



Universiteit
Leiden
The Netherlands

Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

Citation

Cunha, E. L. T. P. (2020, March 19). *Contributions to the computational processing of diachronic linguistic corpora*. LOT dissertation series. LOT, Amsterdam. Retrieved from <https://hdl.handle.net/1887/133504>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/133504>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/133504> holds various files of this Leiden University dissertation.

Author: Cunha, E. Landulfo Teixeira Paradela

Title: Contributions to the computational processing of diachronic linguistic corpora

Issue date: 2020-03-19

Contributions to the
Computational Processing of
Diachronic Linguistic Corpora

Published by

LOT

Kloveniersburgwal 48

1012 CX Amsterdam

The Netherlands

phone: +31 20 525 2461

e-mail: lot@uva.nl

<http://www.lotschool.nl>

Cover illustration: “Circuito verbal”, by Lorenza Lourenço (2020)

ISBN: 978-94-6093-343-1

NUR: 616

Copyright © 2020 Evandro Landulfo Teixeira Paradela Cunha. All rights reserved.

Contributions to the
Computational Processing of
Diachronic Linguistic Corpora

Proefschrift

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op donderdag 19 maart 2020
klokke 10.00 uur

door

Evandro Landulfo Teixeira Paradela Cunha

geboren te Belo Horizonte, Brazilië
in 1986

Promotores: Prof.dr. Willem F. H. Adelaar
Prof.dr. Virgilio A. F. Almeida
(Universidade Federal de Minas Gerais &
Harvard University)

Co-promotor: Dr. Søren K. Wichmann

Promotiecommissie: Prof.dr. Marian A. F. Klamer
Dr. Jelena Prokić
Dr. Peter Burger
Prof.dr. Bart de Boer
(Vrije Universiteit Brussel)
Prof.dr. Martin Hilpert
(Université de Neuchâtel)

This dissertation is part of a cotutelle agreement and was presented to Universiteit Leiden in partial fulfillment of the requirements for the degree of Doctor in Linguistics, and to Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science. It was financially supported by CAPES and CNPq (Brazil), and LUCL (the Netherlands).

To my brother,
Voor mijn broer,
Para meu irmão,
Rafael

Contents

List of figures	11
List of tables	13
Acknowledgements / Agradecimientos	15
1 Introduction	17
1.1 Linguistics and computer science: an overview . . .	17
1.2 What has computer science to offer corpus linguistics?	23
1.3 Doing diachronic corpus linguistics	25
1.4 Corpus linguistics and the digital humanities	28
1.5 About this dissertation	30
2 Building a diachronic corpus of comments extracted from news portals and websites	35
2.1 Introduction	35
2.2 Comments in news portals	38
2.3 General description of the resources	43
2.3.1 The Web scraper	43
2.3.2 The corpus	49
2.4 Research possibilities	57
2.5 Concluding remarks	69

3	Establishment and obsolescence of linguistic items in a diachronic corpus	73
3.1	Introduction	73
3.1.1	Related work	75
3.2	Defining establishment and obsolescence as binary notions for diachronic corpus linguistics	78
3.3	The algorithm	80
3.3.1	Requirements	80
3.3.2	Rules for a naive algorithm	82
3.3.3	Proposed algorithm	83
3.4	Case studies	89
3.4.1	Case 1: Statistics on established and obsolete words	93
3.4.2	Case 2: Characteristics of established and obsolete words	95
3.4.3	Case 3: Lexical heritage from past decades	96
3.4.4	Case 4: Lost words	98
3.4.5	Case 5: Short-lived words	102
3.5	Concluding remarks	104
4	Diachronic corpora and quantitative approaches to the lexicon: the case of the term <i>fake news</i>	109
4.1	Introduction	109
4.1.1	Research question	113
4.1.2	Related work	114
4.2	Data sources	124
4.2.1	English-written news corpus	124
4.2.2	Brazilian news corpus	128
4.3	Analyses and results	131
4.3.1	Web search behavior	131
4.3.2	Co-occurring named entities	135
4.3.3	Semantic fields of the surrounding vocabulary	139

4.3.4	Co-occurrence networks	142
4.3.5	Topics addressed in the contexts	146
4.3.6	Polarity	149
4.3.7	Summary of results	151
4.4	Concluding remarks	152
5	Conclusions	157
5.1	Summary of the dissertation	159
5.2	Major contributions	161
A	Lexical heritage from past decades (data)	165
B	Empath categories	171
C	Papers and presentations	177
	Bibliography	183
	Summary	215
	Samenvatting	217
	Resumo	219
	Curriculum vitae	221

List of figures

1.1	Intersections between computer science, linguistics, cognitive science and artificial intelligence	19
1.2	Simplified workflow of the activities involved in the research employing computer-based corpora for language studies	24
1.3	Position of synchronic, diachronic, contemporary and historical data in reference to the present time	27
2.1	Comments posted by readers in a Yahoo! News article	40
2.2	Operating diagram of the Web scraper Xereta . . .	47
2.3	Number of comments per month in the second version of the corpus Xereta	52
2.4	Fragment of the corpus Xereta displayed in Calc . .	54
2.5	Concordance lines of the corpus Xereta at AntConc	59
2.6	Fragment of the corpus Xereta highlighting comments' temporal (date and time) information	60
2.7	Positively rated comments in the UOL news portal	62
2.8	Fragment of the corpus Xereta highlighting positive rating ("likes") information	63
2.9	Fragment of the corpus Xereta highlighting commentators' usernames	64

2.10	Replies to a comment in the UOL news portal . . .	66
2.11	Fragment of the corpus Xereta highlighting the field “Resposta a...” (<i>reply to...</i>)	67
3.1	Percentage of words that became established and ob- solete per decade	94
3.2	Average length of words that became established and obsolete in given decades	96
3.3	Percentage of different parts of speech among words that became established and obsolete in different decades	97
4.1	Output of the query for <i>fake news</i> in the Google Books Ngram Viewer	111
4.2	Numbers of academic publications per year returned by the query for <i>fake news</i>	115
4.3	Contexts including <i>fake news</i> in the NOW Corpus .	126
4.4	Map highlighting the countries considered in the English-written news corpus, grouped by region . .	127
4.5	Normalized volumes of searches for the expression <i>fake news</i> on Google Search from 2010 to 2018 . . .	132
4.6	Percentage of words in each semantic field repre- sented by an Empath category	141
4.7	Co-occurrence networks before and after elections in the English-written corpus	143
4.8	Co-occurrence networks before and after elections in the Brazilian corpus	144
4.9	Average polarity of the contexts in each region before and after elections	151
5.1	Distribution of the terms <i>cozinha</i> and <i>petralha</i> in the corpus Xereta across time	164

List of tables

2.1	Columns available in the output file of the Web scraper Xereta	46
2.2	Number of comments per period in the second version of the corpus Xereta	52
2.3	Most frequent words in the second version of the corpus Xereta	53
2.4	Basic information regarding different corpora of news comments	58
3.1	Number of deviations in each position according to the proposed algorithm for two fictitious examples .	86
3.2	Outcomes of attempts to detect established/obsolete words using a first/last attestation approach, an algorithm following naive rules and the proposed algorithm	92
3.3	Lists of common words in previous decades that became obsolete in the corpus in a particular decade (1850s-1920s)	100
3.4	Lists of common words in previous decades that became obsolete in the corpus in a particular decade (1930s-1980s)	101

3.5	Words classified as short-lived considering the period between the 1860s and the 1950s	104
4.1	Number of contexts containing the term <i>fake news</i> in the English-written news corpus according to the geographical origin of the news media	129
4.2	Number of contexts containing the term <i>fake news</i> in both English-written and Brazilian news corpora according to the year and period of publication . .	130
4.3	Countries with the highest proportions of searches for <i>fake news</i> on Google Search before and after 2016 US election	133
4.4	Most frequent search terms related to <i>fake news</i> on Google Search before and after elections	134
4.5	Most mentioned named entities in the periods before and after elections	137
4.6	Most mentioned named entities in the periods before and after US election, considering the geographical origin of the corresponding news media	138
4.7	Main topic for each region and period according to the LDA output	148

Acknowledgements / Agradecimentos

Completing a PhD is not a task that can be accomplished alone. For this reason, I feel obliged to thank a number of people who directly contributed to the writing of these pages.

First, to my advisors, who were instrumental throughout the development of this work: Søren Wichmann, for all his support and willingness to contribute, offering valuable guidance and feedback; Virgilio Almeida, from whom I receive essential lessons and advice on academic career paths; and Willem Adelaar, for the trust and for accepting the duty of being my formal advisor. I am very proud to have the names of these three great scientists printed in my dissertation. I also wish to acknowledge the always accurate work of the administrative staff at the institutions involved, especially Jurgen and Merel (LUCL), and Adriana, Linda and Sônia (PPGCC), and to the Brazilian funding agencies CAPES and CNPq for the financial support during the whole doctorate.

From a personal perspective, a very special thank you to those who welcomed me in Leiden and made my stays more enjoyable: my hosts, Wes and Marcel, for the amusing breakfast conversations; Natasha (and Natik!), for the evening teas and dinners; and my colleagues at LUCL, who greeted me with admirable cordiality and readiness to help. At UFMG, I had the privilege of sharing the

lab with very good friends, from whom I learned a lot: Douglas, Gabriel, Josemar, Matheus, Raphael and Yuri. Special thanks go to Douglas for his support during the PhD qualifying exams.

Por fim – mas não menos importante –, agradeço às pessoas que torcem por mim e sempre estiveram ao meu lado. Aos meus pais, Heloísa e Jorge, por terem valorizado desde cedo os meus estudos e me mostrado o valor do conhecimento: este título é consequência direta das longas horas que, durante anos, ambos passaram ao meu lado (muitas vezes, confesso, a contragosto meu) me ensinando toda a sorte de conteúdo escolar. Ao meu irmão, Rafael, que dialoga comigo profundamente sem a necessidade de palavras e com quem divido todas as minhas conquistas, agradeço por todo o apoio, carinho e encorajamento que me dá diariamente. À vovó Nilza, sou grato pelo aconchego, pela doçura e por todos os “causos” compartilhados; e a meus avós que já não estão mais aqui, Jacinto, Naná e Nêgo, pela proteção e pela “saudade que eu gosto de ter”. Aos meus tios e tias, primos e primas, amigos e amigas, pelos bons momentos e incentivos constantes. À Lorenza, agradeço pelo cuidado, compreensão, ternura e paciência, mas principalmente por dividir comigo os melhores momentos da vida.

Tudo o que já foi, é o começo do que vai vir,
toda a hora a gente está num cômputo.
João Guimarães Rosa, *Grande Sertão: Veredas*

All that has been is the beginning of what is to be –
we are forever at a crossroads.
João Guimarães Rosa, *The Devil to Pay in the Backlands*
(in James L. Taylor and Harriet de Onís' translation)

CHAPTER 1

Introduction

1.1 Linguistics and computer science: an overview

Over the past decades, the relationship between linguistics and computer science has intensified increasingly. Fields such as formal language theory, automata theory and artificial intelligence, just to mention a few, are traditionally situated in the intersection between these disciplines, which frequently share concepts, terminology and methods. Gross (1972) mentions that “[b]oth the theory and the technology of computation involve concepts that are relevant to the study of language” (p. 5). Also, we could maybe consider that the very use of the metaphorical (and perhaps oxymoronic) terms *computer language* and *programming language* is an example of this relationship: while only humans have evolved natural languages, we still talk about computer and programming languages as an “analogy between symbol systems for instructing computers and natural

human languages” (Baron, 1994, p. 663).

According to Mitkov (2005), computational linguistics “has expanded theoretically through the development of computational and formal models of language”, and, during this process, “it has vastly increased the range and usefulness of its applications” (p. ix). Due to developments that started in the last decades of the 20th century, it is possible to assert that, in some situations, natural language can now “be used as a medium for communication between man and machine” (Mellish, 1994, p. 672). Indeed, a myriad of natural language processing (NLP) and language technology applications – including machine translation, information retrieval, speech recognition and several others – are becoming progressively more popular, making the relationship between linguistics and computer science clearer to the general public. All of this has been fueled by the ever-increasing processing, speed, memory and data storage capacity of machines, as illustrated by Moore’s law¹. Figure 1.1 illustrates some of the possible intersections between computer science, linguistics, and two additional related fields – cognitive science and artificial intelligence –, and demonstrates the wide range of possibilities at the crossroads of these areas of study.

The points of confluence between these disciplines lie also in the development and use of computational methods and tools to assist investigations in traditional fields of linguistics. In some cases, and especially until a few decades ago, the use of computers in this context is simply “due to [their] ability to store and retrieve large amounts of information”, relying not “on an understanding of linguistic structures”, but rather “on little more than [their] ability to store and manipulate sequences of symbols and electronic signals” (Mellish, 1994, p. 672). Nowadays, however, it is possible

¹ “Moore’s law is the *empirical* observation that component density and performance of integrated circuits doubles every year, which was then revised to doubling every two years” (Thompson and Parthasarathy, 2006, p. 21).

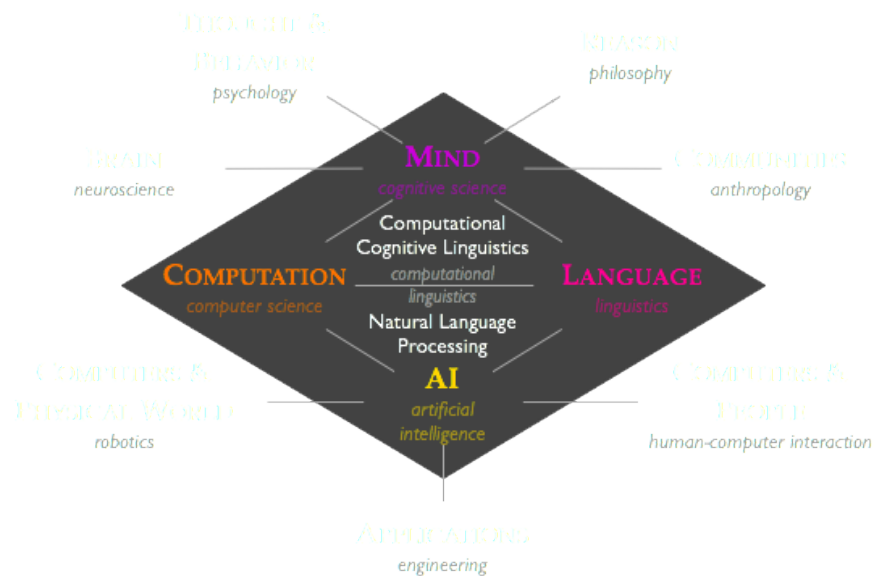


Figure 1.1: Intersections between computer science, linguistics, cognitive science and artificial intelligence. Extracted from <http://www.cs.cmu.edu/~nschneid/index2.html?view=plain> (Prof. Nathan Schneider’s academic webpage) on August 2019.

to mention emerging or already established fields such as computational dialectology (e.g. Heeringa and Prokić, 2017), computational forensic linguistics (e.g. Woolls, 2010), computational historical linguistics (e.g. Rama, 2015), computational lexicology (e.g. Byrd et al., 1987), and computational sociolinguistics (e.g. Nguyen et al., 2016), to name a few, that use computers to gain new insights from large amounts of digital data by employing new methodology, like machine learning or Bayesian phylogenetic inference. In addition, computational methods have been adopted for research on topics as varied as the modelling and simulation of linguistic phenomena (including dialect diffusion (e.g. Prokić, 2017), language acquisition (e.g. Wintner, 2010), language change (e.g. Nettle, 1999), language origin and evolution (e.g. Cangelosi and Parisi, 2002), and

birth, survival and death of languages (e.g. Schulze et al., 2008)), endangered languages documentation, preservation and revitalization (e.g. Bird and Simons, 2003), and language assessment (e.g. Brown, 1997). This ample spectrum of possibilities illustrates the great potential for the use of computational methods and tools in modern linguistics.

Furthermore, the emergence of the Web 2.0 (O'Reilly, 2007), the rise of computer-mediated communication (CMC) and the development of human-computer interaction (HCI) established new frontiers in the relationship between linguistic and computational studies. Crystal (2011), for instance, proposes *Internet linguistics* as an umbrella term referring to the research on the communicative functions of the Internet, and investigates topics that range from the language of Twitter to cybercrime, passing through personalized advertising campaigns and the use of emoticons. We ourselves have conducted research on related themes, including linguistic behavior in online social networks (Cunha and Rocha, 2008; Cunha et al., 2013, 2014b), evolution of Twitter hashtags (Cunha et al., 2011; Cunha, 2012), gendered linguistic styles on the Web (Cunha et al., 2012, 2014a; Las Casas et al., 2014) and hateful, violent and discriminatory language in YouTube videos and comments (Ottoni et al., 2018). All these works could be classified in the fields of social computing (sometimes *social informatics*) or computational social science, the broad areas of computer science interested in the intersection between human behavior and computational systems, including the Web.

The importance of the studies in Internet linguistics is mainly due to the growing relevance of the Web in people's lives in most of the world. According to Saliés and Shepherd (2013), more than 1,000 languages are represented on the Web, which makes it a crucial space for linguistic documentation, preservation and revitalization. Also, the Internet is home of different media, such as blogs,

personal websites, online social media, wikis, news portals, instant messaging tools, each one containing its own particularities of use and linguistic behavior. In addition, in many cases new models of analysis must be developed for linguistic research in these environments, given that “computer-mediated communication (CMC) has challenged the dichotomy between speech and writing even further” (Degand and Van Bergen, 2018, p. 47).

Corpus linguistics² (or, less frequently, *corpus-based linguistics*) is another very rich point of convergence between linguistics and computer science. In a broad sense, corpus linguistics is the empirical study of “real life” language expressed in any collection of written or spoken texts deliberately gathered together and organized. Indeed, according to earlier definitions, a corpus is simply a set of real utterances to be linguistically analyzed (cf. Dubois et al., 1986): Sebba and Fligelstone (1994), for example, define *corpus* as “a body of language material assembled with a view to extracting linguistic information from it” (p. 769), while Bussmann (1996) describes it as “[a] finite set of concrete linguistic utterances that serves as an empirical basis for linguistic research” (p. 106). Conforming to these definitions, the most crucial feature of a corpus is (and will always be) the fact that it relies on concrete, naturally-occurring language data taken from actual written or spoken sources³.

² It is not consensual whether corpus linguistics should be considered a branch of linguistics, a method for doing linguistics or something else. As put by Taylor (2008), “[i]n terms of what corpus linguistics ‘is’, not only have various definitions been offered, but alternatives have been explicitly addressed and rejected. These include (...): *corpus linguistics* is a *tool*, a *method*, a *methodology*, a *methodological approach*, a *discipline*, a *theory*, a *theoretical approach*, a *paradigm* (theoretical or methodological), or a combination of these” (p. 180). However, it is not my goal to address the peculiarities involved in this discussion here. For more on this debate, see Chapter 1 of McEnery and Wilson (1996).

³ For those not familiar with linguistic studies, it may come as a surprise that *naturally-occurring language data* should be granted a special label. It must be remarked, however, that the Chomskyan formal/generative view, “that

In the more specific sense used in modern linguistics, however, the term *corpus* (plural *corpora*) tends to convey additional connotations, “among them machine-readable form, sampling and representativeness, finite size, and the idea that a corpus constitutes a standard reference for the language variety it represents” (Lüdeling and Kytö, 2008, p. v). In addition, when we talk about corpus linguistics, we are usually referring to computer-aided investigations of large digitised corpora containing not only text samples, but often also including linguistic meta-information (e.g. part of speech annotation) and/or extra-linguistic meta-information (e.g. speaker’s gender, text genre) (Zinsmeister, 2015), which might be manually or automatically added. It is interesting to note that it was not always so. In the words of Sampson and McCarthy (2005),

[c]orpus linguistics today is so thoroughly dependent on computers that it would be easy to suppose that the discipline only began after computers had become available to linguists. That is by no means true. (...) [T]he man who really inaugurated the modern corpus-linguistics tradition was Charles Fries⁴, who worked in the 1950s – a time when digital computers were primitive machines familiar only to a scattering of the world’s mathematicians. (Sampson and McCarthy, 2005, p. 9)

Nowadays, however, doing corpus linguistics without the aid of

has dominated linguistics in the 20th century” (Backus, 2014, p. 92), is usually not interested in real examples of attested language (i.e., *performance*), but in sentences obtained through introspection in order to explore the underlying ability to use language (i.e., *competence*). Crystal (2010) explains that, pursuant to this perspective, real samples “were inadequate because they could provide only a tiny fraction of the sentences it is possible to say in a language; they also contained many non-fluencies, changes of plan, and other errors of performance” (p. 433).

⁴ Although other works related to corpus linguistics had already been carried out since at least the 19th century.

computers is unthinkable. Considering this, I will now discuss what computer science has to offer corpus linguistics. The answer to this question will serve as a motivation for the rest of this dissertation.

1.2 What has computer science to offer corpus linguistics?

Thirty years ago when this research started it was considered impossible to process texts of several million words in length. Twenty years ago it was considered marginally possible but lunatic. Ten years ago it was considered quite possible but still lunatic. Today it is very popular. (Sinclair, 1991, p. 1)

The previous passage, taken from John Sinclair’s seminal work on corpus analysis entitled *Corpus, concordance, collocation*, illustrates quite well the amount of possibilities opened by the use of computer power to process large quantities of language data. Sinclair considers that these possibilities were already popular at the time of his book’s publication, in 1991. Thirty years later, at the moment of publication of this dissertation, an even greater body of work concerns the employment of computational methods and tools in the study of the most varied linguistic phenomena observable in corpora.

In the words of Bowker and Pearson (2002), a modern corpus “can be described as a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria” (p. 9). According to this definition, a corpus must be large⁵ and in electronic format for systematic processing, thus requiring the aid of computational processing methods for its exploration. In this

⁵ Although subjective, I understand that, in this context, the use of the adjective *large* indicates an amount of text that cannot be manually analyzed in its entirety.

way, it becomes clear that the use of computational procedures is inherent to modern corpus linguistics. To illustrate this, Figure 1.2 depicts a simplified workflow of the activities involved in the research employing computer-based corpora for language studies. We can observe that computerized activities are involved in most of the steps: in the transformation of real-world language data into a corpus (through compilation, digitising and pre-processing steps), in its quantitative characterization, in the process of automatic annotation and, most of the times, in the exploration of the corpus itself.

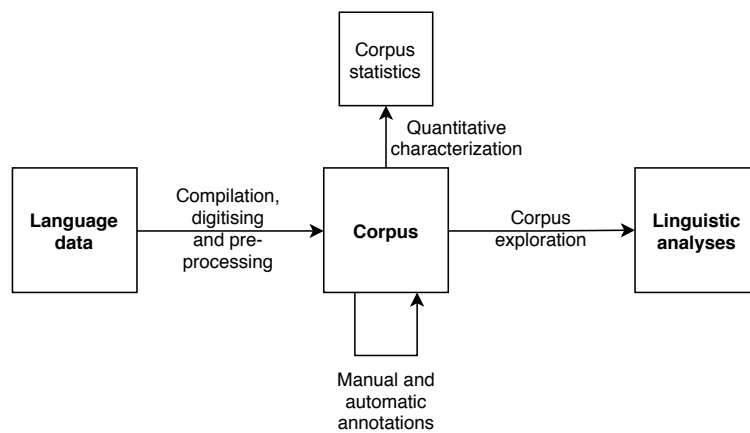


Figure 1.2: Simplified workflow of the activities involved in the research employing computer-based corpora for language studies. This figure is roughly inspired in the “[a]bstract workflow for corpus building” presented by Glaznieks et al. (2014).

Within a functionalist framework⁶, the benefits brought by the use of electronic corpora for linguistic studies are extensive, mak-

⁶ But not necessarily within a formalist one, less interested in the study of real linguistic utterances, in real contexts.

ing it possible to analyze frequencies, probabilities and patterns of occurrence in large corpora (Krishnamurthy, 1997). In addition to this, Mello (2012) mentions that among the advantages offered by the use of electronic corpora (when compared to non-electronic data compilations) are the opportunities of public access to systematized linguistic data and to computational tools available for data treatment and analysis, which enables reproducibility, consistency and re-use with high levels of representativeness and reliability.

Even though I focus here on the question *what has computer science to offer corpus linguistics?*, I would like to mention in passing that the reverse question is also worth asking: *what has corpus linguistics to offer computer science?* Among the possible answers for this question, one that is especially relevant is that high-quality linguistic datasets and corpora are extremely useful for research and technological development of natural language processing applications (Mello et al., 2012), in particular those based on machine learning and artificial intelligence methods. This justifies the need for quality and usability evaluation of these resources.

1.3 Doing diachronic corpus linguistics

The distinction between *synchrony* and *diachrony* is probably one of the most remarkable linguistic principles expounded by Ferdinand de Saussure, as reported in his *Cours de linguistique générale* (Saussure, 1916). Put simply, a synchronic approach aims at analyzing language at a specific point in time, while a diachronic approach considers language change through time. Therefore, a corpus might be considered diachronic if it contains a temporal component that allows us to investigate linguistic phenomena that evolve over time, regardless of its duration.

Although the term *historical corpus* is frequently used as a syn-

onym of *diachronic corpus*⁷, I argue that a terminological distinction should be drawn at this point. In a historical corpus, there should be a (somewhat arbitrary) temporal distance between the present time and the era represented by the data, but not necessarily a temporal factor *inside* the corpus. Conversely, in a diachronic corpus, there should be a well-documented temporal sequence in the data, but the time series recorded must not necessarily be distant from the present. Accordingly, an example of a historical corpus that is not diachronic is a particular corpus of documents from the 16th century that lack additional temporal information, while a corpus made of posts published in an online social media platform containing timestamps is an example of a diachronic corpus that is not historical. This position is shared with other scholars, such as Svensén (1993)⁸ and Krishnamurthy (2003), who consider it important to distinguish between the oppositions *synchronic vs. diachronic* and *contemporary vs. historical*. Krishnamurthy (2003) adds that

[t]he terms [synchronic and diachronic] are relevant not so much with reference to the period of time within which the corpus texts were produced, but rather to the way in which the texts can be accessed. If the corpus can be accessed only as a single entity, then it is functionally *synchronic*, whether the component texts were produced on the same day, within the same year, or even within the same century, because there is no possibility of studying the development of language during that day, year or century. If the corpus texts are held in such a way that texts from a particular period of

⁷ See, for example, the oft-cited corpus typology of Hunston (2002), which does not differentiate these concepts, and designates “[a] corpus of texts from different periods of time” (p. 14) as both *historical* and *diachronic*.

⁸ Who is mostly interested in the distinction between *synchronic*, *diachronic*, *historical* and *contemporary* in terms of dictionaries.

time can be accessed as a separate and discrete group, then the corpus is functionally *diachronic*. We can compare April texts with November texts, or texts from the first decade of the century with texts from the final decade. Crucially, we can observe and comment on language change. (Krishnamurthy, 2003, emphases in original)

Figure 1.3 illustrates these dichotomies, depicting a timeline that shows the position of synchronic, diachronic, contemporary and historical data in reference to the present time.

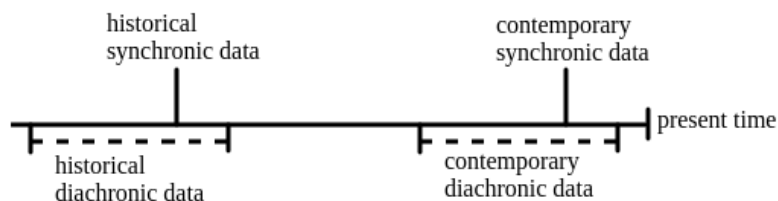


Figure 1.3: Timeline showing the position of synchronic, diachronic, contemporary and historical data in reference to the present time.

The expanded use of computational tools and methods in recent decades provided wider opportunities for researchers interested in corpus-assisted diachronic linguistics. Diachronic corpora have been used for research from a variety of perspectives, ranging from discourse analysis (e.g. Partington, 2010; Baker, 2013) and sentiment detection (e.g. Buechel et al., 2016) to historical sociolinguistics (e.g. Nevalainen and Raumolin-Brunberg, 1996) and register analysis (e.g. Biber and Finegan, 2001), providing unparalleled opportunities for the empirical investigation of language variation and change. In addition, these resources also allow researchers to investigate societal shifts reflected in language, especially when the employed corpora include metadata on speakers, topics and conversation situations – sometimes making it possible to even evaluate

the role of individual speakers in language variation and change. Also, Hilpert and Gries (2016) show how quantitative methods hold considerable potential for diachronic corpus analysis. The authors mention the use of different approaches and tools for this type of investigation and conclude that “quantitative analytical methods can make phenomena visible that would otherwise not be open for inspection” (p. 52).

In this dissertation, I explore these opportunities by presenting a set of contributions related to the computational processing of diachronic corpora, offering insights to three of the multiple stages of the research in this field.

1.4 Corpus linguistics and the digital humanities

Corpus linguistics has also been holding a very prolific relationship with the field of knowledge broadly known as *digital humanities*. In the words of Frischer (2011), “digital humanities can be defined as the application of information technology as an aid to fulfill the humanities’ basic tasks of preserving, reconstructing, transmitting, and interpreting the human record” (p. 28). Presner and Johanson (2009) define digital humanities as “an umbrella term for a wide array of practices for creating, applying, and interpreting new digital and information technologies”, and add that the field

is a natural outgrowth and expansion of the traditional scope of the Humanities, not a replacement or rejection of humanistic inquiry. In fact, the role of the humanist is critical at this historic moment, as our cultural legacy migrates to digital formats and our relation to knowledge, cultural material, technology, and society is radically re-conceptualized. (Presner and Johanson, 2009, p. 2)

It is not my goal here to discuss the definition of digital humanities in detail or what may (or may not) be considered part of this field, but rather to mention its relationship with corpus linguistics. Jensen (2014) evidences this relationship and states that “it is undeniable that a mutual exchange of a scholarly nature has begun between DH [digital humanities] and CL [corpus linguistics]” (p. 131). Some examples are the investigations that use corpora for digital literary studies, such as the one carried out by Culpeper (2014), who employs corpus linguistics methods to analyze different characters in *Romeo and Juliet*, and the one accomplished by Saccenti and Tenori (2012), who study Dante’s *Divina Commedia* using word frequencies as style markers for statistical analysis. I also mention the Graphic Narrative Corpus (Dunst et al., 2017), a digital corpus of graphic novels, memoirs and non-fiction written in English, which was specifically compiled for the use in digital humanities projects.

On this relationship between corpus linguistics and the digital humanities, Jensen (2014) concludes his article with the following interesting statement:

[a]t the end of the day, the goals of CL [corpus linguistics] are not all that different from those of DH [digital humanities]: both seek to shed light on one or more aspects of the human experience, and neither is afraid to explore the opportunities offered by digital technology. (Jensen, 2014, p. 131)

As will be shown in the next section, several of the contributions of this dissertation touch on digital humanities’ areas of interest, especially when the topic of analysis is the human behavior on the Web and the relationship between new media and society.

1.5 About this dissertation

As mentioned above, the main goal of this dissertation is to offer contributions to three of the stages of the research involving the computational processing of diachronic linguistic corpora. More specifically, my focus is placed on the following tasks: (a) corpus building and compilation; (b) designing of tools and algorithms for data exploration; and (c) data analysis for linguistic, cultural and historical research. I do not propose here a single main research question, since this dissertation is the outcome of three independent (although related) projects that share the interest in the application of computing power to diachronic corpus linguistics. These contributions are presented respectively in Chapters 2, 3 and 4, which are summarized below.

Chapter 2: Building a diachronic corpus of comments extracted from news portals and websites

Comments published by readers of news portals and online newspapers and magazines – such as Yahoo! News and The New York Times, in English; G1, Terra and UOL, in Brazilian Portuguese; NU.nl in Dutch; among several others – are expressions of a text genre that grows along with the expansion of Internet use in the world. The analysis of this type of text is relevant for professionals from various fields of knowledge, including social scientists and journalists concerned with the public perception of news and the relationship between media and society. In the context of linguistics, analyses of comments published in news portals allow the study of issues related to this genre in the most varied domains, including lexical, morphosyntactic and pragmatic. These texts might also be useful for the investigation of phenomena such as language variation and change in the online world, and the relationship between language and technology. It is necessary, therefore, to develop tools to

assist the collection and organization of such data, as well as to compile corpora that comprise this text genre. In Chapter 2, we present two useful resources in this regard: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal. The scraper has a simple format, allowing users with no technical background and limited computing knowledge to simultaneously collect all the comments published in a given set of news by simply informing a list of URLs. The corpus brings not only the texts of the comments themselves, but also important meta-information such as dates and times of publication, commentators' usernames and numbers of positive evaluations ("likes") received by the comments. Here we discuss results of the efforts for the elaboration of these two resources, and present the main conceptual and computational challenges and limitations faced during their elaboration. In addition, we provide a general characterization of the compiled corpus. We also mention some possibilities for research employing both the scraper and the corpus presented here, hoping that these ideas might be of some value to researchers wishing to use these resources. As far as we are concerned, besides being a contribution to corpus approaches in a language other than English, ours is the first large and general available corpus of on-line news comments in Portuguese language. We believe that these contributions are of special interest to those involved in the analysis of real-world language data and in methodologies of corpus compilation.

Chapter 3: Establishment and obsolescence of linguistic items in a diachronic corpus

When exploring diachronic corpora, it is often beneficial for linguists to pinpoint not only the first or the last attestation dates

of certain linguistic items, but also the moments in which they become more strongly established in the corpus or, conversely, the moments in which they, despite still being part of the language, become obsolete. In Chapter 3, we propose an algorithm to assist the identification of such periods based on the frequency of items in a corpus. Our simple and generalizable algorithm can be used for the investigation of any linguistic item in any corpus which is divided into time frames. Our idea is to facilitate the discovery of trends and patterns in the dynamics of a language as we process big amounts of text computationally. The proposed algorithm receives as input the frequency of the items in each time frame of the corpus, and may be employed for the analysis of any collection of linguistic items – be it lexical, phonological or morphosyntactical –, regardless of language or historical period. We also demonstrate the applicability of our method using lexical data from the Corpus of Historical American English (COHA), providing case studies on the statistics and characteristics of words that appear in or disappear from this corpus in different periods. We show, for example, that the percentage of established words among all words across decades fluctuates without showing a specific upward or downward trend, and that the proportion of adjectives among new words has increased steadily over the past two centuries, mostly mirrored by a decrease in the proportion of new verbs. We also identify words that became established in different decades and are still frequent in the 2000s; words that, inversely, were previously frequent but got lost over the decades; and short-lived words – words that flared up in the corpus for some time and then disappeared. We hope that these case studies provide some new insights to the field of quantitative diachronic linguistics and to the study of the lexicon, and also motivate future studies employing the method presented here.

Chapter 4: Diachronic corpora and quantitative approaches to the lexicon: the case of the term *fake news*

The term *fake news*, linked to misinformation and manipulation – especially in online environments –, gained popular attention over recent years, particularly during and after the 2016 presidential election in the United States of America and, in Portuguese, during and after the 2018 general election in Brazil. In Chapter 4, we analyze the use of this expression in the traditional media and provide a quantitative investigation on how this term has been conceptualized in the news in the second decade of the 21st century. Our goals are to present a series of computationally-driven analyses performed on diachronic data and to show how this kind of investigation is able to reveal changes in the semantic framing of a given expression. We study the perception and conceptualization of the expression *fake news* in the traditional media using data collected from Brazilian and English-speaking media sources based in 20 different countries. Our outcomes corroborate previous indications of a high increase in the usage of the expression *fake news* and indicate contextual changes around this term after the 2016 US presidential election. Among other results, when comparing different periods of time (before and after the elections), we observe changes in the vocabulary and in the mentioned entities around the term *fake news*, in the topics related to this concept and in the polarity of the texts around it, as well as in Web search behavior of Google Search users interested in this concept. These results suggest that this expression underwent a change in perception and conceptualization after 2016 and 2018, and expand the understanding on the usage of the term *fake news*, which helps to comprehend and more accurately characterize this relevant social phenomenon. More than just analyzing an isolated case, however, our aim is also to present a framework of analysis that can be applied and replicated in other situations, based on the diachronic investigation of vocabulary through differ-

ent and complementary approaches that allow the understanding of semantic change of a lexical item from diverse perspectives.

In short, this dissertation contributes to the scholarship on diachronic corpus linguistics⁹ by: (a) supplying an open source and free Web scraper of comments posted on news websites, available both for download and for online use (Chapter 2); (b) presenting a diachronic corpus containing more than 200,000 comments (plus meta-information) collected from a major Brazilian news portal (Chapter 2); (c) proposing a simple method to assist in the identification of specific time periods of establishment and obsolescence of linguistic items in diachronic corpora (Chapter 3); (d) releasing a series of case studies based on real data concerning two centuries of the dynamics of the American English lexicon (Chapter 3); (e) suggesting a framework to study diachronic changes in the conceptualization of a term through replicable computational and quantitative methods (Chapter 4); and (f) providing a series of observations regarding changes in the conceptualization of the expression *fake news* in English-written and Brazilian news articles (Chapter 4). It is pertinent to mention that all of these contributions, albeit related, are self-sufficient, meaning that it is not necessary to read these chapters in any particular order to get all the information. Since these three chapters are based on a set of five papers published (or accepted/submitted for publication) independently¹⁰, repetitions of definitions, arguments and related works may occur. This also implies that the background needed for each chapter (including the citation of related studies) is provided in the corresponding introduction section.

⁹ And, in part, also to the scholarship on digital humanities.

¹⁰ See Appendix C for detailed information on these papers and on other research activities carried out during the period of doctoral studies.

CHAPTER 2

Building a diachronic corpus of comments extracted from news portals and websites¹

2.1 Introduction

Within the context of corpus linguistics, an imperative task is the development of tools and resources capable of assisting researchers in the process of collecting and organizing material for analysis.

¹ This chapter reproduces with minor changes the article “Xereta: A Brazilian corpus of online news comments” (Cunha et al., under review), submitted for publication and presented (under the title “Contribuições para a coleta e a compilação de um corpus de comentários de portais de notícias”) at the *XI International Conference of the Brazilian Linguistics Association (ABRALIN 2019)*, held in Maceió, Brazil, in May 2019. It is an extended and updated version of the paper “A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias” (Cunha et al., 2017), published in the *Proceedings of the X International Conference of the Brazilian Linguistics Association (ABRALIN 2017)* and presented at this conference, held in Maceió, Brazil, in March 2017. See Appendix C for more information.

Despite the large amount of corpora of all types available (written and spoken, general and specialized, synchronic and diachronic, historical and contemporary etc.), in certain cases researchers face situations in which it is impossible to perform the desired study due to the lack of adequate material in already collected corpora. From our experience, this seems to be a recurrent issue when considering online content (such as social media posts, blog entries, online forum discussions and the like): even though this kind of data is often and increasingly useful for researchers from various fields of study, those with little or no programming skills are sometimes prevented from obtaining the desired data for their research due to the lack of suitable and accessible tools and resources².

Various types of content available on the Internet – from personal conversations in online chat rooms to media texts aimed to attract broad audiences – may be of interest to corpus linguists, since the character of these texts and the language represented electronically are greatly diverse. Kilgarriff (2005), in a controversial statement, argues that “it is the Web that presents the most provocative questions about the nature of the language” (p. 473); Crystal (2004) adds that “there are good grounds for viewing the arrival of the Internet as an event which is as revolutionary in linguistic terms as it has been technologically and socially” (p. 65).

One of the new text genres that have been growing in use as Internet access increases across the world is the genre *comment in news portal*. This type of text is frequent in webpages that publish news articles, such as online newspapers and magazines, as a means of interaction between readers and media producers. Indeed, the

² This is a particularly pertinent situation in resource-limited parts of the world where higher proportions of the population lack advanced digital skills: technical and technological obstacles to working with certain types of data might increase the gap between developed and developing countries (as well as between rich and poor institutions) regarding research results and quality of scientific outcomes.

sections dedicated to readers' comments in news portals might be roughly understood as modern versions of the traditional readers' letters sections in print media.

