tag sequence, word sequence, number of words per PP*). Following the standard process of the parser, all PPs are nested, with their constituents hierarchically arranged.

Figure 3.15: PP output

the other hand most genetic data regarding H. gemonensis							
A	B	C	D	E	F	G	
Corpus File	Sentence #	PP extracted from parser	Phrase tag sequence	POS tag sequence	Word sequence	Number of words/PP	
ZS.2015.08.txt	20	(PP (IN On) (NP (NP (DT the) (JJ other) (NN hand)) (,) (NP (NP (JJS most) (JJ genetic) (NNS data)) (PP (VBG regarding) (NP (NNP H.) (NNS gemonensis))))))	NP NP NP NP PP NP	DT JJ NN JJS JJ NNS V	the other hand m	9	
ZS.2015.08.txt	26	(PP (IN among) (NP (NNS reptiles)))	NP	NNS	reptiles	1	
ZS.2015.08.txt	28	(PP (IN in) (NP (DT all)))	NP	DT	all	1	

3.4.2 Initial NP categorization

The extracted NPs were first organized by type and size, partially following Berlage's (2014) study and categorization level of NP complexity. A spreadsheet was designed to house the extracted NPs, which were manually entered into five individual tabs, each for a different type of NP realization, organized following the assumption that independent NPs will show a maximum of 5 items. This assumption was made based on a quick read of the NP patterns in the data output files. NP categorization was broken down as follows:

- **NP 1-5:** independent NPs – not hierarchically dependent on other phrases, with columns 1-5 referring to the number of items in the NPs;
- **PP NP:** NPs hierarchically connected to PPs;
- **NP COORD:** NPs with coordinated heads;
- **NP VP:** NPs that contain VPs;
- **NP THAT:** NPs immediately followed by a 'that' clause.

Figure 3.16: tabulated NP categorization

NPs with...	2 items	3 items	4 items	
1 item				
NP 1	NP 2	NP 3	NP 4	NP 5
account	Our data	the existing divergence	a relaxed clock model	
place	Specific names	the major cladogenesis	a typical parapatric distribution	
Italy	the Messinian	their ecological features	a comparative multidisciplinary study	
they	the Quaternary	a single species	the major dextral clades	
we	our results	the two species		
it	this allele	a landmark-based approach		
Llorente	the Balkans	several taxonomic revisions		
species	phylogeographic studies	a monotypic species		
we	Hierophis vindiflavus	different evolutionary lineages		
it	life-history differences	a population genealogy		

☰ NP 1-5 ▾ PP NP ▾ NP COORD. ▾ NP VP ▾ NP THAT ▾

NPs were categorized per output file, that is, per batch of ten files for each publishing year of each academic journal covered in the CoBRA. Figure 3.16 exemplifies how the NPs were organized in categories in separate spreadsheets. It refers to the RAs penned in 2015 and compiled from *Zoologica Scripta*. Data output organization and categorization were carried out manually so that the extracted NPs could be individually verified.

3.4.3 Issues

The same procedures undertaken in Queiroz (2019) were enlisted to extract the NPs. However, because Queiroz's research dealt with a corpus of much smaller size (51,187 words, or \approx 5% of the research corpus compiled for this thesis), with shorter sentences, issues were likely to arise during data processing, specifically with parsing.

Given the robustness of the CoBRA, I asked Mr. Borges to extract the data per batch of 10 RAs at a time, instead of retrieving the NPs from the entire corpus at once. Still, even with a much lower number of words, the parser required some adjustments in its parameters, in order to successfully process, analyze, and extract longer strings of data.

NP extraction with Stanford Core NLP is an automated process, with minimal human intervention. The NPs from the CoBRA, however, seem to be requiring more human power than ideally needed. Not only was it necessary to tweak the script to accommodate the corpus data, but some NPs took much longer to be parsed, considerably more than usually expected. More importantly, the output showed parsing issues. Examples 1 and 2 reflect those issues.

Example 1

Introduction Speciation mechanisms and intraspecific differentiation processes

(NP (NP (NN Introduction) (NN Speciation) (NNS mechanisms)) (CC and) (NP (JJ intraspecific) (NN differentiation) (NNS processes)))

Example 2

The power to link evolutionary events and the relative taxonomic implications to the geological history, geomorphologic features and climate changes of the study area

(NP (DT the) (NN power) (S (VP (TO to) (VP (VB link) (NP (NP (JJ evolutionary) (NNS events)) (CC and) (NP (DT the) (JJ relative) (JJ taxonomic) (NNS implications)))) (PP (TO to) (NP (NP (DT the) (JJ geological) (NN history)) (, .) (NP (JJ geomorphologic) (NNS features)) (CC and) (NP (NP (NN climate) (NNS changes)) (PP (IN of) (NP (NP (DT the) (NN study) (NN area))

Examples 1 and 2 represent an inconsistent and unreliable output (see Attachment D for more). We can see that while some NPs are correctly extracted (example 1), others are not. In Example 2, the entire period is labeled and extracted as an NP. Such issues pointed to the need for a solution to properly handle the size of the NPs. In the next subsection, this solution is discussed.

3.5 NP chunking revisited

For a second and final attempt to extract the NPs, an NLP Ph.D. candidate from the university was enlisted to perform the NP chunking task exclusively on Stanford Core NLP. This meant not using NLTK in Python. After discussing with Ms. Evelyn Amorim the output needed, we reached a decision as to how the data should be formally and computationally handled to yield the desired output⁵⁷.

Decisions as to how the output should be presented were also made differently. Firstly, instead of saving the files in CSV format (comma-separated value), the output was displayed in TSV (tab-separated value). This solved some of the reported issues. In particular, it allowed the output to be organized in tabs. Figure 3.17 displays the output from this second attempt at NP chunking, as seen on the following page.

Another decision that had to be taken refers to whether the NPs should be nested. As we noticed in the previous extraction, complications arose when long NPs had to be extracted. The ideal solution would have been to devise an algorithm. Given the time frame of this thesis, such an alternative would not have been attainable⁵⁸. A viable solution to this issue was to extract the NPs un-nested. This meant that all NPs would be retrieved in a flat manner, without being hierarchically linked to other phrasal nodes.

For instance, an NP such as *the cold-adapted Anatolio-Balkan genus* produced in this sentence: (...) *we evaluate the genealogical history of the cold-adapted Anatolio-Balkan genus Anterastes especially to test the possible effects* (...) would be automatically retrieved from the CoBRA without any connection to the prepositional head that hierarchically governs this NP in this particular context.

Not nesting the NPs would then require looking at the context for confirmation of any hierarchical links. Although this may seem time-consuming or inconvenient, it actually turned out to be productive. One of the main advantages of this solution is that by extracting all NPs in this way, a more comprehensive landscape of NP use in the corpus was produced. Whether or not they are part of other phrases, such as PPs, the NPs extracted reflect the dimension with which this corpus employs constructions of nominal nature.

⁵⁷ Other extractions were made, namely NER and PPs. These, however, are not the focus of the present research and thus will not be discussed in this chapter. It should be noted that familiarity with Linguistics played a very important role in the second attempt. Ms Amorim is a PhD candidate in NLP with a focus on Lexical Semantics. She is co-advised by Dr. Marcia Cançado, from the same graduate program where this thesis was developed.

⁵⁸ This would require training a machine and making improvements to the algorithm, which, at this stage, would also require specialized assistance from NLP professionals.

Choosing to focus the NP chunking on Stanford Core NLP made a large difference in data processing and output. It seems as though many of the previous issues stemmed from the NLTK library and the type of parsing selected. In NLP, NP chunking is understood as shallow or partial parsing and requires IOB tagging, a machine learning-based approach (JURAFSKY; MARTIN, 2019). A training set was created and tested at Stanford Core NLP. The output was satisfactory and consistent, at a +90% accuracy rate.

Figure 3.17: data extraction output

	A	B	C	D	E
1	Corpus file	NP	NP (POS tag set	NP (word sequence)	Number of words
2	../data/entrada/E	(NP (NN apple)	(NP NN CC NN N	apple and hawthorn flies	4
3	../data/entrada/E	(NP (DT This) (JJ	NP DT JJ JJ NN	This striking genome-wide similarity	4
4	../data/entrada/E	(NP (ADJP (JJ e;	NP ADJP JJ CC	experimental and natural populations	4
5	../data/entrada/E	(NP (NNP Noor)	NP NN CC NP N	Noor & Feder 2006	4
6	../data/entrada/E	(NP (NNP Barret	NP NN CC NP N	Barrett & Schluter 2008	4
7	../data/entrada/E	(NP (NNP Noor)	NP NN CC NN C	Noor & Feder 2006	4
8	../data/entrada/E	(NP (NNP Noor)	NP NN CC NP N	Noor & Feder 2006	4
9	../data/entrada/E	(NP (DT the) (NP	NP DT NP NN C	the direction and magnitude	4
10	../data/entrada/E	(NP (DT the) (NN	NP DT NN NN NI	the fly <i>Rhagoletis pomonella</i>	4
11	../data/entrada/E	(NP (NNP Coyne	NP NN CC NN C	Coyne & Orr 2004	4
12	../data/entrada/E	(NP (JJ numerou	NP JJ ADJP RB	numerous geographically overlapping taxa	4
13	../data/entrada/E	(NP (JJ many) (J	NP JJ JJ NN NN	many new host plants	4
14	../data/entrada/E	(NP (JJ several) (NP JJ JJ NN NN	several different plant families	4
15	../data/entrada/E	(NP (DT The) (AC	NP DT ADJP RB	The most recent example	4
16	../data/entrada/E	(NP (PRP\$ its) (NP PRP JJ NN N	its native host hawthorn	4
17	../data/entrada/E	(NP (DT the) (JJ	NP DT JJ NN NN	the eastern United States	4
18	../data/entrada/E	(NP (JJ Genetic)	NP JJ CC NN NN	Genetic and field studies	4
19	../data/entrada/E	(NP (DT the) (VB	NP DT VB JJ NN	the hypothesised initial stage	4
20	../data/entrada/E	(NP (DT a) (JJ gr	NP DT JJ NN NN	a gross migration rate	4
21	../data/entrada/E	(NP (DT a) (JJ fa	NP DT JJ JJ NN	a facultative pupal diapause	4
22	../data/entrada/E	(NP (DT The) (JJ	NP DT JJ JJ NN	The earlier fruiting time	4
23	../data/entrada/E	(NP (DT the) (JJ	NP DT JJ NN NN	the ancestral hawthorn race	4
24	../data/entrada/E	(NP (DT the) (JJ	NP DT JJ NN NN	the sympatric host races	4
25	../data/entrada/E	(NP (DT the) (JJ	NP DT JJ JJ NN	the total genome-wide impact	4
26	../data/entrada/E	(NP (DT the) (NN	NP DT NN NN NI	the within-generation selection experiment	4

4. RESULTS AND DISCUSSION

4.1 Overview

This chapter is concerned with the results obtained from the automated NP extraction, which generated a total of 240,801 NP chunks⁵⁹. More specifically, this chapter focuses on the 72,877 deep-cleaned NPs **initiated by content words only, which may include hyphenated compounds**, such as *one-gigabyte compact flash cards*⁶⁰. The NPs selected for analysis range from two to five items, by focusing on **NPs that contain hyphenated premodifiers**, such as *pollinator-friendly management practices* and *host-related divergence*.

In an exploratory fashion, this chapter reports and discusses the frequency, behavior, and distribution of the NP chunks retrieved by looking into **the use of hyphenated elements in premodification**, since this thesis is mainly focused on **premodification in complex NPs in scientific writing**⁶¹. The sets analyzed have **approximately 5,800 complex NPs**⁶² **initiated by content words, with at least one hyphenated item** (e.g., *spatially-explicit generalised additive models*, *pike-like carnivores*).

The NPs carrying hyphenated premodifiers⁶³ were selected from wordlists of content word-initiated NP sets generated in both Sketch Engine (KILGARIFF et al., 2010, 2014) and Notepad++ (HO, 2019). As explained in chapter three, in order to avoid overlaps, the final list of extracted NPs is un-nested.

This means that the extracted NPs may be embedded within VPs or PPs. This was a methodological decision taken during the NP extraction as a way to better comprehend the full extent of the NPs in the corpus. **Examples may thus include NPs that, in their original context, are part of other phrases.**

⁵⁹ *NP chunking* is the process of automatically extracting NPs via NLP tools (cf. JURAFSKY; MARTIN, 2019). Chunking involves shallow parsing by design, meaning that all NP chunks are un-nested.

⁶⁰ Hyphenated compounds may be initiated by functional words. Computationally-wise, hyphenated compounds are read as a single unit, which is why no difference is made between content and functional words in these NPs.

⁶¹ As a punctuation device, hyphenation is characteristic of writing and may be signalled by prosody and stress in speech, as far as compounds are considered - hyphenated or not (cf. HUDDLESTON; PULLUM, 2002).

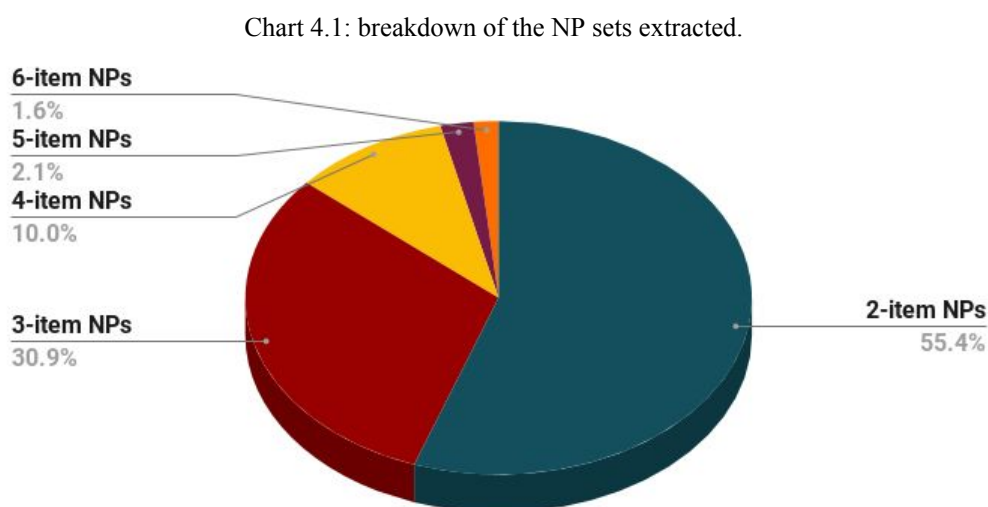
⁶² A distinction is made between hyphenated compounds (*genome-wide consequences*) and words with hyphens and prefixes (*semi-natural conditions*) only as far as the present analysis is concerned. Frequency-wise, the 5,800 NPs contain hyphenated compounds and words with hyphens and prefixes.

⁶³ The primary reason for examining hyphenated premodifiers in complex NPs stems from the fact that in writing hyphenation is a device meant to be used strategically to aid comprehension. Hence, a deeper look at the use of hyphenated premodifiers seemed fitting, especially given their recurrent employment in the CoBRA.

4.2 NP chunks

4.2.1 Overall counts

240,801 NPs were automatically extracted via Stanford Core NLP (MANNING et al, 2014). As expected, 2-item NPs are more frequent, accounting for 55.4% of the extracted NP chunks, with 133,410 NPs. In contrast, longer NPs are gradually less recurrent: the 3-item set comprises 74,353 NPs, 30.9% of the extractions. The 4-item NP set covers 10%, with 24,099 NPs, the 5-item set has 5,095 NPs, 2.1% of the chunks, and the +6-item set comprises 3,822 NPs, or 1.6% of the NP extractions. Pie chart 4.1⁶⁴ illustrates these percentage breakdowns.



4.2.2 Extracted and cleaned NPs

A first cleaning was required to remove a few errors and several undesired elements. Despite having been seemingly tagged correctly as NPs via Stanford Core NLP (MANNING et al., 2014), nominal arrangements such as *parameter D*, *the genus Nostoc spp.*, and *Fig. 1b* are not typical examples of pre-modified NPs. In fact, these NPs are comprised of two heads, but this software reads these strings as a single NP since it cannot process context-sensitive information effectively⁶⁵.

⁶⁴ Unless otherwise stated, tables, graphs/charts displayed in this chapter are my own creation. Pie charts include percentages. Whenever attainable, bar charts accompany a goodness-of-fit trendline with frequency R^2 calculated.

⁶⁵ This is the standard computation process for context-free grammars (Phrase/Constituency Grammar), which are “based on a purely declarative formalism” and therefore do not “specify *how* the parse tree for a given sentence should be computed”, as explained in Jurafsky and Martin (2019, para. 2, italics in the original). This means that the corpus to be parsed should be as devoid of errors as possible, seeing that sentences riddled with grammatical errors may not be parsed (cf. JURAFSKY; MARTIN, 2019). For learner and spoken corpora, issues

The same rationale can be applied to highly pre-modified NPs such as *the neotropical direct-developing genus Eleutherodactylus* (Example 4.1), which actually contains two heads: *the neotropical direct-developing genus* (N-head1 highlighted in gray) and *Eleutherodactylus* (N-head2 highlighted in yellow), respectively. In constituency-based parsing, however, such heads are not read separately, as shown in the parse structure following Example 4.1.

Example 4.1

→ (...) such as the cement gland and a coiled gut, which are absent in **the neotropical direct-developing genus Eleutherodactylus** (...)

(CoBRA file ZS.2015.02.txt)

Constituency-based parsing:

```
(ROOT
  (NP
    (NP (DT the) (NN neotropical))
    (NP (JJ direct-developing) (NNS genus))
    (NP (NNP Eleutherodactylus))))
```

As seen in this representation, the Stanford Core NLP (MANNING et al., 2014) constituency-based parser reads *the neotropical direct-developing genus Eleutherodactylus* as a single NP, with *Eleutherodactylus* as the head noun. In a dependency-based parser, however, *Eleutherodactylus* is no longer the only head. Figure 4.1 showcases the dependency structure⁶⁶ for *the neotropical direct-developing genus Eleutherodactylus*, in which both *genus*, as well as *Eleutherodactylus*, are heads, with *Eleutherodactylus* classified as an appositional modifier⁶⁷, <np-close>⁶⁸, and as the <nhead> subsequent to the *genus* <nhead>.

may arise, which is why such types of corpora need to be cleaned and/or annotated for errors and disfluencies, so that they resemble written, error-free texts that can be properly computationally processed. In general and specialized written corpora, texts often undergo numerous edits and grammar checks pre-publication, therefore they tend not to present major grammatical errors that may prevent successful parsing.

⁶⁶ Created with Bick’s (2020) English VISL (Visual Interactive Syntax Learning), with dependency rules based on Karlsson et al.’s (1995) Constraint Grammar (CGr), a methodological paradigm in Computational Linguistics, in which context-dependent rules are written by linguists to be compiled into a grammar that generates tags such as derivation, syntactic function, and dependency, among others. As Bick and Didriksen (2015, p. 31) explain, CG can provide “a framework for expressing contextual linguistic constraints allowing the grammarians to assign or disambiguate token-based, morphosyntactic readings”. It is therefore different from context-free grammars.

⁶⁷ From Zeman et al (2019): “An appositional modifier of an NP is an NP immediately to the right of the first NP that serves to define or modify that NP.”, as based on de Marneffe et al. (2006, 2013, 2014), de Marneffe and Manning (2008), Petrov et al. (2012), and Zeman (2008).

⁶⁸ <np-close> is a secondary attachment marker devised by Bick (2005, 2008, 2009), and applied to post-nominal modifiers. This dependency is also signalled by the tag @N<, which assigns to *Eleutherodactylus* the syntactic function of post-nominal dependent (BICK, 2010).

Figure 4.1: dependency structure for *the neotropical direct-developing genus Eleutherodactylus*.

```
</β>
the [the] <def> ART S/P @>N #1->4
neotropical [neotropical] <jbio> <DA:neotropisk?> <DL:bio> ADJ POS @>N #2->4
direct-developing [direct-developing] <pcp1> <heur> <DL:bio> ADJ POS @>N #3->4
genus [genus] <DA:genus> <ac-cat> <second> <DL:bio> <def> <nhead> N S NOM @NPHR #4->0
Eleutherodactylus [Eleutherodactylus] <*> <Aamph> <asisprop> <DL:bio> <def> <np-close> <nhead> N S NOM @N< #5->4
</β>
```

A considerable number of similar NPs were either deleted or relocated. For instance, *the cold-adapted Anatolio-Balkan genus Anterastes* had *Anterastes* deleted, thus making it a 4-item NP, instead of a 5-item NP. In Example 4.2, we can see that this string consists of two heads: *the cold-adapted Anatolio-Balkan genus* is the first and *Anterastes* is the second. Using the word *genus* followed by the genus' scientific name is fairly standard in Biology scientific writing, as evidenced by the high amount of discarded NPs showing this pattern.

Example 4.2

→ Here, we evaluate the genealogical history of **the cold-adapted Anatolio-Balkan genus Anterastes** especially to test the possible effects (...)

(CoBRA file ZS.2015.03.txt)

This cleaning also removed NPs containing in-text author references (*Barton 1983*), symbols (*50+ times*), and foreign names (*Corporación Regional Autónoma de Caldas*), even when those were initiated and/or coordinated by a word in English, as in *the Università degli Studi di Napoli Federico II* or in *Colobus guereza and Theropithecus gelada*. This was carried out semi-automatically in Notepad++ (HO, 2019) or Python whenever possible. Sentences are shown in Examples 4.3 to 4.7, for each NP set, as retrieved from the CoBRA.

Example 4.3

→ (...), however, behaviours are viewed as **correlated traits** that can generate trade-offs (...)

(CoBRA file OE.2016.08.txt)

Example 4.4

→ We thus used it as another index of **male mate preference** in addition to the usual mate-copying index which is based on copulation (...)

(CoBRA file AB.2018.07.txt)

Example 4.5

→ In **coastal mangrove nesting habitats**, Recovery Project staff monitor YSBL nests in natural substrates (...)

(CoBRA file BC.2016.06.txt)

Example 4.6

→ **The normalised relative Dipteracin expression** and translation activity are ratios and were therefore log-transformed before analysis.

(CoBRA file EL.2015.02.txt)

Example 4.7

→ We also use genome skimming to recover complete or near-complete sequences of **nuclear 18S and 28S RNA genes** (...)

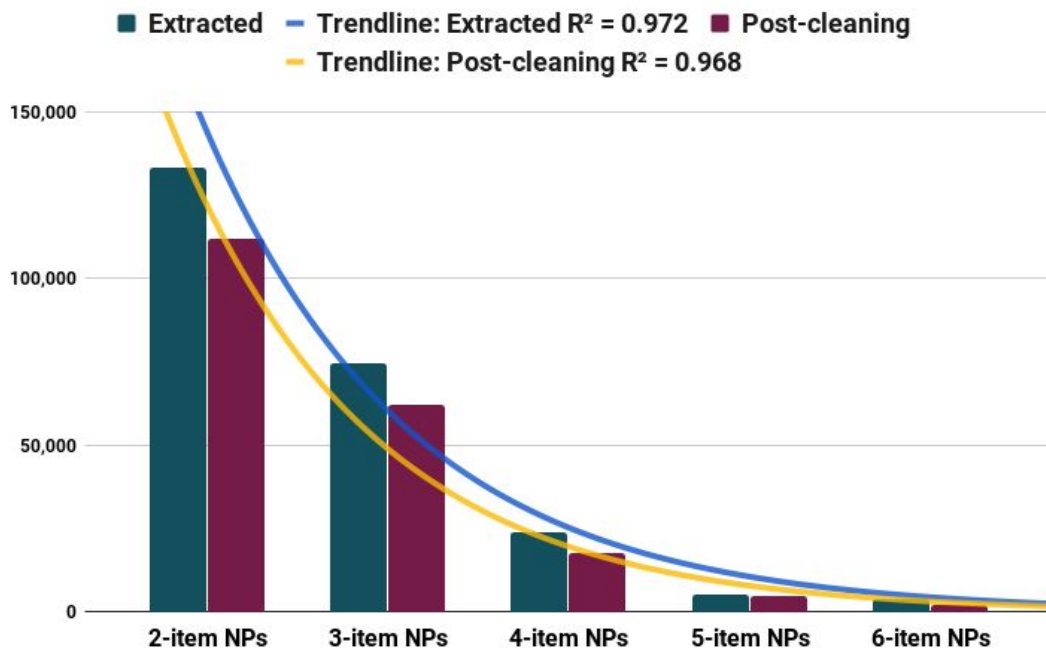
(CoBRA file ZS.2017.07.txt)

Table 4.1 shows the number of NPs extracted per set (2-item, 3-item NPs, etc.) and the amount of NPs post-cleaning. Bar chart 4.2 below displays the figures from Table 4.1, with an automatically-calculated trendline for

Table 4.1: number of NPs extracted and cleaned.

	2-item NPs	3-item NPs	4-item NPs	5-item NPs	+6-item NPs	Total
Extracted	133,410	74,353	24,099	5,095	3,822	240,779
Post-cleaning	112,072	61,980	17,643	4,553	1,980	198,228

Chart 4.2: number of NPs extracted and cleaned, with trendlines.



The patterns illustrated in Chart 4.2 serve as exemplifications of the *Principle of Least Effort*, more commonly known as *Zipf's Law* (ZIPF, 2016 [1949]), in which dramatic drops in quantity contrast with the increasing length of the NPs sets. This classic, fundamental law is thus understood as follows: in a large sample of words, the frequency of any word is inversely proportional to its rank in frequency. Computationally tested and validated by Ferrer i Cancho and Solé (2003, p. 790), the Law is viewed as “the outcome of the nontrivial arrangement of word-concept associations adopted for complying with hearer and speaker needs”.

The extracted NPs represent approximately **19%** of the total token count in the corpus, meaning that **nearly one-fifth of the corpus consists of NPs ranging from two to ten items**. These encompass simple (*the benefits*) and gradually more complex NPs (*habitat type*, *Mean monthly air temperatures*, *current and future forest regeneration trajectories*), in a wide range of morphosyntactic and semantic traits (*the photosynthetically active upper 5 mm green tips*, *a more negative seasonally integrated $\delta^{13}C$ value*), including the recurrent use of acronyms (*a lower final SMR* and *Large newly designated SPAs*) and hyphenation (*routinely-derived EO-products*, *moisture-driven environmental proxies*), as seen in Examples 4.8 to 4.17.

Example 4.8

→ A bird that can combine **the benefits** from stored food supplies with an ability to enter hypothermia at night will be well adapted (...)

(CoBRA file OE.2017.09.txt)

Example 4.9

→ (...) we expect that individual variation in habitat selection can also occur regarding **habitat types** that appear to be used in proportion to their availability at the population level (...)

(CoBRA file OE.2016.08.txt)

Example 4.10

→ **Mean monthly air temperatures** recorded at Bellingshausen (...) and Vernadsky stations (...) between 2000 and 2012 fall in the range of $-9^{\circ}C$ to $+2^{\circ}C$.

(CoBRA file OE.2016.09.txt)

Example 4.11

→ Model predictions estimating regeneration probability are particularly valuable for estimating **current and future forest regeneration trajectories** (...)

(CoBRA file BC.2015.04.txt)

Example 4.12

→ (...) **the photosynthetically active upper 5 mm green tips** are likely to go through substantially more wetting-drying cycles than the denser, more protected, non-photosynthetic tissue (...)

(CoBRA file OE.2016.09.txt)

Example 4.13

→ (...) a higher proportion of the seasonal net assimilation will occur under high discrimination conditions, which is reflected in **a more negative seasonally integrated $\delta^{13}\text{C}$ value.**

(CoBRA file OE.2016.09.txt)

Example 4.14

→ (...) a wider range of **routinely-derived EO-products** for the UK, including vegetation productivity (...)

