1
Introduction

This textbook aims to teach you how to analyse and interpret language data in written or orthographically transcribed form (i.e. represented as if it were written, if the original data is spoken). It will do so in a way that should not only provide you with the technical skills for such an analysis for your own research purposes, but also raise your awareness of how corpus evidence can be used in order to develop a better understanding of the forms and functions of language. It will also teach you how to use corpus data in more applied contexts, such as e.g. in identifying suitable materials/examples for language teaching, investigating socio-linguistic phenomena, or even trying to verify existing linguistic theories, as well as to develop your own hypotheses about the many different aspects of language that can be investigated through corpora. The focus will primarily be on English-language data, although we may occasionally, whenever appropriate, refer to issues that could be relevant to the analysis of other languages. In doing so, we'll try to stay as theory-neutral as possible, so that no matter which 'flavour(s)' of linguistics you may have been exposed to before, you should always be able to understand the background to all the exercises or questions presented here.

The book is aimed at a variety of readers, ranging mainly from linguistics students at senior undergraduate, Masters, or even PhD levels who are still unfamiliar with corpus linguistics, to language teachers or textbook developers who want to create or employ more real-life teaching materials. As many of the techniques we'll be dealing with here also allow us to investigate issues of style in both literary and non-literary text, and much of the data we'll initially use actually consists of fictional works because these are easier to obtain and often don't cause any copyright

Practical Corpus Linguistics: An Introduction to Corpus-Based Language Analysis, First Edition. Martin Weisser.

issues, the book should hopefully also be useful to students of literary stylistics. To some extent, I also hope it may be beneficial to computer scientists working on language processing tasks, who, at least in my experience, often lack some crucial knowledge in understanding the complexities and intricacies of language, and frequently tend to resort to mathematical methods when more linguistic (symbolic) ones would be more appropriate, even if these may make the process of writing 'elegant' and efficient algorithms more difficult.

You may also be asking yourself why you should still be using a textbook at all in this day and age, when there are so many video tutorials available, and most programs offer at least some sort of online help to get you started. Essentially, there are two main reasons for this: a) such sources of information are only designed to provide you with a basic overview, but don't actually teach you, simply demonstrating how things are done. In other words they may do a relatively good job in showing you one or more ways of doing a few things, but often don't really allow you to use a particular program independently and for more complex tasks than the author of the tutorial/help file may actually have envisaged. And b) online tutorials, such as the ones on YouTube, may not only take a rather long time to (down)load, but might not even be (easily) accessible in some parts of the world at all, due to internet censorship.

If you're completely new to data analysis on the computer and working with – as opposed to simply opening and reading – different file types, some of the concepts and methods we'll discuss here may occasionally make you feel like you're doing computer science instead of working with language. This is, unfortunately, something you'll need to try and get used to, until you begin to understand the intricacies of working with language data on the computer better, and, by doing so, will also develop your understanding of the complexity inherent in language (data) itself. This is by no means an easy task, so working with this book, and thereby trying to develop a more complete understanding of language and how we can best analyse and describe it, be it for linguistic or language teaching purposes, will often require us to do some very careful reading and thinking about the points under discussion, so as to be able to develop and verify our own hypotheses about particular language features. However, doing so is well worth it, as you'll hopefully realise long before reaching the end of the book, as it opens up possibilities for understanding language that go far beyond a simple manual, small-scale, analysis of texts.

In order to achieve the aims of the book, we'll begin by discussing which types of data are already readily available, exploring ways of obtaining our own data, and developing an understanding of the nature of electronic documents and what may make them different from the more traditional types of printed documents we're all familiar with. This understanding will be developed further throughout the book, as we take a look at a number of computer programs that will help us to conduct our analyses at various levels, ranging from words to phrases, and to even larger units of text. At the same time, of course, we cannot ignore the fact that there may be issues in corpus linguistics related to lower levels, such

as that of morphology, or even phonology. Having reached the end of the book, you'll hopefully be aware of many of the different issues involved in collecting and analysing a variety of linguistic – as well as literary – data on the computer, which potential problems and pitfalls you may encounter along the way, and ideally also how to deal with them efficiently. Before we start discussing these issues, though, let's take a few minutes to define the notion of (linguistic) data analysis properly.

1.1 Linguistic Data Analysis

1.1.1 What's data?

In general, we can probably see all different types of language manifestation as language data that we may want/need to investigate, but unfortunately, it's not always possible to easily capture all such 'available' material for analysis. This is why, apart from the 'armchair' data available through introspection (cf. Fillmore 1992: 35), we usually either have to collect our materials ourselves or use data that someone else has previously collected and provided in a suitable form, or at least a form that we can adapt to our needs with relative ease. In both of these approaches, there are inherent difficulties and problems to overcome, and therefore it's highly important to be aware of these limitations in preparing one's own research, be it in order to write a simple assignment, a BA dissertation, MA/PhD thesis, research paper, etc.