The investigation of these readers' comments is relevant for researchers interested in linguistic and textual characteristics of Internet genres, since analyses of these texts allow the study of issues related to this genre in the most varied domains, including lexical, morphosyntactic and pragmatic. They might also be useful for the investigation of phenomena such as language variation and change in the online world, and the relationship between language and technology, while discourse analysts may profitably use this kind of text to examine ideology, bias and representation on the Internet, for instance. Professionals from several other fields of study, including social scientists and journalists concerned with the public perception of news and the relationship between media and society, might also make good use of this kind of material.

It is necessary, therefore, to develop tools to assist the collection and organization of such data, as well as to compile corpora that comprise this text genre. In this chapter, we present two useful resources in this regard: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal. The scraper has a simple format, allowing users with no technical background and limited computing knowledge to simultaneously collect all the comments published in a given set of news by simply informing a list of URLs. The corpus brings not only the texts of the comments themselves, but also important meta-information such as dates and times of publication, commentators' usernames and numbers of likes received by the comments. Both resources have been given the name *Xereta*³ – a colloquial word

³ Pronounced as [ʃe'retə] – or, roughly, as *shea-RE-ta*.

meaning “busybody, nosy” in Brazilian Portuguese.

In the next section, we present related work and the main characteristics of texts posted by readers as comments on news portals. Then we discuss results of the efforts for the elaboration of the two resources introduced here, and present the main conceptual and computational challenges and limitations faced during their elaboration. In addition, we provide a general characterization of the compiled corpus. In Section 2.4, we mention some possibilities for research employing both the scraper and the corpus presented here, hoping that these ideas might be of some value to researchers wishing to use these resources. Finally, we conclude the chapter by presenting future steps that should be followed in order to maintain the project and expand its reach.

2.2 Comments in news portals

News portals and websites are currently responsible for much of the volume of the world Internet traffic. As an illustration, as of August 2019, seven (14%) of the top fifty websites ranked by SimilarWeb⁴ in the world were classified in the category “News and Media”. When visualizing data regarding the Netherlands, this number drops to six (12%); however, data from Brazil show that, in this country, the traffic of news sites is extremely high, with the impressive number of twelve (24%) webpages among the top fifty ranked.

At the end of the second decade of the 21st century, the online news territory, that “began as the simple provision of news using websites on the Internet”, has become “an environment of multi-

⁴ SimilarWeb is a company that provides Web analytics services. As of the time of writing of this dissertation, its ranking could be accessed on <https://www.similarweb.com/top-websites>. The ranking provided by SimilarWeb “calculates the number of monthly unique visitors together with the number of page views across desktop and mobile traffic” (SimilarWeb, 2016).


ple digital platforms and products and numerous ways of accessing news content” (Küng et al., 2016, p. 443). Also according to Küng et al. (2016), at least two eras in the development of online news services are identifiable: while the period from 1993 until approximately 1999 may be called the “era of digital publishing”, the following period can be termed the “era of participation and multimedia”. From the 2000s, with the advent of Web 2.0⁵ (O’Reilly, 2007; Cunha, 2012), news portals began to incorporate readers’ participation: on the same page in which a certain news is published, readers are able to expose their opinion and make it available to others interested in that specific piece of news – thus creating not only content, but, above all, conversations and communities (Amoris et al., 2012). In the words of Milioni et al. (2012), “[t]he popularization of web 2.0 has signaled a new era in audience participation, one that is interactive and allows users to produce and publish their content online”.

Comments posted by readers of news portals and online newspapers and magazines – such as Yahoo! News and The New York Times; or G1, Terra and UOL, in Brazil; or NU.nl, in the Netherlands⁶ – are expressions of a text genre that grows along with the expansion of Internet use in the world. Figure 2.1 illustrates some of the comments posted by readers in a Yahoo! News article. It is interesting to note that popular articles may reach a considerable number of comments, sometimes thousands of them – as is the case shown in the figure, which reached (at least) 6,684 comments. Some portals allow comments to be posted in response to other users’ comments and, in some cases, readers are able to positively



⁵ “Web 2.0 refers to the social use of the Web which allow people to collaborate, to get actively involved in creating content, to generate knowledge and to share information online” (Grosbeck, 2009, p. 478).


⁶ Respectively: <https://news.yahoo.com/>, <https://www.nytimes.com/>, <https://g1.globo.com/>, <https://www.terra.com.br/>, <https://www.uol.com.br/> and <https://www.nu.nl/> .


or negatively evaluate comments.

 **6,684 reactions**





[Sign in to post a message.](#)


Top Reactions   1,178 viewing

13 people reacting 





 **Marie** 20 hours ago ...


Disneyland & Disney World ban smoking, large strollers, ice cubes....and park fees will increase by 12%. Thank you, and enjoy your visit!

 Reply  Replies (270)  3,455  197





 **lbfedu** 19 hours ago ...


Doesn't really matter. Working families can no longer go to Disney World due to huge price increases which Disney conveniently leave out of announcements. Do not like being around smoke either but seems a contradiction to serve alcohol which can lead to issues as well. But then Disney give a rip about anything but revenue.

 Reply  Replies (280)  2,070  222

 **Christian** 20 hours ago ...

So what are they going to do to facilitate guest flow at the places where people line up to get free ice?

 Reply  Replies (97)  2,175  92

 **realist** 19 hours ago ...

Mortgage your home to afford tickets for a family of four and be sure to memorize the park rules so the Disney security team doesn't boot you from the park. And if you want to eat, liquidate the kids college funds and enjoy! Welcome to the Happiest Place on Earth, which is nothing but a Mickey Mouse operation!





 Reply  Replies (70)  1,143  89

Figure 2.1: Comments posted by readers in a Yahoo! News article. In this example, comments are called *reactions* and may receive replies and positive/negative evaluations by other readers.

In the words of Henrich and Holmes (2013), “[c]omments have the potential to increase our understanding of public opinions, how the public makes decisions and how beliefs are formulated” (p. 1). However, according to Potthast (2009), comment boards on news sites “are often flooded with all kinds of junk and spam, which may be a reason why research has widely neglected comments as

a source of information” (p. 724). Indeed, the scholarship on the topic of comments in online news portals is still somewhat limited. Nonetheless, interesting results have already been published, as we show in the next paragraphs. Some of these results are especially relevant for social and media scientists.

One case of recent scholarship is the work of Lee (2012), who surveyed South Korean individuals in order to investigate whether user-generated comments on Internet news sites affect other readers’ perception on the bias of the news report itself. Among her results, we highlight the finding that, in some cases, “people might misattribute the opinions expressed in others’ comments to the news article” (p. 32), demonstrating the influence of comment boards on news reception. Related to this, Milioni et al. (2012) explore the assumption that the possibility of publishing their own comments on news pages gives audience a greater power over influencing the very activity of news production – that is, a power of influencing journalists and media outlets. Their findings suggest that, at least in the analyzed (Greek) context, even though commentators often challenge journalistic viewpoints, “this type of audience participation is not likely to render audiences co-producers of news content in significant ways” (p. 21). Erjavec and Kovačič (2012), using a critical discourse analysis framework, investigate hate speech on Slovenian news websites’ comments, finding that most of the hate speech producers “share characteristics of an authoritarian personality” (p. 899) and are mostly motivated by thrill and fun.

There are also a number of studies that focus on the analysis of this type of Web comments in specific contexts, such as in the political sphere (e.g. Maia et al., 2015). Particularly important for us is the investigation carried out by Rossini (2017), the first user of the scraper presented in this chapter. In her extensive analysis of informal political conversation on the Web, the author employs a systematic content analysis approach to investigate 12,797 comments

extracted from news portals and websites, examining incivility and intolerance in this environment. Among the results obtained, we highlight the one suggesting a normalization of incivility – which is usually employed as a rhetorical resource in situations of disagreement – in these online spaces. The outcomes of her study are relevant and contribute to the understanding of how technology can affect the political and social dynamics of contemporary democratic societies.

The previously mentioned studies are based on surveys and/or qualitative research usually conducted through the individual analysis of each comment. Some other studies have used larger selections of comments extracted from Web portals as objects of analysis, thus taking more quantitative approach. Potthast (2009), for instance, investigates the descriptive nature of this type of comments, revealing that “10 comments suffice to expect a high similarity between the comments and the commented text; 100-500 comments suffice to replace the commented text in a ranking task” (p. 724). Additionally, Reyes et al. (2010) evaluate humorous features on Web comments with the goal of automatically distinguishing between implicit funny comments from not funny ones.

From a computational perspective, Hsu et al. (2009) propose a machine learning approach to rank comments based on quality – which may be useful, for instance, to promote high-quality comments and filter out low-quality ones, including spams –, while Potthast and Becker (2010) present a tool to help summarize and visualize expressed opinions in the form of comments on the Internet. Their motivation is that popular items often receive thousands of comments (see Figure 2.1 as an example), thus making “visitors read (...) only the newest comments and hence get an incomplete and possibly misleading picture of the overall opinion” (p. 668). Moreover, named-entity recognition in news comments on the Web is addressed by Wan et al. (2011), who propose a method for iden-

tifying person, location and organization names in comments.

All of the above-mentioned works were made possible by the availability of real and authentic data collected from news websites. In small-scale studies, in which the amount of data required is more limited, we observe that in most cases the data collection was performed manually (i.e., researchers visited the news pages and manually copied the comments available). In studies using larger numbers of comments (from a few thousands upwards), data collections were certainly accomplished by computational tools. In other cases, already available corpora of news comments were used. A comparison between different corpora of news comments published up to the writing of this dissertation is presented in Section 2.3.2.

In the next section, we introduce two relevant resources for the research on the text genre *comment in news portal*, hoping that they will be especially useful for researchers who, for any reason, are not able to develop their own scripts for this type of data collection.

2.3 General description of the resources

Here, we introduce the two resources presented in this chapter: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal.

2.3.1 The Web scraper

In Web science, a *scraper* refers to an automated agent used to extract data from targeted sources (Upadhyay et al., 2017), thus capturing specific content from webpages. The purpose of the scraper presented here is to return a file containing all comments posted on a given list of news articles published in a news portal. One of its

main features is its simple operation, allowing its use by individuals with only basic computational knowledge.

First, it is important to note that this tool was developed as open source and as being free for use, modification and distribution. In short, it means that anyone can read the code, modify it, contribute, and use it in their own projects. Our Web scraper can be downloaded and used free of charge, without any registration, and if programming-skilled users want to, they might modify it to better meet their goals. The scraper Xereta is licensed under the GNU GPL (General Public License), which guarantees the freedom of running the program for any purpose, studying how it works, adapting it to one's needs, redistributing copies, and ultimately refining it and release improvements that might benefit the whole community. By acting in this way, we are contributing to the community by lowering costs and providing greater flexibility for researchers (Corrado, 2005).

The scraper presented here was developed in Python and is available in two different formats: for download and for online use. The code available for download is the same as the one available for online use. The downloadable version is intended primarily for those who might want to see the code and modify it, while the online version is best suited for general and quick use, especially by less programming-skilled users. The online version of the scraper is available at <http://xereta.herokuapp.com/>⁷. To use it from the website, the user must simply enter in the available box the list of URLs from which comments should be collected⁸. The automati-

⁷ If at any time this page is down, it is recommended to look for instructions at <https://www.dcc.ufmg.br/~evandrocunha/>.

⁸ The list of URLs to be inserted must, of course, be made up of articles of interest to the user. There are some ways to automate the process of obtaining these URLs. One of them is the automated collection from `robots.txt` files and sitemaps. This process is described in the next section, where we detail the collection of the corpus Xereta.

cally downloaded file might be opened in a spreadsheet editor (such as LibreOffice Calc or Microsoft Excel).

The main challenge in designing a Web scraping tool for gathering comments from news sites is that page structures of news portals might completely differ from each other. Given the technology available during the execution of this work, the most reliable way to collect comments from different websites (e.g. The New York Times, Yahoo! News etc.) is to develop a specific piece of code explicitly designed for each one of them. In addition, page structures of different sections (e.g. entertainment, sports, politics etc.) on the same website may also differ. This means that the code used to collect comments from one portal cannot be entirely reused to collect comments from another portal: basically, it is necessary to generate a modified module for each news portal to be collected, which means that developers need to visit different websites (and different sections of the same website) to identify their structural characteristics and the extent to which they differ from each other.

Moreover, the meta-information available on each portal also varies. For example, while some news sites allow comments to be rated positively and negatively, others allow only positive ratings (“likes”), while others do not even offer this possibility. With user convenience in mind, we decided to generate a single output file that compiles all comments from the given URLs, regardless of the source portal. Thus, it was necessary to define the set of meta-information fields to be included in the output file. In the version available at the time of this publication, the output file contains the ten columns mentioned in Table 2.1.

At the time of publication of this work, our Web scraping tool processes pages from two Brazilian portals: Folha de S. Paulo and UOL. However, the idea is to expand this project and make it able to gather content from more websites, especially from Brazilian, Dutch and international portals.

Table 2.1: Columns available in the output file of the Web scraper Xereta.

Column title (in Portuguese)	Content
<i>ID</i>	Comment ID, provided by the news portal
<i>Resposta a...</i>	“Reply to...”: ID of the comment to which the comment replies (if the comment is a reply to another comment; if it is not a reply, this field is filled with “0”)
<i>Título da notícia</i>	Title of the news article
<i>Data da notícia</i>	Publication date of the news article
<i>Usuário</i>	Username provided by the commentator
<i>Comentário</i>	The comment itself
<i>Data do comentário</i>	Publication date of the comment
<i>Hora do comentário</i>	Publication time (hour and minute) of the comment
<i>Aval. positivas</i>	Number of positive ratings received (“likes” or “thumbs up”)
<i>URL</i>	URL of the news article

Figure 2.2 displays the operating diagram of the Web scraper Xereta. We observe that the scraper first receives as input a single file with a list of all URLs to be visited and, at the end, it generates as output a single file containing all the comments (and respective meta-information) gathered from all those URLs. Internally, the scraper’s first task is to consume the file containing the URLs and store them in a list. The program then iterates over this list, visiting one URL at a time. As mentioned above, the algorithm

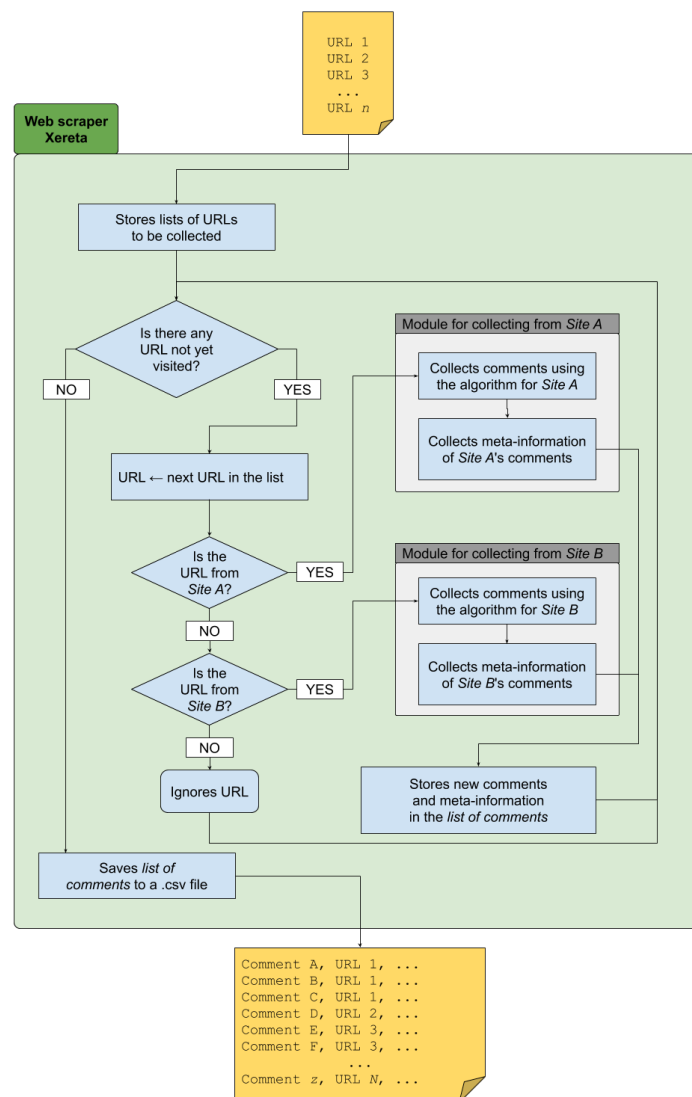


Figure 2.2: Operating diagram of the Web scraper Xereta. The labels *Site A* and *Site B* represent news portals and websites from where the scraper is able to collect comments. Additional modules can be inserted at any time.

for extracting comments and their meta-information is different for each portal, making it necessary to create a specific module for each site. The scraper then checks the URL string itself to identify the source portal and hence which module to use. It is important to note that, despite the need to develop individual modules, our Web scraper is flexible enough to allow the inclusion of additional modules for other portals simply by implementing them using the same format as the already available ones. After comments are extracted from a given URL, they are included in a list where all comments collected are already stored. At the end, when the entire list of URLs is traversed, the program saves the list of comments in the output file.

The developer interested in inserting new modules to collect comments posted on different websites must follow the steps below. First, it is required to inform the scraper which module to use (since a different module is called for each different website). In the version described in this study, this is done by simply analyzing the URL, as follows:

```

if "folha.uol.com.br" in url:           #Check if URL is from Folha
    comentarios = coleta_folha(url)    #Use module "Folha"
elif "uol.com.br" in url:             #Check if URL is from UOL
    comentarios = coleta_uol(url)      #Use module "UOL"
else:                                  #If URL is unknown...
    comentarios = []                   #...do not call any module
    print("URL inválida")              #Inform that URL is invalid

```

In the case of adding a new module, it is necessary to inform the scraper which module will be called depending on the format of the URL by inserting a piece of code such as:

```

(...)
elif "nu.nl" in url:                   #Check if URL is from NU.nl
    comentarios = coleta_nunl(url)    #Use module "NU.nl"
(...)

```

Then it is needed to implement the module itself, which receives as input the string of the URL to be collected and should return a list of comments. In general, a manual analysis of the page structure is required to identify the URL of the page where the comments are stored. The programmer should then look for the fields that correspond to the columns to be output. It is necessary to implement a site-specific data scraping task, which can be performed in a number of ways (e.g. through regular expression, HTML XPath extracting, CSS selector, API request etc.)⁹. In the version of the scraper described here, the list of comments returned should contain the information available in Table 2.1, in that exact order. If it is not possible to collect a certain field or if a given information does not apply for the website (e.g. if the website does not provide the feature of positively rating a comment), the module should fill this information with a null value.

As far as we are concerned, this is the first freely available resource specifically built for the collection of comments published in news portals.

2.3.2 The corpus

Using the architecture of this Web scraper, we collected a corpus containing comments (plus corresponding meta-information) posted in news articles published in the Brazilian major news portal UOL. The UOL website¹⁰, whose acronym stands for *Universo Online*, was chosen because it is one of the most accessed and traditional Brazilian webpages, being active since 1996 (Moreira et al., 2018). In August 2019, it was ranked by SimilarWeb in the seventh position among the most visited websites in Brazil and second in the category “News and Media”, with an average of approximately

⁹ Examples using regular expressions are available in the open-source code of the Web scraper Xereta itself.

¹⁰ <https://www.uol.com.br/> .

650 million visits per month¹¹. UOL publishes news on the most varied topics, including politics, economics, entertainment, sports and science, to name a few.

In 2017, we collected and made available a first version of the corpus Xereta. This version contains 23,455 unique comments posted in news articles published in 2014 and can be called our “pilot corpus”. In 2019, we released a second version, now containing 202,541 unique comments – almost ten times more than in the pilot corpus – posted in news articles published from January 2016 until December 2018, thus comprising three full years of data collection. These three years are especially interesting for the analysis of news portals comments since they are among the most heated times in Brazil’s history, particularly due to the impeachment of (or coup d’état against) President Dilma Rousseff in 2016 and to the 2018 extremely turbulent general balloting that resulted in the election of far-right candidate Jair Bolsonaro. In this chapter, we are referring to this second version of the corpus.

As shown in Figure 2.2, to perform the scraping it is first necessary to provide a list of URLs to be used as input. To obtain this list for our corpus compilation, we followed the steps below:

1. We accessed the `robots.txt` file¹² available at `https://noticias.uol.com.br/robots.txt`;
2. From this file, we obtained the URL of the sitemap¹³ available at `https://noticias.uol.com.br/sitemap/index.xml`;
3. The XML sitemap obtained included a list of additional

¹¹ Data obtained from `https://www.similarweb.com/website/uol.com.br`.

¹² A `robots.txt` file is a text file placed on sites’ root directory giving instructions to search engine robots.

¹³ *Sitemaps* are lists containing the URLs of the pages in a website, usually intended to help search engine bots to explore, crawl and index site’s webpages.

sitemaps, each one corresponding to one month – from January 2016 until the time of the corpus compilation (July 2019);

4. We gathered all the URLs available in the sitemaps corresponding to the period between January 2016 and December 2018, thus obtaining two full years of data.

It is important to notice that each section of the UOL website has its own `robots.txt` file and, consequently, its own sitemaps. We gathered webpages available in the sitemap of the section *notícias* (*news*). Other corpora might be collected from sections such as *economia* (*economics*), whose `robots.txt` file can be found at <https://economia.uol.com.br/robots.txt>, or *carros* (*cars*), whose `robots.txt` file can be found at <https://www.uol.com.br/carros/robots.txt>, for instance¹⁴. This is one of the tasks that we intend to accomplish in future work to make our corpus more comprehensive.

After having obtained the list of URLs to be collected, we proceeded to the corpus compilation using the scraper described in the previous section. As mentioned above, 202,541 unique comments were collected, contemplating 5,112 news stories. Table 2.2 displays the number of comments per year and per semester in the corpus. It is clear that this version of the corpus is somewhat imbalanced per year, since the number of comments from 2017 (89,269) is considerably higher than those from 2016 (53,147) and 2018 (60,099)¹⁵. Also, Figure 2.3 shows the number of comments per month in the corpus, where the temporal imbalance is again evident. However, when partitioning the data into semesters, this imbalance slightly

¹⁴ These URLs were accessed on August 2019 and might change at any time.

¹⁵ There are also 26 comments posted in 2019, due to the fact that some news published in the last days of 2018 were still receiving comments in the first days of 2019.

decreases, with only two semesters out of the range of 27,000-37,000 comments. In future releases, we expect to correct these issues and publish more temporally balanced corpora.

Table 2.2: Number of comments per period (year and semester) in the second version of the corpus Xereta.

Year	2016		2017		2018	
Comments	53,147		89,269		60,099	
Semester	1st	2nd	1st	2nd	1st	2nd
Comments	36,149	16,998	27,124	62,145	32,861	27,238

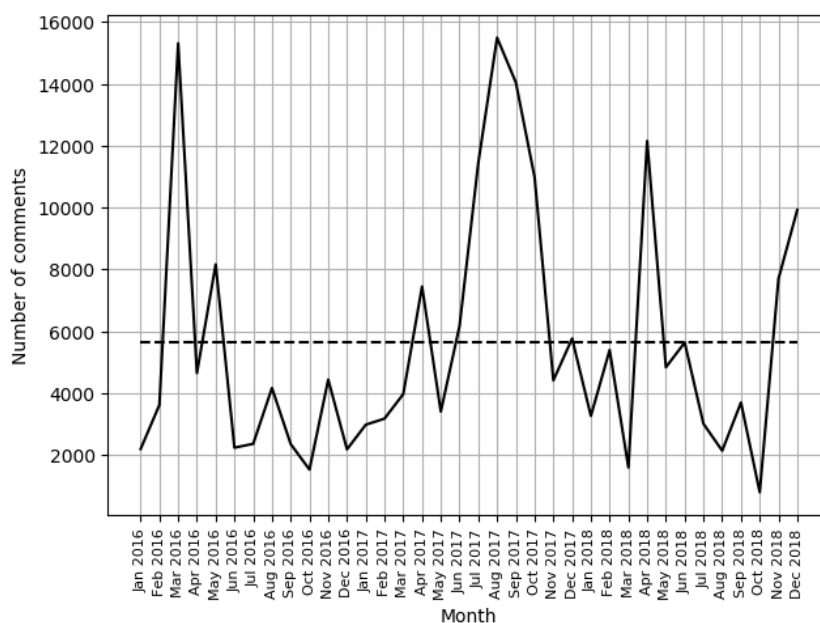


Figure 2.3: Number of comments per month in the second version of the corpus Xereta. The dashed line indicates the mean value (5,626.14).

The corpus includes 6,689,806 word tokens, averaging 33.2 words

per comment, and 109,849 word types. The top ten frequent types in the corpus are listed in Table 2.3. As expected, the most frequent types are all short functional words or high frequency verbs, such as “o” (both a definite article and a pronoun), “que” (both a conjunction and a pronoun), “não” (meaning both “no” and “not”) and “é” (present tense, third-person singular form of the verb “ser”, which means “to be”). The high frequency of these words does not say much about the corpus. However, Table 2.3 also shows the most common nouns in the corpus. This list displays several words linked to Brazilian politics, including the name of two former presidents (Lula and Temer) and a political party (PT).

Table 2.3: Most frequent words (total) and nouns in the second version of the corpus Xereta.

Word	Frequency	Noun	Frequency
o	229,836	brasil	25,627
que	227,927	lula	20,618
de	212,571	país	19,352
e	200,940	povo	18,351
a	188,481	pt	15,451
não	114,138	governo	11,788
é	110,055	anos	11,009
do	98,838	temer	10,821
para	74,328	dinheiro	10,511
se	71,322	presidente	10,402

In total, the comments present in the corpus received 407,250 positive evaluations (“likes” or “thumbs up”), averaging a mere 2.01 per comment. Interestingly, the distribution of positive evaluations among comments appears to follow a “rich-get-richer” pattern: the most liked comment received 297 thumbs up, while 109,603 comments did not receive any positive evaluation.

A	B	C	D	E	F	G	H	I	J	K
ID	Resposta n.º	Título da notícia	Data da notícia	Usuário	Comentário	Data do comentário	Hora do comentário	Comentário Curtidas	URL	
1	26357090	0	2014-01-10	Heber Smeil	Eu adoro ciência, mas acho que a falta de intimidade de cienti	2014-01-10	12:03	13	http://noticias.uol.com.br/ciencia/ultimas	
2	27603970	0	2014-01-29	marciotfeliz	Que vergonha de ler os mais absurdos comentários... as pesso	2014-01-29	14:52	13	http://noticias.uol.com.br/ciencia/ultimas	
3	26357090	0	2014-01-10	Heber Smeil	Eu adoro ciência, mas acho que a falta de intimidade de cienti	2014-01-10	12:01	13	http://noticias.uol.com.br/ciencia/ultimas	
4	26357090	0	2014-01-29	marciotfeliz	Que vergonha de ler os mais absurdos comentários... as pesso	2014-01-29	14:52	13	http://noticias.uol.com.br/ciencia/ultimas	
5	2634432	0	2014-01-08	Marcelo Pimentel	Acto interessante esse tribunal... Torna uma ação imediata para	2014-01-08	18:31	11	http://noticias.uol.com.br/cotidianou/ultim	
6	26359566	0	2014-01-09	Regi Lucas	Deixei de acreditar no Brasil depois de ler o PT mais o para	2014-01-09	17:11	11	http://noticias.uol.com.br/cotidianou/ultim	
7	26359566	0	2014-01-09	Regi Lucas	Deixei de acreditar no Brasil depois de ler o PT mais o para	2014-01-09	17:11	11	http://noticias.uol.com.br/cotidianou/ultim	
8	27489481	0	2014-01-21	Marcelo Bauab	Após ler esta matéria cheguei à conclusão de que os políticos	2014-01-22	17:01	10	http://noticias.uol.com.br/ciencia/ultimas	
9	27489481	0	2014-01-21	Marcelo Bauab	Após ler esta matéria cheguei à conclusão de que os políticos	2014-01-22	17:01	10	http://noticias.uol.com.br/ciencia/ultimas	
10	26294821	0	2014-01-04	Cleio L. V.	Ué...indo não tira tudo o que precisa da floresta? Comida? el	2014-01-04	20:20	10	http://noticias.uol.com.br/cotidianou/ultim	
11	26294821	0	2014-01-04	Cleio L. V.	Ué...indo não tira tudo o que precisa da floresta? Comida? el	2014-01-04	20:20	10	http://noticias.uol.com.br/cotidianou/ultim	
12	26318052	0	2014-01-07	Danielo Silva	O que precisa ser feita e uma intervenção no Maranhão para q	2014-01-07	12:42	10	http://noticias.uol.com.br/cotidianou/ultim	
13	26343441	0	2014-01-09	FCURRAL	Acorda Brasil!!! Onda só o que está acontecendo... O cara vai	2014-01-09	11:58	10	http://noticias.uol.com.br/cotidianou/ultim	
14	26343441	0	2014-01-09	FCURRAL	Acorda Brasil!!! Onda só o que está acontecendo... O cara vai	2014-01-09	11:58	10	http://noticias.uol.com.br/cotidianou/ultim	
15	26359389	0	2014-01-09	Rosana Leite	Quem não quer o Brasil? Quem não quer o Brasil? Quem não	2014-01-10	07:02	10	http://noticias.uol.com.br/cotidianou/ultim	
16	26359389	0	2014-01-09	Rosana Leite	Quem não quer o Brasil? Quem não quer o Brasil? Quem não	2014-01-10	07:02	10	http://noticias.uol.com.br/cotidianou/ultim	
17	26356130	0	2014-01-10	Breno SP	Trabalhar ninguém quer...após morar em SP todo mundo que	2014-01-10	10:39	10	http://noticias.uol.com.br/cotidianou/ultim	
18	27634673	0	2014-01-10	Anonimidade Garant	Gostaria de aproveitar esse espaço para MANIFESTAR meu r	2014-01-10	10:27	9	http://noticias.uol.com.br/cotidianou/ultim	
19	26292046	0	2014-01-04	waimitiro	Isso chama-se fadiga pois nada e eterno, assim como o PT. F	2014-01-31	11:28	9	http://noticias.uol.com.br/ciencia/ultimas	
20	26292046	0	2014-01-04	waimitiro	Isso chama-se fadiga pois nada e eterno, assim como o PT. F	2014-01-31	11:28	9	http://noticias.uol.com.br/ciencia/ultimas	
21	26319731	0	2014-01-07	Ilesta	A família Smeil, aliadíssima de Lula e do PT (até 2002 eram	2014-01-04	16:46	9	http://noticias.uol.com.br/cotidianou/ultim	
22	26319731	0	2014-01-07	Ilesta	A família Smeil, aliadíssima de Lula e do PT (até 2002 eram	2014-01-04	16:46	9	http://noticias.uol.com.br/cotidianou/ultim	
23	26318376	0	2014-01-07	Fa Salvo	os verdadeiros criminosos estão no poder	2014-01-07	15:02	9	http://noticias.uol.com.br/cotidianou/ultim	
24	26351148	0	2014-01-09	Hobolipo	O Maranhão de hoje é o resultado de 40 anos de vida pública d	2014-01-09	15:02	9	http://noticias.uol.com.br/cotidianou/ultim	
25	26350987	0	2014-01-09	Afro	Na TV, Roseana diz que atos "cívicos" estão sendo punidos	2014-01-07	13:05	9	http://noticias.uol.com.br/cotidianou/ultim	
26	26350987	0	2014-01-09	Afro	Na TV, Roseana diz que atos "cívicos" estão sendo punidos	2014-01-07	13:05	9	http://noticias.uol.com.br/cotidianou/ultim	
27	26356676	0	2014-01-10	Rocasa	E pensar que esta mulher quase se tornou Presidente da Repub	2014-01-09	21:37	9	http://noticias.uol.com.br/cotidianou/ultim	
28	26356676	0	2014-01-10	Rocasa	E pensar que esta mulher quase se tornou Presidente da Repub	2014-01-09	21:37	9	http://noticias.uol.com.br/cotidianou/ultim	
29	26356676	0	2014-01-10	Rocasa	E pensar que esta mulher quase se tornou Presidente da Repub	2014-01-09	21:37	9	http://noticias.uol.com.br/cotidianou/ultim	
30	26354925	0	2014-01-10	Ermano 7.1	Quem já viajou pelo Maranhão sabe o quanto esta família nefes	2014-01-10	07:40	9	http://noticias.uol.com.br/cotidianou/ultim	
31	26357786	0	2014-01-10	Arquibuteus	Via fazer que com estes 2 livros humanos, matar não pode, se	2014-01-10	10:56	9	http://noticias.uol.com.br/cotidianou/ultim	
32	27499724	0	2014-01-22	montinho	Um presidente que se preocupa com o povo, não acompanhar	2014-01-10	10:39	9	http://noticias.uol.com.br/cotidianou/ultim	
33	27499724	0	2014-01-22	montinho	Um presidente que se preocupa com o povo, não acompanhar	2014-01-10	10:39	9	http://noticias.uol.com.br/cotidianou/ultim	
34	27499724	0	2014-01-22	Micky Oliver	Um presidente que se preocupa com o povo, não acompanhar	2014-01-10	10:39	9	http://noticias.uol.com.br/cotidianou/ultim	
35	27499724	0	2014-01-22	Micky Oliver	Um presidente que se preocupa com o povo, não acompanhar	2014-01-10	10:39	9	http://noticias.uol.com.br/cotidianou/ultim	
36	27499724	0	2014-01-22	Micky Oliver	Um presidente que se preocupa com o povo, não acompanhar	2014-01-10	10:39	9	http://noticias.uol.com.br/cotidianou/ultim	
37	27499724	0	2014-01-22	Micky Oliver	Um presidente que se preocupa com o povo, não acompanhar	2014-01-10	10:39	9	http://noticias.uol.com.br/cotidianou/ultim	
38	27604961	0	2014-01-29	Corone Antônio	Simples, que se fala um levantamento da origem dessas pesq	2014-01-11	09:12	9	http://noticias.uol.com.br/cotidianou/ultim	
39	27604961	0	2014-01-29	Corone Antônio	Simples, que se fala um levantamento da origem dessas pesq	2014-01-11	09:12	9	http://noticias.uol.com.br/cotidianou/ultim	
40	27491259	0	2014-01-21	Renato - Juridial	Sugestões para resolver a criminalidade no Brasil: 1- prisão per	2014-01-11	00:21	9	http://noticias.uol.com.br/cotidianou/ultim	
41	27491259	0	2014-01-21	Renato - Juridial	Sugestões para resolver a criminalidade no Brasil: 1- prisão per	2014-01-11	00:21	9	http://noticias.uol.com.br/cotidianou/ultim	
42	26291704	0	2014-01-04	Corone Antônio	FOI PASSEAR COM OS OUTROS? PE' TRALHAS NOS ALPES?	2014-01-22	19:03	8	http://noticias.uol.com.br/blogg-e-columa	
43	26291704	0	2014-01-04	Corone Antônio	FOI PASSEAR COM OS OUTROS? PE' TRALHAS NOS ALPES?	2014-01-22	19:03	8	http://noticias.uol.com.br/blogg-e-columa	
44	26291704	0	2014-01-04	Corone Antônio	FOI PASSEAR COM OS OUTROS? PE' TRALHAS NOS ALPES?	2014-01-22	19:03	8	http://noticias.uol.com.br/blogg-e-columa	
45	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
46	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
47	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
48	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
49	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
50	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
51	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
52	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
53	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
54	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
55	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
56	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
57	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
58	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
59	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
60	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
61	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
62	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
63	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
64	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
65	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
66	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
67	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
68	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
69	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
70	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
71	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
72	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
73	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
74	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
75	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
76	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
77	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
78	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
79	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
80	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
81	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
82	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
83	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
84	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
85	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo; essa gente está conseguindo destuir o B	2014-01-22	19:43	8	http://noticias.uol.com.br/blogg-e-columa	
86	26319736	0	2014-01-07	Marcos Gonçalves	Mais é o que tu digo;					

The number of different usernames in the corpus is 47,393, averaging 4.27 comments per username. To comment on UOL news, readers must register in the website with a unique username – that is, it is not possible for more than one reader to share the same username. This ensures that, in principle, all the comments assigned to the same username were posted by the same reader. On the other hand, it is possible that a same user registered herself more than once, using two different accounts. Again, there seems to be a “rich-get-richer” pattern: the most active reader in the corpus is the writer of no less than 948 comments, while 27,221 users interacted only once. Lastly, we highlight the fact that 40,764 (20.13%) of the comments are actually replies to other comments, which suggests a remarkable interaction between UOL’s readers.

As an illustration, Figure 2.4 depicts a fragment of the corpus Xereta displayed in the spreadsheet editor LibreOffice Calc.

Comparison with other corpora of comments from news portals

The corpus Xereta is not the first corpus of news portals comments ever made available. Here we mention and compare other relevant corpora that contain this type of text. We do not cite data collections that were performed only for the purpose of a specific study and were not made available in corpus format to the scientific community.

At the time of publication of this dissertation, the SFU Opinion and Comments Corpus (SOCC)¹⁶ (Kolhatkar et al., 2019) is probably the larger corpus of comments available. It is focused on English-written Canadian media sources and includes 663,173 comments, comprised in a five-year period (from January 2012 to December 2016). A small portion of the corpus (1,043 comments)

¹⁶ Available at <https://github.com/sfu-discourse-lab/SOCC> .

is also annotated for constructiveness, toxicity, negation and appraisal. Differently from the corpus Xereta, however, the articles considered by SOCC are all opinion articles, not hard news articles.

The Yahoo News Annotated Comments Corpus (YNACC)¹⁷ (Napoles et al., 2017) is another large and well-curated corpus of news readers comments written in English. It contains 521,608 comments posted in response to Yahoo! News articles. Among these, 9,160 are annotated for sentiment, persuasiveness and tone. To illustrate how laborious and expensive the process of annotation of such a corpus is, the authors report that this task was performed by 26 professional trained editors and 495 untrained crowdsourced workers. Despite being a large corpus, the YNACC is not suitable for diachronic analysis, since it only includes comments for articles published in April 2016. Another corpus of comments from news portals in English is the SENSEI Social Media Annotated Corpus (Barker et al., 2016). This corpus is not comparable to the previous ones (including Xereta), since it contains only 1,850 comments – all of them annotated – collected from a small set of 18 articles from the British news portal The Guardian.

Cotterell et al. (2014) present an Algerian Arabic-French code-switched corpus¹⁸ that contains 339,504 comments, from which 1,000 are annotated for word level language identification. Fišer et al. (2018), as part of the Janes project¹⁹ – which aims to develop language resources and tools for Slovene user generated content –, describe a corpus containing 299,219 comments from Slovene news portals and encompassing a period of eight years (from March 2007 to January 2015)²⁰. This corpus is entirely tokenised, sentence seg-

¹⁷ Available at <https://github.com/cnap/ynacc> .

¹⁸ Available at https://github.com/ryancotterell/arabic_dialect_annotation .

¹⁹ Available at <http://nl.ijs.si/janes> .

²⁰ The news comment corpus Janes-News 1.0 is available at <https://www> .

mented, word normalised, morphosyntactically tagged, lemmatised and annotated with named entities, which probably makes it the largest annotated corpus of comments up to the publication date of this dissertation.

Besides Xereta, at least two corpora of online news comments in Portuguese have already been compiled. SentiCorpus-PT²¹ (Carvalho et al., 2011) is an European Portuguese annotated corpus of comments to political debates. It includes 2,800 comments from a short period of time (ten days in September 2009). ComentCorpus (Pedro, 2018) is composed by 6,185 comments manually collected from 90 news articles related to the impeachment process of Dilma Roussef, in Brazil. It contains comments from between January and July 2016. Since its content is entirely related to politics, it cannot be considered a corpus of general comments. SentiCorpus-PT and ComentCorpus are focused on opinion mining and irony, and, therefore, are annotated with semantic-discursive information.