(CoBRA file BC.2019.07.txt)

Example 4.15

→ Peat accumulation rates, mineral composition and extent of humification have been interpreted as **temperature- and moisture-driven environmental proxies.**

(CoBRA file OE.2016.09.txt)

Example 4.16

→ **Large, newly designated SPAs** present new opportunities to tackle threats within important foraging habitat.

(CoBRA file BC.2015.09.txt)

Example 4.17

→ (...) individuals that reduced both their SMR and ACT to a greater extent during the experiment, and therefore had **a lower final SMR** (...)

(CoBRA file OE.2016.02.txt)

4.2.3 Observations about +6-item NPs

As shown in Table 4.1, 3,822 NPs with six or more items were extracted. This figure falls to 1,980 NPs post-cleaning. +6-item NPs include cases such as *endangered Sacramento River winter-run Chinook salmon *Oncorhynchus tshawytscha**. Initially classified as an 8-item NP, this NP had *Oncorhynchus tshawytscha* removed to form a 6-item NP, since it has two heads: *salmon* and *Oncorhynchus tshawytscha*.

In the context in which this long NP occurs, shown in Example 4.18, *Oncorhynchus tshawytscha* is between parentheses, hence it is another NP. Albeit semantically connected to the previous NP, specifically to the head *salmon*, *Oncorhynchus tshawytscha* is a new NP and carries information that narrowly identifies the preceding noun head.

Example 4.18

→ (...) the impacts of elevated water temperatures on **endangered Sacramento River winter-run Chinook salmon (*Oncorhynchus tshawytscha*)** in the Central Valley of California

(CoBRA file ZS.2015.03.txt)

The dependency structure below (Figure 4.2) reveals that *Oncorhynchus tshawytscha* is semantically connected to *salmon* (#7->5), with a syntactic function of apposition (@APP). This follows the same rationale behind Example 4.1 seen in subsection 2.2. In *the neotropical direct-developing genus Eleutherodactylus*, both *genus* and *Eleutherodactylus* are heads, and *Eleutherodactylus* is semantically subordinated to *genus*, just as *Oncorhynchus tshawytscha* is to *salmon*. The concrete noun and the scientific name are also heads.

Figure 4.2: dependency structure for *endangered Sacramento River winter-run Chinook salmon (Oncorhynchus tshawytscha)*, made with English VISL- Visual Interactive Syntax Learning.

```

</β>
endangered [endanger] <DA:true> <vq> <mv> <nosubj> <v-quote> V IMPF &headline @FS-STA #1->0
Sacramento River [Sacramento=River] <complex> <*> <Proper> <top> <pre-long> N S NOM @>N #2->5
winter-run [winter-run] <heur> <pre-long> ADJ POS @>N #3->5
Chinook [Chinook] <*> <DA:chinook> <ling> <Hnat> <comp1> <ncomp> N S NOM @>N #4->5
salmon [salmon] <Aich> <food-m> <comp2> <comp2> <idf> <nhead> N P NOM @<ACC #5->1
( [ [ PU @PU #6->0
Oncorhynchus tshawytscha [oncorhynchus=tshawytscha] <insertion> <complex> <*> <heur> <nhead> N S NOM @APP #7->5
] ] PU @PU #8->0
</β>

```

Another +6-item NP that makes for a very intriguing case is *a typical nitrogen-based Nutrient-Phytoplankton-Zooplankton-Detritus NPZD plankton model*. Categorized as a 7-item NP in the output, it contains an acronym (NPZD) and two hyphenated items, one of which can be segmented. *Nutrient-Phytoplankton-Zooplankton-Detritus* refers to a specific four-component model (see DAEWEL; SCHRUM; MACDONALD, 2019), linked to a name, which makes this 7-item NP a 6-item one. Example 4.19 shows the NP in its original context.

Example 4.19

→ The ecosystem model was built on a typical **nitrogen-based, Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) plankton model**
(CoBRA file ZS.2015.03.txt)

The longest complex NP identified displays ten words: *a Thermo Finnigan Delta Plus XP isotope ratio mass spectrometer*. As can be seen, five of the ten items in this NP refer to a proper name. This pattern is also found in other long, complex NPs, as in the 8-item NP *four simultaneous Markov chain Monte Carlo (MCMC) chains*, in which an acronym is present, or the 6-item NP *the logistic Ricker and Beverton-Holt models*, with two items/three elements as names (*Ricker* and *Beverton-Holt*). Examples 4.20 to 4.22 showcase the NPs in context.

Example 4.20

→ Nitrogen isotope ($\delta^{15}\text{N}$) values of sparrow tissues and plants were measured on a Costech 4010 elemental analyser coupled to **a Thermo Finnigan Delta Plus XP isotope ratio mass spectrometer** (...)

(CoBRA file OE.2019.08.txt)

Example 4.21

→ For phylogenetic analysis, two independent runs with **four simultaneous Markov chain Monte Carlo (MCMC) chains** (one cold and three heated), each with random starting trees, were conducted (...)

(CoBRA file ZS.2015.06.txt)

Example 4.22

→ For example different assumptions about the mechanism of competition lead to **the logistic, Ricker and Beverton-Holt models**.

(CoBRA file EL.2016.03.txt)

These longer complex NPs also tend to contain coordinating conjunctions (*or/and*) as premodifiers, as in ***the rarely collected and poorly known species*** and ***individual or collective anti-parasite defences*** (Examples 4.23 and 4.24 for in-context sentences). This propensity for longer NPs to contain coordinating structures had already been attested in Biber et al. (1999) and is more recently verified in Dutra et al. (2020) for specialized, written corpora in the areas of Chemistry and Applied Linguistics. We may thus see a pattern in the use of coordination in longer premodified NPs in specialized, densely informative registers.

Example 4.23

→ (...) we devoted special efforts to sample ***the rarely collected and poorly known species*** inhabiting North America and Eastern Asia (...)

(CoBRA file ZS.2017.09.txt)

Example 4.24

→ Social species, in particular, may be under strong selection for ***individual or collective anti-parasite defences*** that reduce infection risk.

(CoBRA file EL.2015.10.txt)

Participles are also spreadly employed, hyphenated or not: *the entire fern-dwelling ant communities*, *the implemented Metropolis-coupled Markov chain Monte Carlo algorithm* and *a widely accepted eukaryotic SSU rRNA secondary structure model*. Examples 4.25 to 4.27 as follows have been retrieved from the CoBRA to illustrate such uses.

Example 4.25

→ (...) in our study, at the level of **the entire fern-dwelling ant community**, spread across multiple ferns.

(CoBRA file EL.2015.05.txt)

Example 4.26

→ Phylogenetic analysis with Bayesian inference (BI) was performed with *mrBayes* 3.2.1 (Ronquist et al. 2012) and **the implemented Metropolis-coupled Markov chain Monte Carlo algorithm**.

(CoBRA file OE.2019.08.txt)

Example 4.27

→ The predicted secondary structure of the SSU rRNA genes of cyrtophorians corresponds to that of *Tetrahymena canadensis*, which is **a widely accepted eukaryotic SSU rRNA secondary structure model**.

(CoBRA file ZS.2016.06.txt)

Longer NPs also tend to present more morpho-syntactically sequential combinations. For example, *two widely distributed and recently expanded allopatric sister lineages* has two concatenated adverb + participial adjective constructions coordinated by the conjunction *and*, while *the permanently structurally and functionally changing reproductive tissues* has three consecutive adverbs ending in *-ly* coordinated by *and*. This sequentiality may render complex NPs more easily processable since it involves a higher degree of grammatical transparency. Examples 4.28 and 4.29 display the in-context sentences in which these NPs occur.

Example 4.28

→ *G. leopoliensis* consists of **two widely distributed and recently expanded allopatric sister lineages** that diverged from the southern ones ca. 4 Ma (...)

(CoBRA file OE.2019.08.txt)

Example 4.29

→ It is however assumed that heat hardening is – similar to frost hardening (Neuner et al. 2013) – limited in **the permanently structurally and functionally changing reproductive tissues**.

(CoBRA file OE.2015.10.txt)

Finally, the vast majority of +6-item NPs are initiated by a function word: *the same density-independent per-capita birth and death rates, the most abundant anthropogenically undisturbed and disturbed habitat types, and an extensive and taxonomically and trophically highly resolved data*. Examples 4.30 to 4.32 show these NPs in context.

Example 4.30

→ (...) where all species have **the same density-independent per-capita birth and death rates**, and the same immigration rates (...)

(CoBRA file EL.2017.04.txt)

Example 4.31

→ Below we use **an extensive and taxonomically and trophically highly resolved data set** on forest soil invertebrates (...)

(CoBRA file OE.2015.01.txt)

Example 4.32

→ (...) because they are **the most abundant anthropogenically undisturbed and disturbed habitat types** (...)

(CoBRA file OE.2016.08.txt)

4.2.4 NP sets initiated by content words

Following Longo, Höfling, and Saad (1997), undesired⁶⁹ NPs were removed from the cleaned sets and labeled as a) names, b) locations, or c) units of measurement. The first group refers to scientific and proprietary names (*Mesochorus gemellus*, *Trovan Ltd*), the second one regards places and locations, such as cities, lakes/streams, and mountains/forests (*New Zealand*, *Rio Branco*, *Lake Eyre*, *Trinidad & Tobago*) and the third corresponds to measurements (*0.22 μm*, *550 km*).

Seeing that function words (determiners, numerals, articles, etc.) are not super-class items (LEMLE, 1984), premodifiers initiated by functional words were also removed. **The analysis undertaken in this thesis is based on NPs that contain hyphenated elements and that are initiated by content words.** However, hyphenated items are computationally read as a single unit and some content NPs may be premodified by hyphenated items with functional words as the initial element (*out-group species*, *one-tailed tests*). To facilitate understanding, NPs that contain hyphenated premodifiers and whose first constituent is a content word are hereby called *hyphenated content NPs*.

The deep and function-word cleanings were carried out in all other NP chunks, that is, the already clean 3-item, 4-item, and 5-item NP sets underwent a second cleaning to remove NPs that exclusively refer to names, locations, or units of measurement, and to delete NPs initiated by function words. The latter could have been relocated to a new set. This, however, would have meant altering their natural configuration, for instance, by eliminating *the* in *the*

⁶⁹ *Undesired* means that such NPs do not amount to important information for the present study. It is therefore not meant to be taken as incorrect or as an extraction error.

first-court index and making it into a 2-item NP, when in fact it is a 3-item NP initiated by a function word, which, in turn, may be viewed as opposing a fundamental principle of Corpus Linguistics: the study of **naturally occurring** text.

A clean list of 2-item NPs resulted in 112,072 occurrences, of which 9,765 (c. 8.6%) correspond to undesired NPs. The deep-cleaned 2-item content NP set has 102,307 NPs, with scientific and proprietary names representing 5,589 tokens, c. 5% of the NP chunks, while the location and place names removed reached 1,372 NPs (c. 1%) and measurement units totaled 2,804 (c. 2,5%). NPs with functional words as the initial premodifying item were deleted, thus leading this set to 54,410 occurrences, of which 3,562 present hyphenated modifiers, c. 7% of the 2-item hyphenated content NPs.

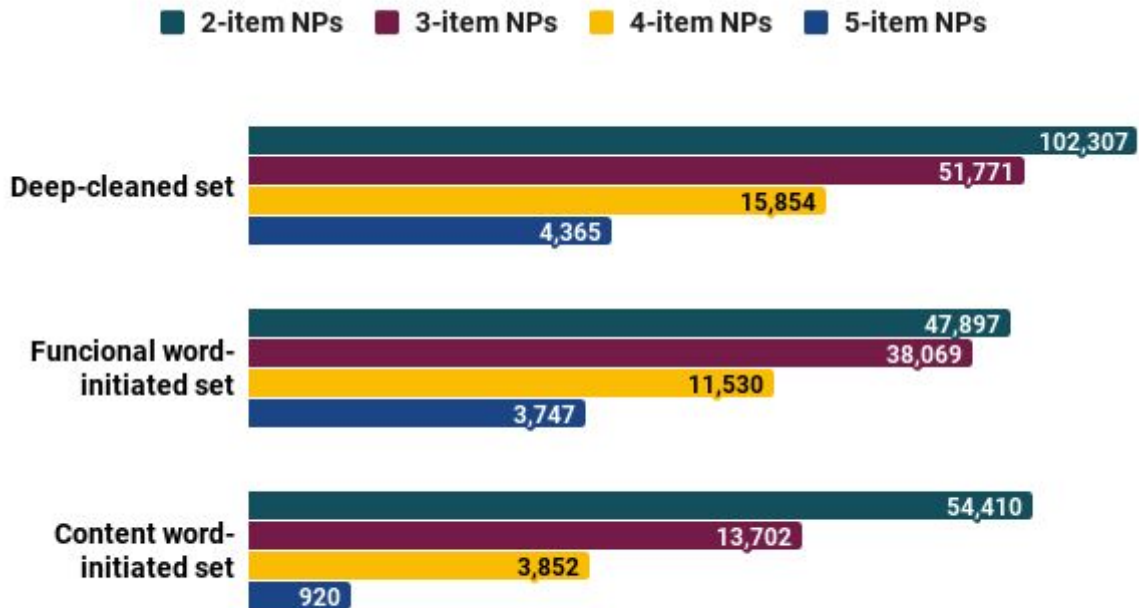
For the 61,980 chunks in 3-item NPs, scientific/proprietary names, locations, and units of measurement totaled 10,209 occurrences (c. 16%) and led to 51,771 3-item NPs. After the functional word cleaning, the final 3-item content NP set totaled 13,702 NPs, c. 26% of the fully-cleaned 3-item NP set. The same procedures were conducted in the 4-item and 5-item NPs, whose counts dropped from 17,643 to 15,854 and from 4,553 to 4,365 NPs, respectively. A list of 3,852 4-item content word-initiated NPs was created (c. 24%) subsequent to sorting out NPs beginning with a function word from those starting with content words. From the 5-item NPs, 920 are content word-initial NPs (c. 21%).

The total amount of content word-initial NPs in the 2-5-item NP range is 72,877. This is approximately 42% of the deep-cleaned set. In comparison with total corpus counts, these sets correspond to 13.3% (deep-cleaned set) and 5.7% (content word-initial set) of all tokens in the CoBRA. Adjusting these frequencies to cover cleaned, complex NPs (i.e. NPs that may contain a functional word as the first constituent, but that are also composed of other content words) translates into an 8.3% share of total corpus counts, c. 8.5% or 109,736 tokens once we add the +6-item NPs in this calculation.

Table 4.2: number of NPs deep-cleaned, initiated by functional and content words.

	2-item NPs	3-item NPs	4-item NPs	5-item NPs	Total
Deep-cleaned set	102,307	51,771	15,854	4,365	174,297
NP initiated by functional words	47,897	38,069	11,530	3,747	101,422
NP initiated by content words	54,410	13,702	3,852	920	72,877

Chart 4.3: number of NPs deep-cleaned, initiated by functional and content words.



As evidenced by the counts in Table 4.2, the longer the NPs, the less frequent they are. This is not only expected but also desired, particularly if one takes into account the *Principle of Economy* (MARTINET, 1955) and *Zipf's Law* (ZIPF, 2016 [1949]). Thus, as far as (natural) language use is concerned, human communication may be seen as an ongoing balancing act between maximum communicative potential and minimal effort (TOBIN, 1990), in a trade-off exercise.

Hence, the cognitive effort exerted to effectively process longer, more structurally and semantically complex NPs, as *the same density-independent per-capita birth and death rates and two widely distributed and recently expanded allopatric sister lineages* (seen in Examples 4.33 and 4.34), must somehow be offset by more transparent morpho-syntactic and semantic relations among the NP items.

Example 4.33

→ (...) where all species have **the same density-independent per-capita birth and death rates**, and the same immigration rates.

(CoBRA file EL.2017.04.txt)

Example 4.34

→ *G. leopoliensis* consists of **two widely distributed and recently expanded allopatric sister lineages** that diverged from the southern ones ca. 4 Ma (...)

(CoBRA file ZS.2018.02.txt)

This rationale should hold true even when language use is directed at a disciplinarily expert, linguistically proficient audience, and even if the genre that propels such complex and compressed constructions allows for the possibility of numerous re-reads for clarity and better comprehension, as is the case with RAs. As for more semantically polysemous, ambiguous or vague expressions, context and target readership expert background knowledge may also help comprehension, as Biber and Gray (2016) argue for longer, compressed, complex NPs used in scientific writing.

4.3 Hyphenated premodifiers

Hyphenated NPs were selected from the above-described content word-initiated NPs by checking wordlists for each NP set in Sketch Engine (KILGARIFF et al, 2010, 2014) and Notepad++ (HO, 2019). In total, 5,789 NPs with hyphenated elements were identified. These hyphenated elements were manually POS-categorized (*genome-wide response*) by sorting out hyphenated prefixed premodifiers (*non-host*)⁷⁰, tri-constituent premodifiers (*state-of-the-art*), and NPs whose hyphenated element is the NP head (*infected pack-mates*). This led to a total of 5,735 NPs with hyphenated premodifiers.

Table 4.3 A: number of NPs with hyphenated premodifiers and hyphenated heads.
Table 4.3 B: number of NPs containing hyphenated premodifiers (*sans* hyphenated heads).

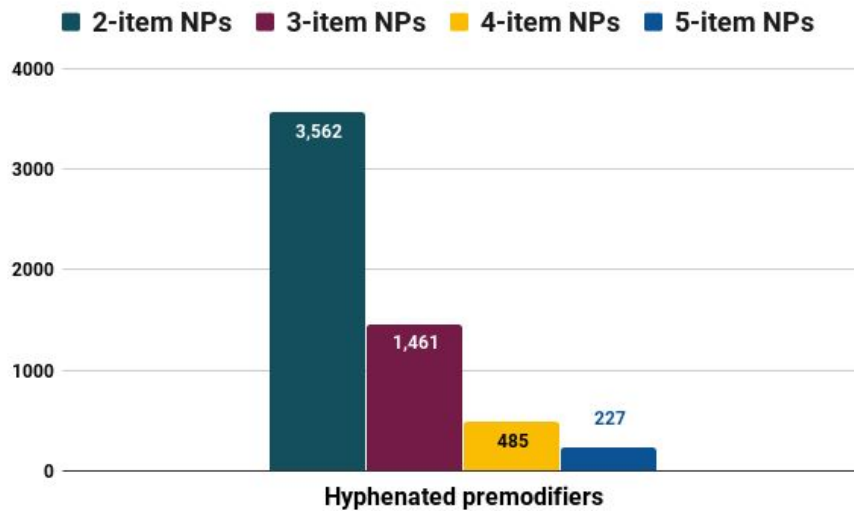
	2-item NPs	3-item NPs	4-item NPs	5-item NPs	Total
A - Hyphenated premods + head	3,609	1,464	487	229	5,789
B - Hyphenated premodifiers	3,562	1,461	485	227	5,735

Table 4.3 shows the breakdown of NPs containing hyphenated premodifiers and heads and NPs displaying hyphenated premodifiers only. Figure 4.3 in the next subsection shows the POS-categorizations for hyphenated premodifiers in the 3-item NP set. As previously stated, the manual POS-categorization covered all content word-initiated NP sets. The only NPs not manually categorized were those belonging to the +6-item NP set, as these longer NPs are not detailedly addressed in the analysis⁷¹.

⁷⁰ Hyphenated prefixed premodifiers were computed for frequency. A discussion is presented in subsection 4.3.

⁷¹ This was a methodological decision made on the grounds of time constraints.

Chart 4.4: number of hyphenated premodifiers.



An interpretation of Table 4.3 and Chart 4.4 shows that frequencies follow a consistent high-to-low trend, in which 2-item NPs concentrate the most tokens while 5-item NPs display the fewest. When the NPs are sorted out for hyphenated elements, we see that hyphenation is overpoweringly favored in premodification, not in head position, with only 54 NPs containing hyphenated tokens taking the NP head (0.93%) in the CoBRA. This could be related to the more strategic use of hyphenation as a compression device, as will be discussed in subsection 4.4.3 towards the end of this chapter.

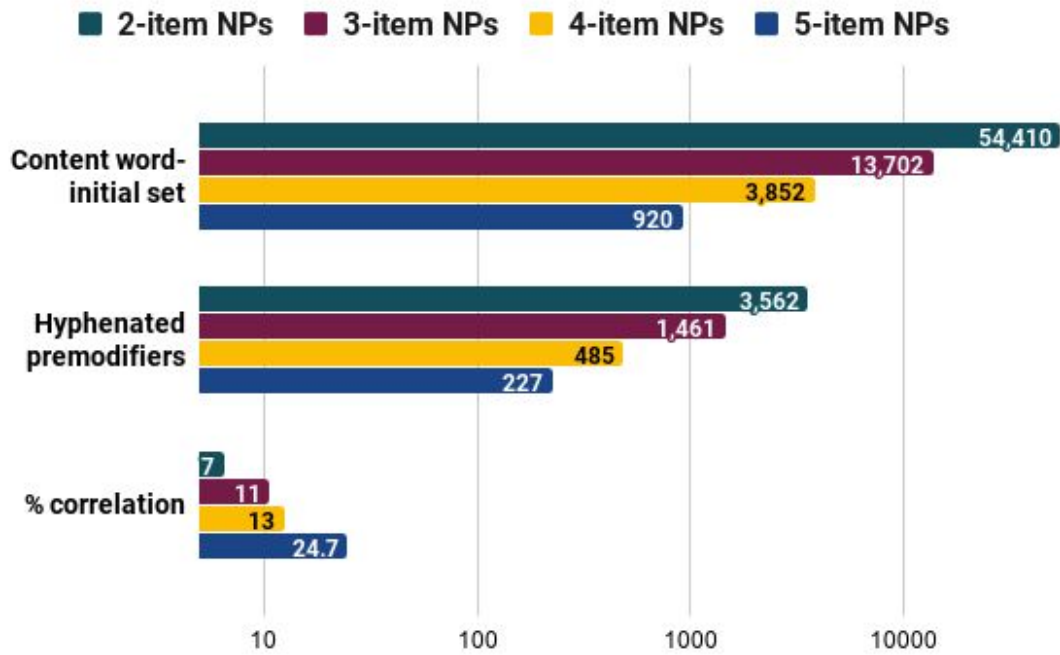
Additionally, hyphenated premodifiers account for a gradually increasing chunk of the content NP sets, reaching nearly 25% of the 5-item NPs. This may be taken as an indication of the compression power of hyphenated compounds in English: the longer the NP, the more use of hyphenated premodifiers. In languages that privilege post-modification, combinatorial patterns, such as the ones discussed, are invariably structured in prepositional phrases and clausal constructions, as shown in subsection 4.5 for Brazilian Portuguese (BP).

It should be stressed that hyphenated elements may occur in a wide range of positions in premodification, from initial, middle, and final slots. The focus of this thesis does not rest on the examination of the premodifiers' internal combinatorial arrangements, but rather on the frequency and distribution of such hyphenated premodifiers in the corpus and their behavior in regard to morpho-syntax and semantics. Table 4.4 and Chart 4.5 display the total number of premodifiers initiated by content words in each NP set.

Table 4.4: number of content NPs and hyphenated premodifiers.

	2-item NPs	3-item NPs	4-item NPs	5-item NPs	Total
Sets initiated by content words	54,410	13,702	3,852	920	72,877
Hyphenated premodifiers	3,562	1,461	485	227	5,735
Percent correlation	6.5%	10.7%	12.6%	24.7%	7.9%

Chart 4.5: number of content NPs and hyphenated premodifiers.



4.3.1 POS-categorization

As already mentioned, the manual POS-categorization of the hyphenated premodifiers was based on the content-word lists generated from the cleaned NP sets. While software such as Sketch Engine (KILGARIFF et al., 2010, 2014) automatically POS-tags hyphenated items, it does so very limitedly. Similar to Stanford Core NLP (MANNING et al, 2014), the software is unable to morpho-syntactically categorize individual elements in hyphenated items, seeing that hyphenated items are processed as a single unit.

Sketch Engine also mislabeled some NPs. For instance, for *model-averaged predicted values*, *model-averaged* was tagged as a noun. The software seems to have read the NP as if it were a clause⁷², in which *model-averaged* is the subject and *predicted* and *values* are the verb

⁷² As based on the extracted NP files, which were uploaded to the software, not on the in-context sentences as originally produced in the corpus.

and the object, respectively composing the predicate. Nevertheless, even if it were possible to automatically POS-tag individual elements in the hyphenated items, this categorization would yield only a partial picture of the phenomenon in question. Example 4.35 displays the NP in context.

Example 4.35

→ **Model-averaged predicted values** for each set of observed values of the independent variables were used for graphical presentation

(CoBRA file BC.2010.10.txt)

As stated in the previous subsection, Figure 4.3 partially reproduces a spreadsheet of the POS-categorizations for the hyphenated premodifiers in the 3-item content word-initiated NP set. Similar spreadsheets were produced for each POS-categorized NP set, except for the +6-item NP set.

Figure 4.3: POS-categorization for 3-item NPs initiated by a content word.

NV	VN	VV	NN
terrestrial gel-encapsulated eggs	foraging-predation risk trade-offs	pierce-sucking feeding modes	genus-level phylogenetic relationships
log-transform ratio characters	linear mixed-effects models	pierce-sucking feeding modes	subtle life-history differences
brain-derived neurotrophic factor	mixed-species net fisheries	Conventional capture-recapture models	separate family-level taxon
gamma-shaped rate variation	mixed-age worker honeybees	conventional capture-recapture modeling	major life-history differences
BI-adjusted sequence trees	Wilcoxon signed-ranks tests	JS mark-recapture models	parallel life-history characteristics
BI-adjusted sequence trees	Wilcoxon signed-ranks tests	spatial capture-recapture models	deep-water whale carcasses
environment-sampled population richness	Wilcoxon signed-ranks tests	localised hunting-related mortality	partial whole-genome sequencing
lineage-based diversification analyses	Wilcoxon signed-ranks tests	large capture-recapture studies	several species-level autapomorphies
alcohol-preserved gill tissue	Wilcoxon signed-ranks tests	warming-induced range shifts	Majority-rule consensus trees
sex-biased gene flow	paired-samples t test	warming-induced range shifts	past sea-level oscillations
ML-based phylogeny inference	paired-samples t test	warming-induced range shifts	past sea-level signals
nuclear protein-coding genes	paired-samples t test		maximum snout-vent length
nuclear protein-coding gene	Flip-Cup Test conditions		individual tree-level analyses
mitochondrial protein-coding genes	Fixed-effect estimates		species-level growth responses
NGS-based genome scans	initial mixed-effects models		host-plant odour experience
other protein-coding genes	adequate protected-area networks		insect-proof mesh cloth
amber-preserved Niphargus fossils	linear mixed-effects models		rest-phase body temperature
ethanol-preserved tissue samples	linear mixed-effects models		rest-phase body temperature
reef-building scleractinian corals	linear mixed-effects models		day-time predation risk
Zanclaea-related skeletal modifications	linear mixed-effects models		low water-use efficiency
taxon-based conservation strategies	linear mixed-effects models		facultative night-time hypothermia
model-corrected genetic distances	linear mixed-effects models		day-time predation risk
log 10-transformed measurements	linear mixed-effects models		adult life-history traits
Phylogenetic lineage-based approaches	linear mixed-effects models		landscape-level water cycling
herbivore-induced plant volatiles	linear mixed-effects models		alternative life-history strategies

Morphosyntactic categorizations tell us that the 3-item complex NP *model-averaged predicted values* contains a hyphenated element composed of a noun and an *-ed* verb form, as well as another item ending in *-ed* followed by the noun *values*. The premodifying items act

as adjectives, meaning that they qualify the noun *values*. Such categorization is important, as it can indicate the type of language item that forms the hyphenated premodifier. This, in turn, may be helpful in understanding the lexico-grammatical patterns of hyphenated compounds in the corpus. In addition, the morpho-syntactic labeling performed is a necessary step prior to manually coding the hyphenated constituents, which can then be used in NLP practices, such as in automated parsing of hyphenated items.

