Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

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 law[1]. 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).

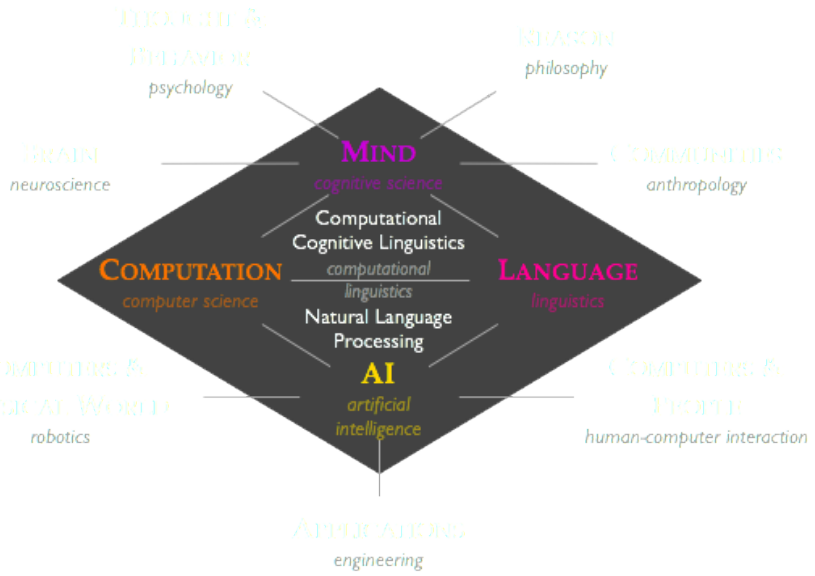

Figure 1.1: Intersections between computer science, linguistics, cognitive science and artificial intelligence. Extracted from `http://www.cs.cmu.edu/~nschneid/index2.html?view=plain` (Prof. Nathan Schneider's academic webpage) on August 2019.

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 linguistics[2] (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 sources[3].

[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 became available to linguists. That is by no means true. (...) [T]he man who really inaugurated the modern corpus-linguistics tradition was Charles Fries[4], 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 large[5] 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.

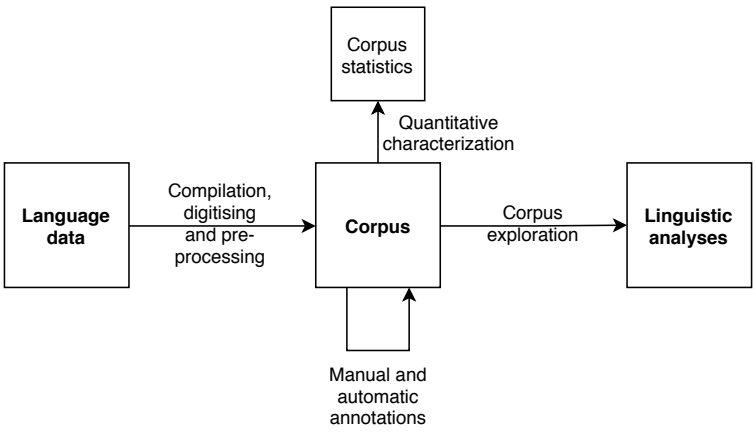

Figure 1.2: Simplified workflow of the activities involved in the research employing computer-based corpora for language studies. This figure is roughly inspired in the "[a]bstract workflow for corpus building" presented by Glaznieks et al. (2014).

Within a functionalist framework[6], 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.

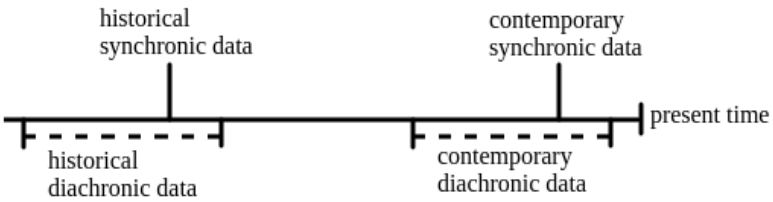

Figure 1.3: Timeline showing the position of synchronic, diachronic, contemporary and historical data in reference to the present time.

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 online 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 linguistics[9] 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) independently[10], 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.