Table 2.4 summarizes the main characteristics of the above-mentioned corpora. It becomes evident that much remains to be done in this domain. In particular, there is a lack of large general and temporally broad corpora of news comments in Portuguese, and the corpus Xereta comes to (partially) supply this demand. In future work, we intend to follow the steps of most of the other corpora and add annotation to at least a small portion of our corpus.

2.4 Research possibilities

The two resources described above hold potential for use in various fields of knowledge, including linguistics, communication and media studies, sociology, political science and digital humanities. In this section, we present an overview of a few studies that could employ

clarin.si/repository/xmlui/handle/11356/1140 .

²¹ Available at <https://github.com/davidsbatista/REACTION-resources>

Table 2.4: Basic information regarding different corpora of news comments. The corpora are ordered according to the total number of comments.

Corpus	Language	Number of comments
SFU Opinion and Comments Corpus	Canadian English	663,173 (1,043 annotated)
Yahoo News Annotated Comments Corpus	English	521,608 (9,160 annotated)
Algerian Arabic-French code-switched corpus	Algerian Arabic/French	339,504 (1,000 annotated)
Janes-News 1.0	Slovene	299,219 (completely annotated)
Xereta	Brazilian Portuguese	202,541 (not annotated)
ComentCorpus	Brazilian Portuguese	6,185 (completely annotated)
SentiCorpus-PT	European Portuguese	2,795 (completely annotated)
SENSEI Social Media Annotated Corpus	British English	1,850 (completely annotated)

our Web scraper and/or our corpus, hoping that these ideas might turn into actual research projects.

Many of the potential studies mentioned in the next paragraphs can be accomplished using the already existing corpora analysis tools. Figure 2.5 shows concordance lines of the corpus Xereta in AntConc, a freeware tool for carrying out corpus linguistics research.



Figure 2.5: Concordance lines (fragment) of the corpus Xereta at AntConc using the search term *Brasil*.

Diachronic research

As mentioned earlier, a diachronic corpus “is a collection of texts including information on the time period to which they relate, e.g. the publication date of a document” (Trevisani and Tuzzi, 2018, p. 130). The corpus Xereta is, by definition, a diachronic corpus, since it provides temporal information concerning the publication of both the news article and the comment. Figure 2.6 displays a fragment of our corpus, highlighting the meta-information regarding date and time of publication of each comment. Also, the files output by our Web scraper contain this information as well.

Comentário	Data do comentário	Hora do comentário
Que vergonha de ler os mais absurdos comentários... as pessoas falam e escrevem o que pensam sem embasa	2014-01-29	14:52
Eu adoro ciência, mas acho que a falta de intimidade de cientistas com bichos, faz com que descubram o óbvio.	2014-01-10	12:01
Acho interessante esse tribunal... Toma uma ação imediata para suspender a implantação de faixas por "suspeita	2014-01-08	18:31
Um absurdo a justiça meter o bedelho. Não sou PT, mas o prefeito Fernando Haddad está na direção correta: a u	2014-01-08	18:28
Pessoal, não se esqueçam que o Lula idolatra essa governadora...Porque será?? Nao sou a favor da pena de mo	2014-01-09	21:14
Após ler esta matéria cheguei à conclusão de que os políticos brasileiros são himens imperfurados, não servem	2014-01-22	17:01
Ué... índio não tira tudo o que precisa da floresta? Comida?, eles caçam, pescam e plantam de tudo (senão não p	2014-01-04	20:20
O que precisa ser feita é uma intervenção no Maranhão para que o Brasil recupere aquele estado que há muito de	2014-01-07	12:42
Acorda Brasil !!! Olha só o que esta acontecendo... O cara veio para criar problemas e não para aerar soluções...	2014-01-09	11:58

Figure 2.6: Fragment of the corpus Xereta highlighting comments’ temporal (date and time) information.

Being a diachronic corpus, Xereta might be employed in research aiming to analyze various phenomena (whether linguistic or not) from a temporal perspective. In the field of linguistics, it is clear that research on language variation and change can benefit from temporal information, which can be used to study how language evolved over a period of time. Since news portals comments are part of a relatively new text genre, any corpus containing such texts will cover a short time span – at most a few years. This obviously puts restrictions on the kinds of research on language variation and change that can be performed. However, previous research (e.g. Cunha et al., 2011; Danescu-Niculescu-Mizil et al., 2013; Eisenstein et al., 2014) shows that language use quickly evolves in social media, spreading across social network connections. Therefore, the investigation of the phenomenon of language variation and change in an

online environment, such as the one presented here, has proved to be a very fertile field, particularly in the domain of the lexicon. One practical example is the analysis of neologisms on the Internet (e.g. Rumšienė, 2004; Zhang et al., 2013).

Besides that, one interesting potential line of research is related to linguistic style accommodation – that is, the fact that participants in interactions “tend to nonconsciously converge to one another’s communicative behavior”, coordinating “in a variety of dimensions including choice of words, syntax, pausing frequency, pitch and gestures” (Danescu-Niculescu-Mizil et al., 2011, p. 745). The authors of the study just cited confirmed the hypothesis of linguistic style accommodation in a large dataset of Twitter conversations, and similar studies could analyze this phenomenon in Brazilian Portuguese, focusing on the genre *comment in news portal* as well.

Regarding fields other than linguistics, we might mention the possibility of using our corpus to investigate changes in readers’ behavior through time. Examples are behaviors such as bullying, incivility, harassment and disrespect, typically present in online discussions and comment boards (see Sarmiento and Mendonça, 2016): do they exhibit the same characteristics across time? How do changes in external circumstances, like political or economic scenarios, influence the way readers interact with news articles and with other readers? Studies have shown a sharp rise in political polarization in Brazil in the second decade of the 21st century (see Hunter and Power, 2019): is it possible to corroborate this from the analysis of news portals comments over time? These are all issues that can be explored using our resources.

Positively and negatively rated comments

In the UOL news portal, readers are able to positively rate a comment by giving it a “like” (or “thumbs up”). Figure 2.7 depicts two comments extracted from an UOL news story, each having been

positively rated by eight readers, while Figure 2.8 highlights positive rating information in a fragment of the corpus. Unlike in other websites (such as Yahoo! News, shown in Figure 2.1), UOL readers, at the time of the writing of this research, do not have the possibility of rating comments negatively, which is the reason why our corpus does not include a column dedicated to the report of “dislikes” (or “thumbs down”). However, a slight change in the code is enough to include a column devoted to this information when collecting from sites that offer this possibility of interaction to their users.

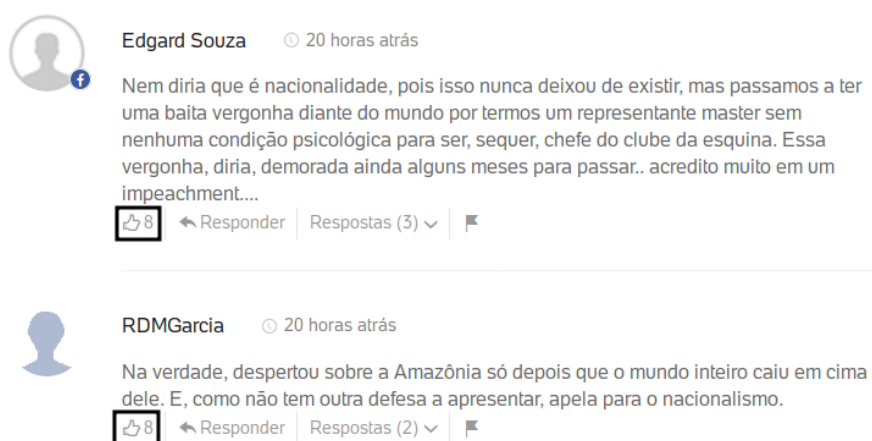


Figure 2.7: Positively rated comments in the UOL news portal. As highlighted in the image, each of these comments received eight “likes” (or “thumbs up”).

Previous studies have investigated linguistic, stylistic, and discursive characteristics of positively and negatively rated replies in social media (e.g. Siersdorfer et al., 2010; Möller et al., 2019) and in question-and-answer platforms such as Stack Overflow (e.g. Calefato et al., 2015) and Yahoo! Answers (e.g. Shah and Pomerantz, 2010), usually employing a computational approach and with the purpose of predicting comment/answer quality. In addition,

Comentário	Data do co	Hora	Curtidas	URL
Você é que tirou nosso amado país do rumo, com tantas riquezas hoje por	2014-06-13	16:28	21	http://noticias.uol.com.br/politica/ultimas
Vai inaugurar até poça d água dizendo que é açude !!!!	2014-07-01	15:17	21	http://noticias.uol.com.br/politica/ultimas
É incrível como alguns PTistas aqui no forum defendem essa presidanta d	2014-06-13	16:28	19	http://noticias.uol.com.br/politica/ultimas
Da mesma forma que educação vem de casa, honestidade também. Não e	2014-06-13	16:27	17	http://noticias.uol.com.br/politica/ultimas
Eu tento evitar comentários, mas é muita mentira, não dá para ficar calad	2014-05-19	14:21	16	http://noticias.uol.com.br/politica/ultimas
Se ela não se importa com xingamentos pq não falou na abertura...se fez	2014-06-13	16:35	16	http://noticias.uol.com.br/politica/ultimas
Isto é reflexo que ela e o Pt criaram pois prometeram o mundo e fundo olh	2014-06-13	16:28	16	http://noticias.uol.com.br/politica/ultimas
Obras não acabadas, mas inauguradas, faz parte da hipocrisia e demagogi	2014-07-01	15:00	16	http://noticias.uol.com.br/politica/ultimas
VENDO A TRAJETORIA DE NOSSA PRESIDENTA !!! FICO MUITO EMOO	2014-01-29	19:39	15	http://noticias.uol.com.br/politica/ultimas
Se a Dilma está certa temos o seguinte. No governo FHC investiu-se 15.7	2014-05-19	14:17	15	http://noticias.uol.com.br/politica/ultimas
OBRIGADO MINISTRO JOAQUIM BARBOSA, POR TUDO. Principalment	2014-07-01	15:29	15	http://noticias.uol.com.br/politica/ultimas
Basta ler os comentários de alguns para se ter uma idéia de como o brasil	2014-06-27	14:44	14	http://noticias.uol.com.br/opiniaio/coluna

Figure 2.8: Fragment of the corpus Xereta highlighting positive rating (“likes”) information.

more fine-grained qualitative studies could be conducted in this area. Since our corpus provides not only the number of positive evaluations received per comment, but also other meta-information such as the headline and the URL of the corresponding article, it is possible to comparatively analyze the ratings of comments in the same news story and examine which features make them gain more (or less) positive evaluations by other readers. A potential research question is: are the characteristics of more/less positively evaluated comments the same in articles mentioning different entities (such as two different politicians) or concerning different topics (sports, technology, entertainment etc.)? This type of research question could be addressed both through linguistic and social methodological approaches.

Anonymity

Usually, news portals readers are free to post comments using freely created usernames, and it is not difficult to observe a great variability among them: while some readers seem to draw upon their own names (in some cases, even their full names), others use eccentric nicknames that entirely prevent them from being identified. In Figure 2.9, which shows a fragment of the corpus highlighting commentators’ usernames, we find cases of readers who apparently use their

own names, such as “Fabiano Palma” and “Rafa Ferreira S”, and others who, on the contrary, prefer completely anonymous usernames, like “comp”, “o predador” (*the predator*) and “Povo” (*People*).

Data da notícia	Usuário	Comentário
2014-07-11	Fabiano Palma	políticos deveriam usar energia que gastam em c
2014-07-11	Luiz Emboabas	Nós brasileiros deveríamos é correr com o PT-PM
2014-07-11	comp	Kassab, se o povo brasileiro se dedicasse a ente
2014-07-11	Nog.	Ótimo artigo! Abordou de forma bem simples um
2014-07-11	Luiz Emboabas	Se todo dia sai um bobo de casa, quer dizer este
2014-07-11	Rafa Ferreira S	E vocês sugerem o que? Ficar assistindo tudo c:
2014-07-11	ferruccio	caros senhores inocentes brasileiros, vou dar um
2014-07-11	Edmundo animal	Para Kassab, voce deveria ter usado sua energia
2014-07-11	Luiz Emboabas	Lacraios do poder...! Fora PSB-PT!
2014-07-11	ferruccio	se o brasileiro tiver vergonha na cara, dedica me
2014-07-11	Nelsonspsp	Se brasileiros se dedicassem mais à politica, po
2014-07-11	Yuri L	"O Brasil ganharia muito se os brasileiros dedica
2014-07-11	ROCK 10	ESSE CABRA SÓ PODE ESTAR DE BRINCADÉ
2014-07-11	GetulioC	Quem terá coragem de perder tempo para ler o q
2014-07-11	GaloGalo	Esse rapazinho tem grande sensibilidade política
2014-07-11	Nespolo	Pois é Kassab, uma das coisas que o eleitor terr
2014-07-11	o predador	a goleada da alemanha representa a vitória da m
2014-07-11	OTal Cinquentaesete	Brasileiros deviam votar em protesto, VOTAR EM
2014-07-11	desnorteado	Políticos devem dedicar à população parte da en
2014-07-11	Povo	cala a boca kassapa....

Figure 2.9: Fragment of the corpus Xereta highlighting commentators’ usernames.

Correa et al. (2015) show that the linguistic differences between anonymous and identified social media content are so significant that automated classifiers can be trained “to distinguish between them with reasonable accuracy” (p. 71). However, as far as we are aware, this phenomenon has not yet been analyzed for texts of the genre *comment in news portal* – certainly not for texts in Brazilian Portuguese. One hypothesis to be investigated is whether users with less identifiable usernames tend to exhibit more aggressive and hostile behaviors than (apparently) less anonymous commentators.

The onomatological analysis of commentators’ usernames also seems like an interesting topic for research in its own right. According to Hooker (2019), social media usernames may be completely

random, but can also be used to refer to attributes or characteristics of the users, their preferences, significant dates or events, or their actual names. The author adds that previous research shows that “participants displayed their identities through the choice of their username” (Hooker, 2019, p. 80). For instance, several usernames found in our corpus are somehow expressing political preferences²². The question then arises how they relate to the behavior of commentators in the platform.

Analysis of conversation threads

Another piece of meta-information present in the Xereta corpus and in the files output from the scraper indicates whether a comment constitutes a reply to a previously posted comment. Sequences of replies to the same comment can trigger true conversations between readers, which can also be analyzed. These can be called *conversation threads* or *conversation sessions*. Figure 2.10 displays a short conversation thread held by three readers of a particular news story at the UOL webpage. De Choudhury et al. (2009) observe that “[p]eople return to a video post that they have already seen and post further comments (say in YouTube) in response to the communication activity, rather than to watch the video again”. This behavior also occurs in the case of online news commentators – and Figure 2.10 itself is an example, since the original commentator reacted to the replies received by her/his own comment.

Figure 2.11 displays how this information is present in the corpus: in the field “Resposta a...” (*reply to...*), the value “0” indicates comments that are not replies to any other comments, while other values refer to the IDs of the replied comments.

Conversation threads pertaining to online social networks have

²² A few selected examples: “Anti PT 2014” (*PT* refers to the Workers’ Party, a major social-democratic political party in Brazil); “Mais a esquerda” (*more to the left*); “Apolitico” (*apolitical*).



Oráculo Aprendiz 20/01/2014 13h44

Os avanços científicos vem dessas maluquices sim. Infelizmente precisamos delas. Mas antes de desvendar os segredos do Universo, o homem deveria desvendar os segredos da sua própria alma, que permanece obscura e sem direção.

0 Respostas (3)



Oráculo Aprendiz 21/01/2014 11h59

Caro Gymno e Marcos A....Se todos os que acreditam em alma e julgamento estiverem errados e no fim das contas tudo não passou mesmo de devaneio, tudo bem...será o simples fim de todos. Mas se for verdade e os incrédulos é que estavam errados...coitados deles. Quem quiser pagar pra ver, que pague!

0



Gymno 20/01/2014 17h31

alma? tem certeza de que há alguma que não seja puro resultado de um devaneio, seja ele religioso ou fóbico (o medo da morte)?

0



Marcos A. 20/01/2014 14h43

Esse negocio de alma é coisa para espírita e não para cientista.

1

Figure 2.10: Replies to a comment in the UOL news portal. In this example, users “Marcos A.” and “Gymno” reply to the original comment of “Oráculo Aprendiz” (the top comment). Then, “Oráculo Aprendiz” replies to both “Marcos A.” and “Gymno” (note that replies are displayed in reverse chronological order).

been the topic of previous research (e.g. Gómez et al., 2008; Ferguson et al., 2014). Caetano et al. (2019), for instance, work on the notion of *attention cascades* in WhatsApp groups, which can also be applied to the case of replies to comments in news portals. According to these authors, “an *attention cascade* begins when a user makes an assertion about a topic in a message to the group” and continues when “[o]ther users join and establish a conversation

ID	Resposta a...	Título da notícia	Data da notícia	Usuário	Comentário
72839088	0	Vou dar corda para Sa	2014-06-24	araka	bando de inguinorante,esse om
72839200	0	Vou dar corda para Sa	2014-06-24	aguia62	Se no brasil tivessem homens ,
72839217	0	Vou dar corda para Sa	2014-06-24	Monteiro70	Ô Eunicio, tu tens raiva do Ama
72839221	0	Vou dar corda para Sa	2014-06-24	PT ForEver	"Antigos espíritos do mal, trans
72839375	0	Vou dar corda para Sa	2014-06-24	pasamu	Demorou, está notícia é muito b
72839557	0	Vou dar corda para Sa	2014-06-24	Paulo Moura	Dá uma corda de cânhamo para
72843906	0	Vou dar corda para Sa	2014-06-24	fugugugugu	Va com Dios
72836130	72836011	Vou dar corda para Sa	2014-06-24	Politico2013	Parece que v. só tem 20 anos...
72836391	72836094	Vou dar corda para Sa	2014-06-24	GALLOOOO	Ela vai pagar a conta fácil, fácil
72836590	72836095	Vou dar corda para Sa	2014-06-24	Tite 108	O povão miseravel do BEIRAD/
72838678	72836095	Vou dar corda para Sa	2014-06-24	Oliver Silva	O Amapa votou nesse verme se
72844605	72836212	Vou dar corda para Sa	2014-06-24	Matrica	Vai e leva todos corrupto junto c
72836577	72836311	Vou dar corda para Sa	2014-06-24	ribeiroed	CONCORDO TOTALMENTE
72837845	72837444	Vou dar corda para Sa	2014-06-24	VimVivenci	Acho que ele deveria ser o futur
72844531	72837888	Vou dar corda para Sa	2014-06-24	aliciq	É verdade riqueza com o sofrim
72844411	72838701	Vou dar corda para Sa	2014-06-24	aliciq	Ainda não morreu porque está p
72844476	72839018	Vou dar corda para Sa	2014-06-24	aliciq	Eu só sinto Saudades do Ferna
72839692	72839056	Vou dar corda para Sa	2014-06-24	Matrica	Coitado do povos do Maranhão
72839637	72839221	Vou dar corda para Sa	2014-06-24	Matrica	São Espírito do mal

Figure 2.11: Fragment of the corpus Xereta highlighting the field “Resposta a...” (*reply to...*). The value “0” indicates comments that are not replies to any other comments, while other values refer to the IDs of the replied comments.

thread by explicitly replying to the root message” (Caetano et al., 2019, p. 29, emphasis in original). The use of the reply feature is then “a signal that the user’s attention was caught by the message she is replying to and that the cascade is a (semi-)structured representation modeling emergent patterns of *collective* attention” (p. 29, emphasis in original). One of the research questions proposed in the previously mentioned study is: “how different are attention cascades in political and non-political groups?” Similar analyses could be performed in the context of news portals comments, since, as far as we know, no studies in this direction have been carried out so far. Also, “[a]s in any real conversation of a group of people, attention may drift to other (possibly weakly related) topics as the conversation goes on” (Caetano et al., 2019, p. 29-30): a possible line of research could be to investigate how far from the news themselves these conversation threads can get.

Cross-corpus research

Cross-corpus research is the comparative use of different corpora to investigate the same (or similar) phenomena. By using this approach, researchers are able to compare different languages, different varieties of the same language or different periods of time, to give some examples. In this case, it seems opportune to employ our corpus to compare news portals comments with texts from other Internet genres or with comments taken from other contexts, such as letters sections in print media. Researchers might be interested, for example, in comparing the use of a specific linguistic item (a word, an expression, a grammatical structure etc.) between genres.

Berber Sardinha (2013) uses multidimensional analysis to compare different Internet and “traditional” (offline) genres, and to understand the variation between these genres and the writing style specific to them. In his work, however, only the genres “email”, “blog post”, “tweet”, “Facebook post” and “webpage” are considered. The availability of the resources presented here now allows the use of the text genre *comment in news portal* also to enter into this type of analysis. One of the observed characteristics of most Internet genres is their proximity to informal oral language. Researchers could compare comments obtained from our resources with pieces of text gathered from different sources on this regard, trying to observe how far/close from oral language these comments are when compared to other Internet genres.

In the above paragraphs, we offered some potential research ideas that may use the resources presented in this chapter. Of course, our intention is not to exhaust all the research possibilities, but rather to contribute by providing ideas and suggestions that could be followed in the future. Our main intention here is to promote and facilitate research on this new text genre, which seems

so interesting and relevant nowadays.

2.5 Concluding remarks

In this chapter, we present and describe two useful resources for research in the fields of corpus linguistics, studies on Internet communication, and digital humanities in general: (a) a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and (b) a freely available corpus composed of comments published at UOL, a major Brazilian news portal. Here, we show that our Web scraper is simple to operate and can be used even by individuals with limited computational skills. This is a feature that we consider very important, given our interest in promoting a broader access to digital content for researchers with little knowledge of programming and Internet data collection, especially in developing countries. We also describe a second (and, at the time of this publication, current) version of our corpus containing more than 200,000 comments (and respective meta-information) posted by readers as reactions to UOL's news articles. This corpus makes it possible to analyze linguistic, textual, and discursive characteristics of the genre of news comments. Lastly, we mention some ideas about projects that could be carried out using the resources presented here, and we expect these ideas to be of interest to academics who wish to promote research on this developing Internet genre.

As far as we are concerned, ours is not only a contribution to corpus approaches in a language other than English, but also the first available large and general corpus of online news comments in Portuguese language. Besides being useful for academic research in linguistics, journalism, social science and other fields, a corpus of online comments may also contribute to the task of Web archival (Day, 2006; Webster, 2017), since it is preserving textual

material that risks being eliminated by companies (due to maintenance costs savings or website redesigns) or simply disappearing together with their news portal hosts in case of website closure or the bankruptcy of a company. In these cases, this corpus could be used, in the future, as a way of accessing human interactions and Web behaviors from past times.

This is of course a long-term and, ideally, an always-ongoing project. In future steps, we plan to keep improving the Web scraper for it to be able to collect comments from additional news portals, be them Brazilian (e.g. G1, Terra), Dutch (e.g. NU.nl) and other (e.g. The New York Times, Yahoo! News). Also important is the maintenance of the presently available scraper, since news portals might eventually modify their page structures, making the current code obsolete²³. Accordingly, we also intend to make available new versions of the corpus, with larger numbers of comments collected from different news portals and in different languages, ultimately enabling cross-language analyses of similarly collected data.

Finally, it is important to mention that the work involved in corpus compilation, preparation and maintenance is extensive, which is a reason why we invite other members of the community interested in this task to participate by improving codes, adapting our Web scraping tool so that it can be used to gather comments from more news portals, and building collections of comments extracted from different sources. All of this would be indispensable for the full development of our resources, and the encouragement for the creation of such a community of developers and users is also one of our goals.

²³ Actually, this very incident occurred during the execution of this work: UOL slightly changed the structure of its news pages and, in 2019, we needed to restructure the module responsible for scraping from this site to get our Web scraper back into operation.

Acknowledgements

We thank Patrícia Rossini, who, while a doctoral student in Social Communication at Universidade Federal de Minas Gerais, motivated us to elaborate the Web scraper of comments and assisted us by testing preliminary versions in her own research, which culminated in her doctoral dissertation on political conversation, incivility and intolerance in digital environments (Rossini, 2017).

CHAPTER 3

Establishment and obsolescence of linguistic items in a diachronic corpus¹

3.1 Introduction

Diachronic and historical corpora are useful tools to study linguistic phenomena that unfold over time, including processes of variation and change. Previous work has employed these kinds of corpora to analyze language change in progress (Hundt and Mair, 1999), to infer cases of variation and change (Bauer, 2002), and to investigate language change using word vector embeddings (Hamilton et al., 2016), to mention a few.

When using diachronic corpora for investigating language variation and change, one of the relevant tasks for researchers is the iden-

¹ This chapter reproduces with minor changes the article “An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus” (Cunha and Wichmann, in press), accepted for publication in *Corpora*. See Appendix C for more information.

tification of specific time periods in which certain linguistic items arise and, conversely, vanish. It is particularly valuable to detect when items (i) are first attested, (ii) become established in the corpus, (iii) become obsolete and (iv) are attested last. Although the detection of the earliest and the latest attestation dates of items in a diachronic corpus is trivial, the same cannot be said about their establishment and obsolescence, because there are no clear and commonly accepted criteria for pinpointing when an item is getting established and when it can be regarded as obsolescent (cf. Tichý, 2018).

The aim of this chapter is threefold: first, to formulate a set of criteria to define binary notions of establishment and obsolescence of items in a diachronic corpus; second, to present an algorithm to assist in the identification of specific time periods of establishment and obsolescence of linguistic items in diachronic corpora according to the previously mentioned criteria; and, finally, to use this algorithm to make a series of general considerations based on real data for the purpose of demonstrating the utility of the methods presented here and for making some observations on two centuries of the dynamics of the American English lexicon that are interesting in their own right. We will observe, among other findings, that the proportion of words established in a given decade is similar across decades and, by studying the words stemming from different decades that are most frequent today, we will get an impression of how the lexical heritage of contemporary American English bears the imprints of salient aspects of life as it was experienced during specific, previous decades.

The algorithm proposed here is simple and generalizable. It can be applied to any corpus that is divided into time frames, regardless of language or historical period, since it only takes as input information on the frequency of the analyzed items in each time frame. Likewise, the nature of the items under analysis is, in principle,

irrelevant to the applicability of our algorithm, so it can also be implemented to examine aspects of language not considered in our case studies, such as phonology or morphosyntax. Moreover, the algorithm, or some derived version, should be generally applicable to the investigation of time series of sociological, anthropological or historical data.

3.1.1 Related work

Previous quantitative investigations on language dynamics have dealt with the notions of birth and death of linguistic items, which are related to the concepts of establishment and obsolescence contemplated here. Petersen et al. (2012), for instance, analyze more than 200 years of data from three different languages with the goal of shedding light on the aggregate dynamics of word evolution in written texts. They investigate variations in the use of words during their lifespans and, among other results, identify a tendency for a peak in word use growth rate to occur around 30-50 years after a word's first attestation in their corpus. Furthermore, the authors find evidence that the dynamics of word evolution might be influenced by historical events, such as wars. This last observation is also made by Bochkarev et al. (2014), who additionally find a relationship between the frequency of a word and its stability in the lexicon of a language, confirming previous results from Pagel et al. (2007). Moreover, Perc (2012) analyzes the evolution of high-frequency English words and phrases, discovering that their lifespan is not uniform across the centuries, and Michel et al. (2011) investigate some patterns in the evolution of English lexicon and grammar. In addition, the work performed by Kerremans et al. (2012) in the scope of the NeoCrawler project² presents a Web crawler that iden-

²Available at <http://www.neocrawler.anglistik.uni-muenchen.de/crawler/html/>.

tifies and retrieves neologisms from the Internet, making it possible to analyze “how words spread in the early stages of their life cycles and how they begin to establish themselves in lexical and semantic networks” (p. 59).

Certainly connected to the concepts of first attestation, establishment, obsolescence and last attestation of an item in a corpus are the studies that use diachronic corpora to investigate language variation and change. Biber and Gray (2011), for instance, analyze the influence of written language on grammatical change, and suggest that new grammatical uses and functions emerge not only in spoken interaction, but also in written registers. Topics such as the variation of the English genitive (Hinrichs and Szmrecsanyi, 2007), the variation of complex prepositions in Brazilian Portuguese (Shepherd, 2014) and the change in the grammar of English verbs (Hilpert and Mair, 2015), to illustrate, have been considered in previous investigations that made use of diachronic corpora. The use of corpora with the aim of investigating creativity in literary and ordinary language, including novel word formation, is scrutinized by Vo and Carter (2010), while Moon (2010), in tackling the question of what corpora can reveal about lexicon, mentions that these tools might contribute to the analysis of the establishment and the institutionalization of new derivations and compounds in a language.

The notion of establishment of linguistic items in a diachronic written corpus from a particular language is not to be confused with the concept of entrenchment of structures in the memory of speakers (Langacker, 1987), which is central in the field of cognitive linguistics. Nevertheless, Schmid (2007) considers that this notion of entrenchment “also applies to language as such and whole speech communities, because the frequency of occurrence of concepts or constructions in a speech community has an effect on the frequency with which its members are exposed to them” (p. 119). As a consequence, it should be possible to talk of a degree of entrenchment

of a linguistic item not only in the memory of individual speakers, but also in a specific language. Indeed, Croft (2000) uses the notion of entrenchment in his proposal of an evolutionary model of language change, advocating for a strong relationship between the perpetuation of a given linguistic structure in the language and the degrees of entrenchment of this particular structure in the grammar of speakers. However, in discussing the relationship between frequency in natural language use and the entrenchment of complex linguistic strings in the minds of language users, Blumenthal-Dramé (2012) argues for a weak version of the so-called *corpus-to-cognition principle* – since, according to her, only “certain corpus-extracted variables may, to some extent, be used as a yardstick for entrenchment in the brain of an average language user” (p. 205). In this study, it is not our goal to contemplate entrenchment in the memory of speakers, nor to elaborate on the relationship between the frequency of linguistic items in a corpus and their entrenchment in the minds of individuals. For this reason, we opted for the use of the term *establishment* and, by not using the loaded term *entrenchment*, we hope to avoid any kind of misinterpretation of the goals of our proposed method.

Regarding the opposite phenomenon – that is, the loss of linguistic items –, Tichý (2018) presents one of the few studies on lexical obsolescence and mortality in English. Using a fine-grained methodology based on the difference between frequency levels in distinct periods of time, the author proposes a method for extracting from large corpora forms that were once common but later became obsolete. Our methodology differs from his in that Tichý is mostly interested in words that were once very common in the language, while our proposed methodology is more flexible in this regard. Also, Tichý’s proposal, being more fine-grained, is more computationally demanding, whereas our methodology is simpler and more straightforward. We consider the approaches complementary and

imagine that they may even be used together in some specific situation.

Finally, the work of Hilpert and Gries (2009) provides several resources for the assessment of frequency changes in multistage diachronic corpora. The authors present suggestions for the analysis of this kind of data, displaying examples and use cases of great value for historical linguists. In particular, we mention the introduction of the *iterative sequential interval estimation* (ISIE), a method that provides a range of expected frequencies for an item in each time period of the corpus. When the frequency “happens to go beyond the expected values, we have detected a change that merits further attention” (Hilpert and Gries, 2009, p. 393).

3.2 Defining establishment and obsolescence as binary notions for diachronic corpus linguistics

Dictionaries and glossaries of neologisms (e.g. Ayto, 1989, 1990, 1999; Tulloch, 1991; Algeo and Algeo, 1993; Knowles and Elliott, 1997, to mention works on English) attempt to record recent additions to the language, but their editors are usually aware that what they characterize as *new words* might not be new at all. In fact, Tulloch (1991) mentions the potential gap between the point in time in which a word enters the language and the moment when the general public becomes aware of it – which is the occasion when the neologism might be included in the most prestigious dictionaries and can be considered “established in the language” (Ayto, 1999, p. iii). Still, most of the past studies mentioned in Section 3.1.1 that analyze time periods in which linguistic items arose and vanished associate birth and death with, respectively, first and last attestations in a corpus. In this study, we argue and show evidence that

the first appearance of an item in a corpus may occur considerably earlier than its establishment in the corpus itself and, conversely, that an item might still appear in the data long after it became obsolete (see Section 3.4). This fact suggests that it may often be convenient to discriminate between first attestation and establishment as well as between last attestation and obsolescence, so as to obtain a more accurate description of the lifespan of a linguistic item.

As pointed out by Widdowson (2000), it is important to emphasize that a corpus is different from a language and, consequently, that the establishment or the obsolescence of an item in a corpus does not necessarily imply its establishment/obsolescence in a language. At most, it might be claimed that a corpus represents part of a language and that a relationship between these two entities exists.

We are interested in defining binary (rather than continuous³) notions of establishment/obsolescence in order to indicate whether a linguistic item may have arisen in or vanished from a diachronic corpus during the period covered by it. This is particularly useful for researchers interested in extracting lists of candidate items for further research (see Section 3.4.3). We stipulate that, in a particular corpus which includes diachronic information, each linguistic item (be it a word, a morpheme, a syntactic structure or other) may usefully be classified as being in one – and only one – of the following possible states in a given period: (a) established; (b) obsolete; (c) permanent; (d) short-lived; (e) random. These states refer to diachronic patterns of appearance of the item through the corpus. The state *established* concerns items that, although not frequent (above a given threshold) in the beginning of the period, rise in

³ In other words, our aim is to provide sets of (candidate) established/obsolete items rather than some sort of “degree of establishment/obsolescence” per item.

frequency at some point and remain frequent until the end of the period covered the corpus. In other words, established items were not part of the language represented by the corpus, but at some point during its time span they flourished and remained frequent afterwards. The state *obsolete*, conversely, refers to items that are frequent (above a given threshold) in the beginning of the period covered by the corpus, but which at some point decrease in frequency. They are, therefore, items no longer in general use by the end of the corpus, although they may linger on as old-fashioned forms or archaisms making occasional appearances. The state *permanent* describes items that are frequent enough through the whole period covered by the corpus. The state *short-lived* regards items that flared up for some time and then, still during the period covered by the corpus, decreased in frequency again. Finally, the state *random* is reserved for items that do not show any of the aforementioned patterns. In the next section, we further develop this categorization by presenting our proposed methodology for classifying items into the above-mentioned classes.

3.3 The algorithm

3.3.1 Requirements

In order to be accessed by our proposed algorithm, a corpus must be divided into time frames. These time frames might delineate any desired period of time, depending on the nature of the data and on the research goals. Each one of these time frames may represent, for instance, a period of several years, or one decade, or one year, or even one day – the latter in the case of research using data from online social media platforms, for example. For methodological reasons, it is to be preferred that time frames are uniform (both in corpus size and duration, whenever possible) across the whole

corpus, but this is not a strict requirement and alternative methods (such as the one proposed by Gries and Hilpert (2008)) could be used to divide the corpus in time stages. Also, our method relies on the use of topically coherent corpora, so as to avoid that changes in sampling across time lead to change in the frequency of linguistic items.

In our method, when the frequency of a given item in a certain time frame is above a definite threshold, it is represented by the digit 1; when this frequency is below this threshold, by 0. For example, in a corpus divided into six time frames, the *diachronic sequence* of an item whose frequency exceeds the threshold only in the last time frame is denoted by 000001, while the sequence of an item whose frequency exceeds the threshold in all but the second and third time frames is denoted by 100111.

We leave the definition of the boundary between assigning a 0 or a 1 in the diachronic sequence as a choice for the researcher who will use our algorithm, since this depends on additional methodological choices and assumptions. We strongly discourage, however, the use of absolute frequencies as thresholds (as they are dependent on the size of the corpus in each time frame) and, conversely, encourage the use of relative frequencies. For example, a 1 might be attributed to a given item in a particular time frame in case its frequency exceeds $n\%$ of the total size of the corpus in that time frame; otherwise, a 0 will be attributed. A simple and useful case is when this boundary is set on a really low relative frequency (e.g. 0,00000001% of the corpus size). In this case, the mere presence of the item in the time frame is enough to assign a 1 to it. This simple situation is convenient, practical and might still give interesting results, such as the ones we display on Section 3.4.

In Section 3.3.2, we introduce the rules regulating a first algorithm aimed at the categorization of linguistic items into one of the previously mentioned states – *established*, *obsolete*, *permanent*,

short-lived or *random*. We begin by stating naive rules that are ultimately not satisfactory for our intentions. In Section 3.3.3, however, an improved version of these rules, more effective for the purposes of the goals declared here, is presented.

3.3.2 Rules for a naive algorithm

A first (and naive) version of an algorithm aiming to solve the task of categorizing a linguistic item into one of the aforementioned states may be based on the following rules:

- Established items: those that are not frequent enough in the corpus before a certain time frame, but from a given point start to exceed the frequency threshold in all of the following time frames, without exception. Example of a diachronic sequence in a corpus containing six time frames: 000111.
- Obsolete items: those that are frequent in the first time frame(s), but from a given point onwards are not frequent enough in any of the following time frames, without exception. Example: 111000.
- Permanent items: those that are frequent in all time frames, without exception. Only possible diachronic sequence: 111111.
- Short-lived items: those that are not frequent enough in the extremes of the period covered by the corpus, but that are consistently frequent during an intermediate period. Example: 00111100.
- Random items: those that do not fit into any of the previous cases. Example: 100101.

It is clear that these rules only work for what we might call “perfect” patterns, in which linguistic items “appear” or “disappear” at a certain point and keep this status until the end of the period covered by the corpus, without fluctuations. According to this method,

an item which, in a corpus divided into ten time frames, exhibits the pattern 0001011111 is considered an example of a random pattern, even though it is obvious for us that it clearly illustrates an item established sometime around the middle of the period covered by the corpus. To solve this issue, an improved version of these rules, allowing for some deviations from perfect patterns, is presented in the next section. Without the allowance of these deviations, the low frequency of an item in a specific time frame would be too severely punished, being enough to disregard the item as an innovation; conversely, the presence of an item in a specific time frame could be enough to disregard it as an obsolete item.

3.3.3 Proposed algorithm

Here, we propose an algorithm that enhances the previous approach by allowing for small deviations from perfect patterns, thus making it possible to include more (and more accurate) data into the lists of established and obsolete items of a corpus. The core idea is (i) to compare the observed (real) diachronic sequences of each item in the corpus with perfect patterns for establishment and obsolescence, and then (ii) to select a specific time frame as representing the time of establishment or obsolescence, using the criterion that it should be the time frame that produces the smallest amount of deviation from these perfect patterns.

Consider the following fictitious example. In a corpus divided into ten time frames, the linguistic item *A* exhibits the diachronic sequence 0001110111 – according to which *A* is not frequent enough in the initial periods of the corpus, but after time frame four it is consistently frequent, with the only exception being time frame seven. Our algorithm inspects each position in between two adjacent time frames, starting from position one (which lies in between the first and the second time frames, as in 0_001110111). The perfect

pattern indicating the establishment of an item in this position is 0_111111111 (i.e., the item is not present before the position and is consistently present after it), while the perfect pattern indicating its obsolescence at this point is 1_000000000. Here, the algorithm investigates the observed sequence for item *A* and counts deviations from the two perfect patterns. By *deviations* we mean differences in particular points of the diachronic sequences: for instance, if, in a given place, a 0 is found in the observed sequence when a 1 is expected according to the perfect pattern, then we detect a deviation⁴.