The main issue with morpho-syntactic categorizations is that they neither provide the meaning nor the relations between the hyphenated premodifiers in the NP. Computationally, such a task would require dependency parsing with semantic-pragmatic annotation. Despite being manually feasible in a larger time frame, semantic annotation is a very time-consuming endeavor, as is dependency parsing, especially because dependency grammar relies on quite different theoretical assumptions, often at odds with phrase grammar (i.e., constituency-based grammar), as discussed in Hudson (2016).

4.3.2 Hyphenated premodifiers in 2-item content NPs

The 2-item set has 3,562 hyphenated premodifier occurrences (tokens), of which 1,470 are unique entries. This computation does not include the 202 2-item content NPs whose head carries the hyphenated element (*infected pack-mates*, seen in Example 4.36). It does, however, include the 15 2-item NPs formed by a hyphenated head as well as a hyphenated premodifier (*life-history trade-offs*, in Example 4.37 below).

Most of the 2-item hyphenated NPs, therefore, present the hyphenated element as their premodifier, with a solid and non-hyphenated word as the head. Table 4.5 shows the twenty most frequent hyphenated premodifiers in the 2-item content NP set (example sentences from the corpus can be found following the table).

Example 4.36

→ We also assessed the burden of mange within the pack by analysing the effect of the number of **infected pack-mates**.

(CoBRA file EL.2015.10.txt)

Example 4.37

→ In the context of **life-history trade-offs** in the evolution of viviparity, we suggest that the extent of correlation between reproductive traits (...)

(CoBRA file OE.2019.03.txt)

Table 4.5: twenty most frequent hyphenated premodifiers in the 2-item content NP set.

Top 20 premodifiers	Freq. (raw)
<i>long-term</i>	60
<i>non-host-infested</i>	51
<i>take-off</i>	48
<i>standard-sized</i>	44
<i>non-native</i>	44
<i>high-value</i>	41
<i>problem-solving</i>	31
<i>non-host</i>	30
<i>within-group</i>	29
<i>semi-natural</i>	29
<i>old-growth</i>	25
<i>between-group</i>	23
<i>yellow-gular</i>	22
<i>temperature-driven</i>	22
<i>large-brained</i>	22
<i>white-gular</i>	21
<i>land-use</i>	20
<i>density-dependent</i>	20
<i>home-range</i>	19
<i>small-scale</i>	19

Many of the twenty most frequent hyphenated premodifiers are familiar expressions in the English language: *long-term* (60 tokens), *take-off* (48 tokens), and *small-scale* (19 tokens). Others make use of prefixes: *non-native* (44 tokens) and *semi-natural* (29 tokens), while four employ regular and irregular past participles: *non-host-infested* (51 tokens), *standard-sized* (44 tokens), *temperature-driven* (22 tokens), and *large-brained* (22 tokens). This corroborates the results found in the comprehensive corpus-based work of Sanchez-Stockhammer (2018), which will be addressed in section five. Examples 4.38 to 4.46 are corpus sentences showing these NPs.

Example 4.38

→ As such, a small addition to the proposed conservation target at no additional cost to local landowners will provide **long-term benefits** to human well-being (...)

(CoBRA file BC.2017.09.txt)

Example 4.39

→ Within the domain of **take-off ability**, birds face a trade-off between **take-off speed** and **take-off angle** (...)

(CoBRA file AB.2015.07.txt)

Example 4.40

→ Monitoring **small-scale fisheries** through observers poses a major challenge due to the large number of vessels (...)

(CoBRA file BC.2018.01.txt)

Example 4.41

→ Our analyses represent a minimum (and conservative) estimate of the importance of **non-native ants**.

(CoBRA file OE.2015.06.txt)

Example 4.42

→ Finally, consistently with previous laboratory experiments, food stressed individuals did not suffer from reduced adult survival also under **semi-natural conditions**.

(CoBRA file OE.2017.05.txt)

Example 4.43

→ Experience with **non-host-infested leaves** on the contrary resulted in a reduced attraction towards **non-host-infested plants** (...)

(CoBRA file OE.2019.04.txt)

Example 4.44

→ Also, small females were attracted by **standard-sized males** but less so than **standard-sized females**.

(CoBRA file AB.2015.04.txt)

Example 4.45

→ **Temperature-driven reversals** in competitive advantage were often linked to analogous reversals in the competitive advantage predicted by the model.

(CoBRA file EL.2018.06.txt)

Example 4.46

→ **Large-brained animals** were marked with a green and red dot on the left and right body side respectively just below the dorsal fin.

(CoBRA file EL.2015.04.txt)

4.3.2.1 Frequency and distribution

In terms of frequency, it could be said that the most frequent hyphenated premodifiers in the 2-item content NPs act analogously to function words in a corpus; that is to say, there are fewer highly frequent hyphenated premodifiers in the 2-item content NP set and these tend to be farther apart from highly infrequent ones (see Table 4.6). This may be seen as a sign of Zipf's Law in place, a least effort reasoning to balance out the morphosyntactic and semantic complexity of longer NPs.

As exhibited in Table 4.6, the six most frequent hyphenated premodifiers in the 2-item NP set occur more than 40 times each, leading to a total of 288 tokens. These premodifiers are *long-term*, *non-host-infested*, *take-off*, *standard-sized*, *non-native*, and *high-value*, as already shown in Table 4.5 in the previous subsection, in which Examples 4.38, 4.39, 4.41, 4.43, and 4.44 display in-context sentences. Example 4.47 corresponds to a CoBRA sentence with the premodifier *high-value*, as seen below.

Example 4.47

→ (...) we explored potential solutions using Marxan for priority areas that proactively conserve **high-value habitats** of these 2 iconic species in currently intact wildlands (...)
(CoBRA file BC.2019.05.txt)

In stark contrast, nearly a thousand hyphenated premodifiers occur only once. In other words, of the 1,470 NPs in this set, 995 are *hapax legomena*, which means that nearly 65% of the hyphenated premodifiers in the 2-item content NPs are single entries. Long-tail graph 4.1 indicates this pattern, as this is yet another example of Zipf's Law. Examples 4.48 to 4.54 refer to some hapaxes from the CoBRA, illustrated as follows.

Table 4.6: number of items and tokens and their frequency range.

No. of NPs	No. of tokens	No. of tokens per NP
6	288	40+
2	61	30-39
10	233	20-29
43	571	10-19
136	759	4-9
271	627	2-3
925	925	1

Example 4.48

→ *Crenicichla* are **pike-like carnivores** with 'ambush and stalk' hunting strategies, they are often sympatric to the guppy and can impose a high predation pressure.
(CoBRA file EL.2015.04.txt)

Example 4.49

→ **Conservation-oriented tools** such as REDD initiatives and Intact Forest Landscapes could also benefit from this approach.
(CoBRA file BC.2018.02.txt)

Example 4.50

→ We consider next an analysis of **soft-sediment macrofauna** described by Ellingsen & Gray (...)

(CoBRA file EL.2015.06.txt)

Example 4.51

→ (...) heat stress behaviours occur at the upper end of the thermal neutral zone, whether demarcated by 40 °C in a desert species such as a zebra finch or by 34 °C degrees in a temperate-zone breeding species such as **white-crowned sparrows**.

(CoBRA file AB.2019.01.txt)

Example 4.52

→ We used the `manyglm` function with **PIT-trap resampling** and a negative binomial distribution, implemented in the `mvabund` package.

(CoBRA file OE.2018.01.txt)

Example 4.53

→ Responses towards HIPVs have been intensively studied in parasitoids and nearly every tested species preferred volatiles from **herbivore-attacked plants** to those from undamaged plants in two-choice assays.

(CoBRA file OE.2019.04.txt)

Example 4.54

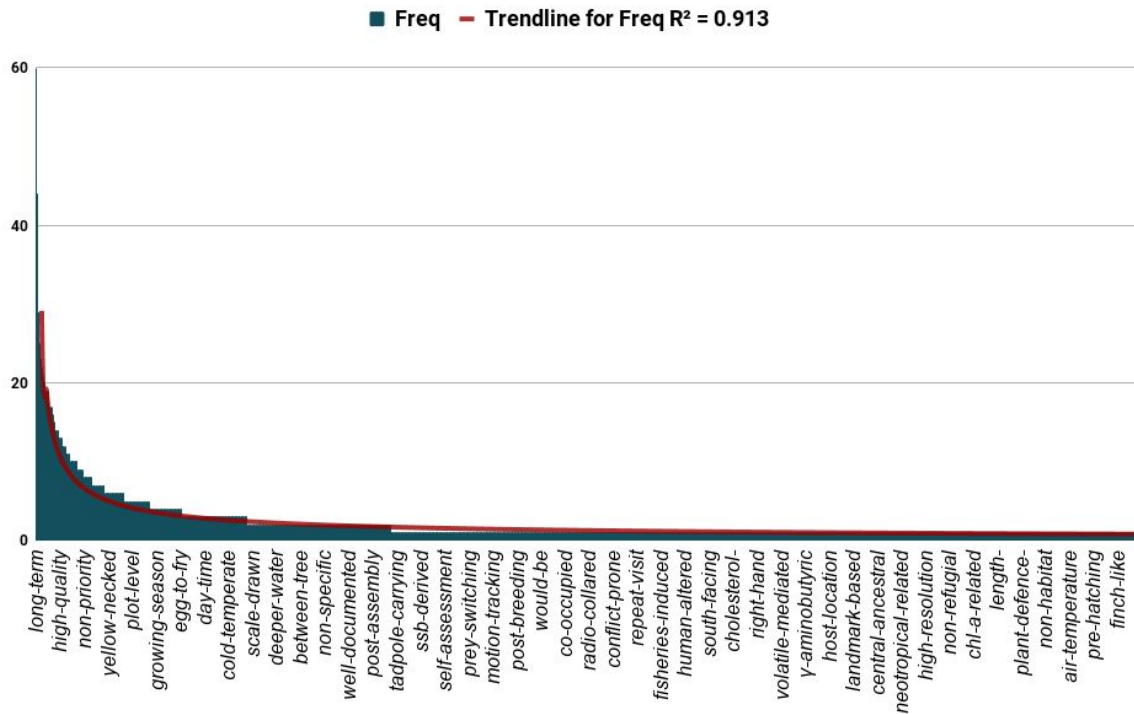
→ Both species likely maintained gene flow between Trinidad and the mainland through **land-bridge connections** at low sea level stands during the Pleistocene.

(CoBRA file ZS.2019.05.txt)

Many of the high frequent hyphenated premodifiers make up complex NPs used only in one or two RAs, which makes for a very restricted distribution across the corpus. However frequent these premodifiers and NPs may be, their high frequency cannot be understood as a pattern in the CoBRA and their frequency would need to be adjusted accordingly. An example is the premodifier *non-host-infested*, used in one text only, by a single author. This means that despite being highly frequent in this set, *non-host-infested* is concentrated in only one article, which is why its high frequency must be reconsidered.

The same applies to *standard-sized* (AB.2015.04.txt), *large-brained* (EL.2015.04.txt), *high-value* (BC.2015.05.txt), *non-host* (OE.2019.04.txt), *problem-solving* (AB.2016.06.txt.), as well as *non-native*, distributed mostly in two articles (BC.2018.08.txt, OE.2015.06.txt). These hyphenated premodifiers were thus adjusted for frequency. As can be verified in the long-tail Graph 4.1 below, they are no longer computed as top frequencies at the head of the graph. **All long-tail graphs have been adjusted for frequency.** Sentence examples follow the Graph for the most frequent premodifier (*long-term*).

Graph 4.1: distribution of the hyphenated premodifiers in the 2-item NP set.



As the most frequent premodifier in this set (and, as will be seen, in other sets as well), *long-term* is primarily employed with the head nouns *loss*, *studies*, *data*, and *effects*, each NP occurring five times, and *funding* and *persistence*, with four occurrences each. Examples 4.55 to 4.60 display these NPs, as retrieved from the CoBRA.

Example 4.55

→ (...) but there is also a need to study **long-term loss**.

(CoBRA file EL.2018.08.txt)

Example 4.56

→ In this zone, **long-term studies** show that sympatric resident and transient bottlenose dolphins differ in behaviour, group size and site fidelity.

(CoBRA file AB.2018.05.txt)

Example 4.57

→ Using **long-term data** on a badger (*Meles meles* Linnaeus, 1758) population naturally infected with *Mycobacterium bovis*, we built an integrated population model (...)

(CoBRA file EL.2016.02.txt)

Example 4.58

→ (...) the indirect paths from treatments (...), support the hypothesis that **long-term effects** of initial host richness and fertilisation on disease are determined by changes in host community structure.

(CoBRA file EL.2019.10.txt)

Example 4.59

→ Further, where **long-term funding** is doubtful, fences are likely to be a waste of both time and money.

(CoBRA file BC.2019.02.txt)

Example 4.60

→ (...) the latter species shows a number of different haplotypes also in northern areas, in different Alpine regions (Fig. 7), suggesting **long-term persistence** of these populations.

(CoBRA file ZS.2015.08.txt)

Attention should be paid to the fact that *long-term* may also be non-hyphenated. From the 213 occurrences for *long-term/long term* in the entire corpus, 77 are non-hyphenated, with four of those referring to the expression *in the long term*. This means that 73 occurrences are not hyphenated (c. 35%) when *in the long term* is discarded⁷³, with *long-term* preferred: 136 occurrences, c. 65% of total uses. Statistically, these differences are significant, as illustrated in Table 4.7 below (chi-squared test used).

Table 4.7: hyphenated vs. solid uses for *long-term*

	Long-term	Long term
Frequency	136	65%
Percent correlation	73	35%
Total	209	100%
Z-score	4.6924	
P-value	< .00001	

Moreover, the distribution of the hyphenated variant is widespread across the corpus, with NPs containing *long-term* found in 58 of the 250 RAs (c. 23%). Examples 4.61 and 4.62 correspond to sentences containing *long term* and *in the long term*, respectively, as retrieved from the CoBRA.

Example 4.61

→ (...) these changes were observed > 20 years after timber extraction ended, supporting previous findings of **the long term impacts** of roads in Central Africa and Amazonia.

(CoBRA file BC.2017.08.txt)

⁷³ As an idiomatic expression, *in the long term* should not be understood as a compound, since it is meant to be read more holistically and because it most frequently does not accept variations (e.g., **in the longest term*, **in a long term*, **in a/the distant term* etc.). See Dressler (2006) and Dressler, Letner and Korecky-Kröll (2010).

Example 4.62

→ (...) it is unlikely that the eradication of Himalayan balsam will negatively impact on the UK honey bee population **in the long term**.

(CoBRA file OE.2017.01.txt)

4.3.2.2 Verifying dispersion

Dispersion plays a very important role in identifying lexico-grammatical patterns in a corpus. Without further understanding how highly frequent items are distributed in a corpus, frequency alone can be misleading (cf. GRIES, 2008; 2010 for detailed discussions). Hence, in order to comprehend the frequency of occurrence and the distribution of the most recurrent hyphenated premodifiers in the 2-item NP set, an examination of their diffusion in the corpus was required and done in AntConc (ANTHONY, 2019) by verifying the concordance plots for each of the twenty most frequent premodifiers listed in Table 4.5⁷⁴.

Figures 4.5 and 4.6 depict partial concordance plots for *density-dependent*, which is relatively well-distributed, and for *land-use*, which makes for a more intriguing case. While *density-dependent* is computed twenty times and is well-dispersed in the corpus, forming only 2-item NPs, *land-use* is computed forty-six times, occurring both more condensedly and more dispersedly in 2-item and 3-item NPs (*land-use composition*, *wider land-use patterns*). The condensed uses correspond to eighteen of the forty-six occurrences and can be traced back to file BC.2019.07.txt. Examples 4.63 to 4.65 respectively contain an NP with *density-dependent* premodifying the head noun *dispersal*, and sentences with *land-use composition* and *wider land-use patterns*.

Example 4.63

→ (...) and such strategies could lead to **density-dependent dispersal**.

(CoBRA file AB.2017.07.txt)

Example 4.64

→ There is evidence that bird taxonomic and functional diversity can increase within HNV farmland in relation to **land-use composition** (...)

(CoBRA file BC.2019.07.txt)

Example 4.65

→ (...) accounting for > 95% of farmed land at each site, mirroring **wider land-use patterns** throughout the Colombian Andes.

(CoBRA file BC.2017.05.txt)

⁷⁴ This procedure was also done for the twenty most frequent premodifiers in the other NP sets.

From the next ten most frequent hyphenated premodifiers in the 2-item content NP set, *old-growth*, *between-group*, *yellow-gular*, *white-gular*, *land-use*, *life-history*, *non-target*, and *small-scale* are concentrated in one/two articles. In other words, only three of the twenty most frequent hyphenated premodifiers in 2-item content NPs are more evenly distributed in the corpus. Examples 4.66 to 4.73 below show the NPs in context.

Example 4.66

→ There was a positive effect of secondary forest age on PD (likelihood ratio test, $P = 0.017$), with forests ~ 20 years reaching **old-growth levels**.

(CoBRA file BC.2017.05.txt)

Example 4.67

→ However, the level of **between-group variability** was not sufficient to warrant incorporating individual as a random effect.

(CoBRA file BC.2015.09.txt)

Example 4.68

→ Similarly, the gular region of **yellow-gular males** differed significantly from that of females in all three colour measures.

(CoBRA file AB.2019.08.txt)

Example 4.69

→ (...) we find two alternative morphs: yellow-gular males and **white-gular males**.

(CoBRA file AB.2019.08.txt)

Example 4.70

→ Threatened native ecosystems are those that currently cover < 1% of Uruguay and are expected to further decline in size due to **land-use change**.

(CoBRA file BC.2017.09.txt)

Example 4.71

→ Thus, we advocate treating **life-history differences** between the sexes as a co-adapted whole.

(CoBRA file OE.2019.10.txt)

Example 4.72

→ Environmental DNA assays reduce missed detections resulting from samples dominated by **non-target DNA** (...)

(CoBRA file BC.2019.01.txt)

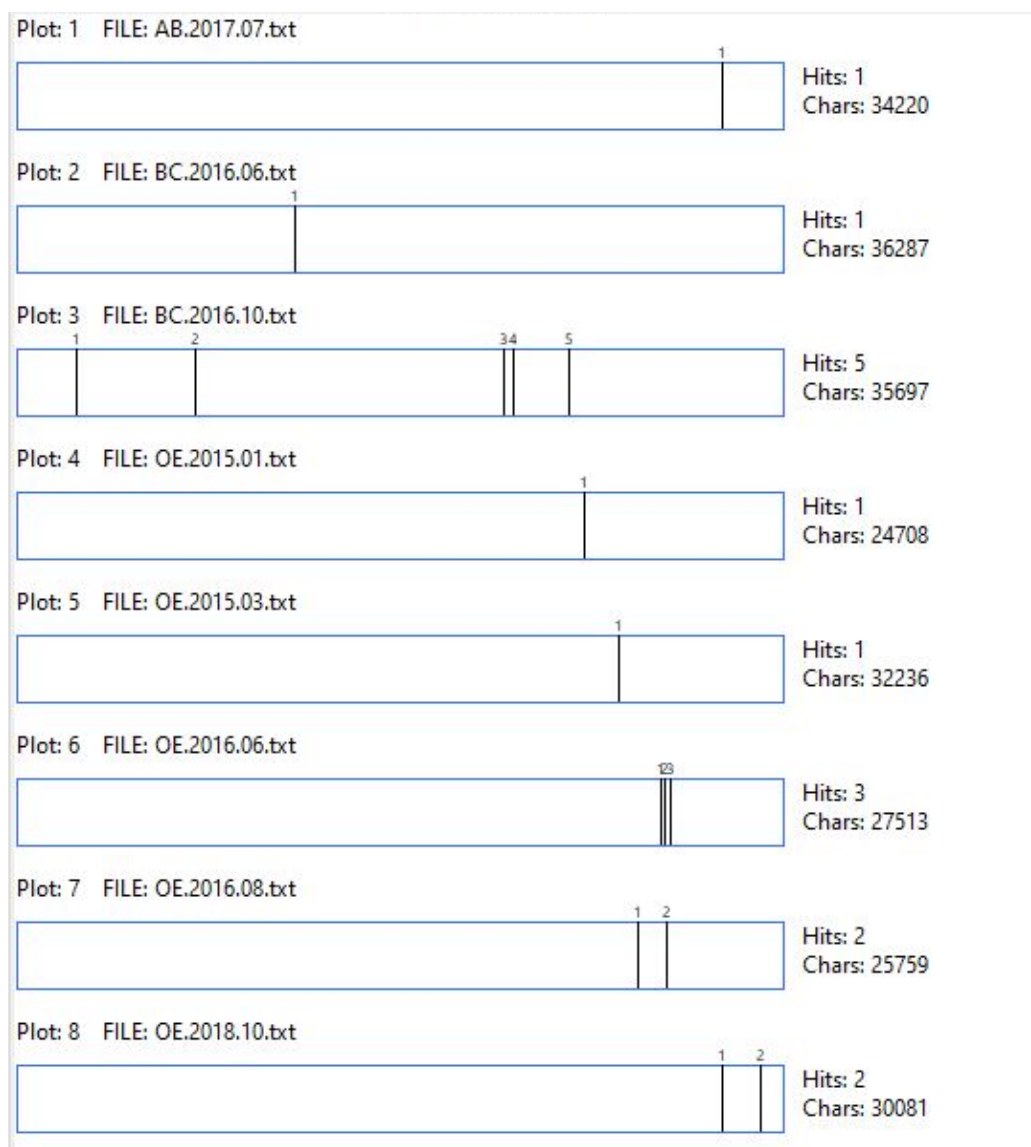
Example 4.73

→ To quantify **small-scale activity** (i.e., activity on the “bush-level”), we calculated the variance of the signal strength (...)

(CoBRA file OE.2018.02.txt)

These results could suggest that hyphenated premodifiers in short NPs are employed parsimoniously, with the most frequent ones referring to grammaticalized and/or lexicalized expressions (*long-term*) or prefixed compounds (*semi-natural*). This may point to higher use of NP items that require less cognitive effort in processing and production, thereby facilitating understanding of the whole complex NP.

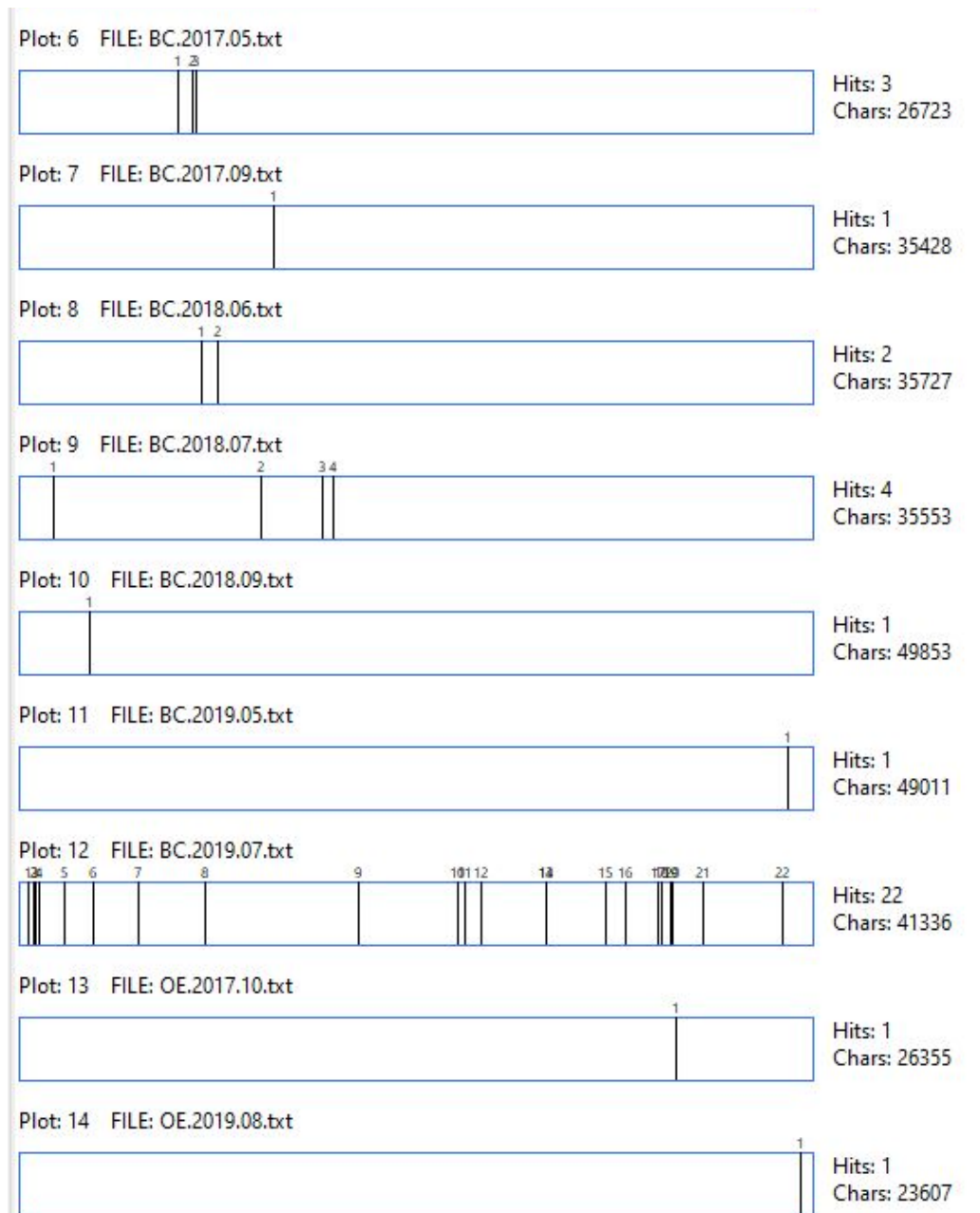
Figure 4.4: partial concordance plot for *density-dependent*.



This, however, does not change the aforementioned trend in frequency, seeing that half of the most frequent hyphenated premodifiers from the 2-item content NPs are more evenly distributed in the corpus. In other words, the few, better distributed, highly frequent

words are still significantly more recurrent than the *hapax legomena* and than those occurring 2-3 times (c. 18% of the 1,470 unique occurrences). This means that the long tail graph still stands and that the highly frequent premodifiers occupy the position of functional items in a corpus, with most of the less frequent items in the short-tail section of the graph, as would be expected of highly infrequent content words.

Figure 4.5: partial concordance plot for *land-use*.



4.3.2.3 Tri-constituent premodifiers

The main surprise in this set is the second most frequent hyphenated premodifier. Not only does *non-host-infested* consist of three elements, but one of such elements (*to infest*) is of verbal morphosyntactic nature. This might create some difficulty for less proficient readers. Still, the fact remains that this 2-item NP is actually composed of five constituents, not two. *Non-host-infested* mostly premodifies *plants* and *leaves* (Examples 4.74 and 4.75 seen below). Segmenting may result in *plants/leaves that are/were not infested by a non-host*, which may be relatively more complex to process for less proficient readers.

