# Contributions to the computational processing of diachronic linguistic corpora

Cunha, E. Landulfo Teixeira Paradela

# 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 independent[1], 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 belong[2]. 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' Party[3], 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 time[4].

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 *coxinha* (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.