Let us return to the example of the sequence 0001110111. At the first position, when the assumption is that the item gets established after that point in time, the algorithm finds three deviations from the perfect pattern (the three 0s in time frames two, three and seven); when the assumption is that the item becomes obsolete, there will be seven deviations from the perfect pattern (the 0 in time frame one and the six 1s in time frames four, five, six, eight, nine and ten), as illustrated below, where arrows indicate deviations:

<table style="border-collapse: collapse; width: 100%;"> <tr> <td style="padding-right: 10px;">Observed sequence</td> <td style="padding-right: 10px;">0_001110111</td> <td style="border-left: 1px solid black; padding-left: 10px; padding-right: 10px;"></td> <td style="padding-right: 10px;">Observed sequence</td> <td>0_001110111</td> </tr> <tr> <td></td> <td style="text-align: center;"> ↓↓ ↓ ↓↓ ↓ </td> <td style="border-left: 1px solid black;"></td> <td></td> <td style="text-align: center;"> ↓ ↓↓ ↓↓ ↓ ↓↓ ↓↓ </td> </tr> <tr> <td style="padding-right: 10px;">Perfect pattern (establishment)</td> <td style="padding-right: 10px;">0_111111111</td> <td style="border-left: 1px solid black; padding-left: 10px; padding-right: 10px;"></td> <td style="padding-right: 10px;">Perfect pattern (obsolescence)</td> <td>1_000000000</td> </tr> </table>	Observed sequence	0_001110111		Observed sequence	0_001110111		↓↓ ↓ ↓↓ ↓			↓ ↓↓ ↓↓ ↓ ↓↓ ↓↓	Perfect pattern (establishment)	0_111111111		Perfect pattern (obsolescence)	1_000000000	
Observed sequence	0_001110111		Observed sequence	0_001110111												
	↓↓ ↓ ↓↓ ↓			↓ ↓↓ ↓↓ ↓ ↓↓ ↓↓												
Perfect pattern (establishment)	0_111111111		Perfect pattern (obsolescence)	1_000000000												

After these results have been obtained for the first segmentation, the algorithm moves to the next position (00_01110111). Here, two deviations from the perfect pattern of establishment (the two 0s in time frames three and seven) and eight deviations from the perfect pattern of obsolescence (the two 0s in time frames one and two, and the six 1s in time frames four, five, six, eight, nine and ten) are

⁴ These deviations might be counted, for example, by employing an edit distance algorithm, such as the Levenshtein distance algorithm, that returns the minimum number of single-character edits required to change one sequence into the other.

found. In the third position (000_1110111), only one deviation from the perfect pattern of establishment is found (the 0 in time frame seven), while nine deviations from the perfect pattern of obsolescence are detected (the three 0s in time frames one, two and three, and the six 1s in time frames four, five, six, eight, nine and ten). In the next step, two deviations from the perfect pattern of establishment (the 1 in time frame four and the 0 in time frame seven) and eight deviations from the perfect pattern of obsolescence (the three 0s in time frames one, two and three, and the five 1s in time frames five, six, eight, nine and ten) are identified in the fourth position (000_1110111). The process continues until all positions⁵ are analyzed, after which the position producing the smallest number of deviations can be found. This position will represent a possible moment of establishment or obsolescence. In the case of item *A*, Table 3.1 shows that the smallest number of deviations is found under the assumption of establishment (rather than obsolescence) and is observed in position three, indicating that this linguistic item might have been established in the corpus immediately after this point – that is, within time frame four.

Let us go on to consider, in the same corpus, a linguistic item *B* exhibiting the diachronic sequence 1111100100. After the inspection of the nine positions in between each two adjacent time frames, the proposed algorithm outputs that the smallest number of deviations from a perfect pattern is found in position five, but now the assumption is that of obsolescence. In this case, the decision implies that the item has become obsolete in the time frame following that position, which corresponds to time frame six, as displayed again in Table 3.1.

⁵ Note that the number of positions to be analyzed equals $tf - 1$, where tf represents the number of time frames in which the inspected corpus is partitioned.

Table 3.1: Number of deviations in each position according to the proposed algorithm for two fictitious examples A and B . In the first example, the smallest number of deviations is observed in position three under the assumption of establishment, indicating that A might have gotten established immediately after that position (in time frame four). In the second example, the smallest number of deviations is observed in position five under the assumption of obsolescence, suggesting that B may have become obsolete immediately after that position (in time frame six).

Item A			Item B		
Position in the diachronic sequence	Deviations (establishment)	Deviations (obsolescence)	Position in the diachronic sequence	Deviations (establishment)	Deviations (obsolescence)
1 (0_001110111)	3	7	1 (1_111100100)	5	5
2 (00_01110111)	2	8	2 (11_11100100)	6	4
3 (000_1110111)	1	9	3 (111_1100100)	7	3
4 (0001_110111)	2	8	4 (1111_100100)	8	2
5 (00011_10111)	3	7	5 (11111_00100)	9	1
6 (000111_0111)	4	6	6 (111110_0100)	8	2
7 (0001110_111)	3	7	7 (1111100_100)	7	3
8 (00011101_11)	4	6	8 (11111001_00)	8	2
9 (000111011_1)	5	5	9 (111110010_0)	7	3

It is worth noting that, for each position, the number of deviations from the perfect pattern indicating establishment plus the number of deviations from the perfect pattern indicating obsolescence equals the amount of time frames in the corpus. This is obviously expected, since each 0 or 1 in the observed sequence is always a deviation from a perfect pattern (either regarding establishment or obsolescence), but never a deviation from both.

The proposed algorithm will always output a smallest number of deviations from the perfect patterns, but this value might be considered excessive in some cases. For this reason, a cut-off point of the number of acceptable deviations from establishment and obsolescence should also be defined, and cases that exceed this threshold should be assigned to the pool of cases of random distributions. This cut-off point must be set by the researcher according to some sensible considerations that will vary according to the type of corpus in question: if it is a lexical corpus of child language acquisition with day-to-day recordings, for example, there might be many deviations since a single child is not expected to exercise its full vocabulary every day; if it is a large historical corpus of texts with yearly time frames, the cut-off point could be set to fewer deviations⁶. Here we must refrain from generalization about such thresholds, but we give an example of how to derive one from the behavior of a specific corpus in Section 3.4. What we mainly want to stress is that the use of an algorithm such as the one described here has the advantage that there has to be such an explicit threshold. Even if it is defined in somewhat *ad hoc* ways in individual cases, it will force researchers to be specific about their choice, enhancing transparency and replicability of a given study.

Finally, in case of ties such that the smallest number of devia-

⁶ It is trivial to observe that the naive rules presented in Section 3.3.2 correspond to the algorithm proposed here when this cut-off point equals to zero.

tions occurs at more than one position, we advocate for choosing the position that includes more time frames with the item being present so as to maximize the amount of positive attestations after minimizing the amount of deviations. An example would be as follows: in a diachronic sequence such as 1111101000, where the smallest number of deviations from a perfect obsolescence pattern (which is one) is achieved both in positions five and seven, we favor choosing the latter (corresponding to the eighth time frame) as the moment of obsolescence; conversely, in a sequence such as 0001011111, we favor choosing the fourth time frame (rather than the sixth) as the moment of establishment.

In conclusion to the present section, we present a summary of the steps made by the algorithm.

Summary of the algorithm:

1. Go to the first position in between two adjacent time frames.
2. Calculate the number of deviations from the perfect patterns of both establishment and obsolescence.
3. If there are unexplored positions in between two adjacent time frames, go to the next position and repeat step 2; otherwise, go to the next step.
4. Compare the value found for the smallest number of deviations S with the maximum threshold for deviations allowed T ;
 - I. If $S > T$, the item is considered neither established nor obsolete.
 - II. If $S \leq T$:
 - i. resolve potential ties by choosing the position that includes most time frames with the item being present;
 - ii. consider the time frame immediately after the corresponding position as the time frame of establishment or obsolescence.

The previously described algorithm is able to identify items classified as *established* or *obsolete* according to our defined criteria, but not items evaluated as *short-lived* – which are classified as *random* by it. In Section 3.4.5, we provide a case study in which we suggest a way of adapting our method for this specific situation.

In the next section, we apply our algorithm to a real corpus, supplying five case studies to illustrate its usage and some of its potential for producing interesting observations.

3.4 Case studies

In order to demonstrate the applicability of the algorithm proposed in Section 3.3.3, we applied it to the Corpus of Historical American English (COHA). This corpus contains more than 100,000 texts from different sources (fiction and non-fiction books, magazines, and newspapers) published in the United States of America from 1810 to 2009 (Davies, 2012), and can be explored online and downloaded from its webpage⁷. In this work, we use the case-insensitive list of unique words⁸ (types), annotated with part of speech (PoS) tags. This list contains the frequency of each pair (word + PoS tag) in each of the twenty decades spanned by the data. In this way, it is often possible to differentiate between homonyms (e.g. *light*, that can be tagged as adjective, noun, verb and others). We also removed all words classified with the tags for “formula”, “proper noun” (neutral for number, singular and plural), “letter of the alphabet” (singular and plural), “foreign word” (such as *arbre*, *bueno* and *deum*) and “unclassified word” (which includes ideophones like *bang-bang*, unrecognizable words such as *carige*, exclamations like

⁷ <https://corpus.byu.edu/coha/>

⁸ Here, we define a *word* simply as a string of characters uninterrupted by a space. It deserves mentioning that the downloadable COHA frequency data excludes words that occur less than three times in total in the corpus.

gotcha and recognizable words whose context is apparently unexpected). In total, we analyze 381,698 pairs of word + PoS tag in this corpus. As mentioned in Section 3.3.1, in these case studies we set the boundary between a 0 and a 1 in a really low relative frequency (0,00000001% of the corpus size) – so, the mere presence of a word in a time frame is enough to assign a 1 to it. By using this straightforward criterion, our goal is to show that even a method based on the simple presence/absence of items in specific time frame is able to rapidly bring interesting and useful results.

Having selected the corpus to work with, we need to decide on the value of T , i.e., the maximum threshold for how many deviations from the perfect patterns we can accept so to advocate for the establishment or the obsolescence of the analyzed items. Although, as mentioned in Section 3.3.3, the decision must to some degree be *ad hoc*, it should at least be backed up by an explicit criterion. Our approach here is to look at the statistics of establishment of words using the perfect pattern (no deviations) as a baseline: if the number of words that get established in different decades allowing for d amount of deviations is consistently proportional to the number of words that get established under the zero-deviation criterion, then the given value of d is acceptable. But how should “consistently proportional” be defined? Here, we look at the time series for the proportion of words that became established in each decade out of all words in the decade using different values of $d > 0$, and correlate these numbers with the corresponding numbers for the zero-deviation curve. If the p -value of a Pearson correlation is below 0.05 for a given value of d , then that amount of deviation is taken to be acceptable. In our case, it is only for $d = 1$ that we find an acceptable correlation: $p = 0.0047$, $\rho = 0.605$; for $d = 2$ we already get $p = 0.0569$ and the correlation goes down to $\rho = 0.4323$. Results continue to get worse as more deviations are allowed for. Thus, it is clear that too much noise would be admitted into any

statistics on the establishment of new words (and presumably on their obsolescence as well) if more than one deviation is considered acceptable in this case. For one deviation, the observations will also contain some noise, but more (and still reliable) data will be included⁹.

As an illustration of how a few words are evaluated by our algorithm in COHA, Table 3.2 displays the outcomes of attempts to detect established/obsolete words using respectively the naive (zero-deviation criterion) approach and our proposed method implementing the one-deviation criterion. The words selected for illustration are all singular common nouns present in COHA. Words are marked with (a) when they represent cases in which their first/last attestation matches the outcomes of both algorithms; with (b) when the naive rules cannot determine their date of establishment/obsolescence and our algorithm finds that the first or last occurrences are, respectively, also the decades of establishment or obsolescence; with (c) when the naive rules again cannot tell their date of establishment/obsolescence and our algorithm now finds that the first or last occurrences are, respectively, *not* the decades of establishment and obsolescence; with (d) when the decade of establishment/obsolescence is considered random by both methods. The (b) and (c) cases are particularly relevant since they illustrate data that would be lost from the purview of a study of lexical establishment or obsolescence if no deviations were admitted.

⁹ We stress that the decision on the value of this maximum threshold of deviations allowed must to some degree be *ad hoc*: since there are no “right” and “wrong” sets of established/obsolete items, this threshold depends on whether the researcher desires to obtain more comprehensive lists or more restricted ones – for the former, a higher threshold could be stipulated; for the latter, a lower value should be set. The point about using correlations and the p -value is that the distribution with one deviation from the perfect pattern is significantly similar to a distribution without any deviation, so the deviation can arguably be ignored.

Table 3.2: Outcomes of attempts to detect established/obsolete words using a first/last attestation approach, an algorithm following naive rules and the proposed algorithm (with a one-deviation criterion) in a selection of words present in the Corpus of Historical American English (COHA). Each time frame represents a decade, ranging from 1810s to 2000s. The examples chosen are all words tagged as “singular common noun”.

Word	Observed diachronic sequence	First attestation	Outcome according to	
			naive rules	proposed algorithm
(a) <i>victrola</i>	00000000001111111111	1910s	established (1910s)	established (1910s)
(b) <i>snatcher</i>	00000000110111111111	1890s	random	established (1890s)
(c) <i>bulldozer</i>	00000010000011111111	1880s	random	established (1940s)
(d) <i>wife-murderer</i>	00001011101100010010	1850s	random	random
Last attestation				
(a) <i>secrecy</i>	11111111111100000000	1920s	obsolete (1930s)	obsolete (1930s)
(b) <i>gratulation</i>	1111111111111010000	1960s	random	obsolete (1970s)
(c) <i>destruction</i>	11100001000000000000	1880s	random	obsolete (1840s)
(d) <i>unfeelingness</i>	00110000110110100100	1980s	random	random

3.4.1 Case 1: Statistics on established and obsolete words

Figure 3.1 shows the percentage of words that became established (left figure) and obsolete (right figure) per decade in COHA according to our algorithm and using the one-deviation criterion. In the left figure, the U-shaped nature of the curve concerning the establishment of words considering the whole corpus is easily explained by two factors that must always be acknowledged by the researcher: first, the proportion of words that had not appeared previously in the corpus is necessarily higher in the first time frames than in the next ones, as a consequence of the phenomenon known as Herdan’s or Heaps’ law (Herdan, 1964; Heaps, 1978), according to which vocabulary size grows slowly compared to the size of the document/corpus; second, the proportion of words arisen in a certain decade that are consistently present in the following ones (i.e., the words considered established conforming to our criteria) is necessarily higher in the last time frames than in the previous ones, because most of these recently established words did not have time to become obsolete yet. To demonstrate these two effects more precisely, we include two additional curves in the graph, corresponding to the percentage of words that became established in a given decade considering only certain time windows (six- and eight-decade windows). In other words, we reduce the corpus to sliding windows of six and eight decades in order to decrease the “advantage” that early and late decades hold compared to middle decades. For these two additional curves, however, we are employing a zero-deviation criterion, since one deviation in a universe of only a few decades might be considered disproportionate. These additional curves do not display such a clear U-shaped nature, even though the one regarding six-decade windows still slightly reflects this pattern especially in its left tail. Also, both exhibit the same shape, suggesting that the

proportion of established words among all words in a given decade is similar across time and that the use of different windows in this case might be no more than a question of how much data one wishes to consider: around 3% for six-decade windows vs. around 2% for eight-decade windows.

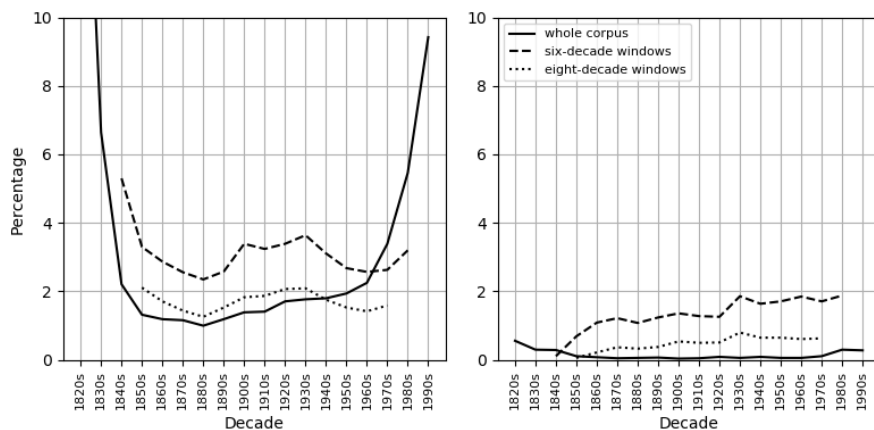


Figure 3.1: Percentage of words that became established (left) and obsolete (right) per decade using a one-deviation criterion and applying six- and eight-decade windows combined with a zero-deviation criterion. Curves comprise different time spans according to the sizes of the sliding windows.

Regarding the right figure, we observe that the proportion of words that became obsolete among all words in a particular decade is also more or less constant, with lower percentages than the ones referring to established words. Additional research must be carried out to more precisely understand the meaning of these results and their implications for language dynamics.

3.4.2 Case 2: Characteristics of established and obsolete words

Investigating certain characteristics of words considered established or obsolete according to our proposed algorithm is also a possible line of study. Figure 3.2, for example, shows the average length (in number of characters) of words that became established and obsolete in given decades. Here, we observe an irregular shape of the curve concerning words that became obsolete, but a consistently positive slope in the curve regarding established words, indicating a persistent increase in the average length of words established in the corpus across time – that goes from around eight and a half characters in the mid-19th century to almost ten characters in the second half of the 20th century. Since COHA is balanced by genre across time (Davies, 2012), this finding should in principle not be attributed to artifacts of the corpus (such as a potential increase in the proportion of scientific literature, for example). Additional investigation should be conducted to better understand this phenomenon, presumably employing other data and associating this results with the abundant previous work on word length (Grzybek, 2007).

Different analyses can be carried out also considering the PoS tags of the words in the corpus. Figure 3.3 depicts the percentage of parts of speech (grouped as “adjective”, “adverb”, “noun”, “verb” and “other”) among words that became established (left figure) and obsolete (right figure) in each decade. Among the established words, we visually notice a descending trend in the proportion of verbs and an ascending trend in the proportion of adjectives across the decades. The other curves are not consistently rising or falling – although, if we consider only the time period starting in the 1960s, we do observe a tendency for the proportion of nouns among the established words to increase. Regarding the words that became

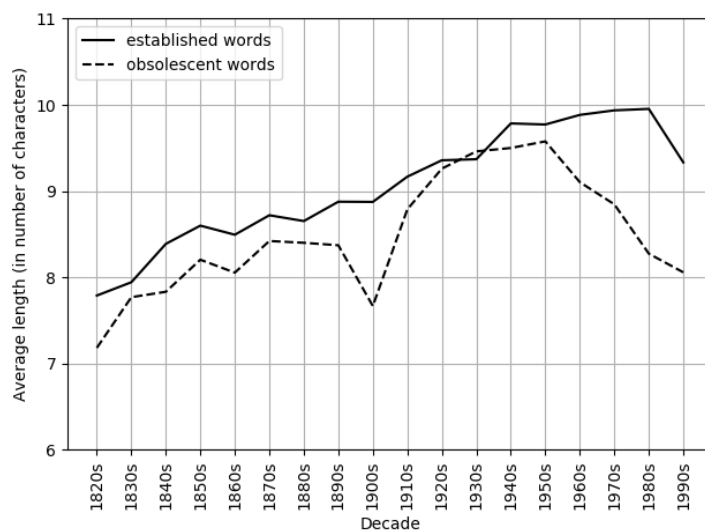


Figure 3.2: Average length (in number of characters) of words that became established and obsolete in given decades using a one-deviation criterion.

obsolete, the curve that represents nouns seems to exhibit a downward trend, while the others show constant fluctuation through time. Again, the fact that COHA is balanced by genre across time suggests that these patterns, in principle, should not be due to artifacts of the corpus, even though additional investigation is needed to better comprehend the phenomena reported here.

3.4.3 Case 3: Lexical heritage from past decades

The lexicon of every language at time t embodies strata from different periods in time during which new words that are still used at t became established. We are now onto a bit of “stratigraphy”, employing our algorithm to generate lists of those words established in different decades that are today the most popular in the corpus. More precisely, we select, from the words established in each

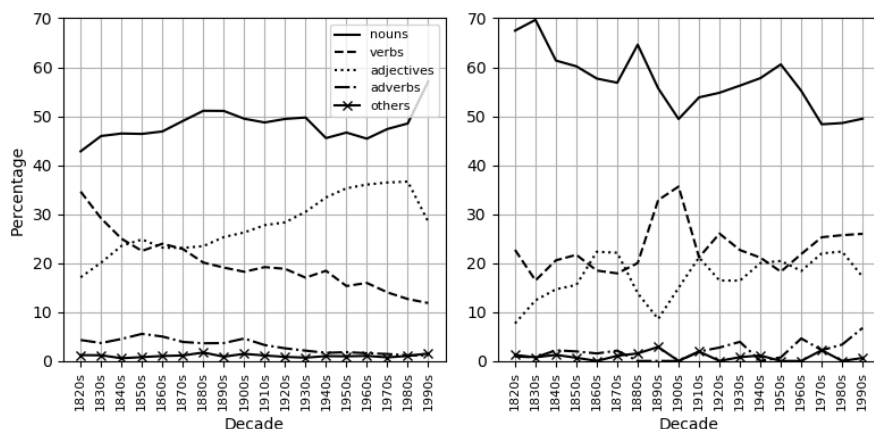


Figure 3.3: Percentage of different parts of speech among words that became established (left) and obsolete (right) in different decades.

decade between the 1850s and the 1980s, the fifty words that are most frequent in the 2000s. This ensures that we capture a portrait of today’s lexical heritage from past decades which is both reasonably detailed and still salient to speakers of American English.

The result of the selection procedure is displayed in Appendix A. After grouping these words into semantic categories (e.g. by using tools like Empath (Fast et al., 2016) or LIWC (Tausczik and Pennebaker, 2010)) or building networks (e.g. by a co-occurrence metric), it would be possible to make some generalizations concerning which semantic domains have been major contributors to these different historical strata or to determine the overall relationship among words established in a given decade¹⁰.

Impressionistically, for instance, it seems that the 1870s gave us

¹⁰ These generalizations, however, should be made carefully. Even if the corpus is balanced by genre across time (which is the case of COHA), the topics covered by the texts themselves might vary systematically (and not nicely randomly) over time. A possible way of mitigating this (potential) issue could be to implement a topic detection method, such as latent Dirichlet allocation (LDA) (Blei et al., 2003), in order to ensure that topics are coherent over time.

much vocabulary relating to the built environment, such as *hallway*, *downtown*, *driveway*, *taxi*, *headlights* and *neon*. The 1880s were big on sports, cf. *golf*, *hockey*, *olympics*, *coaching*, *scoring*. The 1890s were innovative in the communication domain, see *movie*, *television*, *wireless*, *phones*. The 1910s opened the written language to include words that would have been considered too obscene to print earlier: *fuck*, *goddam*, *dick*. The 1920s introduced several relatively abstract concepts relating to workflow: *coordinator*, *feedback*, *processing*, *implementation*, *operational*. The list of the 2000s most frequent words stemming from the 1940s does not reveal that a big war happened; instead, for instance, we see elements of such an everyday affair as food consumption: *supermarket*, *microwave*, *fridge*, *burgers*, *yogurt*. The 1950s show nascent environmental concerns: *pesticides*, *recycling*, *environmentally*, *pollutants*. Apart from giving us the concept of *lifestyle(s)* in general, the 1970s also showed news in different domains of lifestyle, such as the food domain, cf. *tofu*, *fast-food*, *sushi*, *veggies*. During the 1980s, the (personal) computer is the single most dominant factor in lexical innovation: *laptop(s)*, *database(s)*, *pcs*, *algorithms*, *download*, *firewall*. As we move closer to the present, it is predictable that some of the lexical legacy encountered is going to be short-lived, and some of the terms from the 1970s and 1980s, indeed, already feel somewhat outdated.

A comprehensive study of the lexical legacy of different periods in current American English could probably use a larger selection and, as mentioned, systematic methods of defining semantic domains and networks. In any case, we show that our algorithm is effective for identifying the lexical material needed for such research.

3.4.4 Case 4: Lost words

Another use of our proposed method is to generate lists of previously popular items that became obsolete in the corpus during

its time span. This is interesting because, unlike innovations, obsolete items are not commonly covered in existing literature (Tichý, 2018). Here, we select, from the words that became obsolete in each decade between the 1850s and the 1980s, the ten with the highest frequency before their obsolescence. In this fashion, we display a vocabulary that was particularly relevant in the past, but that has lost terrain in American English after some decades.

The lists of the once common words that became obsolete in COHA in a particular decade are shown in Tables 3.3 (1850s-1920s) and 3.4 (1930s-1980s). In the second half of the 19th century, it is possible to encounter typos and spelling mistakes that ceased to appear in the 20th century (maybe partially due to the development of more accurate typing and printing techniques), such as *had'nt* (1870s), *do'nt* (1880s), *hav'nt* (1890s), *was'nt* (1890s) and *did'nt* (1890s). There are also several other words that are still easily recognizable but have obsolete or semi-obsolete spellings, including *errour* (1850s), *pennyless* (1870s), *musquitoes* (1880s), *negociation* (1890s), *villany* (1930s), *reconnoissance* (1940s), *trowsers* (1950s) and *persistency* (1960s), to name a few. In some cases, the obsolete spelling is more faithful to the etymology of the word, as in *holydays* (1880s), which became “holidays”, and *cocoa-nut* (1930s), which became “coconut”. Further, the lists exhibit spellings that are still present in the corpus, but no longer with a specific syntactic function, such as *under* as a comparative adjective (1900s), *itself* as a singular common noun (1900s) and *notwithstanding* as a subordinating conjunction (1970s).

Of particular interest is the illustration provided by these lists of the phenomenon of the historical spelling change of English compounds. According to Shertzer (1996), “[t]he usual sequence is for the words to be written separate at first, then to become hyphenated, and finally to be written solid” (p. 109). We observe that several compounds that are nowadays usually written

Table 3.3: Lists of common words (+ PoS tags) in previous decades that became obsolete in the corpus in a particular decade (1850s-1920s). Words are ordered according to their frequency before their obsolescence. When the word ranked in the eleventh position has the same frequency as the one in the tenth position, we include both. The meaning of each PoS tag is explained in Appendix A.

1850s	1860s	1870s	1880s
errour_nn1	copy-right_nn1	had'nt_vv0	phrensy_nn1
scymetar_nn1	hazle_nn1	ancke_nn1	sassacus_nn1
almanzor_nn1	do'st_vv0	pennyless_jj	ancks_nn2
pedrillo_nn1	pannels_nn2	wo-begone_jj	cotemporary_jj
musquetry_nn1	phrensied_jj	inartificial_jj	mosquitoes_nn2
inquietudes_nn2	choaked_vvd	wrapp_nn1	afford_nn1
renegado_nn1	fire-side_jj	teaze_vvi	holydais_nn2
errours_nn2	barb'rous_jj	rivalships_nn2	gallopped_vvd
zegri_nn1	famish_jj	a'nt_vv0	apalachian_jj
broad-street_nn1	fann_nn1	returnless_jj	do'nt_vv0
potawatamies_nn2	incommunicative_jj		vanquish_jj
1890s	1900s	1910s	1920s
merchandize_vv0	shakspeare_vv0	shakspeare_nn1	the_nnt1
rivalship_nn1	immoveable_jj	eend_nn1	desponding_jj
hav'nt_vv0	under_jjr	did'st_vv0	flag-staff_nn1
had'st_vv0	say'st_vv0	creatur_nn1	sportively_rr
guarantied_vvn	xve_nn1	deth_vvz	befel_vv0
intenseness_nn1	itself_nn1	piano-forte_nn1	stopt_vv0
was'nt_vv0	wall-street_nn1	saidst_vv0	discomposed_vvn
negociation_nn1	pedee_nn1	thou'st_nn1	enginery_nn1
cretur_nn1	xve_vv0	applauses_nn2	school-fellows_nn2
did'nt_nn1	sdeath_nn1	knitting-work_nn1	sarvice_nn1
		see'st_vv0	

in a solid form are present in the corpus as hyphenated compounds and that these became obsolete at some point – probably around the time when their corresponding solid form were gaining popularity. This is the case of *copy-right* (1860s), *fire-side* (1860s), *wo-begone* (1870s) (now most commonly written “woe-begone”), *piano-forte* (1910s) (now mostly encountered as just “piano”), *flag-staff* (1920s), *school-fellows* (1920s), *dew-drops* (1930s),

Table 3.4: Lists of common words (+ PoS tags) in previous decades that became obsolete in the corpus in a particular decade (1930s-1980s). Words are ordered according to their frequency before their obsolescence. When the word ranked in the eleventh position has the same frequency as the one in the tenth position, we include both. The meaning of each PoS tag is explained in Appendix A.

1930s	1940s	1950s
villany_nn1	csar_nn1	trowsers_nn2
prison-house_nn1	new-comer_nn1	school-master_nn1
nuther_vv0	custom-house_nn1	mantel-piece_nn1
wofully_rr	custom-house_jj	despatch_vvi
unbiased_jj	bethink_vv0	hill-top_nn1
dew-drops_nn2	reconnaissance_nn1	aliment_nn1
cocoa-nut_jj	hill-tops_nn2	corner-stone_nn1
log-house_nn1	prayer-meetings_nn2	leipsic_nn1
can't_vv0	school-boys_nn2	exhaustless_jj
palm-tree_nn1	sketch-book_nn1	self-complacency_nn1
1960s	1970s	1980s
acquirements_nn2	sich_vv0	intrusted_vvn
inclosure_nn1	ball-room_nn1	arm-chair_nn1
persistence_nn1	now-a-days_rt	fellow-men_nn2
state-room_nn1	frying-pan_nn1	quitted_vvd
upon_nn1	notwithstanding_cs	with_nn1
intrenchments_nn2	hesitating_jj	common-place_jj
snuff-box_nn1	reprobation_nn1	fitly_rr
strifes_nn2	banditti_nn2	unwearied_jj
guard-house_nn1	by-gone_jj	small-pox_nn
heart-strings_nn2	plighted_jj	inclosed_vvn

new-comer (1940s), *corner-stone* (1950s), *state-room* (1960s), *ball-room* (1970s), *now-a-days* (1970s), *arm-chair* (1980s), *common-place* (1980s) and various others that can be recognized in the table. Nonetheless, a few compounds, such as *wall-street* (1900s) and *knitting-work* (1910s), seem to have taken the opposite direction, now being more commonly written as separate words. A comprehensive study aiming to analyze this phenomenon in a quantitative fashion could benefit from our proposed method to obtain these lists

of obsolete items per time frame and investigate how different factors (e.g. time, accumulated frequency, sudden frequency rise/fall) act and impact this process of orthographic variation and change.

3.4.5 Case 5: Short-lived words

The method described in Section 3.3.3 is able to assist in the identification of items classified as *established* or *obsolete*, but not of items evaluated as *short-lived*. Here, we provide a short case study in which we suggest a way of adapting it for this specific purpose. Our goal is to find words that flared up in the corpus for some time and then, still during the period covered by the corpus, disappeared. According to our previously mentioned criteria, these words are considered neither established (since they are already gone) nor obsolete (since they are not part of the corpus in its initial period), but in some cases it might be interesting to analyze them in order to investigate the process of lexical variation and change in more detail.

A possible way of adapting our method to the case of short-lived items is by applying the proposed algorithm to selected intermediate subcorpora. One solution would be to look for items whose diachronic sequences hold only 0s in their extreme time frames, such as in 0001111000, then cut off the extremes of the corpus (say, the n_1 time frames in the beginning and the n_2 time frames in the end of the time span covered by the corpus) and, finally, apply the algorithm only to the remaining intermediate sequences, looking for established, obsolete and permanent items in these subcorpora.

For the present exploratory purposes we adapted our method to handle cases of words that did not appear in COHA before the 1860s and disappeared again no later than the 1950s – in other words, these items are present neither in the five first nor in the five last time frames of the corpus ($n_1 = n_2 = 5$). We then applied

our algorithm considering just this subsection of the corpus. We extracted words evaluated as permanent – which are, of course, perfect cases of short-lived words, presenting the diachronic sequence [00000]1111111111[00000]¹¹. We also gathered other not-so-short-lived words evaluated as established and obsolete in the subsection of the corpus, but only those that appeared in at least eight decades and with no deviations allowed¹².

The words that emerged from this analysis are listed alphabetically in Table 3.5. The vast majority of them are compounds (either hyphenated or solid), short-lived spelling variants and bona fide words that came and went. Among the hyphenated compounds, we find words such as *farm-lands*, *hair-pin* and *saddle-bag* – all of them more commonly written in a solid form nowadays. These data are useful for the study of the historical spelling change of English compounds mentioned in Section 3.4.4. Words such as *comp'ny*, *yisterday* and *s'posin* are examples of short-lived spelling variants. The comparative adjective *humaner* (meaning *more humane*) and the nouns *leisureliness* (*leisurely* + *-ness*) and *stereopticon* (an old type of slide projector) are interesting examples of short-lived items found here: when searching on another source, the Google Books Ngram Viewer¹³, we find that all of them exhibit a similar frequency pattern, peaking around the 1920s.

These results are just an illustration of the kind of content that can be obtained from such an analysis. It is important to notice that looking for short-lived items is not, in principle, one of the goals of the method introduced in this chapter, and that the adaptation presented in this case study is just a workaround. The main pitfall of this adaptation is that it depends on the selection of spe-

¹¹ The 0s in between square brackets correspond to the extremes of the corpus that were cut off.

¹² That is, those which, for the period of the subcorpus studied, presented the diachronic sequences 0111111111, 1111111110, 0011111111 and 1111111100.

¹³ <https://books.google.com/ngrams/> .

Table 3.5: Words (+ PoS tags) classified as short-lived according to the adaptation of our method and considering the period between the 1860s and the 1950s. Words are alphabetically ordered. The meaning of each PoS tag is explained in Appendix A.

a-beatin_nn1	crep_nn1	ha'r_nn1	race-track_nn1
a-laughin_nn1	dilapidated-looking_jj	hair-pin_nn1	rose-petals_nn2
a-puttin_nn1	dish-towels_nn2	hay-wagon_nn1	s'posin_nn1
a-quiver_vv0	dust-heap_nn1	hereinbefore_rr	sabe_vvi
a-sittin_nn1	ear-drums_nn2	herse'f_nn1	saddle-bag_nn1
all-rail_jj	earnin_nn1	hez_vv0	spoilin_nn1
alongshore_nn1	east-bound_jj	high-tariff_jj	staff-officer_nn1
baggage-man_nn1	farm-hands_nn2	humaner_jjr	station-master_nn1
bath-chair_nn1	farm-lands_nn2	ice-floe_nn1	stereopticon_nn1
bird-shot_nn1	field-glass_nn1	idealizing_jj	street-cars_nn2
black-fringed_jj	field-glasses_nn2	jumping-jack_nn1	talesmen_nn2
bodder_vvi	fitten_vvn	leisureliness_nn1	tek_vvi
bofe_nn1	food-supply_nn1	lucile_nn1	trades-union_nn1
bread-winner_nn1	foregathered_vvd	myse'f_nn1	unfoldment_nn1
broncho_nn1	forehanded_vvn	pack-train_nn1	up-train_nn1
burled_vvn	four-bit_jj	pay-rolls_nn2	w'at_nn1
catchee_nn1	full-armed_nn1	pepsin_nn1	w'en_jj
chromos_nn2	garden-party_nn1	play-actin_nn1	water-bottle_nn1
coat-sleeves_nn2	glarin_nn1	pony-cart_nn1	weazened_vvd
comp'ny_jj	groceryman_nn1	prohibitionist_jj	wedding-bells_nn2
consul-general_jj	grouped_jj	pulse-beats_nn2	yisterday_nn1

cific subsections of the corpus to be analyzed by the researcher. A possible goal for future work is to design and develop a specific and more effective method for finding short-lived items in diachronic corpora.

3.5 Concluding remarks

In the field of corpus linguistics, the analysis of diachronic corpora with the goal of explaining diverse phenomena in human languages is becoming increasingly widespread. In this context, we need methods and procedures aiming to discover trends and patterns in the dynamics of a language as we process big amounts of text com-

putationally. With the present contribution, we hope to specifically generate more interest in the birth and death of components such as words, expressions and grammatical constructions in corpora that span over time.

Here, we introduce the notions of *establishment* and *obsolescence* as complementary to the trivial concepts of first and last attestations of linguistic items in diachronic corpora. Subsequently, we propose an algorithm to identify the time period of establishment and obsolescence of linguistic items based on their frequency in a diachronic corpus. This algorithm may be employed for the analysis of any linguistic item, be it lexical, phonological or morphosyntactical. The method proposed here is, of course, only one of the numerous possibilities for the achievement of similar goals. Other methods, including more mathematically sophisticated ones, could be evaluated as well. Alternatives that look promising for further consideration are approaches that would model *probabilities* of establishment and obsolescence. Such approaches would have the double advantage of allowing more accurate estimates of when a probability of occurrence exceeds a given threshold, and of allowing to make such estimates with fewer arbitrary parameters (e.g. lengths of periods, occurrence thresholds within a period, which patterns to consider as indicating what kind of event etc.). In this work, our focus is to demonstrate a simpler and easier-to-implement method, but we plan to discuss more sophisticated approaches in future studies.

We demonstrate the applicability of our proposed algorithm using a real corpus spanning 200 years of data and supplying case studies concerning the character of words that got established and obsolete in American English in different periods. Among the outcomes of these case studies is the observation that the percentage of established words among all words across decades fluctuates without showing a specific upward or downward trend. We also found

that the proportion of adjectives among new words has increased steadily over the past two centuries, mostly mirrored by a decrease in the proportion of new verbs. Then, we provided a sketch study of the lexical heritage in American English, identifying words that became established in different decades and are still frequent in the 2000s. We also looked at obsolescent vocabulary – vocabulary that was previously frequent but has been getting lost over the decades. Finally, we briefly investigated whether the method could be adapted to find short-lived words – words that flared up in the corpus for some time and then disappeared. These sketch studies are mainly presented with the goal of motivating future studies employing the method presented here.

It may be obvious but still it is necessary to recall that a corpus is different from a language. As a consequence, when we consider the establishment or the obsolescence of a linguistic item in a *corpus*, we are not necessarily referring to the establishment or the obsolescence of this item in a *language*. This distinction is particularly relevant when we deal with corpora based on written texts (like COHA itself or the Google Books corpus) – since, for instance, an item might be used for a long time in the oral language before it gets established in the written register. When considering the whole language, it is clear that the algorithm can only identify the decade during or *before which* (*ante quem*) a word became established or the decade during or *after which* (*post quem*) a word became obsolete. This situation is of course due to the fact that “it is much simpler to prove that something exists (...) than to prove that something does not exist” (Tichý, 2018, p. 82). This fact becomes even clearer if we think about the application of our method to domain-specific corpora (consisting of academic, legal, medical etc. texts): the results will of course reflect the specificity of the analyzed data.

Regarding our case studies, it is important to remember that words are pairings of form and function. Words not always start

their lives with a meaning and get lost with that same meaning, since in real-life diachronic lexical change there are also forms that come into being with a particular connotation but at some point lose that connotation, while still living on with a completely different one; and such words occur alongside words that live on with their original meaning. This must be taken into account when the researcher employs our (or any other) method to automatically obtain lists of forms that get established or become obsolete in a corpus.

As stated by Hilpert and Mair (2015), it is imperative to demonstrate “how the use of corpus data allows researchers to go beyond the mere statement that a grammatical change happened, and to address the questions of **when** and **how** something happened” (Hilpert and Mair, 2015, p. 199, emphases in original). With our theoretical discussion, our proposed algorithm, and the case studies that were presented here, we hope to have taken a step in this direction.