Example 4.74

→ We expected that **non-host-infested plants** would confuse the foraging parasitoids.

(CoBRA file OE.2019.04.txt)

Example 4.75

→ Experience with **non-host-infested leaves** on the contrary resulted in a reduced attraction towards **non-host-infested plants**.

(CoBRA file OE.2019.04.txt)

Preposition-looking particles make up the majority of the 3-constituent premodifiers, as indicated in Table 4.8. Of the 18 occurrences, half contain *to*, which may convey a sense of movement, direct correlation, and/or transformation to the premodifiers. The only exception is *hard-to-identify*, in which *to* is bound to the verb *identify*, as this is used in its infinitive form ‘to identify’. Examples 4.76 to 4.79 right below Table 4.8 illustrate such uses.

Table 4.8: tri-constituent hyphenated premodifiers.

NPs containing prepositional particles			
to	of	by	through
<i>c-to-nutrient</i>			
<i>summer-to-winter</i>			
<i>hard-to-identify</i>			
<i>snout-to-vent</i>	<i>out-of-phase</i>		
<i>depth-to-water</i>	<i>out-of-group</i>	<i>genotype-by-environment</i>	<i>diversity-through-time</i>
<i>wall-to-wall</i>	<i>proof-of-concept</i>	<i>case-by-case</i>	
<i>seed-to-seedling</i>	<i>colony-of-origin</i>		
<i>nitrogen-to-carbon</i>	<i>young-of-year</i>		
<i>egg-to-fry</i>			
<i>year-to-year</i>			

Example 4.76

→ With our approach, we produced **wall-to-wall forest cover change maps** from 1990 to 2014 for the full territory of Madagascar.

(CoBRA file BC.2018.06.txt)

Example 4.77

→ We provided **proof-of-concept evaluations** of species-specific qPCR assays applied to three potential survey techniques.

(CoBRA file BC.2019.01.txt)

Example 4.78

→ (...) the role of *X. laevis* in these environments requires **case-by-case investigation**.

(CoBRA file BC.2015.01.txt)

Example 4.79

→ **Diversity-through-time plots** produced by simulating species accumulation (...) show that Galápagos land birds have not attained an equilibrium diversity (...)

(CoBRA file EL.2015.09.txt)

Other premodifiers composed of three constituents often refer to time (e.g., *6-day-old*) or measurements (*7-mm-square*), with 17 and 6 items, respectively. An interesting case is the use of adjectives in final position in the 3-constituent premodifiers: *more-than-additive*, which modifies *effects*, and *gc-ead-active*, which pre-modifies *peaks*. These are lone occurrences in the set. Example sentences 4.80 to 4.83 display NPs containing these premodifiers.

Example 4.80

→ We normalized fatty acid compositions to dry weight (dwt) in neonates and **6-day-old animals** of all clones and calculated mean values for the three species.

(CoBRA file OE.2016.04.txt)

Example 4.81

→ The mussels were deployed to each cage inside plastic mesh bags (with **7-mm-square openings**) to limit movement (...)

(CoBRA file EL.2016.06.txt)

Example 4.82

→ The physiological effects of these ecological stress factors often exacerbate each other (...), having **more-than-additive effects** on individuals, populations and communities.

(CoBRA file EL.2015.02.txt)

Example 4.83

→ **GC-EAD-active peaks** were identified using the Kovats' retention indices (KI) from GC-EAD, GC-MS and published records (...)

(CoBRA file EL.2019.07.txt)

On the other end of the spectrum are tri-constituent hyphenated premodifiers that have nouns or verbs as their final constituent, in a hierarchical relation, with the final constituent as the ‘head’, such as in *low-light-acclimated*, in which *low-light* functions as the premodifier to *-acclimated*. These are subordinate compounds (cf. BAUER; HUDDLESTON, 2002), just as *herbivore-attacked plants*. In Example 4.84 we can observe this premodifier in the NP, which actually includes a further adjective + coordinate conjunction subordinate to *-acclimated*.

Example 4.84

→ Comparing **high- and low-light-acclimated chlorophytes**, one might find a 2- to 20-fold difference in the concentration of these complexes

(CoBRA file EL.2016.09.txt)

Non-hierarchical combinations are also found: *glucose-methanol-choline* is one such case. A minor category in English, *dvandva compounds* are “coordinative in that the bases are of equal status instead of being in a relation of subordination”, which is very typical of proper names, as explained by Bauer and Huddleston (2002, p. 1648). Example 4.85 shows the sole NP with this tri-constituent, coordinative compound.

Example 4.85

→ We targeted GDH, whose encoded product is a member of the superfamily of **glucose-methanol-choline oxidoreductases** (GMCs) that include enzymes known to (...)

(CoBRA file EL.2018.09.txt)

Tri-constituent hyphenated premodifiers in which two of the three constituents form names and present a ‘head’ were not considered, such as in *glyceraldehyde-3-phosphate* and *chl-a-related*, seen in Examples 4.86 and 4.87.

Example 4.86

→ DNA sequence data were also collected for five nuclear loci: the nuclear exon (...), **glyceraldehyde-3-phosphate dehydrogenase** (...)

(CoBRA file ZS.2018.01.txt)

Example 4.87

→ We used these data in separate principal components analyses (PCA) for key environmental variables (...), and this process was repeated for **chl-a-related variables** (...)

(CoBRA file EL.2016.06.txt)

4.3.3 Hyphenated premodifiers in 3-item content NPs

From the 13,760 3-item NPs formed solely by content words, there are 1,461 instances of 3-item NPs with hyphenated premodifiers, roughly 11% of the content-word list. There are 797 items, of which 572 are single entries. Because 3-item content NPs are composed of two premodifiers to the noun head, a closer inspection into the NPs was required to sort out 3-item content NPs composed of two hyphenated premodifiers. Table 4.9 displays the twenty most frequent hyphenated premodifiers in this set, accompanied by an annexed column with the *matches*⁷⁵ between the 3-item and 2-item NP sets.

Table 4.9: twenty most frequent hyphenated premodifiers in the 3-item content NP set.

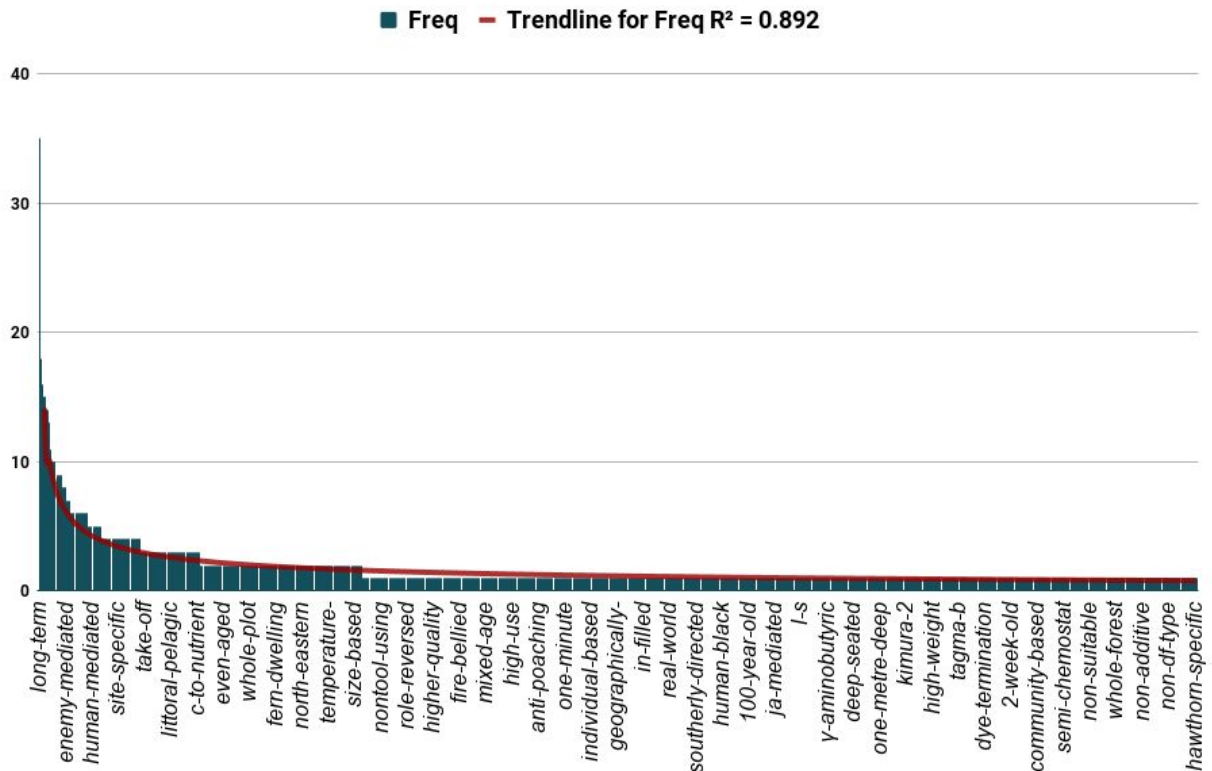
Matches	Top 20 premodifiers	
2-item NPs	Hyphenated premodifier	Freq. (raw)
	<i>among-individual</i>	46
✓	<i>long-term</i>	35
	<i>between-individual</i>	18
✓	<i>old-growth</i>	16
✓	<i>high-value</i>	15
✓	<i>semi-natural</i>	15
✓	<i>non-native</i>	14
	<i>life-history</i>	14
	<i>herbivore-induced</i>	13
✓	<i>within-group</i>	11
✓	<i>land-use</i>	10
	<i>species-specific</i>	10
✓	<i>density-dependent</i>	10
	<i>mixed-effects</i>	9
	<i>between-sex</i>	9
	<i>long-distance</i>	9
	<i>large-scale</i>	9
✓	<i>home-range</i>	8
	<i>temperature-size</i>	8
	<i>enemy-mediated</i>	8

Interestingly, nine of the top twenty hyphenated premodifiers in the 3-item NP set are also highly frequent in the 2-item NP set. This possibly indicates a usage pattern, especially if we consider that some of the more recurrent premodifiers are just as frequent in the 4-item set.

⁷⁵ Premodifiers that coincide across the sets. For instance, *long-term* is a premodifier in 3-item **and** 2-item NPs, but it **does not overlap** (*long-term* occurrences from the 2-item NP set **are not nested** within the 3-item NP set).

Because *among-individual* occurs mostly in one article only, its frequency was adjusted, thus making *long-term* the most frequent hyphenated premodifier in this set and matching the trend observed in the previous set. Examples 4.88 to 4.92 indicate the five most frequent premodifiers, as adjusted for frequency, following Graph 4.2 below.

Graph 4.2: distribution of the hyphenated premodifiers in the 3-item NP set.



Example 4.88

→ In addition, the ingested hooks and their decomposition products may cause **sub-lethal long-term consequences**.

(CoBRA file BC.2016.08.txt)

Example 4.89

→ This study applies DNA profiling to investigate the sources and vectors of new propagules, to detect **illegal human-mediated translocations**.

(CoBRA file BC.2016.04.txt)

Example 4.90

→ In the range of **littoral-pelagic niche use** we also found a positive relationship between the species' trophic position and the range of littoral-pelagic foraging.

(CoBRA file OE.2015.07.txt)

Example 4.91

→ Moreover, when **food C-to-nutrient** (especially C:P) **ratios** are low, growth rates can be inhibited, but due to excessive rather than limiting nutrients

(CoBRA file EL.2017.07.txt)

Example 4.92

→ On the other hand, many broad-ranging species, (...), are typically best conserved through landscape-scale policy mechanisms, rather than **site-specific conservation approaches**.

(CoBRA file BC.2016.07.txt)

Only two NPs have two hyphenated premodifiers in this set: *real-world conservation decision-making*, a single entry, and *alternative life-history trade-offs*, which occurs twice, as seen in Examples 4.93 and 4.94 below. If not for the presence of *real-world* and *life-history*, these premodifiers would have been discarded.

Example 4.93

→ (...) new datasets on biodiversity and ecosystem services are becoming increasingly available at the resolution needed to inform **real-world conservation decision-making**.

(CoBRA file BC.2017.09.txt)

Example 4.94

→ Oviparous and viviparous populations can occur in sympatry in the same environment, making this a unique system for investigating **alternative life-history trade-offs** (...)

(CoBRA file OE.2019.03.txt)

4.3.4 Hyphenated premodifiers in 4-item content NPs

Hyphenated premodifiers in this set largely match the previous ones in frequency. It is surprising to see that double hyphenated premodifiers in a single NP string are not found in the 4-item set. Seeing that these NPs carry four items and may display even more constituents internally, it would be expected to see double hyphenation. However, this was only observed in the 5-item and +6-item NP sets.

Example 4.95

→ **Female-biased regime focal males** spent 56% of their time mounting female targets.

(CoBRA file AB.2018.06.txt)

Example 4.96

→ (...) using range maps digitized from **recent species-specific assessment reports** available at the Public Registry for Species at Risk.

(CoBRA file BC.2019.08.txt)

Example 4.97

→ In order to compare EMF communities in overmature planted set-asides and **ancient old-growth forest communities**, locations were selected that paired forest types in a randomised block design.

(CoBRA file BC.2016.01.txt)

Example 4.98

→ Instead, our results showed that males from **Male-biased experimental evolution regimes** that had evolved under stronger opportunities for sexual selection engaged in less SSB than males that evolved under Female-biased ratios.

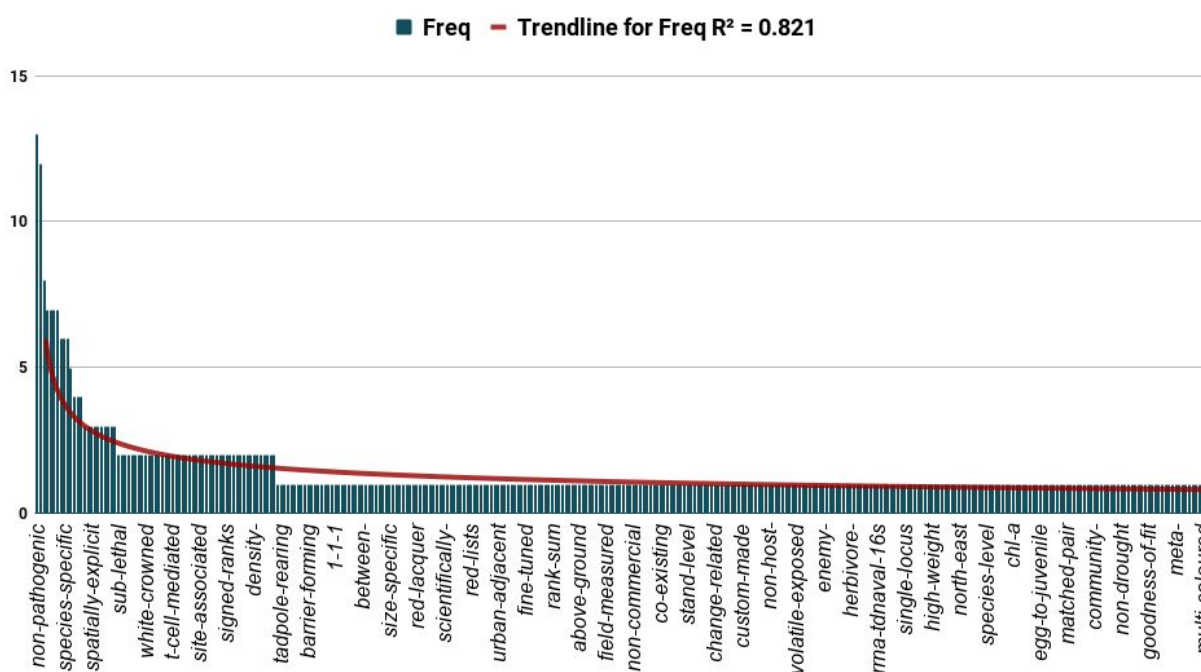
(CoBRA file AB.2018.06.txt)

Table 4.10: twenty most frequent hyphenated premodifiers in the 4-item content NP set.

Matches		Top 20 premodifiers	
2-item NPs	3-item NPs	Premodifier	Freq. (raw)
		<i>non-pathogenic</i>	13
✓		<i>standard-sized</i>	12
✓	✓	<i>high-value</i>	8
		<i>female-biased</i>	7
✓	✓	<i>old-growth</i>	7
		<i>live-aboard</i>	7
		<i>male-biased</i>	7
✓	✓	<i>non-native</i>	6
	✓	<i>species-specific</i>	6
✓	✓	<i>long-term</i>	6
		<i>log-transformed</i>	5
		<i>post-release</i>	4
	✓	<i>among-individual</i>	4
		<i>short-term</i>	4
		<i>two-tailed</i>	3
✓	✓	<i>home-range</i>	3
		<i>spatially-explicit</i>	3
✓	✓	<i>semi-natural</i>	3
		<i>sex-specific</i>	3
		<i>context-specific</i>	3

Irregular nouns and plurals are the most frequent head nouns of this set, with *species* at the top, followed by *rates*, *models*, *data*, *traits*, and *populations*. Semantically speaking, the plural noun heads can welcome a wide range of conceptual information, as if they were some sort of canvas on which an array of semantic loads can be combined. Nouns such as *models* or *rates* require a specification, which we can see by the employment of nouns and adjectives in Examples 4.99 to 4.101, as illustrated in Graph 4.3

Graph 4.3: distribution of the hyphenated premodifiers in the 4-item NP set.



Example 4.99

→ Currently, gaps in research exist for country-wide analyses at a fine resolution that encompass the full set of **biological and socio-economic data** needed to inform conservation decision-making in developing countries.

(CoBRA file BC.2017.09.txt)

Example 4.100

→ These changes have stoichiometric implications: **body carbon-specific ingestion rates** of nauplii are high and typically exceed 100% of body C per day (...)

(CoBRA file OE.2018.07.txt)

Example 4.101

→ We do not provide management, conservation, or recovery plans for specific populations of caribou or grizzly bears, which should rely on **fine-tuned quantitative habitat models** with higher accuracy that are typically limited to smaller geographic areas.

(CoBRA file BC.2019.05.txt)

4.3.5 Hyphenated premodifiers in 5-item content NPs

As expected, the amount of multi-item complex NPs is drastically reduced in this set, as is the number of hyphenated premodifiers. The 918 occurrences initiated by content words, with at least one hyphenated premodifier, have as their second, third, and fourth most frequent hyphenated premodifiers a past participle compound in so-called ‘head’ position: *area-based* and *soil-borne*, for instance. Example sentences 4.102 and 4.103 exhibit these premodifiers.

Example 4.102

→ **Other effective area-based conservation approaches**, as referred to in Aichi Target 11, may be important to ensure effective conservation of species in complex socio-ecological landscapes.

(CoBRA file BC.2016.07.txt)

Example 4.103

→ Here, we hypothesized that plants can distinguish between volatiles of **pathogenic and non-pathogenic soil-borne fungi**.

(CoBRA file OE.2019.09.txt)

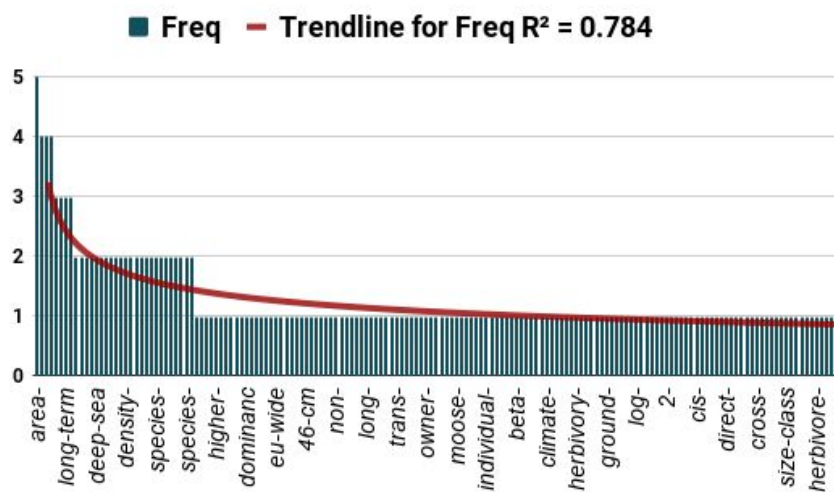
Table 4.11: twenty most frequent hyphenated premodifiers in the 5-item content NP set.

Matches			Top 20 premodifiers	
2-item NPs	3-item NPs	4-item NPs	5-item NPs	Freq. (raw)
			pre-critical	11
			area-based	5
	✓	✓	standard-sized	4
			soil-borne	4
		✓	non-pathogenic	4
			present-day	3
			area-specific	3
✓	✓	✓	long-term	3
			fern-dwelling	3
✓			yellow-gular	2
✓			white-gular	2
			wall-to-wall	2
✓	✓	✓	semi-natural	2
			deep-sea	2
		✓	male-biased	2
			same-sex	2
			sub-meter	2
			mid-elevation	2
	✓		long-distance	2
✓	✓		density-dependent	2

Table 4.11 indicates that one quarter (five premodifiers) of the twenty most frequent hyphenated premodifiers are matches for hyphenated premodifiers in the 2-, 3-, and 4-item content word-initiated NPs. Frequency patterns also match the previous long-tail distribution trend from the other NP sets, as shown in Graph 4.4. This may be interpreted as a sign that the use of premodifiers follows a consistent pattern in the CoBRA, in a mirroring effect in which the most frequent premodifiers tend to coincide across the NP sets.

Because *pre-critical* is used primarily in a single document, its frequency was adjusted in Graph 4.4. This means that while comparatively very reduced in number of tokens, as far as the other NP sets are concerned, the 5-item set has as its most frequent premodifiers three past participles: *area-based*, *standard-sized*, and *soil-borne*. Frequency-wise, the premodifiers are very much reduced in the CoBRA, almost as if they were tiny specs in the greater picture of this corpus. Nevertheless, their importance lies in the fact that they may point to a tendency of complex, compressed NPs to have syntactic compounds as premodifiers.

Graph 4.4: distribution of the hyphenated premodifiers in the 5-item NP set.



Examples 4.104 to 4.107 illustrate four complex NPs from the 2-5-item NP sets, with *long-term* acting mostly in predication, in the capacity of an object. Example 4.108 displays *long-term* in a 6-item content NP, for the sake of comparison.

Example 4.104

→ Our methods can be extended to study cases where **long-term relaxation** occurs via neutral drift.

(CoBRA file EL.2018.08.txt)

Example 4.105

→ These factors can influence **long-term community dynamics**.

(CoBRA file OE.2016.05.txt)

Example 4.106

→ (...) groups of interconnected haplotypes that are isolated from each other could represent cryptic species or **long-term genetically isolated populations**.

(CoBRA file ZS.2018.02.txt)

Example 4.107

→ In this study, we aimed to investigate the interplay between **controlled long-term perceived predation risk**, body size and boldness in the fathead minnow, *Pimephales promelas*.

(CoBRA file AB.2019.04.txt)

Example 4.108

→ **Strong, equitable and long-term social bonds** in the dispersing sex in Assamese macaques.

(CoBRA file AB.2016.05.txt)

4.4 Exploratory observations

4.4.1 The past participle

When dealing with hyphenated compounds in the CoBRA, it seemed rather difficult to not address the use of past participle-formed hyphenations in the NP sets. Not only are these constructions typical of scientific writing (BIBER; GRAY, 2016; BAUER; HUDDLESTON, 2002), they are also fairly productive in the CoBRA (cf. Table 4.12), particularly the regular participial forms.

The idea of viewing *-ed* adjectival suffixes as compounds is not new. Fabb (1998), for instance, calls these constructions *synthetic/verbal compounds*, in which the compound’s head is a derived word formed by a verb and one or more affixes. *Expert-tested* is one such case, as discussed in Jurida (2018), in which *to test* is the verb and *-ed* is the past participle suffix. In Bauer and Huddleston (2002), these are called *verb-centered compound adjectives*, which can be active or passive, semantically speaking. For a full review of synthetic compounds, please see Ackema and Neeleman (2010).

Table 4.12: past participle in hyphenated premodifiers in the 2-5 range of NP sets.

Past participle	2-item NPs	3-item NPs	4-item NPs	5-item NPs	Total
Regular (-ed)	830	296	145	51	1,322
Irregular	64	23	4	5	96
Total	894	319	149	56	1,418

Irregular past participles are naturally included in this group, although as seen in Table 4.12 and Chart 4.6, the regular past participle is remarkably more favored in the CoBRA. This difference, however, is not statistically significant, with a *p*-value of 0.197989 (chi-square test used).

Chart 4.6: past participle in hyphenated premodifiers in the 2-5 range of NP sets.

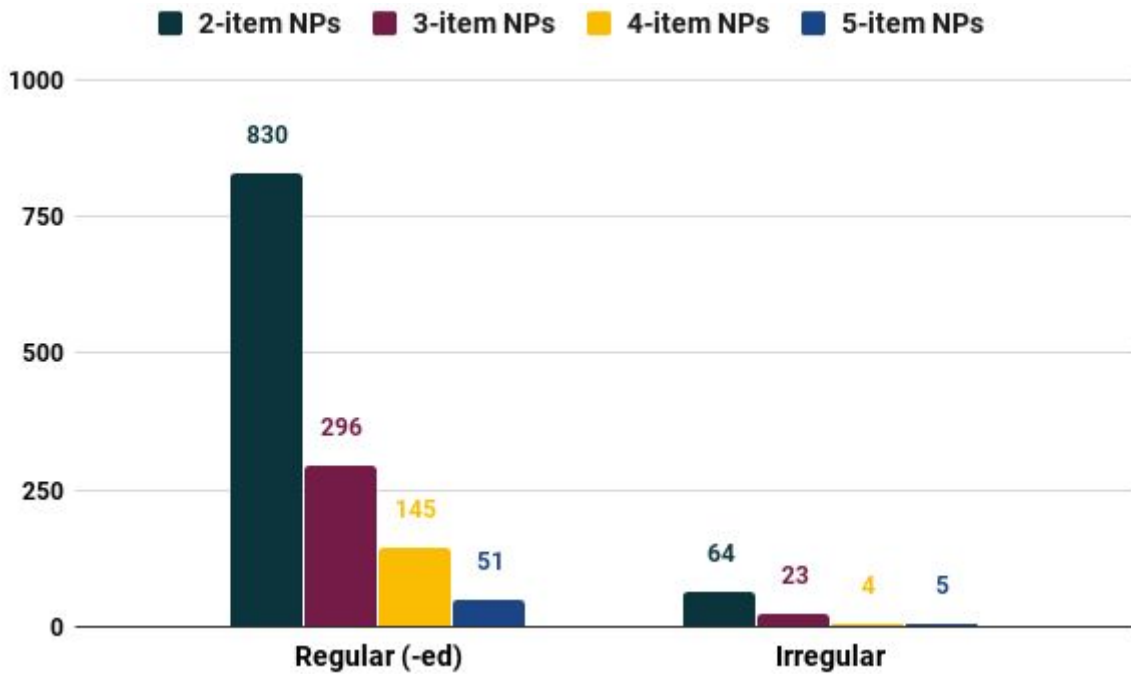
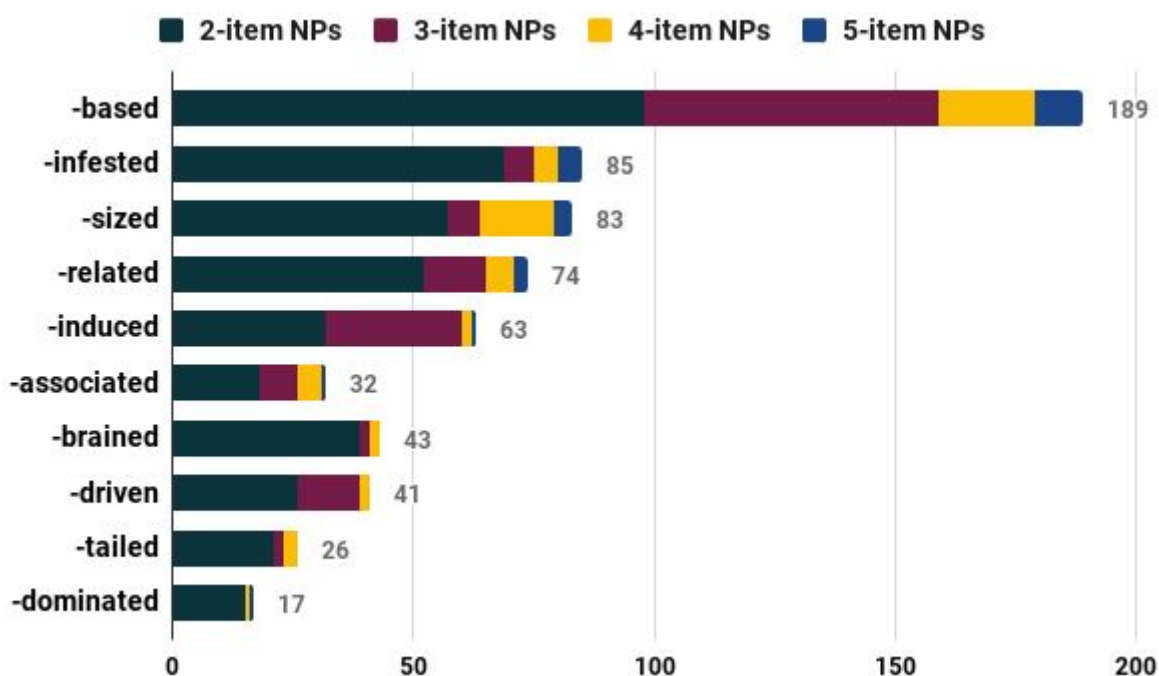


Table 4.13: ten most frequent hyphenated past participles in the 2-5 range of NP sets.