CHAPTER 4

Diachronic corpora and quantitative approaches to the lexicon: the case of the term *fake news*¹

4.1 Introduction

The term *fake news*, defined by Collins Dictionary as “false, often sensational, information disseminated under the guise of news reporting”, gained so much attention that it was named the English word of the year 2017 by both Collins Dictionary (2017) and American Dialect Society (2018). Even though the concept of news arti-

¹ This chapter reproduces with minor changes the article “Quantifying the conceptualization of the term ‘fake news’ in Brazilian and English-speaking media sources” (Cunha et al., under review), submitted for publication. This article is, in its turn, an extended version of the paper “Fake news as we feel it: Perception and conceptualization of the term ‘fake news’ in the media” (Cunha et al., 2018), published as a chapter of *Social informatics* (eds. Steffen Staab, Olessia Koltsova and Dmitry I. Ignatov) and presented at the *10th International Conference on Social Informatics (SocInfo 2018)*, held in Saint Petersburg, Russia, in September 2018. See Appendix C for more information.

cles aimed to mislead readers is by no means new (Standage, 2017), there seems to exist a relationship between the very expression *fake news* with the 2016 presidential election in the United States of America: Davies (2017a), using data from the NOW Corpus, shows that “there is almost no mention of ‘fake news’ until the first week of November 2016 and then it explodes in Nov 11-20, and has stayed very high since then”. The author adds that the reason “why people all of the sudden started talking about something that had really not been mentioned much at all until that time” was “the US elections, which were held on November 9, 2016” (Davies, 2017a). Data from the Google Books Ngram Viewer², however, shows that the use of the term *fake news* had already peaked in earlier periods. Figure 4.1 reproduces the output of the query for *fake news* in this tool – which, at the time of writing of this text, only includes data until the year 2008. We observe that the relative frequency of this expression in the Google Books corpus experienced an increase in the 1910s/1920s, then peaked around 1940, and then started to rise again in the first decade of the 21st century. Since the data provided by this tool does not reach the 2010s, we cannot compare these frequencies with those from 2016 onwards. In any case, there is evidence that the relative frequencies of this expression are even higher in the years after the 2016 presidential election in the United States of America.

Despite its recent success, the widespread use of the term *fake news* has received much criticism. Members of the British Parliament recommended in a report that the Government rejects this expression, since it “is bandied around with no clear idea of what it means, or agreed definition”, and it “has taken on a variety of meanings, including a description of any statement that is not liked or agreed with by the reader” (Parliament of the United Kingdom,

² The Google Books Ngram Viewer is a searchable interface for Google Books, available at <https://books.google.com/ngrams> .

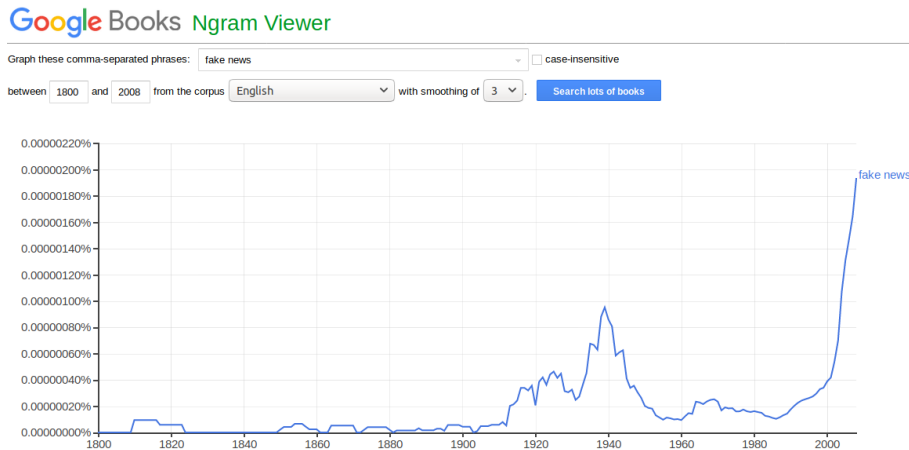


Figure 4.1: Output of the query for *fake news* in the Google Books Ngram Viewer. The chart shows relative frequencies from 1800 to 2008.

2018). It has already been suggested that the expression *fake news* is doing great harm (Habgood-Coote, 2018) and that it should be retired (Sullivan, 2017). In addition, the dissemination of the term *fake news* went beyond the English language. The Commission for the Enrichment of the French Language, for example, declared that “[t]he Anglo-Saxon expression ‘fake news’ (...) has rapidly prospered in French” (BBC News, 2018). The same phenomenon happened in other languages, including in Brazilian Portuguese, in which the borrowed term seems to have spread mainly during 2018 – once again in a context of national elections. Lees (2018) considers that *fake news* “has become an international political catchphrase”, since, at the time of the publication of her report, “more than 20 political leaders worldwide³, from authoritarian regimes to European

³ “Over the past year, political leaders in Burma, Cambodia, China, Egypt, France, Germany, Hong Kong, Hungary, Kuwait, Libya, Malaysia, the Philippines, Poland, Russia, Singapore, Somalia, Syria, Tanzania, Thailand, Turkey, the USA and Venezuela have publicly accused journalists of reporting, or being,

democracies, have used the term to accuse reporters of spreading lies as a way to discredit journalism they do not like” (p. 88).

The sudden popularization of an already existing term (that is, *not* a neologism) in a language poses interesting questions regarding how this term itself is perceived by the speakers of that language. We might ask, for instance: what has changed (if anything) in terms of conceptualization of this expression after its boom? Was there any kind of shift in its meaning when it became widely employed? If so, was this shift uniform across different varieties of the language? Similar questions might be asked regarding the sudden popularization of loanwords and expressions adopted from foreign languages. These are some of the issues of interest in *lexicology*, the area of linguistics focused on the study of the lexicon, that has been fostered thanks to advances in the use of big diachronic real-world corpora in the study of language.

The specific goal of this chapter is to provide a closer look at how newspapers and magazines across the world shaped the term *fake news* – which is a relevant social phenomenon linked to misinformation and manipulation, and that has been facilitated by the rise of the Internet and online social media – in the second decade of the 21st century. We investigate the perception and the conceptualization of this expression through the quantitative analysis of two corpora of news published in 21 countries from 2009 to 2018, thus making it possible to examine not only the diachronic development of this term, but also its synchronic usage in different parts of the world. We complement our investigation with data collected from online search queries that help us to measure how the public interest in the expression *fake news* and in the concepts around it changed over time in different places.

fake news” (Lees, 2018, p. 88). Brazil is not on the list only because far-right Jair Bolsonaro took power in January 2019, i.e., after the publication of Lees’ article.

Our general goal, however, goes beyond the investigation of an individual expression. By studying the diachronic change in the conceptualization of a term through replicable computational and quantitative methods, we initially propose a framework that can be applied to other cases. In this framework, we opt to establish analogies between concepts and means of expression that exceed the strictly linguistic analysis of the lexicon. To achieve this goal, we employ already established analytical methods which, when put together, are able to delineate the semantic framing of a linguistic item. It is interesting to note that, to the best of our knowledge, this is the first attempt to merge these specific methods into one framework that aims to investigate the diachronic change in the conceptualization of an expression. In addition, we also contribute to the research on diachronic corpus linguistics in Brazilian Portuguese – which, although far from being a low-resource language, is much less studied than English in this domain.

4.1.1 Research question

Our main research question here is: was the rise of the public interest in the term *fake news* accompanied by changes in its conceptualization and in the perception about it? Based on sociolexicological theories that defend the existence of a considerable relationship between linguistic and extralinguistic factors with regards to the vocabulary of a language (Matoré, 1953; Cambraia, 2013), our hypothesis is that the change of interest in the phenomenon *fake news* might have altered the general usage of the expression referring to it. Indeed, the results obtained in our investigations indicate, in general, a positive answer to our research question. Among other findings, we show modifications in the related vocabulary and in the mentioned entities accompanying the term *fake news*, in addition to changes in the topics associated with this concept and in the overall

contextual polarity of the pieces of text around this expression in English-written media articles after 2016 and in Brazilian articles after 2018.

This chapter is structured as follows: first, we mention previous works related to the concept of *fake news* and to the usage of large language datasets to investigate social phenomena; in Section 4.2, we present the process of acquisition and preparation of the data sources used in our investigations; in Section 4.3, we describe our analyses, present the results found and discuss their implications; finally, in Section 4.4, we summarize the outcomes of our study and conclude this chapter by discussing possible future outlooks.

4.1.2 Related work

In the years prior to the publication of this study, the amount of scholarly papers mentioning the term *fake news* has grown dramatically. To illustrate this trend, we show in Figure 4.2 the number of academic publications per year returned by the query for *fake news* on Scopus and Web of Science databases⁴. From respectively ten and eight articles in 2016, the numbers increase to 221 and 160 in 2017, and then to 551 and 367 in 2018. These values were obtained through queries performed on October 8, 2019, which is the reason why there is a small drop in the numbers relative to 2019. Still, the amount of publications shown in the graph in this year are likewise substantial. Of course, the fact that the numbers of articles returned by these queries have greatly increased after 2017 does not mean that the research on disinformation and on “yellow journalism” has started in this period, but only that the use of the term *fake news* has increased in this context.

⁴ Scopus (<https://www.scopus.com/>) and Web of Science (<https://www.webofknowledge.com>) are two citation databases that provide information on published scientific articles, journals and conference proceedings.

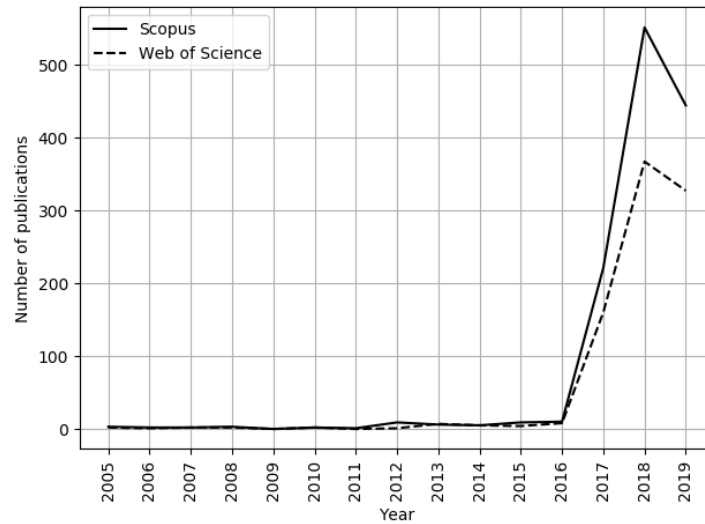


Figure 4.2: Numbers of academic publications per year returned by the query for *fake news* on Scopus and Web of Science databases. These values were obtained through queries performed on October 8, 2019.

On the phenomenon of fake news

The focus of this study is definitely not on the *phenomenon* of fake news, but rather on the use of this very expression in the language. Still, we cite some works on the phenomenon of fake news for the sake of contextualization, since the use of an expression is intimately related to the entity that it represents.

Van Hout and Burger (2015) allude to the boom of satirical fake news sources in the years before the publication of their study. In Sections 4.3.2 and 4.3.4, we show how this is confirmed by our data, since satirical TV shows and hosts often co-occur with the term *fake news* in the period before the 2016 presidential election in the United States of America. Allcott and Gentzkow (2017) study the dissemination of fake news in the particular case of this elec-

tion and allege that “[f]ollowing the 2016 election, a specific concern has been the effect of false stories – ‘fake news,’ as it has been dubbed – circulated on social media” (p. 212). They analyze, from an economic perspective, the consumption of fake news before and during this election, identifying an important role of social media in this context and confirming “that fake news was both widely shared and heavily tilted in favor of Donald Trump” (Allcott and Gentzkow, 2017, p. 212). As we show in the next sections, these results are also corroborated through our corpus-based methods. This fact gives robustness to our methodology, since it shows that it is able to replicate observations obtained from more fine-grained analyses. Vosoughi et al. (2018), using a big dataset collected from Twitter, empirically demonstrate that fake news diffuse “significantly farther, faster, deeper, and more broadly than the truth in all categories of information” (p. 1146). They also show that the effects of misinformation are “more pronounced for false political news than for false news about terrorism, natural disasters, science, urban legends, or financial information” (Vosoughi et al., 2018, p. 1146).

The task of fighting fake news is raised by Lazer et al. (2018), who call for a multidisciplinary effort to address the problem. The authors identify two categories of interventions that might be effective to combat fake news and their influence: “(i) those aimed at empowering individuals to evaluate the fake news they encounter, and (ii) structural changes aimed at preventing exposure of individuals to fake news in the first instance” (Lazer et al., 2018, p. 1095). While the former concerns essentially fact checking, the latter is mostly performed through platform-based detection and intervention via algorithms and bots. Burger et al. (2019), in their turn, discuss a specific element in the junk news (i.e. low-quality news) ecosystem: junk news (part of which may be fake) that are commercially motivated – “i.e. money-driven, highly shareable clickbait

with low journalistic production standards” (p. 1). By studying the Dutch case, they show that, during the period from 2013 to 2017, the average number of user interactions with junk news significantly exceeded that with mainstream news. They also show that more than half of the Dutch Facebook users interacted with a junk news post at least once in this period.

Some studies focus on the scenarios of developing countries, which bring their own characteristics. For illustration, here we cite Pate and Ibrahim (2019), who analyze the impacts of fake news on the Nigerian democratic system; Glowacki et al. (2018), who target political news and information shared over Twitter and Facebook during the 2018 Mexican presidential election; and Arnaudo (2017), who investigates the use of automated accounts (bots) that spread misinformation in three Brazilian political moments: the 2014 presidential elections, the impeachment of former president Dilma Rousseff, and the 2016 municipal elections in Rio de Janeiro. It is also worth mentioning the project entitled *Eleições Sem Fake*⁵ (i.e., *Elections Without Fake*), focused on the development of computational systems to fight against the spread of fake news during Brazilian elections. The cases of other countries, including Romania (Bârgăoanu and Radu, 2018) and South Africa (Wasserman, 2017), have been studied as well.

A significant number of other computationally-driven studies on misinformation on the Web have already been performed, especially on the topics of fake news characterization (e.g. Rashkin et al., 2017; Arif et al., 2018) and its automatic detection (e.g. Conroy et al., 2016; Rubin et al., 2016; Shu et al., 2017; Tschatschek et al., 2018; Reis et al., 2019). Zannettou et al. (2017), for instance, analyze news published on three online platforms (4chan, Reddit and Twitter) in order to identify and characterize the flow of mainstream and fake news between them, shedding some light

⁵ Available at <http://www.eleicoes-sem-fake.dcc.ufmg.br/> .

on the important topic of cross-platform misinformation spread. Vicario et al. (2019) present a framework for identifying polarizing content on social media and, thus, predicting potential targets for hoaxes and fake news. Ruchansky et al. (2017) propose a model for misinformation detection that captures three already observed common characteristics of fake news: (a) low quality of the text, including mismatches between the headline and the body of the article; (b) response to provocation, since “fake news often contains opinionated and inflammatory language, crafted as click bait or to incite confusion” (p. 797), which motivates responses with a high emotional content; and (c) doubtful source, i.e., lack of credibility of the URL, media source and author that published the news story. Despite the good results achieved by the model presented in their study, the classification of fake news still remains, at the writing of this dissertation, a challenging problem with many open questions.

On the expression *fake news*

More related to the object of this study are the investigations on the use of the very expression *fake news*. However, as put by Gelfert (2018), “[w]hile much ink has been spilled, by academics and pundits alike, on [the] disruptive potential and deceptive nature [of the fake news phenomenon], somewhat less attention has been paid to analyzing and defining the term ‘fake news’ ” (p. 85). According to this author, “[i]t is (...) quite natural that a term as recent and controversial as ‘fake news’ should be used in a variety of (sometimes conflicting) ways, thereby making conceptual analysis more difficult” (Gelfert, 2018, p. 85). Gelfert recognizes that the term *fake news* has evolved rapidly and argues that “it should be reserved for cases of deliberate presentation of (typically) false or misleading claims as news, where these are misleading *by design*” (p. 84,

emphasis in original)⁶.

Nielsen and Graves (2017), after analyzing data from focus groups and surveys, found that

[w]hen asked to provide examples of fake news, people identify poor journalism, propaganda (including both lying politicians and hyperpartisan content), and some kinds of advertising more frequently than false information designed to masquerade as news reports. (Nielsen and Graves, 2017, p. 1)

They also observe that people are aware that *fake news* is often employed as “a politicized buzzword used by politicians and others to criticize news media and platform companies” (Nielsen and Graves, 2017, p. 1). This fact reinforces the observation that the term has been carrying an imprecise definition, which is one of the reasons why Habgood-Coote (2019) argues that academics and journalists should stop using it⁷. Tandoc Jr. et al. (2018) contribute to this discussion by reviewing 34 academic articles in order to understand how studies from between 2013 and 2017 have used this expression. The authors classify the found definitions into six categories: news satire, news parody, news fabrication, photo manipulation, advertising and public relations, and propaganda. Even though the use of this term is not analyzed from a diachronic perspective (such as the one proposed in this chapter), Tandoc Jr. et al. (2018) acknowledge that “[e]arlier studies have applied the term to define related but distinct types of content, such as news parodies, political satires, and news propaganda”, but that “it is currently used

⁶ The author then continues: “[t]he phrase ‘by design’ here refers to systemic features of the design of the sources and channels by which fake news propagates and, thereby, manipulates the audience’s cognitive processes” (Gelfert, 2018, p. 84).

⁷ Of course, we assume that Habgood-Coote (2019) is not opposing studies like the one presented here, i.e., that investigate the very use of the term.

to describe false stories spreading on social media” (p. 138). This observation sustains at least one of our results, that suggests a stronger link between the term and parody/satire before the 2016 presidential election in the United States of America, and an association with social networking sites (particularly Facebook, Twitter and WhatsApp) during and after the election campaign.

The specific use of the term *fake news* by Donald Trump has been the subject of previous works too. Ross and Rivers (2018) investigated a corpus containing 1,416 tweets posted by Trump through comparative keyword analysis and show that this term has been employed by him as a pejorative label used to ridicule the critical mainstream media and “to position himself as the only reliable source of truth” (p. 1). Also, Holan (2017) compares the media’s definition of *fake news* with Donald Trump’s definition, arguing that

[w]hen PolitiFact fact-checks fake news, we are calling out fabricated content that intentionally masquerades as news coverage of actual events. When President Donald Trump talks about fake news, he means something else entirely. Instead of referring to fabricated content, Trump uses the term to describe news coverage that is unsympathetic to his administration and his performance, even when the news reports are accurate. (Holan, 2017, p. 121).

Horta Ribeiro et al. (2017) shows that this behavior (that is, labeling as *fake news* any opinions or facts with which one disagrees) is not restricted to Donald Trump and to other populists from around the world. The authors use the suggestive title “Everything I disagree with is #FakeNews” in their article showing that Twitter users also employ the term *fake news* (and other related words and expressions) when designating political content with which they disagree.

Tambini (2017) asks two central questions related to the rise of the expression *fake news*: (a) why have politicians and the media suddenly started talking about fake news? And (b) who benefits from using this concept? According to his view, the three main beneficiaries are the “new populists”, who “use the notion of ‘fake news’ to undermine legitimate opposition, and resist fourth estate accountability”; the “historical losers”, who “claim that political changes result from misinformation”; and the “legacy media”, that “want to discredit the ‘wisdom of crowds’ and aim for a return to trusted news brands” (Tambini, 2017, p. 9).

Finally, we cite Brummette et al. (2018), who use social network analysis, content analysis and cluster analysis to explore the use of the term *fake news* on Twitter. Similarly to what we propose here, the authors investigate the prevalent discussions surrounding this expression through the analysis of elements like the most frequently co-occurring words and hashtags. Our approach, however, is innovative for the study of this term not only because it employs different methods and tools, but mainly because it is guided by a diachronic perspective, which allows comparisons between distinct moments in the history of the studied expression.

On the methodology of this study

From a theoretical viewpoint, this chapter is partially inspired by the seminal studies on social lexicology proposed by Matoré (1949, 1953) and further developed as sociohistorical lexicology by Cambraia (2013). These authors argue for the use of models for lexical analysis that take into account social and extralinguistic factors, and address the link between the lexicon and social transformations. Cambraia (2013) considers that important questions in a sociohistorically-based lexicology are, for example: what makes a lexical item earn or lose a sense (or a meaning)? Or what drives a speaker to create a new word for a concept for which another

word already existed? According to this approach, the answers to these questions must go beyond the analysis of strictly linguistic factors, but should also contemplate extralinguistic elements. From a methodological perspective, the studies influenced by Cambraia (e.g. Guedes and Mendes, 2016; Dores and Toledo, 2018; Rafael and Simião, 2019) propose an in-depth analysis of specific lexical items, and the textual and social contexts in which these items are inserted are investigated. Here we use these previous works as an inspiration for the proposal of a framework to the analysis of the conceptualization of a given item. Differently from them, however, we employ tools and resources from corpus linguistics and natural language processing that are not considered in the original Cambraia’s proposition.

In 2011, Michel et al. (2011) coined the term *culturomics*, meaning a method to study human behavior, cultural trends and language change through the diachronic quantitative analysis of texts, including of digitised books provided by the project Google Books. Several studies explore this method to investigate topics such as the dynamics of birth and death of words (Petersen et al., 2012), semantic change (Gulordava and Baroni, 2011), emotions in literary texts (Acerbi et al., 2013) and general characteristics of modern societies (Roth, 2014), to name a few. Nevertheless, many criticisms arose regarding limitations of inferences derived from the analysis of Google Books due to factors that range from optical character recognition errors and overabundance of scientific literature (Pechenick et al., 2015) to the lack of metadata in the corpus (Koplenig, 2017).

Leetaru (2011) proposes a somewhat complementary approach that he calls *culturomics 2.0*, which employs computational analysis of large text archives composed of historical news data (instead of books) and can, according to the author, “yield intriguing new understandings of human society”. The author performed sentiment

mining and full-text geocoding in order to offer “new insights into how the world views itself and the ‘natural civilizations’ of the news media” (Leetaru, 2011). In the same vein, Flaounas et al. (2010) analyze the European mediasphere and the writing style, gender bias and the popularity of particular topics (Flaounas et al., 2013) in large corpora of news articles. Lansdall-Welfare et al. (2014), also using a large dataset of media reports, observe a change of framing and sentiment associated with nuclear power after the Fukushima nuclear disaster from 2011. The authors detected effects on attention, sentiment, conceptual associations and in the network of actors and actions linked to nuclear power following the accident. Our work draws a lot of inspiration from this study, as some of the methods (such as the analyses of textual polarity and co-occurring named entities) are similar. Also, Lansdall-Welfare et al.’s investigation contains a temporal element as well, since it compares the media coverage of nuclear power before and after the Fukushima disaster. Nonetheless, the focus of our proposal is the analysis of specific linguistic items, which is the reason why we consider our framework a contribution to the diachronic study of the lexicon from a corpus linguistics perspective.

As mentioned earlier, this chapter extends a previously published work (Cunha et al., 2018) in which we investigate the conceptualization of the term *fake news* in English (i.e., not considering Brazilian news sources as well). After this publication, we also used this incipient framework to analyze news articles that mention the name of the software application WhatsApp in Brazil and in parts of the English-speaking world (Caetano et al., 2018). Among the results obtained, we show that WhatsApp started to be linked to misinformation, politics and criminal scams in 2018. The use of the methodology proposed here for another case study, which also corroborated observations provided by other researchers, shows that our framework is robust enough to be employed in a variety of

different contexts. As far as we are concerned, our investigations are the first studies that use a combination of already established methods and tools from corpus linguistics and natural language processing in order to quantitatively examine the history of relevant terms related to technology and online social media, thus helping us to better understand social trends in a fast-changing world.

4.2 Data sources

We use two diachronic corpora of news articles in this study: the first one comes from the Corpus of News on the Web (NOW Corpus), while the second one is a collection of news gathered from Brazilian online newspapers and magazines.

4.2.1 English-written news corpus

The Corpus of News on the Web (NOW Corpus) contains articles written in English and published from 2010 to the present time⁸ in online newspapers and magazines based in 20 different countries (Davies, 2013, 2017b). At the time of writing of this study, this corpus is available for download and online exploration⁹. Our analyses are relative to a version of the corpus available in the month of April 2018, containing around six billion words of data.

Using the NOW Corpus exploration tool, we searched for all the occurrences of the term *fake news*. For each occurrence, the online tool provides a concordance line, or *context* – that is, a piece of text of approximately 20-30 words around (before and after) the

⁸ As of the writing of this dissertation, the NOW Corpus is updated daily. It is, thus, a *monitor corpus* (also called a *dynamic corpus*), i.e., a corpus that is “continually growing over time, as opposed to a **static corpus**, which does not change in size once it has been built” (Baker et al., 2006, p. 64, emphasis in original).

⁹ At <https://www.english-corpora.org/now/> .

searched term. For example, for a certain news article published in July 25, 2017 in the Kenyan newspaper Daily Nation, the context around the term *fake news* is: (...) of social media and a study that said 90 per cent of Kenyans had encountered **fake news**. WhatsApp and Facebook are the two leading sources of misinformation, often (...). For illustration, Figure 4.3 shows a small selection of contexts¹⁰ including *fake news* in the NOW Corpus online exploration tool. All of our analyses were performed in these contexts, since words immediately surrounding a key term are more relevant to the conceptualization of this term than words further away from it, though in the same text – as put by Baker et al. (2006), “concordances provide information about the ‘company that a word keeps’ ” (p. 43). Wynne (2008) adds that the main reason for using keywords in context (KWICs) in corpus linguistics is that “interesting insights into the structure and usage of a language can be obtained by looking at words in real texts and seeing what patterns of lexis, grammar and meaning surround them” (p. 711).

The total number of occurrences of *fake news* extracted from the NOW Corpus in April 30, 2018 is 41,124. These occurrences encompass news articles published in all the 20 countries represented in the corpus, that were then grouped into the following six regions based on their geographical locations: Africa, British Isles, Indian subcontinent, Oceania, Southeast Asia and the Americas. Some countries with very different histories and cultures were included in the same group – for example, Nigeria and South Africa were grouped together under the label *Africa*, as were India and Pakistan under the label *Indian subcontinent*. This is obviously a result of the simplification needed to carry out the study proposed here. Researchers interested in further analyzing a specific part of the world should of course take into account the differences between

¹⁰ A small selection of random concordance lines is also called a *thinned concordance* (Baker et al., 2006).

The screenshot shows the NOW Corpus interface with a search for 'fake news'. The results table is as follows:

SEARCH	FREQUENCY	CONTEXT	OVERVIEW
FIND SAMPLE: 100 200 500 1000 PAGE: << < 1/786 > >>			
CLICK FOR MORE CONTEXT [?] SHOW DUPLICATES			
1	17-02-16 CA	CTV News	A B C water. He speculated on a nuclear holocaust. He blamed journalists for reporting "fake news" and suggested that, in another life, he would've r
2	17-12-28 US	Washington Examiner	A B C Trump's request, we are holding a contest to name the 2017 KING of Fake News. And we want to hear from you," the email from the
3	19-04-18 US	The New York Times	A B C the political spectrum openly debating whether the hacking and leaking of emails -- and the fake news that spread like a wildfire on social media
4	18-04-18 MY	Daily Express	A B C to be the first case in Sabah of anyone being investigated for creating or distributing fake news under the recently gazetted Anti-Fake News Act
5	17-01-15 SG	The Straits Times	A B C Meanwhile, a myriad of websites with names such as rapefugees.net or nooh.info are spreading fake news on existing European politicians; a fa
6	18-09-18 NG	SaharaReporters.com	A B C 40043785 Dele Momodu Is A Shameless Purveyor Of Fake News By Churchill Okonkwo # In the ensuing furor, someone called Chief Dele Mom
7	17-04-17 CA	The Chronicle Journal	A B C Age ", # The " Dust Bowl " man made, that's " fake news ". # Humans have no control over this Planet's Climate. The
8	18-03-28 SG	The Straits Times	A B C # This, he said, could be an additional weapon in the fight against fake news. # At the public hearing on Wednesday (March 28), he
9	19-03-06 SG	The Independent	A B C . # Remember: One of the simplest, yet most effective ways to stop fake news in its tracks is to pause for a moment before forwarding a piece o
10	19-06-05 SG	Yahoo Singapore News	A B C and right-wing media. # Like Trump, 62% of Republicans and Republican-leaning independents said fake news is a big problem, compared with
11	17-04-16 MY	Daily Express	A B C about the matter by calling us at 088-520100," she added. # The fake news that was circulated widely on WhatsApp messaging application was
12	19-03-21 NG	The Punch	A B C . " INEC said. 40759362 Buhari on fake news: Digital space difficult to regulate # President Muhammadu Buhari says it is difficult
13	18-10-24 US	Fox News	A B C # At previous rallies, Trump has derided CNN and other media organizations as " fake news ." and mocked prominent Democrats like California
14	18-10-24 MY	The Edge Markets MY	A B C and WhatsApp -- the four most-used apps by Indonesians -- as key to stamping out fake news. # And while there will be penalties for companies
15	17-12-02 US	The Daily Herald	A B C It underpins Trump's nonstop attacks on non-Foxolinjonsonian media, promoting the myth of " fake news " to defame reporters brave enough to
16	19-04-26 ZA	WeeTracker Media	A B C . Journalists are today required to douse the flames of this dangerous wildfire. # Fake news has negatively impacted the credibility of journalism
17	19-05-07 ZA	Malawi24	A B C only 14 days to go before the May 21 Tripartite Elections, the rise in fake news on Facebook and WhatsApp about the Electoral Body, private org
18	17-04-11 IN	Times of India	A B C 17905034 Fake news: New DAVP guidelines make it harder for media organisations to indulge in fraudulent

Figure 4.3: A small selection of contexts including *fake news* in the NOW Corpus online exploration tool.

its specific regions. In any case, this grouping was motivated by the fact that offline and online news outlets tend to give preference to local and national news, to domestic news about other countries, and to reflect imbalanced information flows between the developed and the developing worlds (Berger, 2009). To illustrate the outcome of this division, we show, in Figure 4.4, a map highlighting the countries considered in the English-written corpus, grouped by region.

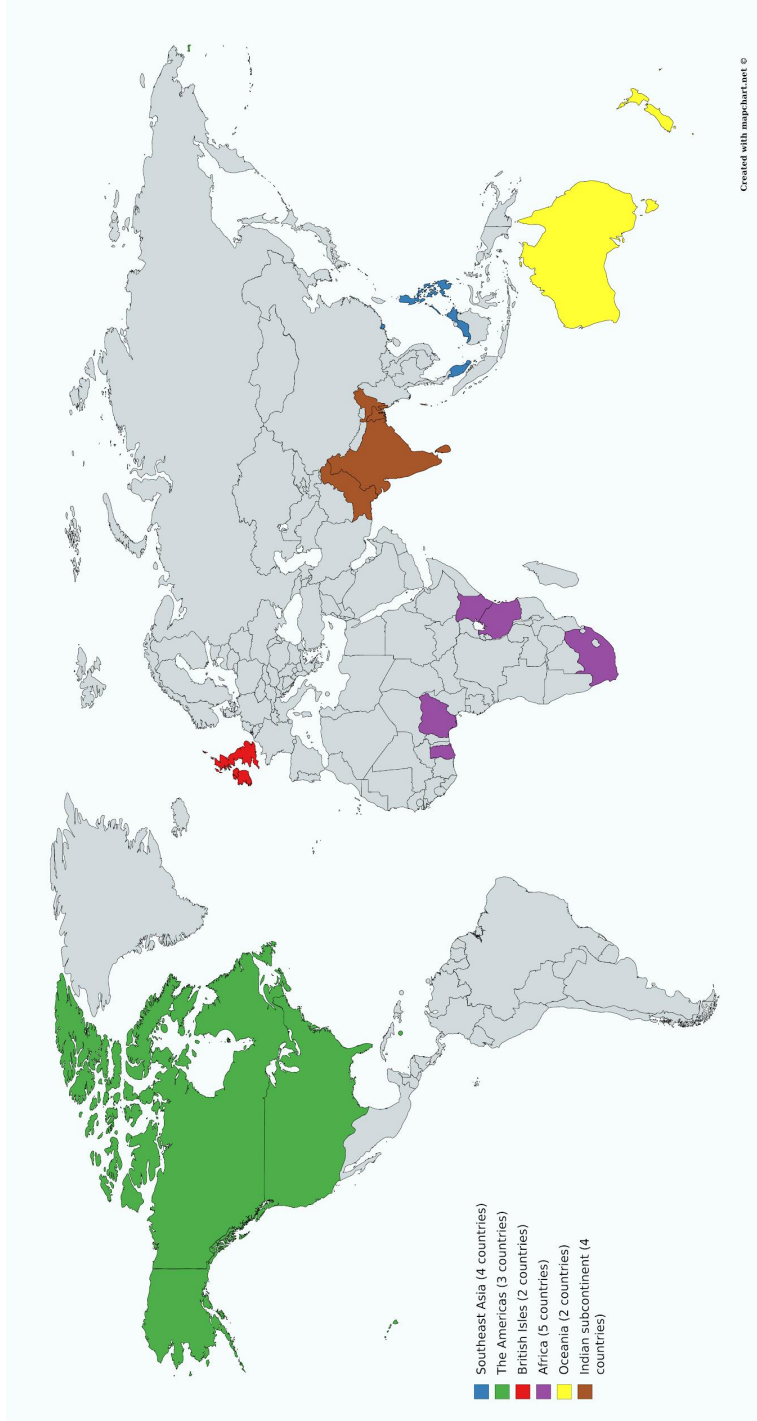


Figure 4.4: Map highlighting the countries considered in the English-written news corpus, grouped by region.

These occurrences also cover each year in the corpus until our data collection (i.e., from 2010 to 2018). Due to the previously observed increase in the usage of the term *fake news* during and after the 2016 presidential election in the United States of America (mentioned in Section 4.1), we categorized the occurrences into two periods: before and after the 2016 US election. The election was held in November, but we set the delimitation date between these periods in the end of the first semester of 2016 (June 30) in order to include the political campaign in the period *after US election*. Table 4.1 and Table 4.2 show, respectively, the number of contexts containing the term *fake news* in this corpus according to the geographical origin of the corresponding news media and to the year and period of publication of the news article.

4.2.2 Brazilian news corpus

Our second data source includes articles collected from Brazilian online newspapers and magazines, all written in Portuguese, also containing the term *fake news*. To collect this material, we used the tool *Selenium*¹¹ to automate exact match searches for the term *fake news* in the following ten major Brazilian news websites: *Exame*, *Folha de S. Paulo*, *Gazeta do Povo*, *G1*, *O Estado de S. Paulo*, *R7*, *Terra*, *Universo Online (UOL)*, *Valor Econômico* and *Veja*. The total number of occurrences of *fake news* extracted from these websites on December 31, 2018 is 4,936. These occurrences appeared in 2,464 unique news articles. Then, we used the Python library *newspaper*¹² to collect the full texts of the news articles containing these occurrences. Finally, we gathered the fifteen words before and after the key term *fake news* to create the contexts/concordance lines that will be analyzed in the next sections.

¹¹ Available at <https://www.seleniumhq.org/>.

¹² Available at <https://pypi.org/project/newspaper/>.

Table 4.1: Number of contexts containing the term *fake news* in the English-written news corpus according to the geographical origin of the corresponding news media.

Region	Country	Occurrences
Southeast Asia	Singapore	3,722
	Malaysia	3,455
	Philippines	3,058
	Hong Kong	171
Total: 25,3% / 10,406		
The Americas	United States	6,775
	Canada	2,960
	Jamaica	124
Total: 24,0% / 9,859		
British Isles	Great Britain	4,213
	Ireland	2,035
Total: 15,2% / 6,248		
Africa	South Africa	2,493
	Nigeria	1,974
	Kenya	1,368
	Ghana	300
	Tanzania	1
Total: 14,9% / 6,136		
Oceania	Australia	3,052
	New Zealand	1,446
Total: 10,9% / 4,498		
Indian subcontinent	India	2,961
	Pakistan	772
	Sri Lanka	147
	Bangladesh	97
Total: 9,7% / 3,977		

Table 4.2: Number of contexts containing the term *fake news* in both English-written and Brazilian news corpora according to the year and period (before or after the 2016 and 2018 presidential elections in the United States of America and Brazil, respectively) of publication of the news article.

English-written news corpus											
Year	2010	2011	2012	2013	2014	2015	2016	2017	2018		
Occurrences	24	43	57	64	89	95	4,766	25,293	10,693		
Period	before US election: 494; after US election: 40,630										
Brazilian news corpus											
Year	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	
Occurrences	1	4	0	0	10	0	3	4	158	4,756	
Period	before BR election: 923; after BR election: 4,013										

Similarly to what was observed in the 2016 US election, the increase in the usage of the term *fake news* in Brazil seems to correspond to the 2018 Brazilian general election (see Section 4.3.1). For this reason, we also categorized the occurrences in this dataset into two periods: before and after the 2018 BR election. The election was held in October, but, again, we set the delimitation date between these periods in the end of the first semester of the year (June 30, 2018) to include the political campaign in the period *after BR election*. Table 4.2 also shows the number of contexts containing the term *fake news* in the Brazilian news corpus according to the year and period of publication of the news article.

It is noteworthy to mention the increase in the frequency of the expression *fake news* in the English-written corpus from 2015 (95 occurrences) to 2016 (4,766 occurrences), and in the Brazilian corpus from 2016 (4 occurrences) to 2017 (158 occurrences) and then to 2018 (4,756 occurrences).

4.3 Analyses and results

In this section, we display and examine the outcomes of our investigations. Each analysis is introduced by a description of how it is able to contribute answering to our research question, followed by the methodology employed, and finally by a presentation and discussion of the results found.

4.3.1 Web search behavior

Before analyzing the data obtained from our English-written and Brazilian news corpora, we investigate whether it is possible to observe changes in Web search behavior regarding the expression *fake news* corresponding to the high increase in its use during and after the 2016 and 2018 elections in the United States of America

and in Brazil, as mentioned in Section 4.1 and observed in Table 4.2.

Data obtained from Google Trends¹³, an online tool that indicates the frequency of particular terms in the total volume of searches in the Google Search engine, informs that, according to this metric, the worldwide public interest in the term *fake news* was approximately constant from 2010 until mid 2016, when it greatly and suddenly increased, as indicated by Figure 4.5a. This corresponds to the period of the campaign for the 2016 US presidential election. When we examine Web searches for the term *fake news* in Brazil over time (Figure 4.5b), we also observe a significant increase. However, in this case, the most noteworthy growth happened in mid 2018, which also corresponds to the period of a major political event: the campaign for the 2018 Brazilian general election.

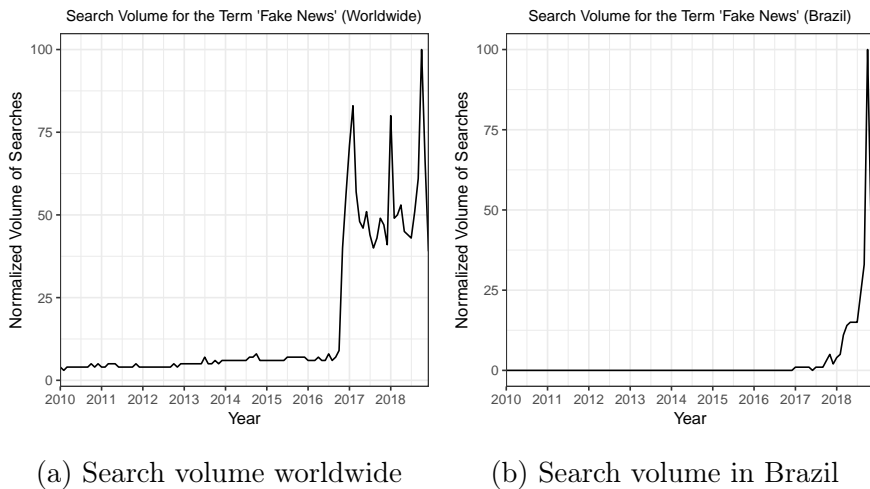


Figure 4.5: Normalized volumes of searches for the expression *fake news* on Google Search from 2010 to 2018. Values represent search volumes relative to the highest point on the chart. A value of 100 is the peak popularity for the term.

¹³ <https://trends.google.com/trends/> .

Google Trends data also show a spatial change regarding queries for the term *fake news*. The ten countries with the highest proportions of searches for *fake news* in each period (before and after the 2016 US presidential election) are listed in Table 4.3. In the period before the 2016 US election, a significant part of the countries with the highest proportions of searches are from the Eastern world (India, United Arab Emirates, Singapore, Qatar, Pakistan). However, after the US election, the proportion of searches for this expression in Western countries increased considerably, especially in Europe (Norway, Denmark, Ireland, United Kingdom, Switzerland). We are not able to provide an explanation for such observation.

Table 4.3: Countries with the highest proportions of searches for *fake news* on Google Search before and after 2016 US election.

Period	Countries
before US election	India, United Arab Emirates, Singapore, United States, Macedonia, Qatar, New Zealand, Canada, Pakistan, Australia
after US election	Singapore, Philippines, United States, Canada, South Africa, Norway, Denmark, Ireland, United Kingdom, Switzerland

A closer look at the data from Google Trends also reveals that the great increase in the public interest for the expression *fake news* coincided with a change in the focus of Web searches. Table 4.4 shows the five most frequent search terms employed by users who also searched for *fake news* in the periods before and after the 2016 US election. We observe that, before the election, searches for *fake news* were generic and regarded terms related to the media industry itself, like *article*, *stories* and *report*; after the election, however,

these searches started to be more focused on political affairs and in the spread of false information, mentioning entities like the elected president of the United States of America in 2016 (Donald Trump), the television news channel CNN (that devotes large amounts of its coverage to US politics) and the online social media Facebook (sometimes considered a major source of fake news on the Internet).

Table 4.4: Most frequent search terms related to *fake news* on Google Search before and after US and BR elections.

Period	Search terms
Worldwide	
before US election	fake news generator, fake news article, fake news stories, make fake news, fake news report
after US election	trump news, the fake news, fake news trump, cnn news, fake news facebook
Brazil	
before BR election	fake news redação, o que fake news, fake news brasil, fake news o que é, fake news significado
after BR election	tse fake news, fake news eleições, fake news kit gay, redação enem 2018, kit gay

Table 4.4 also displays the most frequent terms queried by users who also searched for *fake news* in Brazil, but now in the periods before and after the 2018 Brazilian general election. Before the election, most of the top searches concerned the meaning of the expression *fake news* itself, such as *o que* and *o que é* (i.e., *what* and *what is it*), and *significado* (i.e., *meaning*). This is understand-

able, since a significant part of the Brazilian population does not speak English. During and after the campaign, however, the focus changed, again, mostly to queries related to politically related terms, such as *tse* (acronym for the Brazilian Superior Electoral Court) and *eleições* (i.e., *elections*), and to campaign controversies (*kit gay*¹⁴).

In this section, we used data obtained from the Google Trends tool. From now on, all of our analyses use the data described in Section 4.2, obtained from the NOW Corpus and from our collection of Brazilian news articles.

4.3.2 Co-occurring named entities

The analysis of *named entities* – that is, real-world entities such as persons, organizations and locations that can be denoted with proper names (Tjong Kim Sang and De Meulder, 2003) – co-occurring with certain terms is an interesting way to contextualize these terms. In our case, by identifying which entities are linked to the expression *fake news* in different periods of time and in different parts of the world, we are able to observe relationships of “who and where” in the recent history of our key term.

In our corpora of news articles, we employed a simple method to identify named entities: we made use of the fact that newspapers and magazines consistently capitalize nouns representing named entities and counted all the words that appear capitalized in the con-

¹⁴ The “gay kit controversy” was one of the most contentious topics during the Brazilian presidential campaign of 2018. In short, far-right candidate Jair Bolsonaro accused center-left candidate Fernando Haddad of planning to distribute “gay kits” in schools – a reference to sexual education materials that, according to him, were aimed to “pervert” youngsters and encourage homosexuality. At some point, the Superior Electoral Court considered this information a piece of fake news and ordered Bolsonaro to remove it from his campaign. For more information on fake news in Brazilian politics, see Harden (2019).

texts; then, we manually analyzed the most frequent capitalized words in each subdivision of the corpora (i.e., representing each region and period) to remove words not relative to named entities (such as *I*, *SMS*, *March* and words capitalized for other reasons) and to merge duplicated entities represented more than once (e.g. *Donald* and *Trump*). This “semi-manual method” proved to be more effective than the use of automatic named entity recognition tools probably because of the lack of completeness of the contexts analyzed – which ignore sentence boundaries and punctuation, and may start and finish in indiscriminate positions of the texts (cf. examples of context in Section 4.2.1).

Table 4.5 shows the five most mentioned named entities in the English-written news corpus in the periods before and after the 2016 US presidential election, regardless of geographical origin of the corresponding news media. Before the US election, it is possible to observe a strong connection between humor and fake news: with exception of Facebook, all the other most mentioned named entities are related to satirical TV shows (The Daily Show, Onion News Network) and hosts (Jon Stewart, Stephen Colbert) based in the United States of America. On the other side, in the period after the US election, there is a movement towards politically related entities (Donald Trump), traditional media sources (CNN) and social networking services (Facebook and Twitter). It is interesting to notice that this shift matches the already mentioned (in Table 4.4) shift of interest towards political affairs and the spread of fake news on the Internet observed in Web searches.

Table 4.5 also displays the five most mentioned named entities in the Brazilian news corpus in the periods before and after the 2018 Brazilian general election. Here, before the election, we observe the presence of entities linked to the upcoming electoral process, such as TSE (the Brazilian Superior Electoral Court) and its former president Luiz Fux. These two entities are highly mentioned

Table 4.5: Most mentioned named entities in the periods before and after US and BR elections.

Period	Entities
English-written news corpus	
before US election	The Daily Show, Jon Stewart, Onion News Network, Facebook, Stephen Colbert
after US election	Donald Trump, Facebook, US, CNN, Twitter
Brazilian news corpus	
before BR election	TSE, Luiz Fux, Facebook, Donald Trump, Brasil
after BR election	TSE, Jair Bolsonaro, Brasil, WhatsApp, Folha de São Paulo

due to Fux’s declaration (in June 2018) concerning the possibility of annulment of the election in case of massive fake news influence (Ramalho, 2018). Donald Trump is also mentioned, probably due to influences of the international scenario. Interestingly, after the start of the campaign period, candidate Jair Bolsonaro takes the place of Donald Trump and the online social service WhatsApp replaces Facebook, as a clear reflection of the Brazilian scenario in 2018, more affected by political fake news disseminated through WhatsApp than through Facebook – as alleged by the fifth most mentioned entity, the major newspaper Folha de S. Paulo (Phillips, 2018).

It is particularly interesting to compare the most mentioned named entities in the English-written news corpus after the US election with the most mentioned named entities in the Brazilian

news corpus after the BR election. In both cases, we observe the presence of: (a) the name of the country (*US* and *Brasil*); (b) the conservative (and winning) candidate (*Donald Trump* and *Jair Bolsonaro*); (c) social networking services (*Facebook* and *Twitter*, and *WhatsApp*); and (d) a traditional media source (*CNN* and *Folha de S. Paulo*). This fact shows that, although in different scenarios and times, many similarities still emerge.

Table 4.6: Most mentioned entities in the periods before and after US election, considering the geographical origin of the corresponding news media.

Region	Period	Entities
Africa	before	PDP, Ekiti, Nigeria
	after	Donald Trump, Facebook, US
British Isles	before	Facebook, The Daily Show, Stephen Colbert
	after	Donald Trump, Facebook, US
Indian subcontinent	before	Shahid Afridi, King Salman of Saudi Arabia, BJP
	after	Facebook, Donald Trump, US
Oceania	before	Twitter, The Daily Show, NBC
	after	Donald Trump, Facebook, US
Southeast Asia	before	Korina Sanchez, US, China
	after	Facebook, Donald Trump, US
The Americas	before	The Daily Show, Jon Stewart, Onion News Network
	after	Donald Trump, Facebook, CNN

When we make this same diachronic comparison (*before vs. after* the elections), but now considering the geographical origin of the corresponding news media in the English-written corpus, we observe a noteworthy phenomenon: the global standardization of the named entities related to *fake news*. Table 4.6 shows the three most mentioned entities in the periods before and after US election in each region, and indicates that local entities are more relevant in the period before the US election, when names of geographical regions (Ekiti), countries (Nigeria, China), local political parties (PDP – People’s Democratic Party of Nigeria, BJP – Bharatiya Janata Party of India) and local personalities (Shahid Afridi, King Salman, Korina Sanchez) appear frequently among the most mentioned entities. In the contexts after the US election, however, Donald Trump, Facebook and US are the three most mentioned entities for nearly all the regions – with the sole exception of the Americas, where CNN replaces US.

4.3.3 Semantic fields of the surrounding vocabulary

Besides the investigation of the named entities that accompany a given key term, the analysis of the general vocabulary co-occurring with it is also valuable. According to Cunha et al. (2014b), “vocabulary is a system of mapping the world, so this kind of investigation reveals how groups perceive reality” (p. 215). In our case, one of the possible methods of performing such analysis is by observing the semantic fields (i.e., groups to which semantically related items belong) of the words co-occurring with the expression *fake news* in our contexts.

For performing this task, we first lemmatized all the words in the contexts by employing the WordNet Lemmatizer function provided by the Natural Language Toolkit (Bird et al., 2009) and using

verb as the part of speech argument for the lemmatization method. By applying this lemmatization, we grouped together the inflected forms of the words so that they could be analyzed as single items based on their dictionary forms (*lemmas*).

Then, we used *Empath* (Fast et al., 2016), “a tool for analyzing text across lexical categories”¹⁵, to classify the lemmatized words according to categories that represent different semantic fields, such as diverse topics and emotions. For every context, we calculated the percentage of words belonging to each semantic field represented by an *Empath* category. Due to the high number of categories pre-defined by *Empath* (194 in total), we selected eight that showed interesting results and are relevant for our discussion: *government*, *internet*, *journalism*, *leader*, *negative emotion*, *politics*, *social media* and *technology*. By way of example, the category *internet* includes 79 words such as *homepage*, *download* and *hacker*, while the category *journalism* contains 69 words, including *report*, *article* and *newspaper*. The complete lists of words that comprise each one of these categories are displayed in Appendix B. Since *Empath* is (at the moment of the writing of this dissertation) available only in English, the corpus containing Brazilian news articles was not included in this analysis. In future work, it might be possible to use alternative tools, such as the multilingual Linguistic Inquiry and Word Count – LIWC (Pennebaker et al., 2015), to also consider the Brazilian news corpus in this analysis.

Figure 4.6 displays the average percentage of words in these categories for all the six regions considered here, both before and after the 2016 US election. By analyzing the graphs presented, we observe interesting differences and trends regarding the quantitative utilization of words from the semantic fields considered. We highlight the high increase in the use of words from the related categories *government*, *leader* and *politics* (and also from the supposedly unre-

¹⁵ <https://github.com/Ejhfast/empath-client> .

lated category *negative emotion*) and the high decrease in the use of words from the categories *internet*, *journalism* and *technology* (but not *social media*) in almost all regions after the US election.

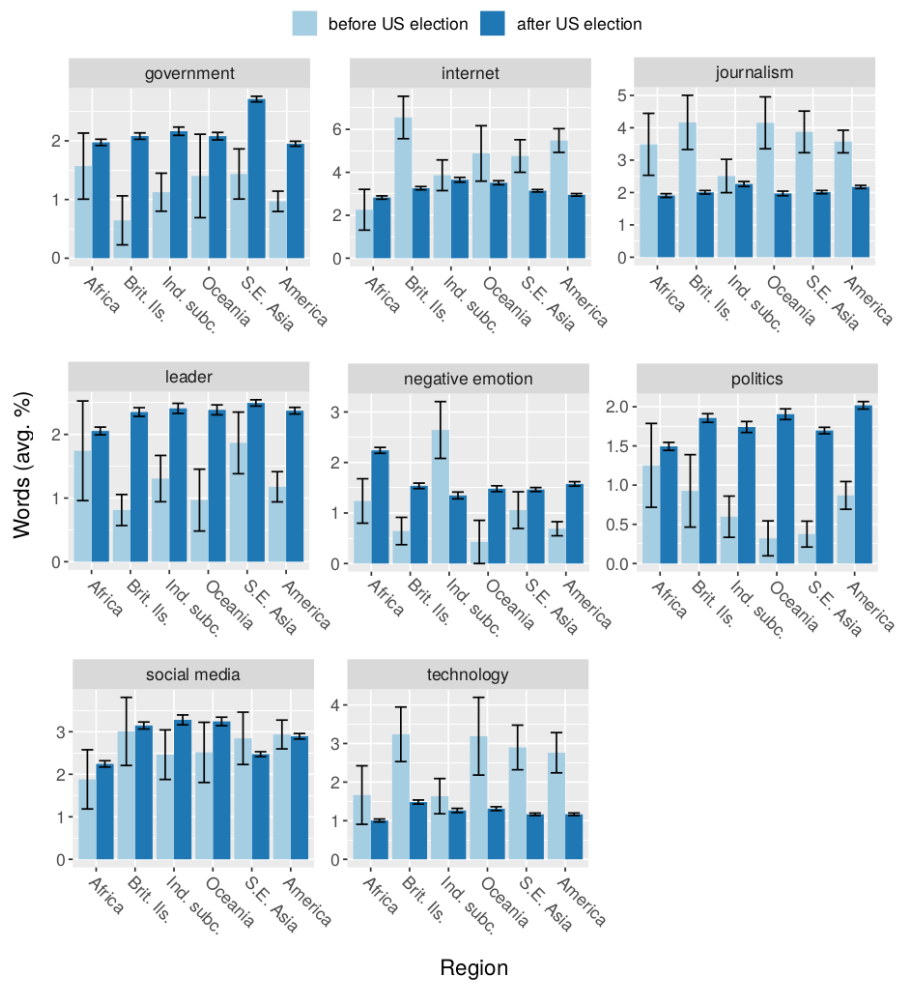


Figure 4.6: Percentage of words in each semantic field represented by an Empath category. Error bars indicate standard errors.

We hypothesize that these results indicate a change in the focus of the news considered here: before the 2016 US election, the term

fake news was probably more mentioned in contexts in which the focus was the *environment* where they occur (Internet, newspapers etc.), sometimes even meta-discussions on the very topic of fake news and its dissemination; during and after the US election, however, the discussion seems to have migrated to themes more close to the *content* of the fake news themselves (politics, elections etc.).

4.3.4 Co-occurrence networks

Another possible method of investigating the vocabulary accompanying a key term in a corpus is through the observation of co-occurrence networks. In our case, this method enables us to visually analyze the words that co-occur with the expression *fake news* in the contexts considered. Here we compare co-occurrence networks between the periods before and after the elections. These networks are represented by graphs, in which each node corresponds to a word and each (weighted) edge corresponds to an association between two given words.

To build our graphs of co-occurring words, we followed the steps below. First, we removed stop words using the lists provided by the Natural Language Toolkit (Bird et al., 2009) for English and Portuguese (the words *fake* and *news* were included in these lists as well, since they are present in all contexts). Then, we extracted the most relevant words from each period by using the *term frequency-inverse document frequency* (*tf-idf*) technique, that reflects how important a word is to a document in a corpus (Rajaraman and Ullman, 2011). We calculated the *tf-idf* for each pair (period, word) and extracted from each period the top 50 words with the highest *tf-idf* scores. In the following step, we counted the number of co-occurrences of each pair of words. For each period, we obtained the list with its 50 most relevant words (according to *tf-idf*) and incremented by one the counter relative to each pair of words in

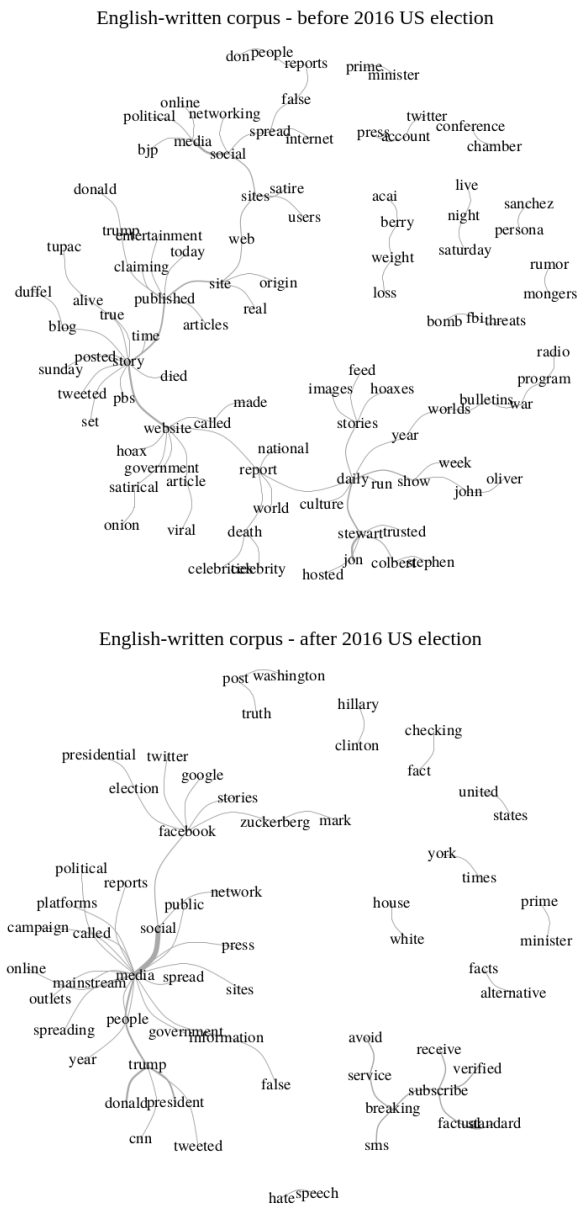


Figure 4.7: Co-occurrence networks before and after the 2016 US election in the English-written corpus.

this list (combination two by two). Instead of using the absolute count of contexts in which two words co-occur, we normalized this value by dividing it by the total number of contexts. At the end of this process, we obtained a graph in which vertices represent words and weighted edges indicate their degree of co-occurrence in the same contexts. Finally, we selected the top 100 edges with the highest weights, calculated the maximum spanning tree out of the remaining graph, and generated trees that depict the most relevant relationships, which are presented in Figures 4.7 (for the English-written corpus) and 4.8 (for the Brazilian corpus).

This method of investigation enables us to make several qualitative observations, which can be further elaborated in specific studies for this purpose. Here we will just draw attention to some clusters that seemed interesting to us. Comparing the two graphs representing the English-written news corpus, we notice, for instance, that before the US election one of the main clusters contains words related to the news industry itself (*articles, published, story*) and to the Internet (*website, tweeted, blog, posted*). Corroborating previous findings (Section 4.3.2), there is also a cluster containing words referring to satirical TV shows and hosts (*daily, show, colbert, oliver, stewart*). There are also a few words related to the political world, including *donald, trump* and *government*, and to minor topics, such as a cluster on the use of açai berry to lose weight. In the graph representing the period after the 2016 US election, we start to observe more terms specifically linked to the US election itself. In the main cluster, *donald, trump* and *president* have an important role; *presidential, election, political* and *campaign* also appear. The thickest edge, however, is between the words *social* and *media*, which appear close to *spread* – suggesting, then, the relationship between social media and the spread of fake news. Some terms that surround meta-discussions about fake news are also present, highlighting relevant related concepts such *alternative facts, fact checking, hate speech*

and *post truth*.

In the Brazilian news corpus, before the 2018 election, most of the relevant co-occurring words in the main cluster are already related to the battle (*combate*) against the dissemination (*disseminação*) of fake news (*notícias, falsas, fabricadas, boatos*), including the battle officially promoted by the justice due to the upcoming elections (*tse, eleitoral, ministro, fux*). The use of bots (*robôs*) in the elections is also mentioned. The scenario in the United States (*estados, unidos*) is also relevant, as seen by the cluster that mentions *presidente, donald* and *trump*. During and after the election, however, the co-occurring words in the main cluster regard mostly the campaign itself. We highlight the names of the main presidential candidates (*jair, bolsonaro, fernando, haddad*) and the focus on social networks (*redes, sociais*), especially on WhatsApp messages (*whatsapp, mensagens*), since this tool was considered the main platform for disinformation during the 2018 elections in Brazil (Bradshaw and Howard, 2018).

4.3.5 Topics addressed in the contexts

In addition to studying the vocabulary around a key term, it is also possible to find the main topics addressed in the pieces of text surrounding the occurrences of the expression *fake news* in our corpora. For this task, we used *latent Dirichlet allocation* (LDA) (Blei et al., 2003), a way of automatically discovering topics in texts. LDA generates summaries of topics in terms of the keywords relevant for each topic, i.e., it returns a set of keywords that illustrates each topic alluded to in the text.

To perform this analysis, we first lowercased and tokenized all the words in both corpora. Then, we removed stop words using the list provided by the Natural Language Toolkit – after having added the words *fake* and *news* to this list, since they appear in all

contexts. Finally, we ran the LDA algorithm using *gensim* (Řehůřek and Sojka, 2010), a Python library for topic modeling. We used topic coherence score (Newman et al., 2010) to choose the optimum number of topics k to be returned by the algorithm. Thus, for each region, we ran the LDA algorithm starting with $k=2$ and ending with $k=20$, and chose the best LDA model, that is, the LDA model with highest topic coherence score. All regions had, respectively, $k=2$ and $k=14$ for the periods before and after the US election, except the Americas, that had $k=8$ and $k=14$, and Brazil, that had $k=16$ and $k=19$. For each region, the LDA returned these k topics containing keywords ordered by importance in the corresponding context, filtered both by region and topic. We then selected the main topic as the representative of each region and period.

Table 4.7 shows the top ranked ten keywords produced by our LDA model that represent the main topic in each region in both English-written and Brazilian media sources, before and after the elections. In this case, the analysis of the LDA output is performed subjectively, by observing and comparing keywords that are representative of each topic. Sometimes, however, the definition of a topic is not very clear. Nevertheless, we can find a few elements that seem to corroborate previous findings of this chapter. Regarding the English-written corpus, for example, we observe, for all regions, a relevant frequency of keywords related to journalism, media and the publishing industry in the period before the US election, like *story*, *website*, *site*, *report* and *article*. In the period after the US election, none of these keywords appears anymore, and we can find examples of keywords linked to politics (like *trump*, *president*, *politics*, *election*, *presidential*, *public* and *government*) and, to a lesser extent, to online social media (like *facebook*, *zuckerberg* and *twitter*). The region that more clearly displays this shift is probably Southeast Asia, whose top keywords in the most relevant topic before 2016 US election are literally *article*, *website*, *story*, *report* and *site*, and the

Table 4.7: Main topic for each region and period. For each topic, ten keywords are presented ordered according to the LDA output.

Region	Period	Main topic keywords
English-written news corpus		
Africa	before	become, world, party, south, leave, week, online, state, give, member
	after	trump, people, spread, president, truth, propaganda, thing, look, show, nigerian
British Isles	before	story, account, real, daily, website, new, show, use, death, state
	after	propaganda, source, russian, american, russia, lie, mean, popular, politics, allegation
Indian subcontinent	before	create, spread, report, death, also, lot, say, not, social, do
	after	facebook, also, problem, user, issue, company, russian, state, work, zuckerberg
Oceania	before	people, story, site, report, website, mortgage, would, fool, year, day
	after	election, influence, media, create, russian, question, policy, discuss, presidential, word
Southeast Asia	before	article, website, story, report, site, celebrity, death, publish, go, viral
	after	public, government, fact, twitter, proliferation, day, official, however, phenomenon, concern
The Americas	before	release, chip, firm, flurry, blue, target, date, breadcrumb, irresponsible, last
	after	facebook, problem, network, company, also, believe, publish, work, policy, russian
Brazilian news corpus		
before BR election		eleição, americana, redes sociais, propaganda, lado, ganhar, evitar, importante, reporter, seriedade
after BR election		imprensa, redes sociais, influenciar, verdade, proliferação, democracia, candidato, procuradora geral, votar, congresso

top keywords in the most relevant topic in the period after the election are *public*, *government*, *fact*, *twitter* and *proliferation*. These observations corroborate our findings from previous sections, where we found, for instance, a change in the focus of the news mentioning *fake news*: the focus changed from journalism and media to politics and online social networks.

In the Brazilian corpus, we observe that keywords related to the US election (*eleição* (i.e., *election*), *americana* (i.e., *american*)) rank high before the BR election, while after the election we note keywords related to information diffusion (*influenciar* (i.e., *to influence*), *proliferação* (i.e., *proliferation*)) and to the elections (*votar* (i.e., *to vote*), *candidato* (i.e., *candidate*)). It turns out, therefore, that the scenario in the United States was an important topic in the news that mentioned *fake news* in Brazil until mid-2016, but then the most important topic became the proliferation of fake news in the Brazilian electoral scenario itself.

To get a better sense of the results that LDA can generate, in future work it might be interesting to explore alternative possibilities for visualizing topics and keywords, and also to consider more than one topic per region and period.

4.3.6 Polarity

Our final analysis explores a different feature of the contexts in which the expression *fake news* appear in our corpora: their *polarities*, that is, whether the expressed opinion in the texts is mostly positive, negative or neutral. Here we performed sentiment analysis¹⁶ (Silva et al., 2016) in each one of the contexts using *SentiStrength*¹⁷ (Thelwall et al., 2010), a tool able to estimate the

¹⁶ “Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic” (Vinodhini and Chandrasekaran, 2012, p. 282).

¹⁷ Available at <http://sentistrength.wlv.ac.uk/> .

strength of positive and negative sentiment in short texts based on their vocabulary. Given a piece of text, this tool returns a score that varies from -4 (negative sentiment) to +4 (positive sentiment). Among the several possible tools for analyzing sentiment, we chose SentiStrength mainly due to two reasons: first, it was developed especially for short texts, like the contexts analyzed here; second, it is available for both English and Portuguese, so we could use the same tool to calculate polarities in our two corpora¹⁸.

The first six graphs in Figure 4.9 depict the average polarity of the contexts in each region of the English-written corpus before and after the 2016 US presidential election, while the rightmost graph shows the average polarity of the contexts in the Brazilian corpus before and after the 2018 Brazilian general election. We first observe a clear dominance of negative polarities in all periods and regions, indicating that the term *fake news* is often related to negative words (Zollo et al., 2015) and sentiments – which is not surprising, since the idea of fake news seems to be strongly associated with negative concepts like misinformation, manipulation and hostility.

In the charts concerning the English-written corpus, we also observe that, in general, the polarity expressed in the contexts in the period after the US election is more negative than before. The only exception is in the British Isles, where the difference of polarity between the periods is not relevant. The main message that we can draw from these results is that media texts on fake news are normally negative, but that the post-election ones are even more. In fact, texts mentioning fake news in politics often involve deception, fraud and accusations. In the graph regarding the Brazilian corpus, the difference of polarity between before and after the 2018 Brazilian general election is almost unnoticeable. Besides that, the

¹⁸ Even though it is not possible to compare polarity values between different languages.

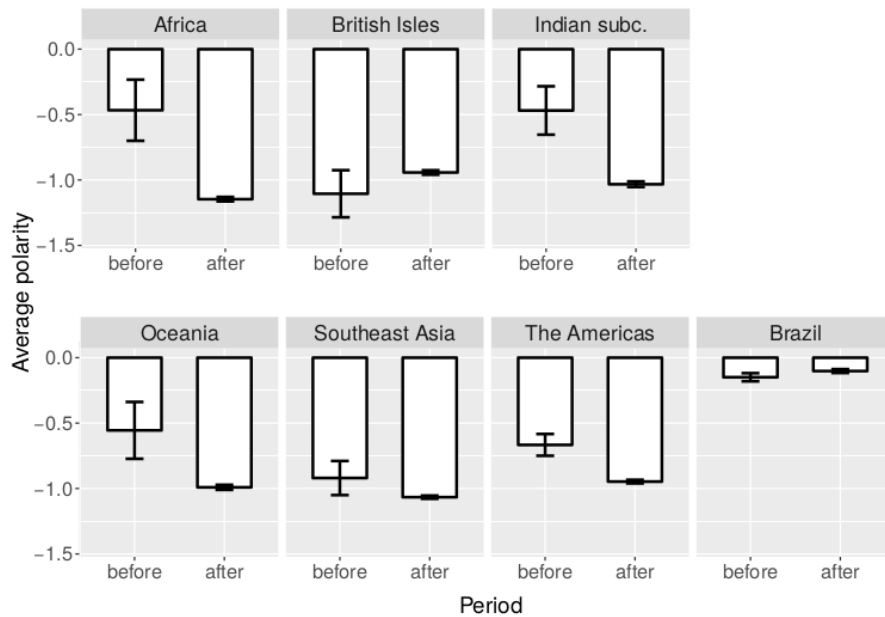


Figure 4.9: Average polarity of the contexts in each region before and after 2016 US election and 2018 BR election. Bars indicate the standard error of the mean.

average polarity in both periods is, albeit also negative, much less negative than in all regions considered in the English-written corpus. The difference between polarities in different languages, however, is not relevant, since sentiment analyses in distinct languages are not completely comparable between themselves.

4.3.7 Summary of results

In this section, we present a series of independent analyses on the texts that accompany the term *fake news* in English-written and Brazilian news. The most relevant outcomes of these analyses can be summarized and integrated as follows:

- globally, the interest for the term *fake news* suddenly in-

creased after the 2016 US election, and in Brazil this interest increased after the 2018 Brazilian election, as indicated by the rise of news mentioning it and of Google Search queries for this expression (Section 4.3.1);

- this growth was accompanied by a change of framing around the term *fake news* – from, for instance, topics regarding the media industry itself to those related to political affairs, both in English and Portuguese (Sections 4.3.1, 4.3.3, 4.3.4, 4.3.5);
- the named entities linked to the expression *fake news* not only changed towards political topics, but also suffered from global standardization after the US election (Section 4.3.2);
- in English, the negativity of the news containing the term *fake news* increased after the US election (Section 4.3.6).

All these results suggest that, as hypothesized in Section 4.1, the rise of public interest in the term *fake news* brought changes to its conceptualization and to the perception about it.

4.4 Concluding remarks

Due to the increased role of the Internet in modern societies, topics regarding misinformation and manipulation in online environments seem to be subject to progressively more public debate and interest, including from the traditional media. Understanding how these topics are viewed through the eyes of opinion leaders is crucial to comprehend how public opinion about them is being shaped in present day.

Here we present a quantitative analysis on the perception and conceptualization of the term *fake news* in two corpora of news articles published from 2009 to 2018 in 21 countries. We investigate how media sources have been reporting topics related to fake news

and whether the rise of the public interest in this very expression during and after the 2016 and 2018 presidential elections in the United States of America and Brazil, respectively, was accompanied by changes of perception and shifts in sentiment about it. We observed changes in the vocabulary and in the mentioned entities around the term *fake news*, in the topics related to this concept and in the polarity of the texts around it, as well as in Web search behavior of Google Search users interested in this concept.

We are also interested in understanding whether the term *fake news* is framed differently across the globe – and, if so, which are these differences. The existence of such variations may result in different shifts in the meanings and in the sentiments around these concepts in various regions of the world, which justifies this study as a way to more clearly understand how the public opinion is being steered in the current context in different countries of the English-speaking world. We understand that, in this way, our study joins the scholarship that “contributes to the nascent debate on the concept of fake news” (Gelfert, 2018, p. 85) both as a linguistic term and, to a lesser extent, as a social phenomenon.

From a journalism studies perspective, our findings come as no surprise: the semantic shift of *fake news* from news satire to political propaganda has already been identified by political scientists and journalism scholars using qualitative methods (see Section 4.1.2). Their findings however, serve to validate the diachronic method employed here. Our study contributes to the literature on the term *fake news* as it is used in news media by adding quantitative data and geographical scope – most of the previous studies focus on the United States of America.

More than just analyzing an isolated case, though, our intention in this chapter is to present an analytical framework that can be applied and replicated in other situations. The basis of our proposal is the diachronic investigation of vocabulary from a “holistic”

view, that is, combining and mixing different approaches in order to understand the phenomenon of semantic change from different perspectives. The same investigations implemented here can be performed in the most diverse contexts, using different items as key terms. As an example, we mention the study carried out by Caetano et al. (2018), which was based on the investigations presented here and used as key term the word *whatsapp*. What is important is that the corpora for analysis have a temporal component, so that different periods can be compared and contrasted with each other, thus suggesting changes (or maintenance) in the conceptualization of that key term over time.

Each of the methods employed here has its own features, advantages and disadvantages. The analysis of the Web search behavior (volume of searches, related queries etc.) regarding a key term (Section 4.3.1) allows the measure of the interest in a particular topic employing users' behavior on search engines as a proxy. The main advantage of this method is its simplicity. However, it is important to remember that search engine users are unlikely to be a statistically adequate sample of a population. Concerning our different approaches to investigate the text co-occurring with a key term, it is interesting to note how each approach complements the others. The analysis of co-occurring named entities (Section 4.3.2) makes it possible to identify people, places and institutions related to the key term. This is part of the issue, but it leaves out non-entities related to the term, which can be covered by the more generic analysis of the semantic fields of the surrounding vocabulary (Section 4.3.3). In addition, the investigation on co-occurrence networks (Section 4.3.4), although less clear from an interpretative point of view, allows the relationships between words to become more evident, and supports different and complementary interpretations that add to the simple analysis of co-occurring word lists. None of these approaches, however, deals with the *topic* of the contexts in

which the key terms are inserted. For this reason, we include topic analysis (Section 4.3.5) – which is, on the one hand, more complex to be performed; but, on the other hand, is able to capture an aspect present at a higher level of analysis than lexical co-occurrence examinations. Finally, the polarity analysis (Section 4.3.6) has the characteristic of being generic and orthogonal to the semantic content of the text, since it can, independently of the topic, evaluate the level of positivity/negativity in the contexts in which the key term is present. It is, thus, the only sentiment-oriented method proposed in our framework, which complements those previously described. The idea of working with all these methods at the same time is precisely to be able to merge their qualities and thus get a broader view of the studied phenomenon.

In this chapter, we analyzed the usage of the term *fake news* in a diachronic perspective, but, for each corpus, only considered two historical moments: before and after a key event in the history of this expression (in English, the 2016 US presidential election; in Brazilian Portuguese, the 2018 Brazilian general election). In the future, we plan to consider a larger spectrum of periods, in order to understand whether (and, if it is the case, when) the conceptualization of *fake news* changed once again. We also intend to add analyses using data from other relevant sources, including Twitter posts and Wikipedia edits, so to observe the use of this term by different actors of the society.

CHAPTER 5

Conclusions

The target of this dissertation, as made explicit by its title, is the computational processing of diachronic linguistic corpora. Therefore, in the previous chapters, I have presented some contributions related to the use of computer power in the field of corpus linguistics, more specifically concerning the processing of diachronic corpora. All of these contributions, although related, are independent¹, and focus on three different stages of the research involving diachronic corpora and their computational processing: (a) corpus building and compilation (on Chapter 2); (b) designing of tools and algorithms for data exploration (on Chapter 3); and (c) data analysis for linguistic, cultural and historical research (on Chapter 4).

For obvious reasons, the contributions presented here are not intended to embrace all stages of corpus linguistics research. For example, I do not address the subject of corpus digitisation – a

¹ These independent contributions are based on a set of five papers published or, at the time of the writing of this text, accepted/submitted for publication (see Appendix C).

task of fundamental importance for the compilation of most historical corpora, especially those based on ancient texts. In this context, one of the steps that has benefited most from computational advances is the task of *optical character recognition* (OCR), that is, “the electronic translation of images of handwritten, typewritten or printed text (usually captured by a scanner) into machine-editable text” (Halabi et al., 2009, p. 24). In the scenario of historical corpus compilation, OCR tools are particularly useful not only for the conversion of old printed texts (gathered from books, newspapers, magazines etc.) into searchable text, but also for *manuscript text recognition* (MTR) (García-Calderón et al., 2018).

Another major topic related to the computational processing of linguistic corpora that was not thoroughly covered in this dissertation is corpus annotation. In the words of Leech (2005), “[c]orpus annotation is the practice of adding interpretative linguistic information to a corpus”. Leech adds that “annotation is a means to make a corpus much more useful – an enrichment of the original **raw corpus**. From this perspective, (...) adding annotation to a corpus is giving ‘added value’ ” (emphasis in original), especially when thinking of possibilities for the automatic analysis of this corpus. One common type of annotation is the addition of part of speech (PoS) tags indicating the word class to which words in the corpus belong². Other types of annotation include semantic annotation (e.g. adding information about the semantic category of words), discourse annotation (e.g. adding information about anaphoric links in a text), stylistic annotation (e.g. adding information about speech and thought presentation), lexical annotation (e.g. adding the identity of the lemma of each word form in a text) – “[i]n fact, it is possible to think up untold kinds of annotation that might be useful for specific kinds of research” (Leech, 2005). Depending on the corpus

² See Chapter 3, in which part of speech tags provided by the Corpus of Historical American English (COHA) are employed in the case studies presented.

and on the type of annotation, they can be manually or automatically added. In Chapter 2, we briefly mention our interest in adding annotations to future versions of our corpus of news comments.

The use of computational approaches in corpus linguistics has been consolidated for decades, and the present work should be regarded as a small set of general contributions to the field. It should be of interest both to linguists concerned with the analysis of corpora and to those more implicated with the development of tools and techniques. It presents several case studies to illustrate the usefulness of the methods and resources introduced here, as well as research ideas to promote and facilitate future investigations employing our contributions. In the next paragraphs, I briefly review each of these contributions before turning to the general conclusions of this dissertation.

5.1 Summary of the dissertation

After an introduction to the research topic and a concise overview of the relationship between linguistics and computer science (in Chapter 1), I report (in Chapter 2) the designing, building and compilation of a Web scraper and a diachronic corpus of comments extracted from news websites. In this chapter, we discuss the significance of the text genre *comment in news portal* within the context of Internet linguistics (Crystal, 2011) and justify the need of developing tools to assist researchers with limited programming knowledge in the task of data collection. As a result, we offer the community an open source and free for use, modification and distribution Web scraper, which is available both for online use and for download. We show that our Web scraper is simple to operate and can be used even by individuals with limited computational skills, which makes it an attractive tool for those interested in conducting research on news portals comments. We also freely provide a cor-

pus composed of more than 200,000 comments published at UOL, a major Brazilian news portal. Our corpus contains not only the comments themselves, but also important meta-information such as dates and times of publication of the comments, commentators' usernames, numbers of likes received by the comments, and information (date and title) of the news stories where these comments were posted. This corpus makes it possible to analyze linguistic, textual, and discursive characteristics of the genre of news comments, and some ideas for future research projects that could be carried out using it are listed in the chapter.

The work presented in Chapter 3 concerns the development of a method for the exploration of diachronic corpora. As stated by Hilpert and Gries (2016), in these “early days for diachronic corpus linguistics” (p. 52), there is a need for the enlargement of the field's toolkit. We offer the description of a simple and generalizable algorithm to assist in the identification of the periods of establishment and obsolescence of linguistic items in any diachronic corpus divided into time frames. Our goal is to provide a method that helps the automatic discovery of trends and patterns in language dynamics. The proposed algorithm uses information on the frequency of items in each time frame of the corpus, and may be employed for the analysis of any collection of linguistic items, regardless of language or historical period. We demonstrate the applicability of this method by supplying case studies on the statistics and characteristics of words that appear in or disappear from the Corpus of Historical American English (COHA) in different periods. Among our results, we highlight findings that concern the proportion of established words among all words across decades, as well as variations in the proportions of different parts of speech over the past two centuries. We also use our algorithm to identify words that became established in different decades and are still frequent, those that were previously frequent but became obsolete, and short-lived

items. In our view, these case studies provide new insights to the field of quantitative diachronic linguistics and to the study of the American English lexicon, and might motivate future studies using the algorithm presented.

Finally, in Chapter 4, we provide an illustration of how computationally-driven analyses performed on diachronic linguistic data are able to reveal changes in the semantic framing of a given expression – in this case, the term *fake news*, that gained popular attention particularly during and after the presidential elections of 2016 and 2018 in, respectively, the United States of America and Brazil. We investigate the lexicon around the expression *fake news* in two diachronic corpora of news articles using a set of quantitative methods and, as a result, we get a picture of how this expression underwent a change in perception and conceptualization after 2016 (in English) and 2018 (in Brazilian Portuguese), helping to comprehend and more accurately characterize this relevant social phenomenon linked to misinformation and manipulation. Our results show changes in the contexts that surround the term *fake news* when the periods before and after the elections are compared. Nevertheless, our major goal is not only to analyze an isolated case, but rather to present a framework of analysis that can be applied to other contexts. This framework is centered on the investigation of vocabulary through a diachronic approach, and employs complementary methods which enable the comprehension of a lexical item’s semantic change from different angles.