Past participle	2-item NPs	3-item NPs	4-item NPs	5-item NPs	Total
<i>-based</i>	98	61	20	10	219
<i>-infested</i>	69	6	5	5	85
<i>-sized</i>	57	7	15	4	83
<i>-related</i>	52	13	6	3	74
<i>-induced</i>	32	28	2	1	63
<i>-associated</i>	18	8	5	1	32
<i>-brained</i>	39	2	2	0	43
<i>-driven</i>	26	13	2	0	41
<i>-tailed</i>	21	2	3	-	26
<i>-dominated</i>	15	0	1	1	17
Total	427	173	61	25	686

Table 4.13 and Chart 4.6 indicate the ten most frequent hyphenated past participles in the 2-5 range of NP sets. As can be seen, the past participle form *-based* is remarkably more productive than any other verb form across all sets in the CoBRA. We can also see only one irregular past participle in the most frequent list: *-driven*.

Chart 4.7: ten most frequent hyphenated past participles in the 2-5 range of NP sets.



4.4.1.2 Decomposition

One way to determine the morpho-syntactic nature of the *-ed* items in the hyphenated premodifiers is to decompose these elements to the smallest units possible to reach the root of the word, stripping it of its affixes. As an example, *referenced*, in *geo-referenced* is reduced to *refer*, with the derivational and inflectional suffixes *-ence* and *-ed* respectively removed. This leads to ‘to refer’ as the basis of the *-ed* constituent. In this light, the hyphenated premodifier could be notationally visualized as (*Pref+V*).

On the other hand, decomposing *encapsulated* in *micro-encapsulated* leads to a noun as the root: *capsule*, which requires removing the prefix *en-* and the verb suffixes *-ate* and *-ed*. This compound can be morpho-syntactically represented as (*Pref+N*) and derivation plays an important part in this process.

As for the prefixed constituent of the hyphenated modifier, should the same rationale apply, then the opposite process would be required to recompose the prefixed constituent to its full form. This has as much to do with pragmatics as with morphology and etymology, since *geo-* may refer to *geography*, *geology*, *geophysics*. Its assigned meaning largely depends on the context and on the words with which it is combined, despite etymologically referring to or relating to the earth (LEXICO, 2020). Without a closer look at the context, it is difficult to determine more precisely what *geo-* specifically refers to, since *geo-* is somewhat vague.

In contrast, *micro-* needs no further semantic analysis other than the knowledge of its meaning: *of reduced or restricted size* (LEXICO, 2020). Either as a prefix or as an adjective, *micro-* is more transparent and less vague than *geo-*, which probably offsets the semantic and morphosyntactic complexity brought about by the prefix *en-* and the suffixes *-ate* and *-ed* in *micro-encapsulated*.

4.4.1.3 Etymological remarks

The manual decomposition of all past participle-ending constituents in the hyphenated premodifiers was assisted by etymological information from three reference works: *Online Etymology Dictionary* (HARPER, 2020), *Oxford Concise Dictionary of English Etymology* (HOAD, 2000), and *A Dictionary of English Etymology* (WEDGWOOD, 1872).

Table 4.14: etymology for the ten most frequent past participles.

Past participle	Etymology
-based	Latin <i>basis</i>
-infested	Latin <i>infestare</i> ⁷⁶
-sized	Old French <i>sise</i>
-related	Latin <i>relatus</i> ⁷⁷
-induced	Latin <i>inducere</i>
-associated	Latin <i>socius</i>
-brained	Old English <i>brægen</i>
-driven	Old English <i>drifan</i>
-tailed	Old English <i>tægl</i>
-dominated	Latin <i>dominatus</i> ⁷⁸

By decomposing the words to their smallest unit possible, it was possible to access the root, as stated above, which provided a route to better understand the morpho-syntactic basis of the word. From this task, the *-ed* participles were classified as verb-based or noun-based, since these were the categories verified in the NP sets. Table 4.13 lists the ten most frequent *-ed* participles identified in the sets, as shown in the previous subsection. Table 4.14 displays the etymological information for each of these participles.

⁷⁶ Meaning ‘to attack, disturb, trouble’, as based on *infestus* (unsafe, hostile, threatening, dangerous). The sense of ‘swarm over in large numbers, attack parasitically’ was first recorded c. 1600 (HARPER, 2020).

⁷⁷ Used as the past participle form of *referre*, meaning ‘to bring back’ in Latin (HARPER, 2020).

⁷⁸ Used as the past participle of *dominari*, meaning ‘to rule, dominate, to govern’ in Latin (HARPER, 2020).

It could be argued that *-based* and *-associated* have the verbs *to base* and *to associate* as their foundation, as these are more readily recognizable than the etymological root. Should one hold *-based* and *-associated* as such, then verb forms make up the overwhelming majority of the *ed*-ending premodifiers. Regardless of root word class, however, segmenting these *-ed* premodifiers will eventually lead to a participial form. For instance, *fossil-based calibrations* may be decomposed as *calibrations that are based on fossils*. A sentence from the corpus can be seen in Example 4.109 for the hyphenated compound.

Example 4.109

→ In addition, based on nuclear data and a revised set of **fossil-based calibrations**, we estimated the timing of major cladogenetic events in *Sorex*.

(CoBRA file ZS.2018.07.txt)

An alternative reading would be *calibrations that have fossils as their basis*. However, this sounds unlikely once we consider that one of the main roles of compounding is economy. Therefore, one could infer that segmentations are expected to be as uncomplicated as possible whenever they are needed. Overcomplicated segmentations would be counterproductive when spelling out the compound for easier understanding.

The relevance of examining the origin of these words is more related to the exercise Brazilian learners/speakers of English have when working with word formation, seeing that a major issue these EAP students face is precisely word formation, as I have witnessed in the past two years. Hence, it might be helpful for them to add etymological examinations to their learning routine as they work on their language skills, by starting with their first language (cf. MATTOS, 2018 for examples). Looking into the etymological aspects of one's language, be it their first or additional language, may bring about more awareness as to its current uses.

4.4.2 Prefixation

The reason for including prefixation in this chapter is simple: prefixation is one way of word formation in English, whereby a lexeme is bound to the base word and cannot be used on its own. *Non-* in *nonresponsive parasitoids* and *non-diapausing pupae* is an example, as is *geo-* in *geo-referenced data* and *geolocator activity data*⁷⁹ (Examples 4.110 to 4.113). Despite

⁷⁹ Bauer and Huddleston (2002) distinguish affixation and combining forms. In this thesis I treat both under 'prefixation', seeing that just like prefixes, most combining forms are generally bound.

entailing a process different from compounding, prefixes may be an element of hyphenated compounds, as one of the compound's constituents (Example 4.114). Prefixes are also used in neoclassical compounds (cf. BAUER; HUDDLESTON, 2002).

Example 4.110

→ Bars represent the percentages of parasitoids that choose each of the two odor sources; N number of parasitoids that make the choice, NR number of **nonresponsive parasitoids**.
(CoBRA file OE.2015.04.txt)

Example 4.111

→ Diapausing individuals had greater pupal fat mass than **non-diapause pupae** across the entire temperature range used.
(CoBRA file OE.2015.08.txt)

Example 4.112

→ (...) but spatial analysis should be applicable to any aerial survey datasets where **geo-referenced data** are routinely gathered, marine or terrestrial.
(CoBRA file BC.2018.03.txt)

Example 4.113

→ Trip activity budgets were determined based solely on **geolocator activity data**.
(CoBRA file BC.2015.09.txt)

Example 4.114

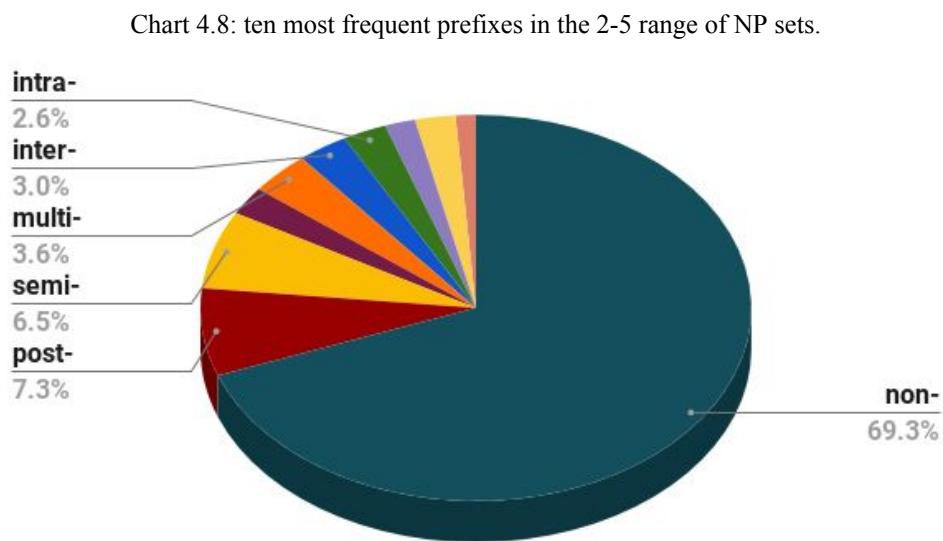
→ Furthermore, it is likely that integrated management strategies within the Mediterranean, incorporating both **area-based and non-area-based measures**, will be required to tackle the range of pressures currently threatening the Balearic shearwater.
(CoBRA file BC.2015.09.txt)

Table 4.15: ten most frequent prefixes in the 2-5 range of NP sets.

Prefixes	2-item NPs	3-item NPs	4-item NPs	5-item NPs	Total
non-	350	82	46	22	500
post-	37	14	10	1	62
semi-	33	21	4	4	62
pre-	12	12	1	13	38
multi-	18	10	4	5	37
inter-	15	4	2	-	21
intra-	13	3	-	1	17
sub-	9	4	4	2	19
eco-	12	1	1	-	14
self-	6	3	-	1	10
Total	505	154	72	49	780

Prefixation is a productive device in the content word-initiated NP sets, with the prefix *non-* overwhelmingly more favored: 500 hyphenated premodifiers rely on it. The difference in frequency between *non-* and the four most recurrent prefixes is considerable, as illustrated in Table 4.11. The content word-initial sets employ a total of 21 hyphenated prefixes/combined forms, from more explicit (*post-*) to more vague ones (*geo-*).

Other prefixation occurrences include the *un-*, as in *un-polluted* [sic] *lakes*, and *anti-* in *anti-predator* and *anti-parasite behaviours* (Examples 4.115 to 4.117). Chart 4.8 depicts the figures from Table 4.15.



Example 4.115

→ Much of the critiques are aimed at the notion that contingent valuation tends to overestimate the value of non-market goods (i.e. clean air, **un-polluted lakes**, etc.).

(CoBRA file BC.2017.07.txt)

Example 4.116

→ Thus, by exhibiting **variable anti-predator behavior**, parasitoids can partly or fully avoid the potential disruptive impact of IGP.

(CoBRA file OE.2015.03.txt)

Example 4.117

→ Adaptations (...) may occur at the individual level, such as increased immune investment (...), or may be expressed at the group-level, including territoriality (Loehle 1995), social sub-structuring (Stroeymeyt et al. 2014) or other **specific anti-parasite behaviours** that benefit the group (...)

(CoBRA file EL.2015.10.txt)

4.5 Discussion

Ultimately, hyphenated premodifiers in complex NPs take the function of adjective by specifying and assigning semantic and pragmatic properties to NP elements and/or heads. In accordance, parsing *risky mate-searching behavior* in Stanford Core NLP (MANNING et al., 2014), renders: *(NP (JJ risky) (JJ mate-searching) (NN behaviour))*), with *mate-searching* correctly functioning as an adjective that modifies *behaviour*. Example 4.118 displays the NP.

Example 4.118

→ However, males maturing at a smaller size as a result of more rapid development could have a distinct advantage when the juvenile period is associated with high mortality rates, as may occur when males undertake **risky mate-searching behaviour**.

(CoBRA file BC.2017.09.txt)

As for measurement, what these results seem to suggest is that, as posited by Berlage (2014), measuring NP complexity based solely on length may be misleading. Short complex NPs composed of one premodifier and a head may in fact contain three or more constituents due to hyphenation, which adds to the already complex interrelationships between the NP constituents. That is, by means of hyphenation, constituents can be added in premodification, condensing and increasing the semantic-pragmatic and syntactic NP loads.

4.5.1 The role of semantic and grammatical transparency

Semantic transparency seems to be very influential in the (somewhat) straightforward semantic-pragmatic character of the hyphenated premodifiers in the content NPs extracted for this thesis. Either as prefixed words (*non-native*), novel compounds (*animal-dispersed plots*) or fixed expressions (*long-term*, *take-off*, *problem-solving*), these highly frequent hyphenated premodifiers are relatively straightforward in terms of semantics and grammar, individually or in association with other words.

Long-term, for instance, contains two content items that exist separately in English. As separately held, *long* and *term* are an adjective and a noun that convey quite simple meanings. As a single unit, *long-term* acquires *adjective* status only, functioning as such in the corpus, as seen in Example 4.105: *These factors can influence long-term community dynamics*. The meaning of the individual items does not change and the presence of the hyphen may indicate to the reader a more transparent semantic association.

Some of these shorter, more transparent compounds may be semantically regarded as a single unit due to lexicalization and grammaticalization. It seems to be the case of *long-term*, which appears to be undergoing a grammaticalization process. The same could be argued for the oft-used *data-driven*, for example, which has become ubiquitous in scientific writing. This might explain the high frequency of *long-term* across the NP sets in the CoBRA and the use of *data-driven* more generally. Example 4.119 below shows an NP with *data-driven* as one of its premodifiers.

Example 4.119

→ Worldwide, > 12,000 IBAs have been identified by the BirdLife International Partnership, using **standardized data-driven selection criteria** based on threat and irreplaceability.

(CoBRA file BC.2016.07.txt)

Because compounds are highly productive in English, novel combinations are easily created. Novel compounds may thus present some difficulty for less proficient language users. Understanding NPs with hyphenated compound premodifiers such as *animal-dispersed plots* or *infection-induced mortality* may require more effort, perhaps by segmentation: *plots which are/were dispersed by animals* and *mortality which is/was induced by infection*, or even by morpho-syntactic decomposition [(N-V)N].

Despite not being crystalized terms, the NPs with hyphenated compound premodifiers in these examples have retained the **core meaning** of the individual constituents. This possibly facilitates comprehension, particularly for Brazilian speakers of English, who might find the similarity in form and meaning reassuring. Examples 4.120 and 4.121 illustrate these NPs. We can see that these hyphenated premodifiers, as shown in their original context, are employed towards the end of the sentence. This may also help less proficient readers (or even readers not entirely familiarized with the content) to make out the meaning of the compounds.

Example 4.120

→ There was more uncertainty in estimates of peak time for growth in **wind-dispersed plots** than **animal-dispersed plots** (...)

(CoBRA file EL.2019.02.txt)

Example 4.121

→ Thus, the loss of intestinal integrity is a good predictor of imminent death in infected flies and explains most of the difference in **infection-induced mortality** (...)

(CoBRA file EL.2015.02.txt)

For Gagné and Spalding (2006, 2010), semantic transparency plays a fundamental role in how compounds are processed. This thesis does not aim to psycholinguistically verify how language users mentally process complex, hyphenated compound NPs, which could only be done with psycholinguistic tests. Still, perhaps we could say that more transparent semantic relations between constituents lead to easier comprehension, provided the constituents retain their core meaning in the NP.

In this sense, the participles *-ed* and *-ing* offer the grammatical transparency needed to more easily process syntactic compounds. The same applies to the prototypical adverb suffix *-ly*. In the NPs discussed in this chapter, the suffix *-ly* may be very helpful to successfully read long, complex NPs with sequential adverbs as premodifiers to a noun head. Such uses may also be indicative of the role morphology has in compounding, as well as the interdependence of form and meaning in language use.

4.5.2 Hyphenation vs. non-hyphenation

Biber and Gray (2016, p. 230-231) suggest that hyphenation in compounds may be in decline in academic registers: “it has recently become common in science writing for these structures to occur without the hyphen – a shift towards an even less explicit representation of the meaning relationship among elements”. Feist (2012) is of the same opinion. However, not much hard evidence is shown in return.

While the examples listed by Biber and Gray (2016) correspond to structures found in the CoBRA, an examination of the content NP sets indicates that non-hyphenated compounds do not outnumber hyphenated ones in the CoBRA. For hyphenated uses of past participles, for instance, Table 4.16 indicates raw counts for hyphenated and solid past participle forms in the entire corpus for the most frequent uses, in which a statistically significant difference is seen (chi-square test used).

Table 4.16: hyphenated vs solid five most frequent past participles.

	Hyphenated	Solid
<i>based</i>	372	73
<i>infested</i>	169	14
<i>sized</i>	147	11
<i>related</i>	81	2
<i>associated</i>	31	2
Total	800	102
p-value		0.000066

The trend discussed in Biber and Gray (2016) might be better translated as *competing variants*, since alternatives may co-occur. Other cases in the CoBRA may also be invoked in support of this view. In the 3-item content word-initiated NPs, for instance, *post-* is employed in hyphenated (*post-nesting migratory pathways*), solid (*posthatching developmental stress*), and open (*post linear disturbance*) compounds. The *competing variants* view is also favored by Sanchez-Stockhammer (2018) in her corpus-based work of English compound spelling.

4.5.3 Dependency links

In regard to *-ed* and *-ing* participles, more specifically, it would be difficult to assess their non-hyphenated uses without a closer look at the dependencies between the participle and its preceding constituent, as well as the semantic-pragmatic properties therein established. For instance, the more clear-cut link between *litter* and *trapping* in *litter-trapping epiphytes* and *rhizobacteria* and *colonized* in *rhizobacteria-colonized plants* does not seem to be readily visible in *static direct sampling bat detectors*.

Length may explain this difference. The shorter NPs contain fewer constituents, which can facilitate dependency identification. However, this is not the only explanation. In Pirrelli, Guevara, and Baroni (2010), compound interpreting, which is the case of most past participle forms in the CoBRA, is viewed as highly contextually-sensitive.

Example 4.122

→ Bird's nest ferns (*Asplenium* spp.) are **litter-trapping epiphytes** that form a two-way by-product mutualism in primary forest with their ant inhabitants (...)

(CoBRA file OE.2015.06.txt)

Example 4.123

→ From the herbivore's perspective, increased levels of these plant nutrients can enrich their diet (...), supporting our observations of increased *M. brassicae* performance in **rhizobacteria-colonized plants**.

(CoBRA file OE.2015.04.txt)

Example 4.124

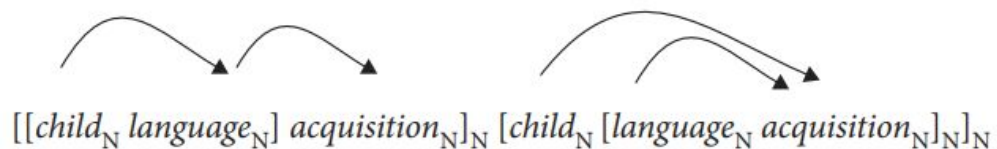
→ Data from **static direct sampling bat detectors** (...) were also incorporated into the bat HSM.

(CoBRA file BC.2015.10.txt)

Computationally speaking, it is dependency, rather than constituency, that governs the process of assigning semantic properties more efficiently, and even then several parsing issues may arise (cf. PIRRELLI; GUEVARA; BARONI, 2010). The crux of dependency

parsing is how the machine understands the dependency links, that is, whether the dependency between the word and the head noun is that of arguments or modifier, as explained in Pirrelli, Guevara, and Baroni (2010). Figure 4.6 reproduces possible links for *child language acquisition*, with *[[child language] acquisition]* read as a premodifier in left-branching, and *[[child [language acquisition]]]* as an argument link in right-branching⁸⁰

Figure 4.6: left- and right-branching dependencies for *child language acquisition*.



Source: Pirrelli, Guevara and Baroni (2010, p. 277)

An attempt was made to dependency-parse the complex NPs using Stanford Core NLP (MANNING et al, 2014), with tweaks to tokenization, a mandatory step prior to parsing and chunking. Unfortunately, this effort did not bear any fruit: the software did not understand the dependency links to their full extent and processed the sample as mostly left-branching. In reality, short, complex NPs and N-N compounds do take left-branching processing, as large corpora indicates, both on computational and cognitive grounds (cf. PIRRELLI; GUEVARA; BARONI, 2010), but left-branching processing does not always fit long NPs.

A second reason for the more straightforward connections in *litter-trapping epiphytes* and *rhizobacteria-colonized plants* is hyphenation, that is, because of the hyphen, it is easier to understand the link between the constituents. Veering towards a more functional outlook, as seen in Kösling and Plag (2009) and Nunberg, Briscoe, and Ruddleston (2002), hyphenation is key in signaling the connection between constituents, which can be particularly useful when dealing with more ambiguous constructions.

We can see from *rhizobacteria-colonized plants*, for instance, that the hyphen signals a link between *rhizobacteria* and *colonized* and directs the interpretation. In this case, the reader is not supposed to read *rhizobacteria* as an agent that *colonized plants*. Evidently, the previous language items in the sentence also help the reader in predicting the meaning of this NP. Still, the hyphen indicates unity between the constituents in a premodifying position.

⁸⁰ Branching direction is often grounded in the notion of headedness. See Polinsky (2012) for a review.

As previously observed, hyphenated words are computationally processed as a single unit, with a series of linguistic and computational factors coming into play in this process (cf. PIRRELLI; GUEVARA; BARONI, 2010). Despite being computationally read as a sole unit, complex hyphenated NPs may not be humanly processed as such. This is a highly debatable issue since no consensus has been reached in the matter, that is, psycholinguists are unsure as to whether or not human processing requires decomposition (for evidence pro- and against, cf. BUTTERWORTH, 1983; LIBBEN, 1998; ZWITSERLOOD, 1994; LIBBEN et al., 2013).

In a pro-decomposition/segmentation perspective, segmenting hyphenated compounds may be viewed as similar to decompressing non-hyphenated premodified complex NPs. *Risky behaviour*, for example, may be decompressed as *behavior that is risky* while *arid zone bird species* can be spelled out as *a species of birds that is native to arid zones*. It is as though the hyphenations were a mini-phrase embedded in a complex NP, in a Russian doll-like effect.

This embedding view is maintained by Pirrelli, Guevara, and Baroni (2010, p. 272) for noun-noun compounds: “the semantic relation between the two constituents of the compound is akin to the modifier-head relation between the two elements of the corresponding phrase”. By looking at the hyphenated premodifiers as mini-phrases in which a modifier-head relation is established, a subordination relation is presupposed, in which the constituents of the so-called mini-phrases are hierarchically interconnected.

4.5.4 Hyphenation as a compression strategy

As shown by the above-reported results, of note in the extracted NP chunks is the use of hyphenated premodifiers in a wide range of morpho-syntactic combinations. For example, 2-item content word-initiated NPs may present content word-initiated patterns: *deeper-water sharks* [(ADJ-N)N] and *land-bridge connections* [(N-N)N], as well as functional and content word patterns: *below-target representation* [(Prep-N)N] and *3-month intervals* [(DT-N)N].

Example 4.125

→ For sharks, highest vulnerability was identified for small coastal and continental shelf species (...) and, surprisingly, **deeper-water sharks**.

(CoBRA file BC.2019.04.txt)

Example 4.126

→ Both species likely maintained gene flow between Trinidad and the mainland through **land-bridge connections** at low sea level stands during the Pleistocene.

(CoBRA file ZS.2019.05.txt)

Example 4.127

→ In the US, previous studies have similarly found **below-target representation** of the country's diverse ecological systems within the protected areas.

(CoBRA file BC.2018.08.txt)

Example 4.128

→ Here, we address the colonization processes at the molecular level (...) in the shallow-water Mediterranean collected over a year at **3-month intervals**.

(CoBRA file ZS.2017.03.txt)

Content words and participial forms also seem to yield productive arrangements, as in *risky mate-searching behavior* [(Adj)(N-V)N] and *forest-dwelling insectivores* [(N-V)N], in which *mate-searching* and *forest-dwelling* function as adjectives modifying the head nouns.

Example 4.129

→ (...) males (...) could have a distinct advantage when the juvenile period is associated with high mortality rates, as may occur when males undertake **risky mate-searching behaviour**.

(CoBRA file OE.2019.10.txt)

Example 4.130

→ The nearby islands of Rota and Saipan have intact bird assemblages when compared to Guam, including **forest-dwelling insectivores** such as Rufous fantails (...)

(CoBRA file OE.2018.01.txt)

These NPs exemplify the morphosyntactic and semantic creative potential hyphenated compounds may offer in compacting information in English. Hyphenated premodifiers may condense even more information than complex NPs already do, especially with abbreviations and acronyms. Examples 4.131 to 4.40 illustrate this rationale:

Example 4.131

→ We used mark-recapture data collected during the annual fisheries assessments to reconstruct growth history of **PIT-tagged Yaqui catfish**.

(CoBRA file BC.2017.06.txt)

Example 4.132

→ The increased reliability of species identifications provided by **qPCR-based DNA analyses** creates opportunities for developing more effective methods for sampling rare carnivores.

(CoBRA file BC.2019.01.txt)

Example 4.133

→ (...) the divergence process may not only be linked with the geographic isolation, but also related with the dynamic colonization during **climate-induced distributional shifts**.

(CoBRA file ZS.2019.09.txt)

Example 4.134

→ **CO2-driven ocean acidification** is considered a major threat to marine species worldwide, and the process is especially accelerated in the CCS.