5.2 Major contributions

In summary, the following major contributions can be drawn from this dissertation:

- an open source and free Web scraper of comments posted on

news websites, available both for download and for online use (Chapter 2);

- a diachronic corpus containing more than 200,000 comments (plus meta-information) collected from a major Brazilian news portal (Chapter 2);
- a simple method to assist in the identification of specific time periods of establishment and obsolescence of linguistic items in diachronic corpora (Chapter 3);
- a series of case studies based on real data concerning two centuries of the dynamics of the American English lexicon (Chapter 3);
- a framework to study diachronic changes in the conceptualization of a term through replicable computational and quantitative methods (Chapter 4);
- a series of observations regarding changes in the conceptualization of the expression *fake news* in English-written and Brazilian news articles (Chapter 4).

Although the chapters of this dissertation are independent (since they result from different research projects), my intention is to integrate them in the future. For example, it would be interesting to apply the method introduced in Chapter 3 to the corpus presented in Chapter 2, so as to find words that got established and obsolete in Brazilian news comments during the time span of the corpus Xereta. Of course, the objectives of an analysis in a three-year corpus are different from those of an analysis in a 200-year one (like the Corpus of Historical American English – COHA). The goal could be the study of Internet neologisms, which have the characteristic of rapidly emerging and disappearing. Then, specific words taken from this corpus could be selected as key terms to be analyzed according

to the framework proposed in Chapter 4, considering the comments in which they appear as their contexts. Just as an illustration, Figure 5.1 shows the distribution of the terms *coxinha* and *petralha* in the corpus Xereta across time. These are both derogatory terms denoting, respectively, conservative individuals and supporters of the Workers' Party³, and spread widely over the Internet during the impeachment process of (or coup d'état against) Dilma Rousseff, in 2016. We can observe that both terms became less and less frequent in the corpus over time, maybe heading towards obsolescence. In the future, we could use the methods presented in Chapter 4 to investigate whether the conceptualization of these terms in Brazilian news portals comments changed over time⁴.

Naturally, as in all research, this work is not without limitations. Most of them are mentioned in their corresponding chapters, but it is worthwhile recalling the major ones. First, our Web scraper, at the time of writing of this dissertation, is only able to extract comments from two news portals (Folha de S. Paulo and UOL), since it depends on the development of a specific module for each website to be collected. We expect to solve this limitation little by little, hopefully counting on external contributors to develop new modules to integrate the scraper. Second, the version of our corpus

³ “A coxinha is literally a type of fried snack, generally filled with chicken, that is common in cafes and bars across Brazil (...). It's difficult to know for sure how this type of person [i.e., conservatives] became associated with the snack, but some (both specialists and laymen) hypothesize that it comes from policemen's association with the snack”; and “[p]etralha is a combination of petista, a supporter of the PT [i.e., the Workers' Party], with metralha, the Brazilian word for the Beagles brothers (Irmãos Metralha), who would continuously attempt to rob Scrooge McDuck in the Disney cartoons” (Freire et al., 2017).

⁴ To perform this analysis, however, additional steps concerning the balance of the corpus should be taken. For example, it is possible that the number of political-related news in the early part of the corpus is higher than in the later one. Further analyses of the phenomenon shall solve this and other issues.



Figure 5.1: Distribution of the terms *cozinha* (above) and *petralha* (below) in the corpus Xereta across time. Each vertical bar represents one occurrence of the term in the corpus.

of news comments presented here is somewhat temporally imbalanced, an issue that we intend to address in future releases of the corpus. Regarding the algorithm presented in Chapter 3, its major limitation is probably linked to its major advantage: the simplicity of the method, which is an asset for facilitating initial extractions of lists of candidate items for further research – but, in some cases, the generic framework of the algorithm has to be supplemented with more fine-grained analyses of frequencies representative of the corpus at hand, especially when dealing with smaller corpora. Finally, we acknowledge that some of the methods employed in Chapter 4 still need to be better consolidated, especially when it regards the visualization and interpretation of the results.

In any case, I believe that this dissertation has achieved its main goal of offering insights into three of the multiple stages of the research involving diachronic linguistic corpora and hope that it has contributed to advance the knowledge on the use of computer power in corpus-assisted linguistics.

APPENDIX A

Lexical heritage from past decades (data)

In Section 3.4.3, we mention the most common words in decade 2000s that, according to the algorithm proposed, were established in the Corpus of Historical American English (COHA) in particular previous decades. In this appendix, we display the lists containing these words (+ PoS tags), which are ordered according to their frequency in decade 2000s.

COHA employs PoS tags from the UCREL CLAWS7 Tagset. The complete description of these tags, including examples, is available at <http://ucrel.lancs.ac.uk/claws7tags.html>. The tags that occur in this table and in the other tables at Chapter 3 are:

- cc**: coordinating conjunction;
- cs**: subordinating conjunction;
- ii32**: general preposition (as part of a sequence);
- jj**: general adjective;
- jjr**: general comparative adjective;
- mc**: cardinal number, neutral for number;

mc2: plural cardinal number;
nn: common noun, neutral for number;
nn1: singular common noun;
nn122: singular common noun (as part of a sequence);
nn2: plural common noun;
nna: following noun of title;
nno: noun, neutral for number;
nnt1: temporal noun, singular;
nnt2: temporal noun, plural;
nnu: unit of measurement, neutral for number;
nnu2: plural unit of measurement;
ra: adverb, after nominal head;
rl: locative adverb;
rr: general adverb;
rt: quasi-nominal adverb of time;
to: infinitive marker;
uh: interjection;
vv0: base form of lexical verb;
vvd: past tense of lexical verb;
vvg: -ing participle of lexical verb;
vvgk: -ing participle catenative;
vvi: infinitive;
vvn: past participle of lexical verb;
vvz: -s form of lexical verb.

1850s	1860s	1870s	1880s
scientists_nn2	photo_nn1	phone_nn1	radio_nn1
experts_nn2	baseball_nn1	programs_nn2	skills_nn2
bike_nn1	photograph_vv0	photos_nn2	researchers_nn2
regional_jj	cigarette_nn1	hallway_nn1	golf_nn1
detective_nn1	options_nn2	long-term_jj	parking_nn1
focused_vvn	cops_nn2	makeup_nn1	ceo_nn1
tablespoons_nn2	typically_rr	downtown_jj	soviet_jj
strategies_nn2	protein_nn1	focus_vvi	networks_nn2
terrorists_nn2	telephone_nn1	classroom_nn1	techniques_nn2
users_nn2	concept_nn1	diabetes_nn1	ratings_nn2
shorts_nn2	dna_nn1	bacteria_nn2	nonetheless_rr
scientist_nn1	genetic_jj	driveway_nn1	korean_jj
technique_nn1	ethnic_jj	racial_jj	consultant_nn1
ongoing_jj	grabs_vvz	immune_jj	hockey_nn1
strategic_jj	backyard_nn1	colorful_jj	viewers_nn2
grill_nn1	cigarettes_nn2	taxi_nn1	olympics_nn2
aluminum_nn1	cd_nn1	scenario_nn1	residential_jj
parked_vvn	concepts_nn2	cultures_nn2	toxic_jj
fake_jj	spectacular_jj	palestinian_jj	subway_nn1
downtown_rl	terrain_nn1	diagnosed_vvn	meaningful_jj
focusing_vvg	focused_vvd	pinta_nn1	sneakers_nn2
photographer_nn1	specialist_nn1	gasoline_nn1	predictable_jj
treatments_nn2	shotgun_nn1	adolescents_nn2	championships_nn2
trailer_nn1	sensed_vvd	evaluation_nn1	binoculars_nn2
bicycle_nn1	worldwide_rl	victorian_jj	foyer_nn1
flipped_vvd	fingertips_nn2	semester_nn1	asset_nn1
gestured_vvd	aging_jj	bartender_nn1	dioxide_nn1
canyon_nn1	underwear_nn1	halloween_nnt1	initiatives_nn2
biology_nn1	clarity_nn1	cardboard_jj	coaching_nn1
ambulance_nn1	optimistic_jj	collaboration_nn1	heck_nn1
variables_nn2	technological_jj	headlights_nn2	orientation_nn1
institutional_jj	emotionally_rr	fingernails_nn2	overseas_rl
slowed_vvd	yanked_vvd	housing_vvg	focuses_vvz
buses_nn2	specialists_nn2	starters_nn2	kilometers_nnu2
providers_nn2	obsession_nn1	neon_nn1	interactive_jj
productivity_nn1	starter_nn1	ramp_nn1	penis_nn1
businessman_nn1	headlines_nn2	output_nn1	unpredictable_jj
interactions_nn2	lethal_jj	interviewed_vvn	technician_nn1
cosmic_jj	zoo_nn1	overweight_jj	touchdown_nn1
overly_rr	raiders_nn2	plasma_nn1	seasonal_jj
vibrant_jj	motors_nn2	shortage_nn1	shack_nn1
ironic_jj	peanut_nn1	microphone_nn1	capitalism_nn1
livestock_nn	innings_nn	format_nn1	hmm_uh
yep_uh	broccoli_nn1	developers_nn2	arthritis_nn1
heartbeat_nn1	awesome_jj	investor_nn1	sweetie_nn1
armored_jj	fictional_jj	outcomes_nn2	entrepreneur_nn1
detectives_nn2	aftermath_nn1	touch_ii32	abs_jj
toddler_nn1	canned_jj	finals_nn2	scoring_nn1
erie_jj	evolutionary_jj	developer_nn1	backdrop_nn1
inning_nn1	vulnerability_nn1	biologist_nn1	comeback_nn1

1890s	1900s	1910s	1920s
movie_nn1	computer_nn1	gon_vvgk	okay_rr
na_to	global_jj	electronic_jj	video_nn1
television_nn1	shit_nn1	aids_nn1	iraqi_jj
environmental_jj	movies_nn2	servings_nn2	airport_nn1
nuclear_jj	weekend_nnt1	sox_nn2	boyfriend_nn1
basketball_nn1	soccer_nn1	pickup_nn1	sexy_jj
garage_nn1	calories_nnu2	agenda_nn1	robot_nn1
ta_to	jazz_nn1	helicopter_nn1	cholesterol_nn1
overall_jj	coverage_nn1	fuck_nn1	workout_nn1
therapy_nn1	genes_nn2	kidding_vvg	bikes_nn2
terrorist_jj	muslim_jj	featuring_vvg	airlines_nn2
clinic_nn1	aircraft_nn	lipstick_nn1	antibiotics_nn2
mommy_nn1	someday_rt	goddamn_jj	and/or_cc
wireless_nn1	quarterback_nn1	teammates_nn2	nonprofit_jj
gym_nn1	technologies_nn2	windshield_nn1	vitamin_nn1
islamic_jj	pm_ra	someplace_rl	hometown_nn1
phones_nn2	rookie_nn1	lineup_nn1	activist_nn1
basically_rr	emissions_nn2	parked_vvd	airline_nn1
scheduled_vvn	suitcase_nn1	nationwide_rl	briefcase_nn1
awareness_nn1	buddy_nn1	part-time_jj	playoffs_nn2
initially_rr	activists_nn2	cafeteria_nn1	wheelchair_nn1
analysts_nn2	hispanic_jj	backseat_nn1	yoga_nn1
tablespoon_nn1	muslims_nn2	prestigious_jj	allegedly_rr
sector_nn1	regulatory_jj	helicopters_nn2	coordinator_nn1
sweater_nn1	expertise_nn1	dick_nn1	footage_nn1
coastal_jj	airplane_nn1	motivated_vvn	insulin_nn1
full-time_jj	jeep_nn1	vitamins_nn2	columnist_nn1
locals_nn2	routinely_rr	sponsored_vvn	creativity_nn1
homework_nn1	diesel_nn1	podium_nn1	feedback_nn1
flashlight_nn1	skiing_nn1	limo_nn1	processing_nn1
weekends_nnt2	grid_nn1	overseas_jj	c'm_vv0
homeland_nn1	researcher_nn1	recordings_nn2	ecological_jj
ok_jj	minimal_jj	airplanes_nn2	hormone_nn1
behaviors_nn2	feminist_jj	consultants_nn2	onstage_jj
hiking_vvg	robots_nn2	postwar_jj	firefighters_nn2
sexually_rr	syndrome_nn1	planners_nn2	pipeline_nn1
bombing_nn1	carbohydrate_nn1	viruses_nn2	implementation_nn1
deadline_nn1	priorities_nn2	lightweight_jj	highlights_nn2
motel_nn1	all-star_jj	protesters_nn2	nazi_jj
motivation_nn1	coconut_nn1	touchdowns_nn2	audio_jj
unemployment_nn1	prostate_nn1	flips_vvz	airports_nn2
hike_nn1	nutrients_nn2	artwork_nn1	operational_jj
trauma_nn1	buddies_nn2	campuses_nn2	slacks_nn2
catalog_nn1	containers_nn2	cleanup_nn1	stairwell_nn1
motorcycle_nn1	entrepreneurs_nn2	paranoid_jj	nylon_nn1
kinda_rr	trillion_nno	highlight_nn1	input_nn1
payroll_nn1	hormones_nn2	breathhtaking_jj	aerobic_jj
spotlight_nn1	artifacts_nn2	aspirin_nn	teammate_nn1
vodka_nn1	featured_vvd	small-town_jj	receptionist_nn1
short-term_jj	psychiatrist_nn1	comics_nn2	workouts_nn2

1930s	1940s	1950s
okay_jj	sidebar_nn1	backpack_nn1
computers_nn2	girlfriend_nn1	infrastructure_nn1
pizza_nn1	t-shirt_nn1	fuckin_rr
fucking_jj	online_rr	ncaa_nn1
israeli_jj	pc_nn1	backup_nn1
wildlife_nn1	teenage_jj	freeway_nn1
fuck_vv0	teenager_nn1	t-shirts_nn2
nba_nn1	teenagers_nn2	ponytail_nn1
sunglasses_nn2	mainstream_jj	girlfriends_nn2
guidelines_nn2	radar_nn1	salsa_nn1
siblings_nn2	upcoming_jj	linebacker_nn1
reportedly_rr	supermarket_nn1	spokeswoman_nn1
laser_nn1	innovative_jj	sunni_nn1
playoff_nn1	asshole_nn1	steroids_nn2
electronics_nn1	workplace_nn1	excerpted_vvd
therapist_nn1	microwave_nn1	pakistani_jj
bullshit_nn1	palestinians_nn2	robotic_jj
racism_nn1	fridge_nn1	pesticides_nn2
mph_nnu	stereo_nn1	superstar_nn1
paperwork_nn1	supportive_jj	hosting_vvg
processor_nn1	desktop_nn1	interface_nn1
fucking_rr	postseason_nn1	parameters_nn2
programming_nn1	sensors_nn2	spacecraft_nn
labs_nn2	monitor_vvi	recycling_nn1
tourism_nn1	israelis_nn2	award-winning_jj
cds_nn2	iraqis_nn1	offseason_nn1
medications_nn2	basics_nn2	surfing_vvg
gop_nn1	breakthrough_nn1	venues_nn2
demographic_jj	fda_nn1	nightstand_nn1
seafood_nn1	irs_nn1	environmentally_rr
condo_nn1	estrogen_nn1	automated_jj
rearview_nn1	predators_nn2	high-risk_jj
saudi_jj	vietnamese_jj	antioxidants_nn2
integrator_nn1	monitoring_vvg	charismatic_jj
behavioral_jj	bikini_nn1	oregano_nn1
low-income_jj	monitoring_nn1	nascar_nn1
adrenaline_nn1	airborne_jj	life-threatening_jj
bam_vv0	cardiovascular_jj	on-site_jj
condom_nn1	antibiotic_nn1	geek_nn1
hp_nn1	graffiti_nn	surreal_jj
homeowners_nn2	staffers_nn2	upbeat_jj
commercials_nn2	health-care_nn1	module_nn1
vinyl_nn1	late-night_jj	weaponry_nn1
plywood_nn1	thai_jj	chiles_nn2
dj_nn1	viewer_nn1	pollutants_nn2
marijuana_nn1	burgers_nn2	addictive_jj
implemented_vvn	predator_nn1	aerospace_nn1
boutique_nn1	yogurt_nn1	clueless_jj
skiers_nn2	targeting_vvg	retro_jj
condoms_nn2	paperback_nn1	multimedia_nn

1960s	1970s	1980s
affordable_jj	online_jj	headnote_nn1
mets_nn2	lifestyle_nn1	high-tech_jj
sustainable_jj	african-american_jj	laptop_nn1
ecosystems_nn2	suv_nn1	videos_nn2
upscale_nn1	parenting_nn1	globalization_nn1
medium-high_jj	genome_nn1	database_nn1
activism_nn1	high-profile_jj	booker_nn1
arguably_rr	workforce_nn1	pcs_nn2
healthcare_nn1	preheat_vv0	cilantro_nn1
mantra_nn1	ceos_nn2	african-americans_nn2
ecosystem_nn1	nonstick_nn1	hip-hop_jj
rehab_nn1	sitcom_nn1	high-end_jj
trendy_jj	jihad_nn1	ppg_nnu
wetlands_nn2	tsp_nnu	sustainability_nn1
lasers_nn2	dumpster_nn1	phd_nna
filmmakers_nn2	recycled_jj	condos_nn2
filmmaker_nn1	networking_nn1	counterterrorism_nn1
autism_nn1	pricey_jj	rapper_nn1
hosted_vvd	carb_nn1	databases_nn2
videotape_nn1	handheld_jj	gdp_nn1
disco_nn1	world-class_jj	biotech_nn1
gays_nn2	tofu_nn1	state-of-the-art_jj
makeover_vv0	sunscreen_nn1	catwoman_nn1
multicultural_jj	deregulation_nn1	mid-level_jj
broadband_jj	fast-food_jj	same-sex_jj
marina_nn1	gurney_nn1	buyout_nn1
attendees_nn2	mid-1990s_mc2	algorithms_nn2
hosted_vvn	ethnicity_nn1	pesto_nn1
boutiques_nn2	updates_nn2	biotechnology_nn1
one-on-one_mc	cardio_nn1	preservice_nn1
entitlement_nn1	shiites_nn2	glynnis_nn1
benchmark_nn1	countertop_nn1	download_vvi
scheer_vv0	sushi_nn2	mentoring_vvg
hologram_nn1	magisterium_nn1	tugger_nn1
fucked_vvn	transnational_jj	laptops_nn2
debuted_vvd	cloning_nn1	stakeholders_nn2
real-time_jj	groundwater_nn1	sweatpants_nn2
dismissive_jj	spokesperson_nn1	wastewater_nn1
fucked_vvd	trailhead_nn1	starbucks_nn2
restructuring_nn1	antidepressants_nn2	wetland_nn1
on-line_jj	hispanics_nn2	minivan_nn1
postmodern_jj	meds_nn2	pager_nn1
real-world_jj	bikers_nn2	gawain_nn1
cuz_nn1	mid-1980s_nn2	basher_nn1
lasagna_nn1	biking_vvg	creationism_nn1
simulations_nn2	recycling_vvg	feel-good_jj
deco_nn122	veggies_nn2	firewall_nn1
cornerback_nn1	midlife_nn1	ipo_nn1
surfers_nn2	superhero_nn1	camcorder_nn1
batmobile_nn1	lifestyles_nn2	downtime_nnt1

APPENDIX B

Empath categories

In Section 4.3.3, we mention that we used lists of words provided by the tool Empath in order to classify the lemmatized words according to categories that represent different semantic fields. We also say that, due to the high number of categories predefined by Empath (194 in total), we selected eight that showed interesting results and are relevant for our discussion: *government*, *internet*, *journalism*, *leader*, *negative emotion*, *politics*, *social media* and *technology*. In this appendix, we list the words that comprise each category used in this dissertation.

Government (113 words): abolish, accordance, accounting, activist, administration, administrative, administrator, advisor, agency, allied, ambassador, amendment, applicable, association, authorities, campaign, capitalist, citizenship, civil, civilian, coalition, commission, communist, congress, consensus, conservation, constitution, contribution, cooperation, cultural, decree, democracy, democratic, dictatorship, economy, elect, election, embassy, empire,

employ, enforce, enforcement, ethical, facility, federal, finance, financial, fund, funding, global, govern, government, homeland, illegally, immigration, implementation, independence, international, involvement, jurisdiction, law, liability, mandate, monetary, nation, negotiate, negotiation, obligation, opposition, organization, organization, parliament, partnership, petition, policy, politician, politics, populace, poverty, program, propaganda, protocol, province, recruiting, regime, regional, regulation, representative, republic, republican, resource, revolution, rule, ruling, safeguard, sector, senate, senator, socialist, societal, society, tax, terrorism, treason, treasury, tyranny, unethical, unified, united, unjust, unlawful, utopian, welfare.

Internet (79 words): access, account, application, archive, article, broadcast, browser, celebrity, chat, chrome, click, coding, compute, computer, cursor, cyber, data, dating, document, download, edit, editing, editorial, edits, email, explorer, facebook, firewall, forum, gamer, gaming, glitch, google, hack, hacker, hacking, header, homepage, icon, info, information, interface, internet, keyword, laptop, link, mobile, multiplayer, networking, online, page, post, posted, profile, program, programming, publish, reporting, research, router, scam, screen, search, segment, server, site, spam, spreadsheet, subscriber, surf, tab, trend, updated, virtual, web, website, wireless, worldwide, www.

Journalism (69 words): administration, analysis, article, assignment, biography, booklet, bookstore, bulletin, campaign, cite, column, composition, copy, corporate, credential, documentary, edition, editor, editorial, essay, excerpt, freelance, gazette, guidebook, handbook, headline, history, homepage, informational, inked, jot, journal, journalism, journalist, leaflet, literary, literature, magazine, newsletter, newspaper, notebook, obituary, outdated, paper, paperwork, paragraph, photo, photographer, photography, print, printed,

printing, printout, publicist, publish, publishing, report, reporter, research, scribbling, showbiz, site, subscription, summarize, tabloid, thesis, typewriter, write, writing.

Leader (98 words): administrative, administrator, admiral, adviser, advisor, agent, ambassador, assist, association, authority, battalion, bishop, boss, briefing, campaign, candidate, captain, chairman, chancellor, chief, command, commandant, commander, commanding, committee, competent, corporate, corporation, council, crew, dictatorship, diplomat, diplomatic, directive, director, duty, elect, election, elite, emperor, employer, enforcer, executive, father, federation, founder, general, government, governor, headquarters, hotshot, in-command, informed, join, leader, leadership, leading, member, mission, negotiator, obey, operation, order, organization, organization, oversee, overthrow, parliament, peacemaker, politician, president, promote, qualified, rank, ranking, rebel, renowned, representative, responsibility, rule, ruler, ruling, senate, senator, sergeant, sir, specifically, strategist, strongest, successor, superior, supervise, supervisor, supreme, syndicate, treasurer, trusted, veteran.

Negative emotion (94 words): accident, afraid, alone, angered, angry, ashamed, bad, badly, beat, beaten, blame, break, care, confused, crazy, crushed, cry, crying, dead, death, depressed, die, dieing, disappointed, drunk, dying, either, fault, fight, fighting, freaked, frightened, fucking, furious, guilty, hard, hate, hated, heartbroken, hell, hit, horrible, hurt, hurting, hurts, insane, kill, killed, killing, last straw, lie, loose, lose, losing, lost, mad, mean, monster, pain, pissed, poor girl, poor guy, punch, raped, react, reason, sad, scare, scared, scary, seeing, shocked, sick, so much pain, stop, stupid, surprised, swear, terrible, terrified, thinking, threatened, tortured, trouble, unthinkable, upset, violent, wanted, weak, worried, worse, worst, worst part, wrong.

Politics (78 words): activist, advisor, advocate, ambassador, amendment, aristocracy, campaign, candidate, citizen, citizenship, committee, communist, community, congress, consensus, conservative, conspiracy, constitution, controversial, controversy, corruption, council, declaration, decree, delegate, democracy, democratic, dictatorship, diplomacy, diplomat, dispute, divided, division, doctrine, elect, election, extremist, federation, fundraising, govern, governor, ideology, influential, involvement, jurisdiction, leadership, liberal, liberation, loyalist, monarch, monarchy, nation, national, nationwide, negotiation, overthrow, parliament, philosophy, policy, politically, politician, politics, presidential, province, provincial, regime, representative, republican, revolutionary, ruling, senate, senator, socialist, society, sovereign, spokesperson, tyranny, unify.

Social media (51 words): bio, browser, camera, chat, comment, computer, contact, delete, email, facebook, fan, follow, followed, follower, following, follows, format, forum, friends, hashtag, homepage, icon, instagram, interaction, message, messaging, messenger, multimedia, network, notification, online, onscreen, password, post, posting, profile, site, snapshot, status, tablet, texts, trend, tweet, twitter, typed, update, video, vine, viral, web, www.

Technology (118 words): 3d, advanced, android, audio, automate, battery, binary, browser, cable, camcorder, cellphone, cellular, cloning, coding, communication, compute, computer, computerized, connector, console, cyber, data, database, desktop, developer, device, digital, download, electronics, engineer, engineering, experimental, firewall, format, futuristic, gadget, gaming, generator, glitch, grid, hack, hacker, hacking, handheld, innovative, install, integrate, interactive, interface, intergalactic, internet, invention, inventor, ipad, keyboard, laboratory, laptop, machinery, mainframe, malfunction, manufacture, messaging, microchip, mo-

bile, module, monitoring, multiplayer, navigation, network, nexus, nuclear, online, operational, optical, outdated, portable, powered, processor, program, programmer, programming, projector, prototype, quantum, radar, research, researcher, robot, router, satellite, scanner, scanning, scientific, scientist, screen, sensor, server, simulator, site, software, solar, spacecraft, spaceship, system, tablet, tech, technical, technician, technological, technology, transmitter, typing, upgrade, virtual, virus, web, website, wireless.

APPENDIX C

Papers and presentations

Papers included in this dissertation

The research presented in this dissertation comprises the following five papers – all published, accepted or submitted for publication at the time of the dissertation defense –, which are reproduced here with minor changes and with consent from co-authors.

Chapter 2: Building a diachronic corpus of comments extracted from news portals and websites

Paper I:

Cunha, E. L. T. P., Magno, G., and Almeida, V. (2017). A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias. In *Anais do X Congresso Internacional da ABRALIN*, pages 764–771. Associação Brasileira de Linguística (ABRALIN).

Paper II:

Cunha, E. L. T. P., Magno, G., and Almeida, V. (under review). *Xereta: A Brazilian corpus of online news comments*. Manuscript submitted for publication.

Chapter 3: Establishment and obsolescence of linguistic items in a diachronic corpus

Paper III:

Cunha, E. L. T. P. and Wichmann, S. (in press). An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus. *Corpora*.

Chapter 4: Diachronic corpora and quantitative approaches to the lexicon: the case of the term *fake news*

Paper IV:

Cunha, E. L. T. P., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (2018). Fake news as we feel it: Perception and conceptualization of the term “fake news” in the media. In Staab, S., Koltsova, O., and Ignatov, D., editors, *Social informatics*, volume 11185 of *Lecture Notes in Computer Science*, pages 151–166. Springer, Cham.

Paper V:

Cunha, E. L. T. P., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (under review). *Quantifying the conceptualization of the term “fake news” in Brazilian and English-speaking media sources*. Manuscript submitted for publication.

Papers not included in this dissertation

Besides the previously mentioned papers, the author of this dissertation contributed to the writing of the following articles published, accepted or submitted for publication during his PhD studies:

Cunha, E. L. T. P. and Cambraia, C. N. (in press). A linguagem de uma fraude: Análise das confissões forjadas em nome dos irmãos Naves. *Fórum Linguístico*.

Cunha, E. L. T. P. and Lourenço, L. “Letteratura di immigrati”: Composições poéticas publicadas na imprensa italiana belo-horizontina no início do século XX. In *Revista da Imigração Italiana em Minas Gerais*, 2019.

Caetano, J. A., Magno, G., **Cunha, E. L. T. P.**, Meira Jr., W., Marques-Neto, H. T., and Almeida, V. (2018). Characterizing the public perception of WhatsApp through the lens of media. In *Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media (RDSM 2018)*.

Otoni, R., **Cunha, E. L. T. P.**, Magno, G., Bernardina, P., Meira Jr., W., and Almeida, V. (2018). Analyzing right-wing YouTube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM Conference on Web Science (WebSci’18)*, pages 323–332. Association for Computing Machinery (ACM). [Best student paper award]

Xue, M., Magno, G., **Cunha, E. L. T. P.**, Almeida, V., and Ross, K. W. (2016). The Right to be Forgotten in the media: A data-driven study. In *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2016(4):389–402. [Also presented at the 16th Privacy Enhancing Technologies Symposium (PETS 2016), Darmstadt, Germany, 2016]

Lourenço, L. and **Cunha, E. L. T. P.** (under review). *Toponímia de influência indígena nos bairros de Belo Horizonte*. Manuscript submitted for publication.

Oral and poster presentations in conferences (as presenting author only)

During his PhD studies, the author of this dissertation presented the following communications and posters in conferences:

- *Contribuições para a coleta e a compilação de um corpus de comentários de portais de notícias*. 11th International Congress of the Brazilian Linguistics Association (ABRALIN), Maceió, Brazil, 2019. [Oral communication]
- *Fake news as we feel it: Perception and conceptualization of the term “fake news” in the media*. 10th International Conference on Social Informatics (SocInfo 2018), Saint Petersburg, Russia, 2018. [Oral communication]
- *A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias*. 10th International Congress of the Brazilian Linguistics Association (ABRALIN), Niterói, Brazil, 2017. [Oral communication]
- *Toponímia de influência indígena nos bairros de Belo Horizonte* (with Lorenza Lourenço). 11th International Congress of the Brazilian Linguistics Association (ABRALIN), Maceió, Brazil, 2019. [Poster presentation]

Talks as an invited speaker

Finally, the author of this dissertation was invited to give the following academic talks during his PhD track:

- *Linguística computacional: Uma área entre dois mundos* [Computational linguistics: An area between two worlds]. Federal Institute Southeastern of Minas Gerais (IF Sudeste MG), Barbacena, MG, Brazil, 2018.

- *I dialetti sul Web: Possibilità di preservazione e rivitalizzazione linguistica?* [Dialects on the Web: Possibility of linguistic preservation and revitalization?]. Settimana della Lingua Italiana nel Mondo. Faculty of Letters of the Federal University of Minas Gerais (FALE/UFGM), Belo Horizonte, MG, Brazil, 2018.

- *Linguística forense: O linguista como investigador e perito* [Forensic linguistics: The linguist as investigator and expert witness]. Faculty of Letters of the Federal University of Minas Gerais (FALE/UFGM), Belo Horizonte, MG, Brazil, 2018.

- *Computadores e sociedade: Novos desafios na era digital* [Computers and society: New challenges in the digital era]. Federal Institute Southeastern of Minas Gerais (IF Sudeste MG), Barbacena, MG, Brazil, 2017.

- *Linguística computacional: Problemas, tendências e desafios* [Computational linguistics: Problems, trends and challenges]. Una University Center, Betim, MG, Brazil, 2015.

- *Generalizations regarding 200 years of lexical evolution of English* (with Søren Wichmann). Work in Progress Series, Max Planck Institute for Evolutionary Anthropology (MPI-EVA), Dept. of Linguistics, Leipzig, Germany, 2015.

Bibliography

- Acerbi, A., Lampos, V., Garnett, P., and Bentley, R. A. (2013). The expression of emotions in 20th century books. *PLOS ONE*, 8(3):e59030.
- Algeo, J. and Algeo, A. S., editors (1993). *Fifty years Among the New Words: A dictionary of neologisms, 1941–1991*. Cambridge University Press, Cambridge.
- Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.
- American Dialect Society (2018). “Fake news” is 2017 American Dialect Society word of the year. Retrieved from <https://bit.ly/2FCfbkX> . Accessed on January 4, 2019.
- Amoris, V., Gollner, A. P., Goulart, E., and Pessoni, A. (2012). Marketing político e redes sociais: Reflexos nas eleições 2010 à Presidência da República. In Queiroz, A. C. F., Tomaziello, P. S., and Macedo, R. G., editors, *Comunicação política e eleitoral no Brasil: Perspectivas e limitações no dinamismo político*, pages 140–158. POLITICOM, Americana.

- Arif, N., Al-Jefri, M., Bizzi, I. H., Perano, G. B., Goldman, M., Haq, I., Chua, K. L., Mengozzi, M., Neunez, M., Smith, H., and Ghezzi, P. (2018). Fake news or weak science? Visibility and characterization of anti-vaccine webpages returned by Google in different languages and countries. *Frontiers in Immunology*, 9:1215.
- Arnaudo, D. (2017). Computational propaganda in Brazil: Social bots during elections. *Computational Propaganda Research Project*, 8.
- Ayto, J. (1989). *The Longman register of new words*, volume 1. Longman, Harlow.
- Ayto, J. (1990). *The Longman register of new words*, volume 2. Longman, Harlow.
- Ayto, J. (1999). *Twentieth century words*. Oxford University Press, Oxford.
- Backus, A. (2014). Towards a usage-based account of language change: Implications of contact linguistics for linguistic theory. In Nicolai, R., editor, *Questioning language contact: Limits of contact, contact at its limits*, volume 1 of *Brill Studies in Language Contact and Dynamics of Language*, pages 91–118. Brill, Leiden.
- Baker, P. (2013). Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English. *Gender and Language*, 4(1):125–149.
- Baker, P., Hardie, A., and McEnery, T. (2006). *A glossary of corpus linguistics*. Edinburgh University Press, Edinburgh.
- Barker, E., Paramita, M., Aker, A., Kurtic, E., Hepple, M., and Gaizauskas, R. (2016). The SENSEI Annotated Corpus: Human summaries of reader comment conversations in on-line news. In

Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL 2016), pages 42–52. Association for Computational Linguistics (ACL).

- Baron, N. (1994). Computer languages. In Asher, R. and Simpson, J., editors, *The encyclopedia of language and linguistics*, volume 2, pages 662–670. Pergamon Press, Oxford.
- Bauer, L. (2002). Inferring variation and change from public corpora. In Chambers, J., Trudgill, P., and Schilling-Estes, N., editors, *The handbook of language variation and change*, pages 97–114. Wiley Blackwell, Hoboken.
- BBC News (2018). Fake news: French language body urges alternative phrase. BBC News. Retrieved from <https://www.bbc.com/news/world-europe-45754756> . Accessed on January 4, 2019.
- Berber Sardinha, T. (2013). Variação entre registros da Internet. In Shepherd, T. G. and Saliés, T. G., editors, *Linguística da internet*, pages 55–75. Contexto, São Paulo.
- Berger, G. (2009). How the Internet impacts on international news: Exploring paradoxes of the most global medium in a time of ‘hyperlocalism’. *International Communication Gazette*, 71(5):355–371.
- Biber, D. and Finegan, E. (2001). Diachronic relations among speech-based and written registers in english. In Conrad, S. and Biber, D., editors, *Variation in English: Multi-dimensional studies*, pages 66–83. Routledge, Oxon/New York.
- Biber, D. and Gray, B. (2011). Grammatical change in the noun phrase: The influence of written language use. *English Language and Linguistics*, 15(2):223–250.

- Bird, S., Loper, E., and Klein, E. (2009). *Natural language processing with Python*. O'Reilly Media Inc.
- Bird, S. and Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Blumenthal-Dramé, A. (2012). *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*, volume 83 of *Topics in English Linguistics*. De Gruyter Mouton, Berlin.
- Bochkarev, V., Solovyev, V., and Wichmann, S. (2014). Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface*, 11.
- Bowker, L. and Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. Routledge, London.
- Bradshaw, S. and Howard, P. N. (2018). Challenging truth and trust: A global inventory of organized social media manipulation. *The Computational Propaganda Research Project*.
- Brown, J. D. (1997). Computers in language testing: Present research and some future directions. *Language Learning & Technology*, 1(1):44–59.
- Brummette, J., DiStaso, M., Vafeiadis, M., and Messner, M. (2018). Read all about it: The politicization of “fake news” on Twitter. *Journalism & Mass Communication Quarterly*, 95(2):497–517.
- Buechel, S., Hellrich, J., and Hahn, U. (2016). Feelings from the past – Adapting affective lexicons for historical emotion analysis. In

Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH), pages 54–61.

- Burger, P., Kanhai, S., Pleijter, A., and Verberne, S. (2019). The reach of commercially motivated junk news on Facebook. *PLOS ONE*, 14(8):e0220446.
- Bussmann, H. (1996). *Routledge dictionary of language and linguistics*. Routledge, New York.
- Byrd, R. J., Calzolari, N., Chodorow, M. S., Klavans, J. L., Neff, M. S., and Rizk, O. A. (1987). Tools and methods for computational lexicology. *Computational Linguistics*, 13(3-4):219–240.
- Bârgăoanu, A. and Radu, L. (2018). Fake news or disinformation 2.0? Some insights into Romanians’ digital behaviour. *Romanian Journal of European Affairs*, 18(1):24–38.
- Caetano, J. A., Magno, G., Cunha, E. L. T. P., Meira Jr., W., Marques-Neto, H. T., and Almeida, V. (2018). Characterizing the public perception of WhatsApp through the lens of media. In *Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media (RDSM 2018)*.
- Caetano, J. A., Magno, G., Gonçalves, M. A., Almeida, J., Marques-Neto, H. T., and Almeida, V. (2019). Characterizing attention cascades in WhatsApp groups. In *Proceedings of the 11th International ACM Web Science Conference (WebSci’19)*, pages 27–36. Association for Computing Machinery (ACM).
- Calefato, F., Lanubile, F., Marasciulo, M. C., and Novielli, N. (2015). Mining successful answers in Stack Overflow. In *Proceedings of the 12th Working Conference on Mining Software Repositories*, pages 430–433. Institute of Electrical and Electronics Engineers (IEEE).

- Cambraia, C. N. (2013). Da lexicologia social a uma lexicologia sócio-histórica: Caminhos possíveis. *Revista de Estudos da Linguagem*, 21(1):157–188.
- Cangelosi, A. and Parisi, D., editors (2002). *Simulating the evolution of language*. Springer-Verlag, London.
- Carvalho, P., Sarmiento, L., Teixeira, J., and Silva, M. J. (2011). Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 564–568. Association for Computational Linguistics (ACL).
- Collins Dictionary (2017). Etymology corner – Collins Word of the Year 2017. Retrieved from <https://www.collinsdictionary.com/word-lovers-blog/new/etymology-corner-collins-word-of-the-year-2017,400, HCB.html> . Accessed on December 22, 2018.
- Conroy, N. J., Rubin, V. L., and Chen, Y. (2016). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.
- Corrado, E. M. (2005). The importance of open access, open source, and open standards for libraries. *Issues in Science and Technology Librarianship*, 42.
- Correa, D., Silva, L. A., Mondal, M., Benevenuto, F., and Gumadi, K. P. (2015). The many shades of anonymity: Characterizing anonymous social media content. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (ICWSM-15)*, pages 71–80. Association for the Advancement of Artificial Intelligence (AAAI).

- Cotterell, R., Renduchintala, A., Saphra, N., and Callison-Burch, C. (2014). An Algerian Arabic-French code-switched corpus. In *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools*, pages 34–37.
- Croft, W. (2000). *Explaining language change: An evolutionary approach*. Pearson Longman, Harlow.
- Crystal, D. (2004). *The language revolution*. Polity Press, Cambridge/Malden.
- Crystal, D. (2010). *The Cambridge encyclopedia of language*. Cambridge University Press, Cambridge, 3rd edition.
- Crystal, D. (2011). *Internet linguistics: A student guide*. Routledge, Abingdon.
- Culpeper, J. (2014). Keywords and characterization: An analysis of six characters in *Romeo and Juliet*. In Hoover, D. L., Culpeper, J., and O'Halloran, K., editors, *Digital literary studies: Corpus approaches to poetry, prose, and drama*, pages 9–34. Routledge.
- Cunha, E. L. T. P. (2012). Etiquetação de micromensagens no Twitter: Uma abordagem linguística. Master's thesis, Universidade Federal de Minas Gerais (UFMG).
- Cunha, E. L. T. P., Magno, G., and Almeida, V. (2017). A elaboração de um coletor e de um corpus de comentários extraídos de portais de notícias. In *Anais do X Congresso Internacional da ABRALIN*, pages 764–771. Associação Brasileira de Linguística (ABRALIN).
- Cunha, E. L. T. P., Magno, G., and Almeida, V. (under review). *Xereta: A Brazilian corpus of online news comments*. Manuscript submitted for publication.

- Cunha, E. L. T. P., Magno, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2012). A gender based study of tagging behavior in Twitter. In *Proceedings of the 23rd ACM Conference on Hypertext and Social Media (HT'12)*, pages 323–324. Association for Computing Machinery (ACM).
- Cunha, E. L. T. P., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (2018). Fake news as we feel it: Perception and conceptualization of the term “fake news” in the media. In Staab, S., Koltsova, O., and Ignatov, D., editors, *Social informatics*, volume 11185 of *Lecture Notes in Computer Science*, pages 151–166. Springer, Cham.
- Cunha, E. L. T. P., Magno, G., Caetano, J., Teixeira, D., and Almeida, V. (under review). *Quantifying the conceptualization of the term “fake news” in Brazilian and English-speaking media sources*. Manuscript submitted for publication.
- Cunha, E. L. T. P., Magno, G., Comarela, G., Almeida, V., Gonçalves, M. A., and Benevenuto, F. (2011). Analyzing the dynamic evolution of hashtags on Twitter: A language-based approach. In *Proceedings of the ACL Workshop on Language in Social Media (LSM 2011)*, pages 58–65. Association for Computational Linguistics (ACL).
- Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2013). A linguistic characterization of Google+ posts across different social groups. In *Notes of the 5th Workshop on Information in Networks (WIN 2013)*. Stern School of Business – New York University (NYU).
- Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014a). He votes or she votes? Female and male discursive strategies in Twitter political hashtags. *PLOS ONE*, 9(1):e87041.

- Cunha, E. L. T. P., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014b). How you post is who you are: Characterizing Google+ status updates across social groups. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT'14)*, pages 212–217. Association for Computing Machinery (ACM).
- Cunha, E. L. T. P. and Rocha, B. (2008). A escrita do jovem usuário de internet em contextos com motivação oral: Comparação com a formação histórica das línguas românicas e dos crioulos de base românica. In *Anais do VII Encontro de Lingüística de Corpus (ELC 2008)*. Universidade Estadual Paulista (UNESP).
- Cunha, E. L. T. P. and Wichmann, S. (in press). An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus. *Corpora*.
- Danescu-Niculescu-Mizil, C., Gamon, M., and Dumais, S. (2011). Mark my words! Linguistic style accommodation in social media. In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, pages 745–754. Association for Computing Machinery (ACM).
- Danescu-Niculescu-Mizil, C., West, R., Jurafsky, D., Leskovec, J., and Potts, C. (2013). No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International World Wide Web Conference (WWW 2013)*, pages 307–318. Association for Computing Machinery (ACM).
- Davies, M. (2012). Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157.
- Davies, M. (2013). Corpus of News on the Web (NOW): 3+ bil-

- lion words from 20 countries, updated every day. Retrieved from <https://corpus.byu.edu/now/> . Accessed on May 4, 2018.
- Davies, M. (2017a). Fake news. Retrieved from <https://corpus.byu.edu/now/help/fake-news.asp> . Accessed on May 4, 2018.
- Davies, M. (2017b). The new 4.3 billion word NOW corpus, with 4-5 million words of data added every day. In *Proceedings of the 9th International Corpus Linguistics Conference*.
- Day, M. (2006). The long-term preservation of Web content. In Masanès, J., editor, *Web archiving*, pages 177–199. Springer, Berlin.
- De Choudhury, M., Sundaram, H., John, A., and Seligmann, D. D. (2009). What makes conversations interesting? Themes, participants and consequences of conversations in online social media. In *Proceedings of the 18th International Conference on World Wide Web (WWW 2009)*, pages 331–340. Association for Computing Machinery (ACM).
- Degand, L. and Van Bergen, G. (2018). Discourse markers as turn-transition devices: Evidence from speech and instant messaging. *Discourse Processes*, 55(1):47–71.
- Dores, M. and Toledo, C. (2018). De “lepra” à “hanseníase”: Uma análise lexicológica de base sócio-histórica. *Diacrítica*, 32(1):179–208.
- Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., and Mevel, J.-P. (1986). *Dicionário de linguística*. Cultrix, São Paulo, 2nd edition.
- Dunst, A., Hartel, R., and Laubrock, J. (2017). The Graphic Narrative Corpus (GNC): Design, annotation, and analysis for the

- digital humanities. In *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 15–20. Institute of Electrical and Electronics Engineers (IEEE).
- Eisenstein, J., O'Connor, B., Smith, N. A., and Xing, E. P. (2014). Diffusion of lexical change in social media. *PLOS ONE*, 9(11):e113114.
- Erjavec, K. and Kovačič, M. P. (2012). “You don’t understand, this is a new war!” Analysis of hate speech in news web sites’ comments. *Mass Communication and Society*, 15(6):899–920.
- Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. Association for Computing Machinery (ACM).
- Ferguson, C., Inglis, S. C., Newton, P. J., Cripps, P. J. S., Macdonald, P. S., and Davidson, P. M. (2014). Social media: A tool to spread information: A case study analysis of Twitter conversation at the Cardiac Society of Australia & New Zealand 61st Annual Scientific Meeting 2013. *Collegian*, 21(2):89–93.
- Fišer, D., Ljubešić, N., and Erjavec, T. (2018). The Janes project: Language resources and tools for Slovene user generated content. *Language Resources & Evaluation*.
- Flaounas, I., Ali, O., Lansdall-Welfare, T., Bie, T. D., Mosdell, N., Lewis, J., and Cristianini, N. (2013). Research methods in the age of digital journalism: Massive-scale automated analysis of news-content – topics, style and gender. *Digital Journalism*, 1(1):102–116.

- Flaounas, I., Turchi, M., Ali, O., Fyson, N., De Bie, T., Mosdell, N., Lewis, J., and Cristianini, N. (2010). The structure of the EU mediasphere. *PLOS ONE*, 5(12):e14243.
- Freire, A., Lloyd, R., and Turgeon, M. (2017). “Seu petralha! Seu coxinha!” – Measuring affective polarization in Brazil. In *Proceedings of the 41o. Encontro Anual da ANPOCS*. Associação Nacional dos Pesquisadores de Ciências Sociais (ANPOCS).
- Frischer, B. (2011). Art and science in the age of digital reproduction: From mimetic representation to interactive virtual reality. *Virtual Archaeology Review*, 2(4):19–32.
- García-Calderón, M. Á., García-Hernández, R. A., and Ledeneva, Y. (2018). Unsupervised multi-language handwritten text line segmentation. *Journal of Intelligent & Fuzzy Systems*, 34(5):2901–2911.
- Gelfert, A. (2018). Fake news: A definition. *Informal Logic*, 38(1):84–117.
- Glaznieks, A., Nicolas, L., Stemle, E., Abel, A., and Lyding, V. (2014). Establishing a standardised procedure for building learner corpora. *Apples – Journal of Applied Language Studies*, 8(3):5–20.
- Glowacki, M., Narayanan, V., Maynard, S., Hirsch, G., Kollanyi, B., Neudert, L.-M., Howard, P., Lederer, T., and Barash, V. (2018). News and political information consumption in Mexico: Mapping the 2018 Mexican presidential election on Twitter and Facebook. *The Computational Propaganda Project: Algorithms, automation and digital politics (COMPROP)*.
- Gómez, V., Kaltenbrunner, A., and López, V. (2008). Statistical analysis of the social network and discussion threads in Slashdot.

In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*, pages 645–654. Association for Computing Machinery (ACM).

Gries, S. T. and Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora*, 3(1):59–81.

Gross, M. (1972). *Mathematical models in linguistics*. Prentice-Hall, Englewood Cliffs.

Grosbeck, G. (2009). To use or not to use web 2.0 in higher education? *Procedia Social and Behavioral Sciences*, 1:478–482.

Grzybek, P. (2007). History and methodology of word length studies. In Grzybek, P., editor, *Contributions to the science of text and language: Word length studies and related issues*, pages 15–90. Springer, Berlin.

Guedes, A. d. S. and Mendes, B. P. (2016). Um estudo lexicológico de base sócio-histórica das formas lexicais “asilo de idosos” e “casa de repouso”. *Raído*, 10(24):38–52.

Gulordava, K. and Baroni, M. (2011). A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics*, pages 67–71. Association for Computational Linguistics (ACL).

Habgood-Coote, J. (2018). The term ‘fake news’ is doing great harm. *The Conversation*. Retrieved from <http://theconversation.com/the-term-fake-news-is-doing-great-harm-100406>. Accessed on January 5, 2019.

- Habgood-Coote, J. (2019). Stop talking about fake news! *Inquiry*, 62(9-10):1033–1065.
- Halabi, Y. S., Sa, Z., Hamdan, F., and Yousef, K. H. (2009). Modeling adaptive degraded document image binarization and optical character system. *European Journal of Scientific Research*, 28(1):14–32.
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501. Association for Computational Linguistics (ACL).
- Harden, C. (2019). Brazil fell for fake news: What to do about it now? Retrieved from <https://www.wilsoncenter.org/blog-post/brazil-fell-for-fake-news-what-to-do-about-it-now>. Accessed on August 12, 2019.
- Heaps, H. S. (1978). *Information retrieval: Computational and theoretical aspects*. Academic Press, New York.
- Heeringa, W. and Prokić, J. (2017). Computational dialectology. In Boberg, C., Nerbonne, J., and Watt, D., editors, *The handbook of dialectology*, pages 330–347. John Wiley & Sons, Hoboken.
- Henrich, N. and Holmes, B. (2013). Web news readers’ comments: Towards developing a methodology for using on-line comments in social inquiry. *Journal of Media and Communication Studies*, 5(1):1–4.
- Herdan, G. (1964). *Quantitative linguistics*. Butterworth, London.
- Hilpert, M. and Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus

linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4):385–401.

Hilpert, M. and Gries, S. T. (2016). Quantitative approaches to diachronic corpus linguistics. In Kytö, M. and Pahta, P., editors, *The Cambridge handbook of English historical linguistics*, pages 36–53. Cambridge University Press, Cambridge.

Hilpert, M. and Mair, C. (2015). Grammatical change. In Biber, D. and Reppen, R., editors, *The Cambridge handbook of English corpus linguistics*, pages 180–200. Cambridge University Press, Cambridge.

Hinrichs, L. and Szmrecsanyi, B. (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language and Linguistics*, 11(3):437–474.

Holan, A. (2017). The media’s definition of fake news vs. Donald Trump’s. *First Amendment Law Review*, 16:121–128.

Hooker, C. (2019). Students and corporate social media: Do college students care about social media usernames? *International Journal of Social Science Studies*, 7(4):79–86.

Horta Ribeiro, M., Calais, P. H., Almeida, V. A., and Meira Jr., W. (2017). “Everything I disagree with is #FakeNews”: Correlating political polarization and spread of misinformation. In *Proceedings of Data Science + Journalism @ KDD 2017 (DS+J’17)*.

Hsu, C.-F., Khabiri, E., and Caverlee, J. (2009). Ranking comments on the social web. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 90–97. Institute of Electrical and Electronics Engineers (IEEE).

- Hundt, M. and Mair, C. (1999). “Agile” and “uptight” genres: The corpus-based approach to language change in progress. *International Journal of Corpus Linguistics*, 4:221–242.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press, Cambridge.
- Hunter, W. and Power, T. J. (2019). Bolsonaro and Brazil’s illiberal backlash. *Journal of Democracy*, 30(1):68–82.
- Jensen, K. E. (2014). Linguistics in the digital humanities: (Computational) corpus linguistics. *MedieKultur: Journal of Media and Communication Research*, 57:115–134.
- Kerremans, D., Stegmayr, S., and Schmid, H.-J. (2012). The NeoCrawler: Identifying and retrieving neologisms from the Internet and monitoring ongoing change. In Allan, K. and Robinson, J. A., editors, *Current methods in historical semantics*, pages 59–96. De Gruyter Mouton, Berlin.
- Kilgarriff, A. (2005). Web as corpus (2001). In Sampson, G. and McCarthy, D., editors, *Corpus linguistics: Readings in a widening discipline*, pages 471–473. Continuum, London.
- Knowles, E. and Elliott, J., editors (1997). *The Oxford dictionary of new words*. Oxford University Press, Oxford.
- Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2019). The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*.
- Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – Reconstructing the composition of the German

corpus in times of WWII. *Digital Scholarship in the Humanities*, 32(1):169–188.

- Krishnamurthy, R. (1997). Keeping good company: Collocation, corpus and dictionaries. In *Cicle de conferències 95-96: Lèxic, corpus i diccionaris*, Barcelona. Institut Universitari de Lingüística Aplicada/Universitat Pompeu Fabra.
- Krishnamurthy, R. (2003). Freeze-frame pictures: Micro-diachronic variations in synchronic corpora. In Jozsef Andor, J. H. and Nikolov, M., editors, *Studies in English theoretical and applied linguistics*, pages 15–31. Lingua Franca Csoport, Pecs.
- Küng, L., Newman, N., and Picard, R. G. (2016). Online news. In Bauer, J. M. and Latzer, M., editors, *Handbook on the economics of the Internet*, pages 443–457. Edward Elgar Publishing, Cheltenham/Northampton.
- Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*, volume 1. Stanford University Press.
- Lansdall-Welfare, T., Sudhahar, S., Veltri, G. A., and Cristianini, N. (2014). On the coverage of science in the media: A big data study on the impact of the Fukushima disaster. In *2014 IEEE International Conference on Big Data*, pages 60–66.
- Las Casas, D., Magno, G., Cunha, E. L. T. P., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014). Noticing the other gender on Google+. In *Proceedings of the ACM Web Science 2014 Conference (WebSci 2014)*, pages 156–160. Association for Computing Machinery (ACM).
- Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R.,

- Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.
- Lee, E.-J. (2012). That’s not the way it is: How user-generated comments on the news affect perceived media bias. *Journal of Computer-Mediated Communication*, 18:32–45.
- Leech, G. (2005). Adding linguistic annotation. In Wynne, M., editor, *Developing linguistic corpora: A guide to good practice*, pages 17–29. Oxbow Books, Oxford. Retrieved from <http://ota.ox.ac.uk/documents/creating/dlc/>. Accessed on August 10, 2019.
- Lees, C. (2018). Fake news: The global silencer. *Index on Censorship*, 47(1):88–91.
- Leetaru, K. (2011). Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space. *First Monday*, 16(9).
- Lüdeling, A. and Kytö, M., editors (2008). *Corpus linguistics: An international handbook*, volume 1. De Gruyter, Berlin/New York.
- Maia, R. C. M., Rossini, P. G. C., de Oliveira, V. V., and de Oliveira, A. G. (2015). Sobre a importância de examinar diferentes ambientes online em estudos de deliberação. *Opinião Pública*, 21(2):490–513.
- Matoré, G. (1949). La lexicologie sociale. *L’Information Littéraire*, 2.
- Matoré, G. (1953). *La méthode en lexicologie: Domaine français*. Didier, Paris.
- McEnery, T. and Wilson, A. (1996). *Corpus linguistics: An introduction*. Edinburgh University Press, Edinburgh.

- Mellish, C. (1994). Computers and language use. In Asher, R. and Simpson, J., editors, *The encyclopedia of language and linguistics*, volume 2, pages 672–673. Pergamon Press, Oxford.
- Mello, H. (2012). Os corpora orais e o C-ORAL-BRASIL. In Raso, T. and Mello, H., editors, *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*, pages 31–54. Editora UFMG, Belo Horizonte.
- Mello, H., Raso, T., Mittmann, M. M., Vale, H. P., and Côrtes, P. O. (2012). Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: Critérios de implementação e validação. In Raso, T. and Mello, H., editors, *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*, pages 125–176. Editora UFMG, Belo Horizonte.
- Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., The Google Books Team, Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., and Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.
- Milioni, D. L., Vadratsikas, K., and Papa, V. (2012). ‘Their two cents worth’: Exploring user agency in readers’ comments in on-line news media. *Observatorio (OBS*) Journal*, 6(3):21–47.
- Mitkov, R., editor (2005). *The Oxford handbook of computational linguistics*. Oxford University Press.
- Möller, A. M., Kühne, R., Baumgartner, S. E., and Peter, J. (2019). Exploring user responses to entertainment and political videos: An automated content analysis of YouTube. *Social Science Computer Review*, 37(4):510–528.

- Moon, R. (2010). What can a corpus tell us about lexis? In O’Keeffe, A. and McCarthy, M., editors, *The Routledge handbook of corpus linguistics*, pages 197–211. Routledge, New York.
- Moreira, T., Brigatto, G., and Rosa, J. L. (2018). Após PagSeguro, UOL busca investidores para outros negócios. Retrieved from <https://glo.bo/2FD1MT1> . Accessed on August 15, 2019.
- Napoles, C., Tetreault, J., Rosato, E., Provenzale, B., and Pappu, A. (2017). Finding good conversations online: The Yahoo News Annotated Comments Corpus. In *Proceedings of the 11th Linguistic Annotation Workshop (LAW-XI)*, pages 13–23. Association for Computational Linguistics (ACL).
- Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, 108:95–117.
- Nevalainen, T. and Raumolin-Brunberg, H., editors (1996). *Sociolinguistics and language history: Studies based on the Corpus of Early English Correspondence*. Rodopi, Amsterdam/Atlanta.
- Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 100–108. Association for Computational Linguistics (ACL).
- Nguyen, D., Doğruöz, A. S., Rosé, C. P., and de Jong, F. (2016). Computational sociolinguistics: A survey. *Computational Linguistics*, 42(3):537–593.
- Nielsen, R. K. and Graves, L. (2017). “News you don’t believe”: Audience perspectives on fake news. Retrieved from <https://reutersinstitute.politics.ox.ac.uk/>

sites/default/files/2017-10/Nielsen&Graves_factsheet_1710v3_FINAL_download.pdf . Accessed on July 02, 2019.

- O'Reilly, T. (2007). What is Web 2.0: Design patterns and business models for the next generation of software. *Communications & Strategies*, 65:17–37.
- Otoni, R., Cunha, E. L. T. P., Magno, G., Bernardina, P., Meira Jr., W., and Almeida, V. (2018). Analyzing right-wing YouTube channels: Hate, violence and discrimination. In *Proceedings of the 10th ACM Conference on Web Science (WebSci'18)*, pages 323–332. Association for Computing Machinery (ACM).
- Pagel, M., Atkinson, Q. D., and Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature*, 449:717–720.
- Parliament of the United Kingdom (2018). Disinformation and ‘fake news’: Interim report. Retrieved from <https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36304.htm> . Accessed on January 2, 2018.
- Partington, A. (2010). Modern diachronic corpus-assisted discourse studies (MD-CADS) on UK newspapers: An overview of the project. *Corpora*, 5(2):83–108.
- Pate, U. A. and Ibrahim, A. M. (2019). Fake news, hate speech and Nigeria’s struggle for democratic consolidation: A conceptual review. In Solo, A. M., editor, *Handbook of research on politics in the computer age*, pages 89–112. IGI Global, Hershey.
- Pechenick, E. A., Danforth, C. M., and Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PLOS ONE*, 10(10):e0137041.

- Pedro, G. W. (2018). ComentCorpus: Identificação e pistas linguísticas para detecção de ironia no português do Brasil. Master's thesis, Universidade Federal de São Carlos (UFSCar).
- Pennebaker, J. W., Booth, R. J., Boyd, R. L., and Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates, Austin.
- Perc, M. (2012). Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface*, 9:3323–3328.
- Petersen, A. M., Tenenbaum, J., Havlin, S., and Stanley, H. E. (2012). Statistical laws governing fluctuations in word use from word birth to word death. *Scientific Reports*, 2(313).
- Phillips, T. (2018). Bolsonaro business backers accused of illegal Whatsapp fake news campaign. The Guardian. Retrieved from <https://www.theguardian.com/world/2018/oct/18/brazil-jair-bolsonaro-whatsapp-fake-news-campaign>. Accessed on January 2, 2018.
- Potthast, M. (2009). Measuring the descriptiveness of Web comments. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*, pages 724–725. Association for Computing Machinery (ACM).
- Potthast, M. and Becker, S. (2010). Opinion summarization of Web comments. In Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., and van Rijsbergen, K., editors, *Advances in information retrieval: Proceedings of the 32nd European Conference on Information Retrieval (ECIR 2010)*, pages 668–669. Springer.

- Presner, T. and Johanson, C. (2009). The promise of Digital Humanities: A whitepaper. Retrieved from <http://humanitiesblast.com/Promise%20of%20Digital%20Humanities.pdf> . Accessed on July 02, 2019.
- Prokić, J. (2017). Quantitative diachronic dialectology. In Wieling, M., Kroon, M., van Noord, G., and Bouma, G., editors, *From semantics to dialectometry: Festschrift in honour of John Nerbonne*, pages 293–301. College Publications, London.
- Rafael, G. C. R. A. and Simião, D. P. (2019). Aidético e soropositivo: Análise sócio-histórica da concorrência entre qualificadores utilizados em referência a portadores do HIV. *Inventário*, 23:45–68.
- Rajaraman, A. and Ullman, J. D. (2011). *Mining of massive datasets*. Cambridge University Press.
- Rama, T. (2015). *Studies in computational historical linguistics: Models and analyses*. PhD thesis, University of Gothenburg.
- Ramalho, R. (2018). Fux diz que Justiça pode anular uma eleição se resultado for influenciado por ‘fake news’ em massa. Retrieved from <https://glo.bo/36JPmVh> . Accessed on June 28, 2019.
- Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMMLP)*, pages 2931–2937. Association for Computational Linguistics (ACL).
- Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

- Reis, J. C., Correia, A., Murai, F., Veloso, A., and Benevenuto, F. (2019). Supervised learning for fake news detection. *IEEE Intelligent Systems*, 34(2):76–81.
- Reyes, A., Potthast, M., Rosso, P., and Stein, B. (2010). Evaluating humour features on Web comments. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).
- Ross, A. S. and Rivers, D. J. (2018). Discursive deflection: Accusation of “fake news” and the spread of mis- and disinformation in the tweets of President Trump. *Social Media + Society*, 4(2):1–12.
- Rossini, P. G. C. (2017). *Conversação política, incivilidade e intolerância em ambientes digitais*. PhD thesis, Universidade Federal de Minas Gerais.
- Roth, S. (2014). Fashionable functions: A Google Ngram view of trends in functional differentiation (1800-2000). *International Journal of Technology and Human Interaction*, 10(2):34–58.
- Rubin, V. L., Conroy, N. J., Chen, Y., and Cornwell, S. (2016). Fake news or truth? Using satirical cues to detect potentially misleading news. In *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*, pages 7–17. Association for Computational Linguistics (ACL).
- Ruchansky, N., Seo, S., and Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management (CIKM'17)*, pages 797–806. Association for Computing Machinery (ACM).

- Rumšienė, G. (2004). Development of Internet English: Alternative lexis, syntax and morphology. *Kalby Studijos (Studies about Languages)*, 6:48–55.
- Saccenti, E. and Tenori, L. (2012). Stylometric investigation of Dante's *Divina Commedia* by means of multivariate data analysis techniques. *International Journal of Computational Linguistics Research*, 3(2):35–48.
- Saliés, T. G. and Shepherd, T. G. (2013). Introdução: Por uma Linguística da Internet. In Shepherd, T. G. and Saliés, T. G., editors, *Linguística da internet*, pages 7–14. Contexto, São Paulo.
- Sampson, G. and McCarthy, D. (2005). From *The structure of English* (1952). In Sampson, G. and McCarthy, D., editors, *Corpus linguistics: Readings in a widening discipline*, page 9. Continuum, London.
- Sarmiento, R. and Mendonça, R. F. (2016). Disrespect in online deliberation: Inducing factors and democratic potentials. *Revista de Ciencia Política (Santiago)*, 36(3):705–729.
- Saussure, F. (1916). *Cours de linguistique générale*. Published by Charles Bally and Albert Sechehave. Payot, Lausanne/Paris.
- Schmid, H.-J. (2007). Entrenchment, salience, and basic levels. In Geeraerts, D. and Cuyckens, H., editors, *The Oxford handbook of cognitive linguistics*, pages 117–138. Oxford University Press, Oxford.
- Schulze, C., Stauffer, D., and Wichmann, S. (2008). Birth, survival and death of languages by Monte Carlo simulation. *Communications in Computational Physics*, 3(2):271–294.

- Sebba, M. and Fligelstone, S. (1994). Corpora. In Asher, R. and Simpson, J., editors, *The encyclopedia of language and linguistics*, volume 2, pages 769–773. Pergamon Press, Oxford.
- Shah, C. and Pomerantz, J. (2010). Evaluating and predicting answer quality in community QA. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 411–418. Association for Computing Machinery (ACM).
- Shepherd, T. M. G. (2014). Changing ‘faces’: A case study of complex prepositions in Brazilian Portuguese. In Berber Sardinha, T. and São Bento Ferreira, T. L., editors, *Working with Portuguese Corpora*, pages 69–88. Bloomsbury Academic.
- Shertzer, M. D. (1996). *The elements of grammar*. Macmillan Publishing Company, New York.
- Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.
- Siersdorfer, S., Chelaru, S., Nejdl, W., and San Pedro, J. (2010). How useful are your comments? Analyzing and predicting YouTube comments and comment ratings. In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, pages 891–900. Association for Computing Machinery (ACM).
- Silva, N. F. F., Coletta, L. F. S., and Hruschka, E. R. (2016). A survey and comparative study of tweet sentiment analysis via semi-supervised learning. *ACM Computing Surveys*, 49(1).
- SimilarWeb (2016). Rank. Retrieved from <https://support.similarweb.com/hc/en-us/articles/213452305-Rank-> . Accessed on August 26, 2019.

- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press, Oxford.
- Standage, T. (2017). The true history of fake news. 1843 Magazine. Retrieved from <https://www.1843magazine.com/technology/rewind/the-true-history-of-fake-news> . Accessed on May 4, 2018.
- Sullivan, M. (2017). It's time to retire the tainted term 'fake news'. The Washington Post. Retrieved from https://www.washingtonpost.com/lifestyle/style/its-time-to-retire-the-tainted-term-fake-news/2017/01/06/a5a7516c-d375-11e6-945a-76f69a399dd5_story.html . Accessed on January 5, 2019.
- Svensén, B. (1993). *Practical lexicography: Principles and methods of dictionary-making*. Oxford University Press, Oxford.
- Tambini, D. (2017). Fake news: Public policy responses. In Tambini, D. and Goodman, E., editors, *LSE Media Policy Project Series*. Media Policy Project, London.
- Tandoc Jr., E. C., Lim, Z. W., and Ling, R. (2018). Defining “fake news”: A typology of scholarly definitions. *Digital Journalism*, 6(2):137–153.
- Tausczik, Y. R. and Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.
- Taylor, C. (2008). What is *corpus linguistics*? What the data says. *ICAME Journal*, 32:179–200.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment in short strength detection informal text. *Jour-*

nal of the American Society for Information Science and Technology, 61(12):2544–2558.

Thompson, S. E. and Parthasarathy, S. (2006). Moore’s law: The future of Si microelectronics. *Materials Today*, 9(6):20–25.

Tichý, O. (2018). Lexical obsolescence and loss in English: 1700–2000. In Kopaczyk, J. and Tyrkkö, J., editors, *Applications of pattern-driven methods in corpus linguistics*, pages 81–103. John Benjamins, Amsterdam.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003*. Association for Computational Linguistics (ACL).

Trevisani, M. and Tuzzi, A. (2018). Learning the evolution of disciplines from scientific literature: A functional clustering approach to normalized keyword count trajectories. *Knowledge-Based Systems*, 146:129–141.

Tschiatschek, S., Singla, A., Gomez Rodriguez, M., Merchant, A., and Krause, A. (2018). Fake news detection in social networks via crowd signals. In *Proceedings of the Web Conference (TheWebConf 2018)*, pages 517–524. Association for Computing Machinery (ACM).

Tulloch, S. (1991). *The Oxford dictionary of new words: A popular guide to words in the news*. Oxford University Press, Oxford.

Upadhyay, S., Pant, V., Bhasin, S., and Pattanshetti, M. K. (2017). Articulating the construction of a web scraper for massive data extraction. In *Proceedings of the 2017 Second International Conference on Electrical, Computer and Communication Technolo-*

- gies (ICECCT)*. Institute of Electrical and Electronics Engineers (IEEE).
- Van Hout, T. and Burger, P. (2015). Mediatization and the language of journalism. *Tilburg Papers in Culture Studies*, 131.
- Vicario, M. D., Quattrociochi, W., Scala, A., and Zollo, F. (2019). Polarization and fake news: Early warning of potential misinformation targets. *ACM Transactions on the Web (TWEB)*, 13(2).
- Vinodhini, G. and Chandrasekaran, R. M. (2012). Sentiment analysis and opinion mining: A survey. *International Journal of Advanced Research in Computer Science and Software Engineering*, 2(6):282–292.
- Vo, A. T. and Carter, R. (2010). What can a corpus tell us about creativity? In O’Keeffe, A. and McCarthy, M., editors, *The Routledge handbook of corpus linguistics*, pages 302–315. Routledge, New York.
- Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.
- Wan, X., Zong, L., Huang, X., Ma, T., Jia, H., Wu, Y., and Xiao, J. (2011). Named entity recognition in Chinese news comments on the Web. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 856–864.
- Wasserman, H. (2017). Fake news from Africa: Panics, politics and paradigm. *Journalism*.
- Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In Brügger, N., editor, *Web 25: Histories from 25 years of the World Wide Web*, pages 179–190. Peter Lang, New York.

- Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1):3–25.
- Wintner, S. (2010). Computational models of language acquisition. In Gelbukh, A., editor, *Computational linguistics and intelligent text processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 86–99. Springer, Berlin and Heidelberg.
- Woolfs, D. (2010). Computational forensic linguistics: Searching for similarity in large specialised corpora. In Coulthard, M. and Johnson, A., editors, *The Routledge handbook of forensic linguistics*, pages 576–590. Routledge, New York.
- Wynne, M. (2008). Searching and concordancing. In Lüdeling, A. and Kytö, M., editors, *Corpus linguistics: An international handbook*, volume 1, pages 706–737. De Gruyter, Berlin/New York.
- Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G., and Blackburn, J. (2017). The Web centipede: Understanding how Web communities influence each other through the lens of mainstream and alternative news sources. In *Proceedings of the 2017 Internet Measurement Conference (IMC 2017)*, pages 405–417. Association for Computing Machinery (ACM).
- Zhang, W., Wu, F., and Zhang, C. (2013). Interpretation of the formation of internet neologisms and their translation from Pound’s perspective of “language energy”. *International Journal of English Linguistics*, 3(2):66–71.
- Zinsmeister, H. (2015). Syntax and corpora. In Kiss, T. and Alexiadou, A., editors, *Syntax – theory and analysis: An international handbook*, volume 3, pages 1912–1941. De Gruyter Mouton, Berlin/Munich/Boston.

Zollo, F., Novak, P. K., Del Vicario, M., Bessi, A., Mozetič, I., Scala, A., Caldarelli, G., and Quattrociocchi, W. (2015). Emotional dynamics in the age of misinformation. *PLOS ONE*, 10(9).

Summary

Computer-assisted corpus linguistics is one of the main points of convergence between linguistic and computational methods. In particular, the use of diachronic linguistic corpora provides opportunities for the quantitative and computational analyses of phenomena concerning language change through time. This dissertation presents the outcome of three independent, although related, projects that share the interest in the application of computing power to diachronic corpus linguistics. Its main goal is to offer contributions to three of the stages of the research involving diachronic linguistic corpora: (a) corpus building and compilation; (b) designing of tools and algorithms for data exploration; and (c) data analysis for linguistic, cultural and historical research. Two useful resources for corpus linguistics in a language other than English are first presented: a Web scraper of comments from news portals and websites, developed as open source and as being free for use, modification and distribution; and a freely available diachronic corpus composed of comments published at UOL, a major Brazilian news portal. These resources are relevant not only for linguists, but also for professionals from other fields, including social scientists and journalists concerned with the public perception of news and the relationship between media and society. Then, we propose a simple

and generalizable method to assist the identification of the periods of establishment and obsolescence of linguistic items in a diachronic corpus based on the frequency of these items in the corpus. This method may be employed for the analysis of any collection of linguistic items, regardless of language or historical period. We also demonstrate its applicability using lexical data from the Corpus of Historical American English (COHA), providing case studies on the statistics and characteristics of words that appear in or disappear from this corpus in different periods. Finally, we describe how diachronic corpora might be used for quantitative linguistic investigation by initially proposing a framework centered on the investigation of vocabulary through a diachronic approach which employs complementary methods from corpus linguistics and natural language processing. Subsequently, the applicability of this framework is demonstrated through the case analysis of the term *fake news*, investigating its perception and conceptualization in the traditional media using data collected from Brazilian and English-speaking media sources. With these contributions, we expect to advance research on diachronic corpus linguistics and on computational methods for linguistic analysis.

Keywords: corpus linguistics; computer-assisted linguistics; diachronic studies.

Samenvatting

Computerondersteunde corpuslinguïstiek is een van de belangrijkste punten van convergentie tussen taalkundige en computationele methoden. In het bijzonder biedt het gebruik van diachronische linguïstische corpora kansen voor de kwantitatieve en computationele analyses van fenomenen met betrekking tot taalverandering door de tijd heen. Dit proefschrift presenteert het resultaat van drie onafhankelijke, hoewel gerelateerde, projecten die dezelfde interesse hebben in de toepassing van rekenkracht op diachronische corpuslinguïstiek. Het belangrijkste doel is bijdragen te leveren aan drie van de fasen van het onderzoek waarbij diachronische taalkundige corpora betrokken zijn: (a) opbouw en compilatie van corpora; (b) ontwerp van instrumenten en algoritmen voor data-exploratie; en (c) gegevensanalyse voor taalkundig, cultureel en historisch onderzoek. Twee nuttige bronnen voor corpuslinguïstiek in een andere taal dan het Engels worden eerst gepresenteerd: een webscraper van commentaren van nieuwsportalen en websites, ontwikkeld als open source en als vrij te gebruiken, te wijzigen en te verspreiden; en een vrij verkrijgbaar diachronisch corpus samengesteld uit commentaren gepubliceerd op UOL, een belangrijk Braziliaans nieuwsportal. Deze bronnen zijn niet alleen relevant voor taalkundigen, maar ook voor professionals uit andere

vakgebieden, waaronder sociale wetenschappers en journalisten die zich bezighouden met de publieke perceptie van nieuws en de relatie tussen de media en de samenleving. Vervolgens stellen we een eenvoudige en generaliseerbare methode voor om de periodes van oprichting en veroudering van linguïstische items in een diachronisch corpus te helpen identificeren op basis van de frequentie van deze items in het corpus. Deze methode kan worden gebruikt voor de analyse van elke verzameling van taalkundige items, ongeacht de taal of de historische periode. We tonen ook de toepasbaarheid aan met behulp van lexicale gegevens van het Corpus of Historical American English (COHA), met casestudies over de statistieken en kenmerken van woorden die in verschillende periodes in dit corpus voorkomen of uit dit corpus verdwijnen. Tot slot beschrijven we hoe diachronische corpora kunnen worden gebruikt voor kwantitatief taalkundig onderzoek door in eerste instantie een kader voor te stellen dat gericht is op het onderzoek van de woordenschat door middel van een diachronische benadering die gebruik maakt van complementaire methoden uit de corpuslinguïstiek en natuurlijke taalverwerking. De toepasbaarheid van dit kader wordt vervolgens aangetoond door de casusanalyse van de term *fake news*, waarvan we de perceptie en conceptualisering in de traditionele media onderzoeken met behulp van gegevens verzameld uit Braziliaanse en Engelstalige mediabronnen. Met deze bijdragen verwachten we onderzoek naar diachronische corpuslinguïstiek en naar computationele methoden voor taalkundige analyse te bevorderen.

Trefwoorden: corpus linguïstiek; computerondersteunde linguïstiek; diachronische studies.

Resumo

A linguística de corpus assistida por computador é um dos principais pontos de convergência entre métodos linguísticos e computacionais. Em particular, o uso de corpora linguísticos diacrônicos oferece oportunidades para a análise quantitativa e computacional de fenômenos relacionados à mudança linguística ao longo do tempo. Esta tese apresenta o resultado de três projectos independentes, embora relacionados, que partilham o interesse na aplicação de métodos computacionais na linguística de corpus diacrônica. Seu principal objetivo é oferecer contribuições para três das etapas da pesquisa envolvendo corpora linguísticos diacrônicos: (a) construção e compilação de corpora; (b) elaboração de ferramentas e algoritmos para exploração de dados; e (c) análise de dados para pesquisa linguística, cultural e histórica. Dois recursos úteis para a linguística de corpus em uma língua diferente do inglês são inicialmente apresentados: um coletor de comentários de portais de notícias, desenvolvido em código aberto e gratuito para uso, modificação e distribuição; e um corpus diacrônico gratuito composto por comentários publicados no UOL, um dos principais portais de notícias brasileiros. Esses recursos são relevantes não apenas para linguistas, mas também para profissionais de outras áreas, incluindo cientistas sociais e jornalistas interessados na percepção

pública de notícias e na relação entre mídia e sociedade. Em seguida, propõe-se um método simples e generalizável para auxiliar na identificação dos períodos de estabelecimento e obsolescência de itens linguísticos em um corpus diacrônico baseado na frequência desses itens no corpus. Esse método pode ser empregado para a análise de qualquer coleção de itens linguísticos, independentemente da língua ou período histórico. Sua aplicabilidade é demonstrada a partir da utilização de dados lexicais do Corpus of Historical American English (COHA), para o qual são fornecidos estudos de caso de caráter quantitativo e qualitativo sobre as palavras que aparecem ou desaparecem desse corpus em diferentes períodos. Finalmente, descreve-se como corpora diacrônicos podem ser usados para a investigação linguística quantitativa, apresentando um método centrado na pesquisa lexical por meio de uma abordagem diacrônica que emprega métodos complementares da linguística de corpus e do processamento de língua natural. A aplicabilidade desse método é demonstrada por meio da análise de caso do termo *fake news*, em que investigam-se sua percepção e conceitualização na mídia tradicional utilizando dados coletados de fontes da mídia brasileira e de língua inglesa. Com essas contribuições, espera-se colaborar para as pesquisas sobre linguística de corpus diacrônica e sobre métodos computacionais para análise linguística.

Palavras-chave: linguística de corpus; linguística assistida por computador; estudos diacrônicos.

Curriculum vitae

Evandro Landulfo Teixeira Paradela Cunha was born in Belo Horizonte, Minas Gerais, Brazil on December 22, 1986. He joined the Universidade Federal de Minas Gerais (UFMG) in 2004 as an undergraduate student at the School of Engineering, before switching to Letters, in which he graduated in 2009.

From 2010 to 2012, he was a master's student in Computer Science at the same university, focused on topics related to computer-mediated communication and language dynamics in the online world. After obtaining his degree, he worked as a researcher at the National Institute of Science and Technology for the Web (InWeb) until 2014, when he joined the Department of Linguistics at the Max Planck Institute for Evolutionary Anthropology (MPI-EVA), in Leipzig, Germany, as an intern within the project "Computational and quantitative methods in historical linguistics".

In 2015, Evandro started his PhD in Linguistics at Universiteit Leiden and in Computer Science at Universidade Federal de Minas Gerais, in a cotutelle agreement, culminating in this dissertation. During his PhD studies, he also served as a lecturer at the Faculty of Letters at Universidade Federal de Minas Gerais in 2016 and 2019, teaching courses on Italian language, Portuguese syntax and composition.