(CoBRA file EL.2016.06.txt)

Example 4.135

→ (...) **scent-based offspring discrimination mechanisms** are known from many vertebrate species (...)

(CoBRA file AB.2016.03.txt)

Example 4.136

→ Taken together these findings suggest that poison frogs may form and use **flexible spatial map-like representations** of their surroundings.

(CoBRA file AB.2016.07.txt)

Example 4.137

→ Food shortage during juvenile development induces **species- and sex-specific adult body size plasticity**.

(CoBRA file AB.2015.04.txt)

Example 4.138

→ (...) the prevailing winds during spring and summer bring **cold, deep, nutrient-rich, high-CO2 seawater** into the nearshore environment in the CCS.

(CoBRA file EL.2016.06.txt)

Example 4.139

→ For example, voice pitch predicts several objective social dominance outcomes in women and men such as political election results (...) and job prestige in **highly stereotypically female-oriented leadership positions**.

(CoBRA file AB.2019.06.txt)

Example 4.140

→ For example, voice pitch predicts several objective social dominance outcomes in women and men such as political election results (...) and job prestige in **highly stereotypically female-oriented leadership positions**.

(CoBRA file AB.2019.06.txt)

4.5.4.1 Implications for EAP teaching

The research conducted does not hold normative implications. It does, however, point to usage patterns as verified in the CoBRA and in other corpus-based, empirical research such as those of Sanchez-Stockhammer (2018) and Dutra et al (2020). Accordingly, the purpose of the observations made in this subsection is to offer some considerations as to how to address the phenomena under investigation in a pedagogical way.

In languages that favor post-modification, such as BP, NP patterns differ from English, and hyphenated premodifiers often form prepositional phrases by retaining the grammatical classes of the premodifiers (*deeper-water sharks* = *os tubarões de águas mais profundas*, *below-target representation* = *representação abaixo da meta*, with two PPs as part of the NP). Table 4.17 illustrates the morphosyntactic notation for examples from the CoBRA and their possible corresponding segmentations in BP, which require rough translations.

Table 4.17: hyphenated premodifiers and their corresponding segmentation in BP.

Notation	Hyphenated complex NP	Corresponding segmentation in BP
[(ADJ-N)N]	<i>deeper-water sharks</i>	<i>tubarões de águas mais profundas</i>
[(DT-N)N]	<i>3-month intervals</i>	<i>intervalos de três meses/trimestrais</i>
[(Prep-N)N]	<i>below-target representation</i>	<i>representação abaixo da meta</i>
[(N-V)N]	<i>forest-dwelling insectivores</i>	<i>insetívoros que vivem na floresta</i>
[(N-ADJ)N]	<i>feature-specific representation</i>	<i>representação de traço específico</i>

As might be inferred from these 2- and 3-item NPs, for Brazilian learners/speakers of English, the process of compounding or compound interpreting, hyphenated or not, may entail a ‘competing’ processability in which the language user is faced with a syntactic pattern quite different from his/her mother tongue: left-branching nominal groups. In Table 4.17 above, the complex NPs in English have their hyphenated premodifiers to the left. BP, on the other hand, displays the opposite pattern as that of English.

English is indeed mostly a right-branching language, with head-initial constructions as *males maturing at a smaller size as a result of more rapid development could have a distinct advantage* (CoBRA file *OE.2019.10.txt*). This language, however, has retained left-branching aspects of its Germanic roots, particularly (but not exclusively) in NPs. The premodification direction in *a smaller size*, *more rapid development*, and *a distinct advantage* indicates left-branching preference.

Romance languages such as BP, quite differently, tend towards more right-branching, post-modification uses, which is why compounding interpreting may require decomposition from Brazilian speakers of English, particularly for less proficient ones. As will be discussed, the pattern favored in BP can be pedagogically contrasted to English language preferences. As such, rough translations may be needed.

4.5.4.2 Trends in hyphenation

Despite the variation mentioned in Biber and Gray (2016) and Feist (2012), and seen in some examples in this thesis, the overall preference when it comes to complex NPs is to employ hyphenation, as verified by Sanchez-Stockhammer (2018). Table 4.18 is an adaptation of this researcher’s list of corpora-verified hyphenation preferences in English compounds, alongside examples from the CoBRA.

Table 4.18: hyphenation preferences in English compounds.

Hyphenation preferences in English compounds	CoBRA examples
<i>Length difference exceeds 1:2 in letters</i>	soft-sediment macrobenthic organisms
<i>Very large frequency difference between constituents</i>	a rock-dwelling species-rich genus red-lacquer coated carbon steel
<i>Compound ends in suffixes -ing, -ed or -er</i>	fisheries- induced mortality
<i>First constituent is a functional word</i>	one -liter eDNA samples, off -highway vehicles
<i>Second constituent: adjective, verb, adverb or functional word</i>	apple- infesting flies extant finch- like species
<i>Most other compounds beginning with the same first constituent are hyphenated</i>	long-term consequences similar long-term data analysis
<i>Most other compounds ending with the same second constituent are hyphenated</i>	strict divergence time- based criteria dissimilarity- based community analysis
<i>constituents are identical</i>	dyadic male-male grooming time
<i>First constituent is a verb (glow-worm)</i>	the six split -plot treatment combinations
<i>First constituent is inflected (broken-down)</i>	Wilcoxon matched -pairs signed -ranks test
<i>content and functional constituents combined</i>	drop-in traps / rain-out effects

4.5.4.3 Pedagogical suggestions

For Brazilian university students who have to read and write in English, especially due to the internationalization growth in publicly-funded higher education institutions (FINARDI; FRANÇA, 2016; SARMENTO et al, 2017), compressed and less explicit language may be very challenging. In addition to specific domain knowledge, Brazilian students and academics also need the necessary English language knowledge to effectively process – or perhaps segment – compressed nominal structures.

Contrastive analysis between hyphenated compounds and complex NPs produced in English and BP may prove helpful when tackling NP complexity at a pedagogical level. One example is contrastive segmentation via rough translations, which may lead learners to reflect upon distinct language patterns in their L1 and L2. Such exercise may also help learners with language awareness.

For these types of pedagogical applications, the main goal is reflecting upon language use, which invariably requires focusing on the more pragmatic, contextual aspects of language in a general sense. This is mainly because, as stated in Pirrelli, Guevara, and Baroni (2010, p. 273), “subtle differences in translating a compound from one language to another” may arise. As these authors contend, effective processing is “often based on recurrent, quasi-lexicalized interpretive schemata” (*ibid*).

Effectively understanding complex, compact expressions, such as the NPs discussed in this research, requires a high degree of familiarity not only with the English language as it is generally employed in academic settings. Perhaps more importantly at a certain proficiency stage is becoming more aware – and, naturally, more familiarized – of the meaning potentials behind complex, compressed structures. In practice, this can be translated as exercising word formation through derivation and compounding.

What the results reported in this chapter possibly mean for language teachers is that, given the challenges scientific writing may present, it might be more productive to work with compound and hyphenated compression as early as the A2 level⁸¹. As indicators of language proficiency, complex NPs and word-formation via derivation are part of the EAP courses held at UFMG (cf. Sá, 2019). Compounding could also be incorporated into EAP teaching.

To that end, this thesis offers a set of practice exercises focused on complex NPs and compounds (Attachment A), in an example of data-driven learning⁸² (DDL). For the sake of clarification, DDL refers to the pedagogical application of corpora through concordancing searches to identify language patterns by observing natural and empirical data. In addition to generating frequency lists from corpora, concordancers such as AntConc (ANTHONY, 2019) allow the search for expressions that can be used as input for classroom teaching activities, as the attached activity shows.

⁸¹ Levels follow the Common European Framework of Reference for Languages – CEFR. www.coe.int/lang.

⁸² For a review of DLL, see Johns (1991), Johns and King (1991), Paker and Ozcan (2017), and Chujo, Anthony and Oghigian (2009). For DDL applications of complex NPs and word formation at the EAP courses at UFMG, see Sá (2019), from which an example has been attached (Attachment B) to this thesis.

Among the reasons for adopting corpus-driven pedagogy is the current availability of large amounts of carefully compiled language data, which can facilitate mapping patterns and exceptions, thus providing language learners and users with the possibility of observing more closely natural language in use. This in turn may avoid analyses based on artificial/biased data (SARDINHA, 2004). As a standard practice, such an approach was also taken in Dutra et al. (2020), following the compilation of specialized written corpora in Chemistry, Biology, and Applied Linguistics.

4.5.5 Limitations

The present thesis focuses on specialized use of language in English writing. As such, it should not be taken as reference for more general uses of this language, nor should it be taken as an exhaustive study of NP complexity, noun premodification, or hyphenation in NPs. While this thesis may offer insights into such phenomena, only a broader study mapping and correlating these phenomena could produce more generalizable statements, particularly if it is done in contrast to general corpora.

Because my primary concern is premodification, post-modification was not examined. Its importance, however, should not be understated. In future research, it may be productive to cross-check pre-modification and post-modification in the CoBRA, isolating the complex NPs that display both modification patterns, and inspecting PPs separately. From this, hyphenated premodifiers could be analyzed syntactically and semantically.

Another interesting possibility would be to establish the dependency links between the premodifiers from 3- to +6-item NPs. Such a task will inevitably require manually tagging or annotating hyphenated constituents, which can then be employed to inform machine learning practices. This means that an algorithm would have to be devised for automated dependency parsing, seeing that current NLP software cannot process these links effectively in standard mode, that is, without more specific input.

As much as possible, this thesis attempted to establish some statistical interpretation of the extracted data. However, more sophisticated statistical tools should be used to provide the data with precise and reliable quantitative verification. Statistical treatments may be done in R (R CORE TEAM, 2018), based on specific packages in the software.

5. CONCLUSION

While reporting on the quantitative part of the NP extraction, hyphenated premodifiers stood out. In an exploratory manner, this thesis attempted to shed light on premodification in complex NPs, particularly in complex, hyphenated elements in premodifying position. Seeing that hyphenated items are computationally (and perhaps even cognitively) read as a single unit, probing into their morphosyntactic and semantic features seemed fitting.

Whether language users process hyphenated compounds as a single unit is beyond the scope of the present research. Psycholinguistic experiments indicate that concatenated, solid compounds (*sunflower*, *offspring*) may or not be processed as a whole⁸³ and that hyphenations are most likely processed differently (BERTRAM et al, 2011; INHOFF et al, 2008; JUHASZ; INHOFF; RAYNER, 2005).

In agreement with Biber and Gray (2016), Gray (2015), Biber et al (1999), Dutra et al (2020), and Pirelli, Guevara, and Baroni (2010), this research confirms the nature of scientific research writing as being more compressed and less explicit, grammatically and semantically. Premodification in complex NPs in the RAs from this corpus often makes use of compression devices such as hyphenation and acronyms.

Overall, 12,747 uses of hyphenation were identified in the CoBRA, of which 5,789 are +2-item content NPs with hyphenated premodifiers, that is, NPs containing two or more items in which the first item is a content word and one of the NP constituents is hyphenated. Some of these hyphenated premodifiers are fixed, crystallized expressions, which tends to facilitate understanding.

In example 5.1, *by-product* can be viewed as one of these crystallized expressions. It is employed as the second premodifier and third item in this 3-item complex NP initiated by a function word. Seeing that this sentence contains four other complex NPs, the use of a more established, low-processing language choice may help readers to fully associate the concepts conveyed in each NP. It may also aid the collective, overall reading of the sentence, that is, the full understanding of the NPs as interconnected pieces in this sentence.

⁸³ For evidence pro- and against decomposition, see: Butterworth (1983), Libben (1998), Libben et al. (2013), Zwitserlood (1994). It is generally accepted that the need for decomposition increases as semantic transparency decreases (LIBBEN; JAREMA, 2006) and that visual representation plays a role in this process.

Example 5.1

→ **Bird's nest ferns** (*Asplenium* spp.) are **litter-trapping epiphytes** that form **a two-way by-product mutualism** in **primary forest** with **their ant inhabitants**.

(CoBRA file AB.2019.06.txt)

Critical in regard to several remarks made in this thesis, though no less important, are Jurida's (2018) observations about compounding. For this author, understanding a word is not contingent on being "aware of how it is constructed, or whether it is simple or complex, (...) whether or not it can be broken down into two or more constituents" (p. 158). Rather, for novel words, this relates more to being "able to use a word which they [we] find new if they [we] learn the new word together with objects or concepts it denotes" (*ibid.*).

In agreement with this view, a minor clarification may therefore be made. As stated in chapter two, while the methodological steps undertaken in this research favored context-free grammars, the perspective to which this thesis subscribes in regard to language use is nothing short of contextualized. In this sense, the use of hyphenated premodifiers in the production or interpretation of complex NPs can only be properly inspected and effectively understood via a closer look at the context in which they occur.

Additionally, the pedagogical suggestions made in chapter four and displayed in both attachments A and B, have as their aim a more 'educational' way of looking at the phenomena approached in this thesis. The suggestions, as already mentioned, are also in line with the CL principles and methodologies for language analysis, which is the ultimate goal of the activities included in this research.

As stated in Leung and van der Wurff (2018, p. 1), corpus data have been increasingly employed in NP-related investigations, thus resulting in "more refined and qualitatively better descriptions and explanations of NP facts", for both synchronic and diachronic examinations. Corpus data and methodology yield studies viewed as "examples of the empirical turn at work in NP-land" (*ibid.*), the importance of which had already been mentioned in Jucker (1993) and is equally stressed in Adamson and González-Díaz (2009).

Because CL research is anchored in large bodies of authentic linguistic data collected and examined via strict methodological procedures (SARDINHA, 2004; SINCLAIR, 2005; McENERY; HARDIE, 2012; GRIES, 2009; LÜDELING; KYTÖ, 2008; DAVIES, 2015), a more reliable picture of how languages are used can be drawn from such studies.

Specifically for noun premodification in complex NPs in RAs, corpus-based analyses such as the one carried out in this thesis may help applied linguists, EAP course designers and teachers in better understanding lexico-grammatical patterns in a highly specialized academic register/genre central to advancement in academia. EAP courses can update course curricula and/or syllabi and enhance classroom and material development practices by adopting CL inquiries of kind herein produced.

As previously mentioned, my motivation for developing this research heavily lies on gaining further insight into complex lexico-grammatical patterns in English scientific writing, especially in regard to more advanced-level constructions that might pose issues to Brazilian learners/speakers of this language. A deeper understanding of such features provides me with better tools to tackle language issues in the classroom, while also equipping me with a deeper, more comprehensive perception of the specificities of research writing.

As Hyland (2016, p. 123) expertly explains it: “the linguistic patterns of texts point to contexts beyond the page or screen, implying a range of social constraints and choices”. That is, the language choices found on the page reflect and/or are guided by socio-cultural practices that quite often transcend the formal properties of the text.

A look into the lexico-grammatical uses of the English language in specific academic contexts is one way of gaining insight into those practices. Analyses focused/ directed at more specific, naturally-occurring specialized texts may provide academics with a deeper and better understanding of the research writing conventions from specific communities of practice, a point made in Reis and Santos (2017).

This was certainly the case while writing this thesis, in which I had the opportunity to actively compare and contrast English language uses in the area of Biology with those of my field of professional/academic expertise. It is precisely this type of linguistic exercise that I hope to develop further in my professional practice, in and outside the classroom. It is also the main takeaway for language teachers and academics who may find their way into this thesis.

REFERENCES

- ACKEMA, P.; NEELEMAN, A. The role of syntax and morphology in compounding. In: SCALISE, S.; VOGEL, I. (eds.) **Cross-disciplinary issues in compounding**. Amsterdam: John Benjamins, 2010, pp. 21-36.
- ADAMSON, S.; GONZÁLEZ-DÍAZ, V. History and structure of the English noun phrase: Introduction. **Transactions of the Philological Society**, v. 107, n. 3, pp. 255–61, 2009.
- ANTHONY, L. **Antconc** (version 3.5.8) [software]. Waseda University, 2019. Available at: <https://www.laurenceanthony.net/software/antconc/>. Last accessed on Jan 5, 2020.
- APPLEBY, Y. *et al.* Who wants to be able to do reference properly and be unemployed? STEM student writing and employer needs. **Journal of Learning Development in Higher Education**, 2012.
- BARONI, M.; EVERT, S. The non-randomness of corpus data and generalised linear models. Statistical Analysis of Corpus Data with R. Unit 8 of **SIGIL: A Gentle Introduction for Computational Linguists**. 2010. Available at: http://www.stefan-evert.de/SIGIL/sigil_R/. Accessed on Dec. 2, 2018.
- BARTNING, I.; ARVIDSSON, K.; LUNDELL, F. F. Complexity at the phrasal level in spoken L1 and very advanced L2 French. **Language, Interaction and Acquisition**, v. 6, n. 2, pp. 181–201, 2015.
- BAUER, L. **English Word-formation**. Cambridge: Cambridge University Press, 1983.
- BAUER, L.; HUDDLESTON, R. Lexical Word-formation. In: HUDDLESTON, R.; PULLUM, G. K. (eds.) **The Cambridge Grammar of the English Language**. Cambridge: Cambridge University Press, 2002.
- BAUGH, A. C.; CABLE, T. **A History of the English Language**. 6th Edition. Pearson, 2013.
- BERBER SARDINHA, T. Using multidimensional analysis to detect representations of national identity. In: BERBER SARDINHA, T.; PINTO, M. V. (eds.) **Multi-Dimensional Analysis**. Research Methods and Current Issues. New York: Bloomsbury, 2019.
- BERLAGE, E. **Noun Phrase Complexity in English**. Cambridge: Cambridge University Press, 2014.
- BERTRAM, R. *et al.* The hyphen as a segmentation cue: It's getting better all the time. **Scandinavian Journal of Psychology**, v. 52, pp. 530-544, 2011.
- BIBER, D. **Variation across Speech and Writing**. Cambridge: Cambridge University Press, 1988.
- BIBER, D. A typology of English texts. **Linguistics**, v. 27, n. 1, pp. 3-44, 1989.
- BIBER, D. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. **Computers and the Humanities**, v. 26, n. 5-6, pp. 331-345, 1992a.
- BIBER, D. On the complexity of discourse complexity: A multidimensional analysis. **Discourse Processes**, v. 15, n. 2, pp. 133-163, 1992b.
- BIBER, D. **University Language: A corpus-based study of spoken and written registers**. John Benjamins Publishing Company, 2006.

- BIBER, D. Corpus-based and corpus-driven analyses of language variation and use. In: HEINE, B.; NARROG, H. (eds.) **The Oxford Handbook of Linguistic Analysis**. Oxford University Press, 2010, pp. 159-191.
- BIBER, D. Grammar change in the noun phrase: The influence of written language use. **English Language & Linguistics**, v. 15, n. 2, pp. 223-250, 2011.
- BIBER, D.; CONRAD, S. **Register, Genre, and Style**. Cambridge: Cambridge University Press, 2009.
- BIBER, D.; CONRAD, S. **Variation in English: Multi-dimensional studies**. Routledge, 2014.
- BIBER, D.; GRAY, B. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. **Journal of English for Academic Purposes**, v. 9, pp. 2-20, 2010.
- BIBER, D.; GRAY, B. Grammar change in the noun phrase: The influence of written language use. **English Language & Linguistics**, v. 15, n. 2, pp. 223-250, 2011.
- BIBER, D.; GRAY, B. **Grammatical Complexity in Academic English: Linguistic Change in Writing**. Cambridge: Cambridge University Press, 2016.
- BIBER, D. et al. Speaking and writing in the university: A multidimensional comparison. **TESOL Quarterly**, v. 36, n. 1, pp. 9-48, 2002.
- BIBER, D.; GRAY, B.; POONPON, K. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? **TESOL Quarterly**, v. 45, n. 1, pp. 5-35, 2011.
- BIBER, D., GRIEVE, J.; IBERRI-SHEA, G. Noun phrase modification. In: ROHDENBURG, G., SCHLÜTER, J. (eds.) **One language, two grammars?: differences between British and American English**. Cambridge: Cambridge University Press, 2010, pp. 182-193.
- BIBER, D. et al. **Longman Grammar of Spoken and Written English**. Essex: Pearson Education Ltd, 1999.
- BICK, E.; DIDRIKSEN, T. Beyond Classical Constraint Grammar. In: **Proceedings of the 20th Nordic Conference of Computational Linguistics - NODALIDA 2015**. Available at: <https://www.aclweb.org/anthology/W15-1807.pdf>. Accessed on: Jan 5, 2020.
- BICK, E. **English VISL - Visual Interactive Syntax Learning**. Institute of Language and Communication (ISK). University of Southern Denmark (SDU). 1996-2020. Available at: <https://visl.sdu.dk/visl/en/>. Accessed on: Jan 5, 2020.
- BIRD S., LOPER, E., KLEIN, E. **Natural Language Processing with Python**. O'Reilly Media Inc, 2009.
- BONELLI, E. T. Theoretical overview of the evolution of corpus linguistics. In: O'KEEFFE, A.; MCCARTHY, M. (eds.) **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, pp. 42-56.
- BULTÉ, B.; HOUSEN, A. Defining and operationalizing L2 complexity. In: HOUSEN, A.; KUIKEN, F.; VEDDER, I. (eds.) **Dimensions of L2 performance and proficiency, complexity, accuracy and fluency in SLA**. Amsterdam: John Benjamins, 2012.
- BURNARD, L. Metadata for Corpus Work. In: WYNNE, M. (ed.) **Developing linguistic corpora: A guide to good practice**. Oxford: Oxbow Books, 2005.

- BUTTERWORTH, G. Structure of the mind in human infancy. **Advances in infancy research**, 1983.
- BYBEE, R. W. What is STEM education? **Science**, v. 329, n. 5995, 2010.
- CARTER, R.; McCARTHY, M. **The Cambridge grammar of English: Spoken and written English grammar and usage**. Cambridge University Press, 2006.
- CARTER, R. et al. **English grammar today: an A-Z of spoken and written grammar**. Cambridge: Cambridge University Press, 2011.
- CHUJO, K.; ANTHONY, L.; OGHIGIAN, K. DDL for the EFL classroom: effective uses of a Japanese-English parallel corpus and the development of a learner friendly, online parallel concordance. **Congresso em Linguística de Corpus (CL 2009)**. Liverpool: Universidade de Liverpool, 2009.
- Collins COBUILD Grammar**. COBUILD English Grammar, 2017.
- CONRAD, S.; MAURANEN, A. The corpus of English as a lingua franca in academic settings. **TESOL quarterly**, v. 37, n. 3, pp. 513-527, 2003.
- CORTES, V. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. **Journal of English for Academic Purposes**, v. 12, n. 1, pp. 33-43, 2013.
- COXHEAD, A. What can corpora tell us about English for Academic Purposes? In: O'KEEFFE, A.; McCARTHY, M. (eds.) **The Routledge Handbook of Corpus Linguistics**. London: Routledge, 2010, pp. 458-470.
- CULLIP, P. Text Technology: The Power-Tool of Grammatical Metaphor. **RELC Journal**, v. 31, p. 76-104, 2000.
- CRESTI, E., MONEGLIA, M. **C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages**. Amsterdam: John Benjamins, 2005.
- CRYSTAL, D. **The English Language**. Second Edition. London: Penguin Books, 2002.
- CRYSTAL, D. **How language works**. Penguin Books, 2005.
- DAEWEL, U., SCHRUM, C., MACDONALD, J. I. Towards end-to-end (E2E) modelling in a consistent NPZD-F modelling framework (ECOSMO E2E_v1.0): application to the North Sea and Baltic Sea. **Geoscientific Model Development**, v. 12, n. 5, pp. 1765-1789, 2019.
- DAHL, T. Textual metadiscourse in research articles: a marker of national culture or of academic discipline? **Journal of pragmatics**, v. 36, n. 10, pp. 1807-1825, 2004.
- DAVIDSE, K. The Structure of the English NP. **Functions of Language**, v. 23, n. 1, 2016.
- DAVIES, M. Corpora: an introduction. In: BIBER, D., REPPEN, R. (eds.) **The Cambridge handbook of English corpus linguistics**. Cambridge University Press, 2015, pp. 11-31.
- de MARNEFFE, M. et al. Generating typed dependency parses from phrase structure parses. In **Proceedings of International Conference on Language Resources and Evaluation - LREC**, 2006. Available at: https://nlp.stanford.edu/pubs/LREC06_dependencies.pdf. Access on: Jan 5, 2020.
- de MARNEFFE, M.; MANNING, C. D. The Stanford typed dependencies representation. In: **COLING Workshop on Cross-framework and Cross-domain Parser Evaluation**.

Available at: <https://nlp.stanford.edu/pubs/dependencies-coling08.pdf>. Access on Jan, 5, 2020.

de MARNEFFE, M. et al. More Constructions, More Genres: Extending Stanford Dependencies. In: **Proceedings of the Second International Conference on Dependency Linguistics** (DepLing), 2013, pp 187-196.

de MARNEFFE, M. et al. Universal Stanford dependencies: A cross-linguistic typology. In: **Proceedings of the 9th International Conference on Language Resources and Evaluation - LREC**, 2014.

DOWNING, A. **English Grammar: A university course**. 2nd edition. Amsterdam/New York Routledge, 2006.

DRESSLER, W. Compound Types. In: LIBBEN, G.; JAREMA, G. (eds.) **The representation and processing of compound words**. Oxford: Oxford University Press, 2006, pp; 23-44.

DRESSLER, W.; LETNER, L. E.; KORECKY-KRÖLL, K. First language acquisition of compounds. In: SCALISE, S.; VOGEL, I. (eds.) **Cross-disciplinary issues in compounding**. Amsterdam: John Benjamins, 2010, pp. 323-344.

DUTRA, D. P.; BERBER SARDINHA, T. A Multi-dimensional section typology of English academic writing. **Arizona Corpus Linguistics Conference** (AZCL). Flagstaff, Northern Arizona University, 2018.

DUTRA, D. P.; QUEIROZ, J. S., ALVES, J. C. Adding information in argumentative texts: a learner corpus-based study of additive linking adverbials. **Revista de Estudos Anglo-Americanos**, v. 46, n. 1, p. 9-32, 2017

DUTRA et al. Adjectives as nominal pre-modifiers in chemistry and applied linguistics research articles. In: RÖMER, U.; CORTES, V.; FRIGINAL, E. (eds.) **Advances in Corpus-based Research on Academic Writing: Effects of discipline, register, and writer expertise**. John Benjamins Publishing Company, 2020, pp. 206–226.

FABB, N. Compounding. In: SPENCER, A.; ZWICKY, A. **The Handbook of Morphology** Oxford: Oxford University Press, 1998, pp. 66-83.

FEIST, J. **Premodifiers in English: Their Structure and Significance**. Cambridge: Cambridge University Press, 2012.

FERRER i CANCHO, R.; SOLÉ, R. V. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, v. 100, n. 3, 2003, pp. 788-791. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/12540826>. Accessed on Dec 6, 2019.

FINARDI, K. R. The slaughter of Kachru's five sacred cows in Brazil and the use of English as an international language. **Studies of English Language Teaching**, 2, pp. 401-411, 2014.

FINARDI, K. R.; FRANÇA, C. O inglês na internacionalização da produção científica brasileira: evidências da subárea de linguagem e linguística. **Intersecções**, v. 19, n. 2, 2016.

GAGNÉ, C. L., SPALDING, T. L. Conceptual combination: Implications for the mental lexicon. In: LIBBEN, G.; JAREMA, G. (eds.) **The representation and processing of compound words**. Oxford: Oxford University Press, 2006, pp; 145-168.

GARNER, B. A. **Modern English Usage**. Oxford: Oxford University Press, 2016.

- GELDEREN, E. **An introduction to the grammar of English**. Revised edition. Amsterdam: John Benjamins, 2010.
- GLASMAN-DEAL, H. **Science research writing for non-native speakers of English**. 1st edition. World Scientific, 2010.
- GLEDHILL, C. The discourse function of collocation in research article introductions. **English for Specific Purposes**, v. 19, n. 2, pp. 115-135, 2000.
- GRAY, B. **Linguistic Variation in Research Articles**: When discipline tells only part of the story. Amsterdam/Philadelphia: John Benjamins, 2015.
- GREENBAUM, S. **The Oxford English grammar**. Oxford: Oxford University Press, 1996.
- GREENBAUM, S., NELSON, G. **An introduction to English grammar**. Pearson Education, 2009.
- GRIES, ST. Dispersions and adjusted frequencies in corpora. **International Journal of Corpus Linguistics**, v. 13, n. 4, 2008, pp. 403-437.
- GRIES, ST. What is Corpus Linguistics? **Language and Linguistics Compass**, v. 3, n. 5, pp. 1225-1241, 2009.
- GRIES, ST. Dispersions and adjusted frequencies in corpora: further explorations. In GRIES, ST, WULFF, S.; DAVIES, M. (eds.) **Corpus linguistic applications**: current studies, new directions. Rodopi, Amsterdam, 2010, pp. 197-212.
- HALLIDAY, M. A. K. **Language as Social Semiotic**: The Social Interpretation of Language and Meaning. London: Edward Arnold, 1978..
- HALLIDAY, M. A. K. **An Introduction to Functional Grammar**. London: Edward Arnold, 1985.
- HALLIDAY, M. A. K.; MARTIN, J. R. **Writing Science**: Literacy and Discursive Power. London: Falmer Press, 1993.
- HARPER, D. **Online Etymology Dictionary**. Available at: <https://www.etymonline.com/>. Accessed on Jan 5, 2020.
- HARRISON, R L.; PARKS, B. How STEM Can Gain Some STEAM: Crafting Meaningful Collaborations Between STEM Disciplines and Inquiry-Based Writing Programs. In: **Writing Program and Writing Center Collaborations**. Palgrave Macmillan, New York, 2017, pp. 117-139.
- HERRERO-ZORITA, C.; SANDOVAL, A. M. Sentence Length and NP Complexity of General and Medical Written Academic and Media Texts. An Analysis Using a Trained Syntactic Parser. VIII International Conference on Corpus Linguistics (CILC 2016), Málaga, Spain, **EPiC Series in Language and Linguistics**, v. 1, 2016.
- HO, D. **Notepad++ Version 7.8.2**. 2019. Available at: <https://notepad-plus-plus.org/>. Access: Jan 5, 2020.
- HOAD, T. F. **Oxford Concise Dictionary of English Etymology**. Oxford: Oxford University Press, 2000.
- HONG, A. L.; HUA, T. K. MENGYU, H. A Corpus-based Collocational Analysis of Noun Premodification Types in Academic Writing. **The Southeast Asian Journal of English Language Studies**, v. 23, n. 1, pp.115-131, 2017.

- HUDDLESTON, R.; PULLUM, G. K. (eds.) **The Cambridge Grammar of the English Language**. Cambridge: Cambridge University Press, 2002.
- HUDSON, R. Dependency Grammar. HIPPISEY, A., STUMP, G. (eds.) **The Cambridge Handbook of Morphology**. Cambridge University Press, 2016.
- HUDSON, G. **Essential Introductory Linguistics**. Oxford: Blackwell Publishers, 2000.
- HUTTER, J. A. **A Corpus Based Analysis of Noun Modification in Empirical Research Articles in Applied Linguistics**. 2015. 90 p. Thesis (Master of Arts in Teaching English to Speakers of Other Languages) – Portland State University, Portland, USA, 2015.
- HYLAND, K. **Academic Discourse: English in a Global Context**. Continuum: London, 2009.
- HYLAND, K. ESP and writing. In: PALTRIDGE, B.; STARFIELD, S. (eds.). **The Handbook of English for specific purposes**. John Wiley & Sons, 2014, pp. 95-113.
- HYLAND, K. Methods and methodologies in second language writing research. **System**, v. 59, pp. 116-125, 2016.
- HYLAND, K.; JIANG, F. Is academic writing becoming more informal? **English for Specific Purposes**, v. 45, pp. 40-51, 2017.
- HYLAND, K.; HAMPS-LYONS, L. EAP: issues and directions. **Journal of English for Academic Purposes**, v. 1, pp. 1-12, 2002.
- INHOFF, A. W., RADACH, R., HELLER, D. Complex compounds in German: Interword spaces facilitate segmentation but hinder assignment of meaning. **Journal of Memory and Language**, v. 42, pp. 23-50, 2000.
- INHOFF, et al. Eye movements during the reading of compound words and the influence of lexeme meaning. **Memory & Cognition**, v. 36, 675-687, 2008.
- JENKINS, J. **English as a lingua franca in the international university: The politics of academic English language policy**. Routledge, 2013.
- JENKINS, J.; LEUNG C. English as a lingua franca. **The companion to language assessment**, v. 4, pp. 1605-1616, 2013.
- JOHNS, T. Should you be persuaded – two samples of data-driven learning materials. **English Language Research Journal**, v. 4, p. 1-16, 1991.
- JOHNS, T; KING, P. Classroom concordancing. **English Language Research Journal**, v. 4, p. 27-45, 1991.
- JUCKER, A. H. The discourse marker well: A relevance-theoretical account. **Journal of Pragmatics**, v. 19, n. 5, pp. 435–452, 1993.
- JUHASZ, B. J.; INHOFF, A.W.; RAYNER, K. The role of interword spaces in the processing of English compound words. **Language and Cognitive Processes**, v. 20, 291-316, 2005.
- JURAFSKY, D; MARTIN, J. H. Constituency Grammars. In: **Speech and Language Processing** (3rd ed. draft), 2019. Available at: <https://web.stanford.edu/~jurafsky/slp3/12.pdf>. Accessed on Dec. 6, 2018.

- JURAFSKY, D; MARTIN, J. H. Constituency Parsing. In: **Speech and Language Processing** (3rd ed. draft), 2019. Available at: <https://web.stanford.edu/~jurafsky/slp3/13.pdf>. Accessed on Dec. 6, 2018.
- JURAFSKY, D; MARTIN, J. H. Dependency Parsing. In: **Speech and Language Processing** (3rd ed. draft), 2019. Available at: <https://web.stanford.edu/~jurafsky/slp3/15.pdf>. Accessed on Dec. 6, 2018.
- JURIDA, S. H. Word Formation in English: Derivation and Compounding. **DHS**, v. 2, n. 5, pp. 157-170, 2018.
- KARLSSON, F. et al. Constraint Grammar: A language independent system for parsing unrestricted text. **Natural Language Processing 4**. Berlin & New York: Mouton de Gruyter, 1995.
- KEIZER, E. **The English noun phrase**: The nature of linguistic categorization. Cambridge: Cambridge University Press, 2007.
- KILGARRIFF, A. et al. The Sketch Engine: ten years on. **Lexicography**, v. 1, pp. 7-36, 2014.
- KILGARRIFF, A. et al. **Sketch Engine**. Available at: <http://www.sketchengine.eu/>. Accessed on Dec. 6, 2018.
- KÖSLING, K., PLAG, I. Does branching direction determine prominence assignment? An empirical investigation of triconstituent compounds in English. **Corpus Linguistics and Linguistic Theory**, v. 5, n. 2, pp. 201-239, 2009.
- KRESS, G. **Multimodality**: a social semiotic approach to contemporary communication. New York: Routledge, 2010.
- KRESS, G., van LEEUWEN, T. **Multimodal discourse**: The modes and media of contemporary communication. London: Arnold Publishers, 2001.
- KUPERMAN, V.; BERTRAM, R. Moving spaces: Spelling alternation in English noun-noun compounds. **Language and Cognitive Processes**, v. 28, n. 7, pp. 939-966, 2015.
- LANGACKER, R. W. **Foundations of cognitive grammar**: Theoretical prerequisites (Vol. 1). Stanford University Press, 1987.
- LEE, D.; SWALES, J. A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. **English for Specific Purposes**. v. 25, pp. 56-75, 2006.
- LEMLE, M. **Análise Sintática**. São Paulo: Ática, 1984.
- LEUNG, A. H-C. van der WURFF, W. Introduction to the noun phrase in English. In: (eds.) **The Noun Phrase in English**: Past and Present. John Benjamins, 2018.
- LEXICO. 2020. Available at: <https://www.lexico.com/>. Accessed on Jan 5, 2020.
- LIBBEN, G. Semantic transparency in the processing of compounds: Consequences for representation, processing, and impairment. **Brain and language**, v. 61, n. 1, pp. 30-44, 1998.
- LIBBEN, G; JAREMA, G. **The Representation and Processing of Compound Words**. Oxford: Oxford University Press, 2006.

- LIBBEN, G. et al. Compound fracture: The role of semantic transparency and morphological headedness. **Brain and Language**, v. 84, pp. 50-64, 2003.
- LINDSTAD, A. M. The Nordic Dialect Database: Mapping Microsyntactic Variation in the Scandinavian Languages. In: JOKINEN, K. BICK, E. (eds.) **Proceedings of the 17th Nordic Conference of Computational Linguistics – NODALIDA 2009**. NEALT Proceedings Series Volume 4, 2009.
- LONGO, O. N., HÖFLING, C.; SAAD, J. C. Os nomes em função adjetiva não predicativa: Contrastes. **Alfa**, 41, pp. 91-107, 1997.
- LORÉS, R. On RA abstracts: from rhetorical structure to thematic organisation. **English for Specific Purposes**, v. 23, pp. 280–302, 2004.
- LU, X. A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. **TESOL Quarterly**, v. 45, n.1, pp. 36-62, 2011.
- LU, X.; AI, H. Syntactic Complexity in College-level English writing: Differences among writers with diverse L1 backgrounds. **Journal of Second Language Writing**, v. 29, pp. 16-27, 2015.
- LÜDELING, A., KYTÖ, M. **Corpus linguistics**. v. 1. Walter de Gruyter, 2008.
- MACEDO, L. D. **Lexical bundles across sections of applied linguistics research articles**. MA thesis. The Federal University of Minas Gerais. February, 2018.
- MAHLBERG, M. **English general nouns: A corpus theoretical approach**. Amsterdam: John Benjamins Publishing, 2005.
- MANNING, C. D. et al. The Stanford CoreNLP Natural Language Processing Toolkit. In: **Proceedings of the 52nd Association for Computational Linguistics Annual Meeting: System Demonstrations**, 2014. DOI: 10.3115/v1/P14-5010.
- MARCUS, M. P.; MARCINKIEWICZ, M. A.; SANTORINI, B. Building a large annotated corpus of English: The Penn Treebank. **Computational linguistics**, v. 19, n. 2, pp. 313-330, 1993.
- MARTINET, A. **Économie des changements phonétiques: Traité de phonologie diachronique**. Francke Bern, 1955.
- MARTÍNEZ, I. A. Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. **Journal of Second Language Writing**, v. 14, n. 3, pp. 174-190, 2005.
- MARTINEZ, C. Unitex/GramLab. Available at: <https://unitexgramlab.org/>. Accessed on Dec 10, 2019.
- MARTÍNEZ-INSUA, A. E.; PÉREZ-GUERRA, J. An open-sesame approach to English noun phrases: defining the NP. **English Language and Linguistics**, v. 15, pp. 201-221, 2011.
- MATTOS, E. **Compound Nouns**. Atividade pedagógica. Programa de Incentivo à Formação Docente - PIFD da Pós-reitoria de Graduação. FALE/UFMG. Novembro, 2019
- MATTOS, E. English as an international language: a study of Brazilian teachers' beliefs and intercultural competence. **TEFL – 7th International Conference on Teaching English as a**

- Foreign Language** – Intercultural language education for increased European identity and cohesion – Nova University, Lisbon, 9-10 Nov. 2018. <https://www.cetaps.com/events/tefl7/>.
- MAURANEN, A.; HYNINEN, N.; RANTA, E. English as the academic lingua franca. In: **The Routledge Handbook of English for Academic Purposes**. Routledge, 2016, pp. 68-79.
- McENERY, T.; XIAO, R. Character Encoding in Corpus Construction. In: WYNNE, M. (ed.) **Developing linguistic corpora: A guide to good practice**. Oxford: Oxbow Books, 2005.
- McENERY, T.; XIAO, R.; TONO, Y. **Corpus-based Language Studies: An advanced resource book**. London/New York: Routledge, 2006.
- McENERY, T.; HARDIE, A. **Corpus linguistics: method, theory and practice**. Cambridge: Cambridge University Press, 2012.
- MELANDER, B.; SWALES, J. M.; FREDRICKSON, K. M. Journal abstracts from three academic fields in the United States and Sweden: National or disciplinary proclivities? **Trends in linguistics studies and monographs**, v. 104, pp. 251-272, 1997.
- MEYER, C. F. **English corpus linguistics: An introduction**. Cambridge University Press, 2002.
- MOGULL, S. A. **Scientific and Medical Communication. A Guide for Effective Practice**. New York: Routledge, 2018.
- NAIR, P. K. R.; NAIR, V. D. **Scientific writing and communication in agriculture and natural resources**. Switzerland: Springer, 2014.
- NINI, A. **Multidimensional Analysis Tagger**. Version 1.3.1. Manual. 2018. Available at: <http://sites.google.com/site/multidimensionaltagger>. Accessed on Jul 16 , 2019.
- NUNBERG, G.; BRISCOE, T.; RUDDLESTON, R. Punctuation. In: HUDDLESTON, R.; PULLUM, G. K. (eds.) **The Cambridge Grammar of the English Language**. Cambridge: Cambridge University Press, 2002.
- ORTEGA, L. **Understanding Second Language Acquisition**. London: Hodder Education, 2009.
- PAKER, T.; ÖZCAN, Y. E. The Effectiveness of Using Corpus-Based Materials in Vocabulary Teaching. **International Journal of Language Academy**, v, 5, n. 1, p. 62-81, 2017.
- PARROTT, M. **Grammar for English language teachers: with exercises and a key**. 2nd edition. Ernst Klett Sprachen, 2010.
- PARKINSON, J.; MUSGRAVE, J. Development of noun complexity in the writing of English for Academic Purposes students. **Journal of English for Academic Purposes**, v. 14, pp. 48-59, 2014.
- PALLOTTI, G. A simple view of linguistic complexity. **Second Language Research**, v. 31, n. 1, pp. 117–134, 2015.
- PAYNE, J.; HUDDLESTON, R. Nouns and noun phrases. In: HUDDLESTON, R.; PULLUM, G. K. (eds.) **The Cambridge Grammar of the English Language**. Cambridge: Cambridge University Press, 2002, pp. 323-524.
- PETROV, S. et al. A universal part-of-speech tagset. In: **Proceedings of LREC**, 2012. Available at: <https://github.com/slavpetrov/universal-pos-tags>. Accessed on Jan 5, 2020.

- PIRRELLI, V.; GUEVARA, E.; BARONI, M. Computational issues in compound processing. In: SCALISE, S.; VOGEL, I. (eds.) **Cross-disciplinary issues in compounding**. Amsterdam: John Benjamins, 2010, pp. 271-286.
- POLINKSY, M. Headedness, again. In: **UCLA Working Papers in Linguistics, Theories of Everything**, v. 17, pp. 348-359, 2012.
- PULLUM, G. K.; HUDDLESTON, R. Preliminaries. In: (eds.) **The Cambridge Grammar of the English Language**. Cambridge: Cambridge University Press, 2002, pp. 1-42.
- QUEIROZ, J. M. S. **The grammatical complexity of English noun phrases in Brazilian learners' academic writing**: a corpus-based study. MA thesis. The Federal University of Minas Gerais. February, 2019.
- QUIRK, R.; GREENBAUM, S.; LEECH G.; SVARTVIK, J. **A Comprehensive grammar of the English language**. New York: Longman, 1985.
- R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, 2018. <https://www.R-project.org>.
- RASO, T.; MELLO, H. **The C-ORAL-BRASIL I: Reference Corpus for Informal Spoken Brazilian Portuguese**. Amsterdam, John Benjamins, 2012.
- RAYSON, P. E. Computational tools and methods for corpus compilation and analysis. In: BIBER, D.; REPPEN, R. (eds.). **The Cambridge handbook of English corpus linguistics**. Cambridge: Cambridge University Press, 2015, pp. 32-49.
- REIS, C. M. B.; SANTOS, W. S. Inglês sem Fronteiras como locus privilegiado de formação inicial de professores de línguas estrangeiras, In: SARMENTO, S.; ABREU-LIMA, D. M.; MORAES FILHO, W. B. (orgs.) **Do Inglês sem Fronteiras ao Idiomas sem Fronteiras: A construção de uma política linguística para a internacionalização**. Belo Horizonte: Editora UFMG, 2017.
- REYNOLDS, J. A. et al. Writing-to-learn in undergraduate science education: a community-based conceptually driven approach. **CBE – Life Sciences Education**, v. 11, n. 1, pp. 17-25, 2012.
- ROSS, J. R. Nouniness. In: AARTS, B. et al. (eds.) **Fuzzy Grammar: A Reader**. Oxford: Oxford University Press, 2004 [1973], p. 351–422.
- ROUSE, M. **Compiler**. Available: <https://whatis.techtarget.com/definition/compiler>. Access on Nov 12, 2019.
- SÁ, E. M. A linguística de corpus no ensino de escrita acadêmica: experiências de uma disciplina de inglês para fins acadêmicos. In: **IV Congresso de Inovação e Metodologias no Ensino Superior**. 2-5 April, 2019. Belo Horizonte, Minas Gerais, Brasil. Available at: <https://congressos.ufmg.br/index.php/congressogiz/IVCIM/paper/view/883>.
- SALAGER-MEYER, F., SEGURA, G. M. L., RAMOS, R. D. C. G. EAP in Latin America. In: HYLAND, K; SHAW; P. (eds.) **The Routledge Handbook of English for Academic Purposes** Routledge, 2016, pp. 133-148.
- SANCHEZ-STOCKHAMMER, C. **English Compounds and their Spelling**. Cambridge: Cambridge University Press, 2018.
- SARDINHA, T. B. **Linguística de corpus**. São Paulo: Editora Manole Ltda, 2004

- SARKAR, D. **Text Analytics with Python**. A Practitioner's Guide to Natural Language Processing. Apress, 2019.
- SARMENTO, S. et al. IsF e internacionalização. In: SARMENTO, S.; ABREU-LIMA, D. M.; MORAES FILHO, W. B. (orgs.) **Do Inglês sem Fronteiras ao Idiomas sem Fronteiras**: A construção de uma política linguística para a internacionalização. Belo Horizonte: Editora UFMG, 2017, cap. 5.
- SEIDLHOFER, B. **Understanding English as a lingua franca**. Oxford: Oxford University Press, 2013.
- SEPP, M. **Phonological constraints and free variation in compounding**: A corpus study of English and Estonian noun compounds (Unpublished doctoral dissertation). City University of New York, New York, 2006.
- SINCLAIR, J. Corpus and text-basic principles. In: WYNNE, M. (ed.) **Developing linguistic corpora**: A guide to good practice. Oxford: Oxbow Books, 2005, p. 1-16.
- SOLLACI, L. B.; PEREIRA, M. G. The introduction, methods, results, and discussion (IMRaD) structure: a fifty-year survey. **Journal of the Medical Library Association**, v. 92, n. 3, 2004.
- SWALES, J. M. **Genre analysis**. English in academic and research settings. Cambridge: Cambridge University Press, 1990.
- SWALES, J. M. **Research genres**. Explorations and applications. Ernst Klett Sprachen, 2004.
- TAYLOR, A.; MARCUS, M.; SANTORINI, B. The Penn treebank: an overview. In: **Treebanks**. Springer, Dordrecht, 2003.
- TESNIÈRE, L. **Eléments de Syntaxe Structurale**. Paris: Klincksieck, 1959.
- THOMPSON, P. **A pedagogically-motivated corpus-based examination of PhD theses**: macrostructure, citation practices and uses of modal verbs. University of Reading. Available at: <http://paulslals.org.uk/thesis.pdf>. Accessed on May 25, 2019.
- THOMPSON, P. Genre approaches to theses and dissertations. In: HYLAND, K.; SHAW, P. (eds.) **The Routledge Handbook of English for Academic Purposes**. London/New York: Routledge, 2016, pp. 379-391.
- TOBIN, Y. **Semiotics and linguistics**. London: Longman, 1990.
- TRIBBLE, C. ELFA vs. Genre: A new paradigm war in EAP writing instruction? **Journal of English for Academic Purposes**, v. 25, pp. 30-44, 2017.
- TRIBBLE, C.; WINGATE, U. From text to corpus – A genre-based approach to academic literacy instruction. **System**, v. 41, pp. 307-321, 2013.
- VENTOLA, E. Abstracts as an object of linguistic study. In: ČMEJRKOVÁ, S.; DANEŠ, F.; HAVLOVÁ, E. (eds.) **Writing vs. Speaking**: Language, text, discourse, communication. Tübingen: Gunter Narr, 1994, pp. 333–352.
- WEISSBERG, R.; BUKER, S. **Writing up research**. Englewood Cliffs: Prentice Hall, 1990.
- WOLFE, J., BRITT, C. ALEXANDER, K. Teaching the IMRaD genre: Sentence combining and pattern practice revisited. **Journal of Business and Technical Communication**, v. 25, n. 2, pp. 119-158, 2011.

WEDGWOOD, H. **A Dictionary of English Etymology**. Trübner, 1872.

WYNNE, M. (ed.) **Developing linguistic corpora: A guide to good practice**. Oxford: Oxbow Books, 2005.

YANG, G. **Grammatical Features of Structural Elaboration and Compression Common in Advanced ESL Academic Writing**. 109 p. Thesis (Master of Arts) – Department of Linguistics and English Language. Brigham Young University. Provo, Utah, 2015.

YOON, H. More than a linguistic reference: The influence of corpus technology on L2 academic writing. **Language Learning & Technology**, v. 12, n. 2, pp. 31-48, 2008.

ZEMAN, D. Reusable Tagset Conversion Using Tagset Drivers. In **Proceedings of LREC**, 2008. Available at: http://lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf. Accessed on Jan 5, 2020.

ZEMAN, D. et al., **Universal Dependencies 2.5**. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, 2019. Available at: <http://hdl.handle.net/11234/1-3105>. Accessed on Jan 5, 2020.

ZIPF, G. K. **Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology**. Ravenio Books, 2016 [1949].

ZWITSERLOOD, P. The role of semantic transparency in the processing and representation of Dutch compounds. **Language and cognitive processes**, v. 9, n. 3, 341-368, 1994.

ATTACHMENT A

Corpus-driven EAP applications (MATTOS, 2019)

COMPOUND NOUNS

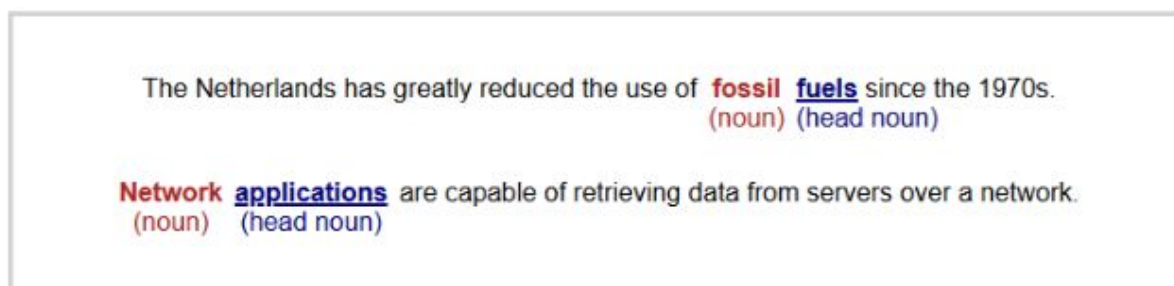
In Brazilian Portuguese “de”, “do”, and “da” are commonly used to form phrases with nouns, as in “copo de plástico”, “tela do computador”, “política da empresa”. In English, noun-noun phrases often replace the need to use “of”. Look at the examples below:

- **plastic cup** --- copo de plástico --- a cup made of plastic
- **computer screen** --- tela do computador --- the screen of the computer
- **company policy** --- política/norma da empresa --- the policy of the company

These are examples of **compound nouns**. Brazilian Portuguese does not typically employ this mechanism to form compound nouns¹, which is why this might be difficult for Brazilian learners of English.

In English, compound nouns are usually formed **combining +2 words**, hyphenated (*check-in*, *sister-in-law*) or not (*computer screen*, *mobile phone*). Compound nouns may also be formed by a **single word** that is the result of two nouns, for example bathroom --- bath + room --- a room to bathe.

The word used before the head noun in compound nouns acts as an adjective because it modifies the head noun. Look at the examples below:



¹ Compound nouns in BP are typically formed by composition, with combination of hyphenated words, such as in “segunda-feira”, with or without “de”, “do”, and “da”, as in “estrela-do-mar”. Compound nouns are also formed by agglutination, for example, “aguardente” --- *água* + *ardente*. However, BP does provide some instances of noun-noun compounds, as in “porco-espinho” and “peixe-espada”, but these are usually hyphenated.

SPECIES		GROUP/CLASS	DEFINING CHARACTERISTICS
A fossil fuel	is	a fuel	that is derived from fossils.
A network application	is	an application	which operates in a network

Compound nouns display different functions. In *plastic cup*, for example, plastic indicates the material the cup is made of. This is why coffee cup and cup of coffee are two very different things. *Coffee cup* is the cup, the container where the coffee is poured into. *Cup of coffee* is a cup with coffee inside.



A coffee cup



A cup of coffee

Compound nouns can be presented in different combinations. Here are some combinations headed by nouns:

Noun	+	Noun	credit card, plane ticket,
Adjective	+	Noun	monthly payment, hot dog
Verb	+	Noun	swimming pool, payphone
Preposition	+	Noun	underground, afterlife

Please note that other combinations are possible, such as *adjective + verb* and *preposition + verb* act as a noun when used together to form compound nouns. For example: *I really appreciate your **input** on my oral presentation. **Public speaking** is not easy for a shy student like me.*

Noun	+	Verb	haircut, sunrise, sunblock, bus stop
Adjective	+	Verb	public speaking, slow dancing
Preposition	+	Verb	input, overrule, outperform
Noun	+	Preposition	night-out, passer-by, lineup
Verb	+	Preposition	takeout, check-in, check-up
Noun	+	PP²	mother-in-law, editor-in-chief

In terms of pronunciation, word stress usually falls on the first part of the compound noun³:

CAR park, BATHroom, WEBSite, BROther-in-law, DOORbell, CHECK-in

However, not all compound nouns follow this rule. Some have spoken stress on the second part, especially in proper names and titles:

Mount EVERest, Prime MINister, New YORK

The functions of compound nouns can be identified by looking at the word(s) that come before the head of the compound noun. For example: **computer** program --- program is the head, **computer** specifies it. What **type** of program? A **computer** program.

In BP, the function is usually identified by looking at the word(s) that come after the noun and we often find a preposition. **Computer** program becomes “programa **de computador**”. Instead of looking to the left of the noun, in BP we have to look to the right.

To express **ownership** in English, **of** or **'s** may be used. When the noun refers to people, animals, dates and organizations, we will almost always use the apostrophe **'s**. For things we mostly use **of**. or example:

- my **father's car**: the car belongs to my father - my father owns the car --- “o carro do meu pai”

² Prepositional phrase, as in “around the world”, “in the morning”, “behind the scenes”, “out of stock” etc.

³ <https://dictionary.cambridge.org/pt/gramatica/gramatica-britanica/nouns-compound-nouns>.

- the **bank's interest rates**: the interest rates 'belong' to a specific bank, this bank has set the interest rates; banks are run by people --- "as taxas de juros do banco"
- **Teacher's Day**: the day 'belongs' to teachers, not to other professions, it's a special day; celebratory days and holidays are set by people --- "Dia do Professor"

The following table⁴ lists the main functional categories of noun compounds.

1. Material What is it composed of?	Copper wire (A wire composed of copper)
2. Mode of Operation How does it work?	Friction brake (A brake that works by means of friction)
3. Purpose What does it do?	Air filter (A filter for cleaning air)
4. Location Where is it used/ found?	Laptop computer (A computer that can be used on a person's lap)
5. Time When is it used?	Summer cottage (= <i>kesämökki</i>) (A cottage that is used in the summer)
6. Shape / form What does it look like?	Disc brakes (Brakes that are shaped like round discs)
7. Inventor / user Who discovered/ uses it?	The Doppler effect (An effect that was proposed by Christian Doppler)
	Passenger car (A car that is used by passengers)

⁴ Retrieved from <http://sana.aalto.fi/awe/grammar/compounds/page2.html>.

EXERCISES

1. Identify the function of the compound nouns below.

1. academic experiences
2. chemical engineering
3. candy industry
4. fish migration
5. Brazilian rivers
6. cosmetics company
7. teaching internship
8. undergraduate course
9. human speech sounds
10. college entrance examination

2. Choose the correct alternative. Justify your answers.

1. humanity's needs / needs of humanity / both
2. purpose's statement / statement of purpose / both
3. River's Ecology lab / River Ecology lab / both
4. the students' performance / the performance of the students / both
5. hydroelectric's impact / hydroelectric impact / both

3. Explain the difference in meaning between *the school's labs* and *the school labs*?

4. Write apostrophe 's or the *of-phrase* for the prompts below.

1. (a glass) wine →
2. (my friend) book →
3. (the teacher) computer →
4. (the number) room →
5. (Brazil) economy →

5. Rewrite the expressions below eliminating the word *of*.

1. a project of Scientific Initiation⁵
2. proposals of energy teaching
3. the development of materials
4. the routine of the laboratory
5. collection of data
6. organization of scientific meetings
7. methods of teaching
8. creation of public policies
9. management of the NGO
10. a student of the universit

⁵ Iniciação Científica would be translated more naturally as: (Undergraduate) Research Mentorship.

ATTACHMENT B

Corpus-driven EAP applications (SÁ, 2019)

3.1 Escrita acadêmica no IFA

Por ser altamente especializada e complexa, acredita-se que a escrita acadêmica não é tão facilmente adquirida (BIBER *et al.*, 2007), razão pela qual universidades em todo o mundo oferecem cursos de redação acadêmica em programas de graduação e pós-graduação. No IFA, esse trabalho é desenvolvido por meio da análise e produção de uma série de gêneros acadêmicos, entre eles o ensaio (*essay*) e o artigo acadêmico (*academic article*). No IFA I, mais especificamente, são abordadas a resenha (*review*), a carta de intenções (*statement of purpose*) e a apresentação acadêmica (em Powerpoint).

A seleção de tais gêneros prevê a construção de conhecimentos que poderão ser aplicados mais ou menos imediatamente na atuação acadêmica dos alunos da UFMG, no Brasil ou internacionalmente. Em outras palavras: ao selecionarmos os gêneros acadêmicos com os quais os alunos terão contato, optamos por explorar aqueles que serão mais úteis em intercâmbios institucionais ou em publicações acadêmicas.

Esse conhecimento da escrita em inglês para fins acadêmicos implica não somente o estudo das condições de produção e recepção dos gêneros escolhidos, mas também a dimensão pragmática, bem como um maior entendimento dos padrões léxico-gramaticais dos registros presentes nos gêneros acadêmicos com os quais os alunos terão contato. Esse é um dos focos das disciplinas IFA, em que as produções escritas são realizadas em até três versões, exatamente para que esses vários aspectos da língua inglesa possam ser desenvolvidos com a devida atenção.

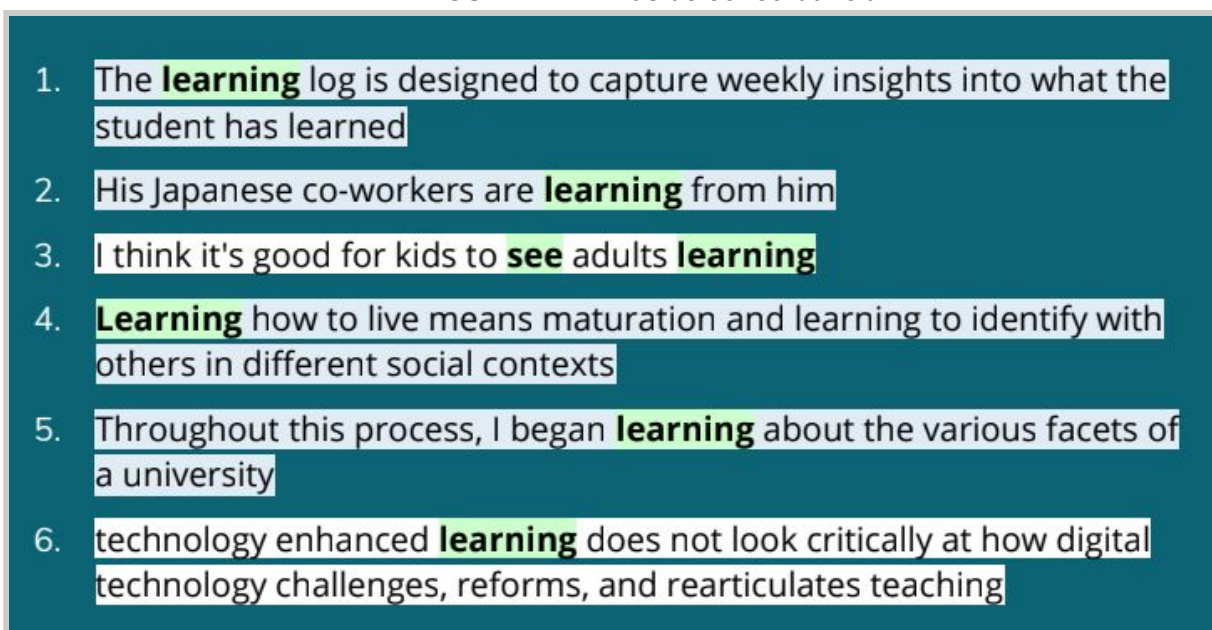
O trabalho com a dimensão léxico-gramatical é inicialmente realizado de forma mais indireta e analítica por meio de atividades do livro didático e/ou baseadas em LC. Ainda seguindo o viés analítico, mas ampliando o escopo contextual, as questões léxico-gramaticais são então identificadas e discutidas em um exemplar do gênero acadêmico a ser produzido. Nesse momento, os alunos precisam justificar o uso dessas estruturas linguísticas. Por fim, chega a hora de colocar esse conhecimento em prática mais diretamente, na produção escrita.

3.2 Linhas de concordância

A seleção de atividades apresentadas serve de exemplo de aplicação da LC no ensino de escrita acadêmica em inglês. As imagens referem-se a atividades especialmente elaboradas e aplicadas na disciplina IFA I em 2018, com foco na formação de palavras por afixação e na análise de sintagmas nominais mais ou menos complexos. Como indicado em Biber *et al.* (2007), Gray (2015) e Biber e Gray (2016), os sintagmas nominais complexos são muito recorrentes na prosa acadêmica. A afixação, por sua vez, é uma das principais formas de desenvolvimento lexical, seja em L1 ou L2.

O trabalho com a terminação *-ing* em diferentes classes de palavras é iniciado na primeira unidade de Hewings, Thaine e McCarthy (2012) e funciona como principal entrada para o estudo das questões gramaticais supracitadas. É também quando os alunos têm o primeiro contato com as linhas de concordância retiradas do COCA, como ilustrado na figura 1 para a palavra *learning*, que serve de base para a visualização da interdependência entre forma e significado das palavras de uma língua. Nesse sentido, linhas de concordância podem ser extremamente úteis, visto que oferecem o contexto imediato dos itens lexicais e referem-se a produções autênticas de escrita.

FIGURA 1 – Linhas de concordância



Fonte: reprodução de Davies (2018).

Por exemplo, na figura 1 as linhas 1 e 6 contêm sintagmas nominais complexos formados com a palavra *learning*, que funciona como adjetivo e substantivo, respectivamente. Ao pedirmos para os alunos lerem as linhas de concordância e identificarem a classe gramatical

de *learning* em cada sentença, estamos levando-os a compreender que o afixo *-ing* não se refere apenas ao presente contínuo do inglês e pode ser utilizado para formar palavras de outras classes gramaticais.

Através desse tipo de atividade, os alunos também têm a chance de observar como sintagmas nominais complexos podem ser formados em inglês. Ao compará-los com seus equivalentes em português, os alunos podem entender melhor diferenças cruciais entre os dois idiomas, isto é, a preferência da língua inglesa pela pré-modificação e do português pela pós-modificação nominal. Esse é um conhecimento tácito que poderá ser aplicado na escrita de resenhas e cartas de intenções.

3.3 Atividades

FIGURA 2 – Sintagmas nominais complexos

Did you know that somewhere between 70 and 90 percent of everything you see outdoors — the plants, trees, flowers — is perpetuated by pollinators?

Flowering plants reproduce when pollen from **a male flower** is carried, usually by an insect, to fertilize **a female flower**.

Plants need help in this process.

Why would insects want to come to their aid? Because the plants provide two incentives:

- First, they make **extra pollen**, so bugs can eat a portion of it (it's full of protein).
- Second, they set out **sugary nectar** to convince bugs to come on by and get covered with that pollen.

That's why many bugs are pollinators besides the **well-known honeybees**, including moths, bumblebees, and butterflies.

Fonte: adaptação de Hewings, Thaine e McCarthy (2012).

Adaptada da segunda unidade de Hewings, Thaine e McCarthy (2012), a figura 2 tem os sintagmas nominais destacados para preparar a análise da terminação *-ing*. Após buscarmos o significado das expressões pelo contexto, os alunos tiveram uma breve discussão sobre a ordem de cada palavra nos termos assinalados, buscando perceber que o padrão do inglês é diferente do português, que tende a preferir elementos pós-modificadores. Para chegar a

essa conclusão, os alunos fizeram uso de estratégias variadas, incluindo-se nesse processo a análise contrastiva e o uso de tradução. *Flowering plant*, por exemplo, foi traduzida como “planta que dá flor” e extra pollen virou “pólen adicional/a mais”.

Além disso, os alunos foram levados a notar as diferenças de pluralização nos dois idiomas, bem como a terminação - *ing* no termo *flowering plant*, cujo significado foi contrastado com sua tradução, procurando mostrar que o prefixo - *ing* não se refere a um verbo, nesse caso, mas a um adjetivo utilizado para pré-modificar o substantivo *plants*. Vemos aqui o raciocínio indutivo colocado em prática a partir dos exemplos selecionados, com foco em dois pontos linguísticos do inglês. Em seguida, com a projeção de linhas de concordância da expressão *flowering* (figura 3), tal como produzida nos registros escritos (acadêmico, jornalístico e literário) do COCA, os alunos leram as frases buscando entender o sentido geral de cada enunciado e passaram às atividades de análise linguísticas (figura 4).

FIGURA 3 – Linhas de concordância

- There are a lot of **flowering** plants, and each species makes its own pollen grain.
- This white **flowering** tree is typical of the sertao.
- South Africa's diverse and beautiful groups of **flowering** plants can bring excellent prices.
- and **flowering** in these plants is influenced by many environmental factors
- legumes are the third-largest family of **flowering** plants
- disrupting the normal **flowering** and fruiting cycle of the plant community
- the West Bank has witnessed a **flowering** of artistic activity, including poetry readings, seminars
- Dobson collected **flowering** plants in the field and transported them in pots
- South Africa is noted for its tremendous diversity and richness of **flowering** plants
- Botanist Roland M. Jefferson has always loved Japanese **flowering** cherry trees.
- an intense **flowering** of art and culture that began in southwestern Germany
- What these two towns shared was an intellectual **flowering** in improbable places
- Plants in the wild look and behave differently, growing at different rates and **flowering** at different times.

Fonte: reprodução de Davies (2018).

FIGURA 4 – Atividades baseadas em linhas de concordância

LANGUAGE STUDY

Does "flowering" have the same function in all sentences?
Does it have the same meaning?

How did you reach these conclusions about the function and the meaning of "flowering"?

Think about other examples of words that display the same form but have different grammatical functions.

Discuss these questions in pairs/small groups

Fonte: acervo próprio.

ATTACHMENT C

Corpus-driven L1 applications (MATTOS, 2018)

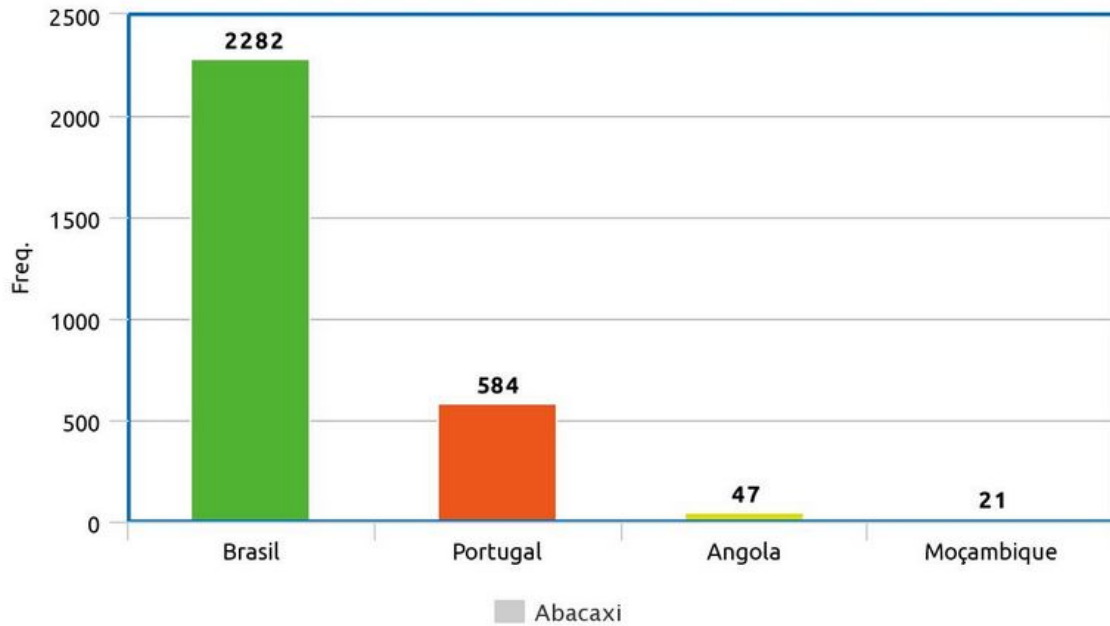
OVERALL ORIGINS AND VARIETIES

- ▶ Adopt an inquisitive approach to local, regional, national, and international cultures and language varieties;
- ▶ Identify salient lexical, phonological, and pragmatic differences;
- ▶ Analyze main cultural products, customs, typical behaviors and attitudes;
- ▶ Investigate underlying social and historical factors;
- ▶ Challenge stereotypes.

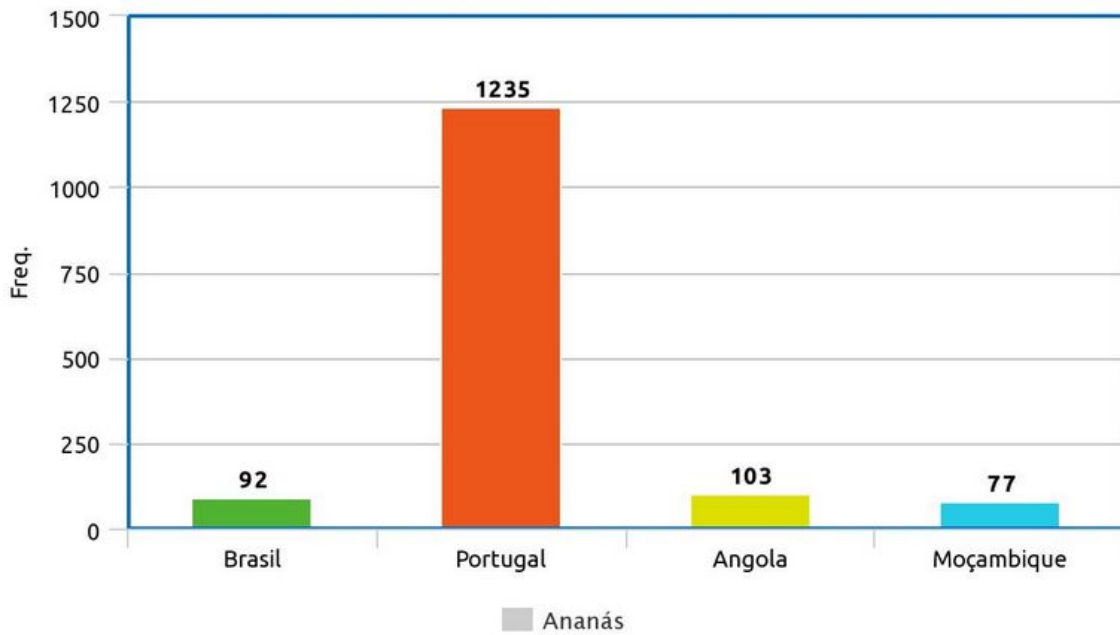
L1 ORIGINS AND VARIETIES

- ▶ Etymological self-investigation: first name, surname, name of hometown or current city etc.;
- ▶ L1 etymology: start with food and place names, move to fixed expressions;
- ▶ L1 varieties: start with more concrete, less abstract concepts;
- ▶ Observe and/or investigate local, regional, national, and transnational variations;
- ▶ Associate differences with social and historical factors.

Abacaxi: distribuição por países
<https://www.corpusdoportugues.org/web-dial/>



Ananás: distribuição por países
<https://www.corpusdoportugues.org/web-dial/>



Expressões idiomáticas

Abacaxi

Vem do tupi e significa "fruta de cheiro forte".

Outros países de língua portuguesa preferem dizer *ananás*, que também vem do tupi e era mais usada no início da colonização europeia.

"Que abacaxi!"

De casca dura e cheia de espinhos, o abacaxi é difícil de descascar.

Por isso chamamos um problema ou situação complicada de "abacaxi" e dizemos que resolver um problema difícil é como "descascar um abacaxi".



ATTACHMENT D

Chunking issues (partial output)

<p>(NP (NP (NP (DT A) (NNS sisters) (POS ')) (NN story)) (: :) (NP (NP (NP (NP (JJ comparative) (NN phylogeography) (CC and) (NN taxonomy)) (PP (IN of) (NP (NP (NNP Hierophis) (NN viridiflavus) (CC and) (NN H.) (NN gemonensis)) (PRN (-LRB- -LRB-) (NP (NNP Serpentes)) (. .) (NP (NNP Colubridae)) (-RRB- -RRB-))))) (JJ Abstract)) (SBAR (S (NP (PRP We)) (VP (VBD used) (NP (DT a) (JJ multidisciplinary) (NN approach)) (S (VP (TO to) (VP (VB infer) (NP (NP (DT the) (NN taxonomy) (CC and)</p>	<p>A sisters ' story comparative phylogeography and taxonomy of Hierophis viridiflavus and H. gemonensis (Serpentes Colubridae) Abstract We used a multidisciplinary approach to infer the taxonomy and historical biogeography of Hierophis viridiflavus and H. gemonensis performing molecular analyses of mitochondrial (16S Cyt-b ND4) and nuclear markers (PRLR) a landmark-based morphometric study and a cytogenetic analysis</p>	55
--	--	----

(JJ historical) (NN biogeography)) (PP (IN of) (NP (NP (NNP Hierophis) (NN viridiflavus) (CC and) (NN H.) (NN gemonensis)) (. .) (VP (VBG performing) (NP (NP (NP (JJ molecular) (NNS analyses)) (PP (IN of) (NP (NP (JJ mitochondrial)) (PRN (-LRB- -LRB-) (NP (NN 16S) (. .) (NN Cyt-b) (. .) (NN ND4)) (-RRB- -RRB-)) (CC and) (NP (JJ nuclear) (NNS markers)) (PRN (-LRB- -LRB-) (NP (NN PRLR)) (-RRB- -RRB-))))) (. .) (NP (DT a) (JJ landmark-based) (JJ morphometric) (NN study)) (CC and) (NP (DT a) (JJ cytogenetic) (NN analysis)))))))))) (. .))		
---	--	--

<p>(NP (NP (DT a) (JJ fundamental) (NN role)) (PP (IN in) (NP (NP (VBG generating) (JJ new) (JJ biological) (NN diversity)) (PRN (-LRB- -LRB-) (NP (NP (NNP Darwin) (CC &) (NNP Wallace)) (SBAR (S (NP (NP (CD 1858)) (: :) (NP (NNP Darwin) (CD 1859)) (: :)) (VP (VB Cook) (NP (NP (CD 1906)) (: :) (NP (NNP Dobzhansky) (CD 1937)) (: :) (NP (NNP Mayr) (CD 1963)) (: :) (NP (NP (NNP White)) (NP (CD 1978))))))))) (-RRB- -RRB-)))))</p>	<p>a fundamental role in generating new biological diversity (Darwin & Wallace 1858 Darwin 1859 Cook 1906 Dobzhansky 1937 Mayr 1963 White 1978)</p>	<p>22</p>
<p>(NP (NP (NP (NNS Results)) (NP (JJ Molecular) (CC and) (JJ phylogenetic) (NN analysis))) (S (NP (PRP We)) (VP (VBD analysed) (NP (NP (CD 85) (NNS sequences)) (PP (IN of) (NP (NN 16S)))))) (. .) (PP (IN for) (NP (NP (DT a) (JJ total) (NN alignment)) (PP (IN of) (NP (CD 511) (NN nucleotide) (NNS positions)))))) (. .)</p>	<p>Results Molecular and phylogenetic analysis We analysed 85 sequences of 16S for a total alignment of 511 nucleotide positions</p>	<p>19</p>

ATTACHMENT E

Chunking script (partial)

```
import java.io.*;
import java.nio.file.Files;
import java.nio.file.Paths;
import java.util.*;

import edu.stanford.nlp.util.CoreMap;
import org.apache.commons.io.FileUtils;
import org.apache.commons.io.FilenameUtils;

import edu.stanford.nlp.ling.CoreAnnotations;
import edu.stanford.nlp.pipeline.Annotation;
import edu.stanford.nlp.pipeline.StanfordCoreNLP;
import edu.stanford.nlp.trees.*;
import org.apache.commons.io.filefilter.TrueFileFilter;

public class Classifier {

    public void tree2conllNP(Tree tree, String output) throws IOException {

        File out = new File(output);
        FileOutputStream fop = new FileOutputStream(out, true);

        if (!out.exists()) {
            out.createNewFile();
        }
        String header = "id\ttoken\tlabel\tconst\n";

        fop.write(header.getBytes());
        fop.flush();

        String label = "O";
        for (Tree t: tree) {

            if (t.isLeaf()) {
                String text = t.toString();

                if (text.equals("-RRB-")) {
                    text = ")";
                    label = "O";
                }

                if (text.equals("-LRB-")) {
                    text = "(";
                    label = "O";
                }
            }
        }
    }
}
```

```

    }

    //System.out.println(t.nodeNumber(tree) + "," + text + "," + label + "," + t);

} else {
    String chunk = t.label().toString();
    if (chunk.equals("NP")) {
        label = "B-NP";
    } else {
        // mudou de chunk? dai eh O
        if (chunk.equals("VP") || chunk.equals("PP")) {
            label = "O";
        } else {
            if (label.equals("B-NP") || label.equals("I-NP")) {
                label = "I-NP";
            }
        }
    }
    if (!chunk.contains("-RRB-") && !chunk.contains("-LRB-")) {
        String line = t.nodeNumber(tree) + "\t\t" + label + "\t" + t + "\n";
        fop.write(line.getBytes());
        fop.flush();
    }
}
}
fop.close();
}

```

```

public void tree2conll(Tree tree, String output) throws IOException {

```

```

    File out = new File(output);
    FileOutputStream fop = new FileOutputStream(out, true);

```

```

    if (!out.exists()) {
        out.createNewFile();
    }

```

```

    String header = "id\ttoken\tlabel\tconst\n";

```

```

    fop.write(header.getBytes());
    fop.flush();

```

```

    String label = "O";
    for (Tree t: tree) {

```

```

        if (t.isLeaf()) {
            String text = t.toString();

            if (text.equals("-RRB-")) {
                text = "";
                label = "O";
            }
        }
    }
}

```

```

        if (text.equals("-LRB-")) {
            text = "(";
            label = "O";
        }

        //System.out.println(t.nodeNumber(tree) + "," + text + "," + label + "," + t);

    } else {
        String chunk = t.label().toString();
        if (chunk.equals("PP")) {
            label = "B-PP";
        } else {
            // mudou de chunk? dai eh O
            if (chunk.equals("VP") || chunk.equals("NP")) {
                label = "O";
            } else {
                if (label.equals("B-PP") || label.equals("I-PP")) {
                    label = "I-PP";
                }
            }
        }
        if (!chunk.contains("-RRB-") && !chunk.contains("-LRB-")) {
            String line = t.nodeNumber(tree) + "\t\t" + label + "\t" + t + "\n";
            fop.write(line.getBytes());
            fop.flush();
        }
    }
}
fop.close();
}

public String readData(String filename) throws IOException {
    File file = new File(filename);
    String text = "";

    BufferedReader br = new BufferedReader(new FileReader(file));

    String st;
    while ((st = br.readLine()) != null) {
        text = text + st;
    }

    return text;
}

public static void main(String[] args) throws IOException {
    Properties props = new Properties();
    props.setProperty("annotators", "tokenize,ssplit,pos,lemma,parse");
    // use faster shift reduce parser
    props.setProperty("parse.model", "edu/stanford/nlp/models/srparser/englishSR.ser.gz");
    props.setProperty("parse.maxlen", "100");
}

```



```

// set up Stanford CoreNLP pipeline
StanfordCoreNLP pipeline = new StanfordCoreNLP(props);

File dirData = new File("/Users/evelin.amorim/Documents/chunking/data/entrada/");
Classifier c = new Classifier();

for (final File fileEntry:FileUtils.listFiles(dirData, TrueFileFilter.INSTANCE,
TrueFileFilter.INSTANCE)) {

    if (fileEntry.getName().endsWith(".txt")) {

        System.out.println(fileEntry.getName());

        // build annotation for a review
        Annotation annotation =
        new Annotation(c.readData(fileEntry.getAbsolutePath()));
        String output = FilenameUtils.removeExtension(fileEntry.getAbsolutePath()) +
        "_pp.tsv";
        String outputNP = FilenameUtils.removeExtension(fileEntry.getAbsolutePath()) +
        "_np.tsv";

        // annotate
        pipeline.annotate(annotation);
        List<CoreMap> ann =
        annotation.get(CoreAnnotations.SentencesAnnotation.class);
        for (CoreMap a: ann) {
            // get tree
            Tree tree = a.get(TreeCoreAnnotations.TreeAnnotation.class);
            c.tree2conll(tree, output);
            c.tree2conllNP(tree, outputNP);
        }
    }
}

System.exit(0);
}
}

```