Before we move on to a more detailed discussion of the different forms of data, it's perhaps also necessary to clarify the term *data* itself a little more, in order to avoid any misunderstandings. The word itself originally comes from the plural of the Latin word *datum*, which literally means '(something) given', but can usually be better translated as 'fact'. In our case, the data we'll be discussing throughout this book will therefore represent the 'facts of language' we can observe. And although the term itself, technically speaking, is originally a plural form referring to the individual facts or features of language (and can be used like this), more often than not we tend to use it as a singular mass noun that represents an unspecified amount or body of such facts.

1.1.2 Forms of data

Essentially, linguistic data comes in two general forms, written or spoken. However, there are also intermediate categories, such as texts that are written to be spoken (e.g. lectures, plays, etc.), and which may therefore exhibit features that are in between the two clear-cut variants. The two main media types often require rather radically different ways of 'recording' and analysis, although at least some of the techniques for analysing written language can also be used for analysing transliterated or (orthographically) transcribed speech, as we'll see later when looking at some dialogue data. Beyond this distinction based on medium, there are of

course other classification systems that can be applied to data, such as according to *genre* , *register* , *text type* , etc., although these distinctions are not always very clearly formalised and distinguished from one another, so that different scholars may sometimes be using distinct, but frequently also overlapping, terminology to represent similar things. For a more in-depth discussion of this, see Lee (2002).

To illustrate some of the differences between the various forms of language data we might encounter, let's take a look at some examples, taken from the Corpus of English Novels (CEN) and Corpus of Late Modern English Texts, version 3.0 (CLMET3.0; De Smet, 2005), respectively. To get more detailed information on these corpora, you can go to https://perswww.kuleuven.be/~u0044428/, but for our purposes here, it's sufficient for you to know that these are corpora that are mainly of interest to researchers engaged in literary stylistic analyses or historical developments within the English language. However, as previously stated, throughout the book, we'll often resort to literary data to illustrate specific points related to both the mechanics of processing language and as examples of genuinely linguistic features. In addition to being fictional, this data will often not be contemporary, simply because much contemporary data is often subject to copyright. Once you understand more about corpora and how to collect and compile them yourself, though, you'll be able to gather your own contemporary data, should you wish so, and explore actual, modern language in use.

Apart from being useful examples of register differences , the extracts provided below also exhibit some characteristics that make them more difficult to process using the computer. We'll discuss these further below, but I've here highlighted them with boxes.

Sample A – from *The Glimpses Of The Moon* by Edith Wharton, published 1922

IT rose for them--their honey-moon--over the waters of a lake so famed as the scene of romantic raptures that they were rather proud of not having been afraid to choose it as the setting of their own.

"It required a total lack of humour, or as great a gift for it as ours, to risk the experiment," Susy Lansing opined, as they hung over the inevitable marble balustrade and watched their tutelary orb roll its magic carpet across the waters to their feet.

"Yes--or the loan of Strefford's villa," her husband emended, glancing upward through the branches at a long low patch of paleness to which the moonlight was beginning to give the form of a white house-front.

Sample B – from: *Eminent Victorians* by Lytton Strachey, published 1918

Preface

THE history of the Victorian Age will never be written; we know too much about it. For ignorance is the first requisite of the historian—ignorance, which simplifies and clarifies, which selects and omits, with a placid perfection unattainable by the highest art. Concerning the Age which has just passed, our fathers and our grandfathers have poured forth and accumulated so vast a quantity of information that the industry of a Ranke would be submerged by it, and the perspicacity of a Gibbon would quail

before it. It is not by the direct method of a scrupulous narration that the explorer of the past can hope to depict that singular epoch. If he is wise, he will adopt a subtler strategy. He will attack his subject in unexpected places; he will fall upon the flank, or the rear; he will shoot a sudden, revealing searchlight into obscure recesses, hitherto undivined. He will row out over that great ocean of material, and lower down into it, here and there, a little bucket, which will bring up to the light of day some characteristic specimen, from those far depths, to be examined with a careful curiosity.

Sample C – from *The Big Drum* by Arthur Wing Pinero, published 1915

Noyes.
[Announcing Philip.] Mr. Mackworth.
Roope.
[A simple-looking gentleman of fifty, scrupulously attired—jumping up and shaking hands warmly with Philip as the servant withdraws.] My dear Phil!
Philip.
[A negligently—almost shabbily—dressed man in his late thirties, with a handsome but worn face.] My dear Robbie!
Roope.
A triumph, to have dragged you out! [Looking at his watch.] Luncheon isn't till a quarter-to-two. I asked you for half-past-one because I want to have a quiet little jaw with you beforehand.
Philip.
Delightful.
Roope.
Er—I'd better tell you at once, old chap, whom you'll meet here to-day.

Sample A is clearly a piece of narrative fiction, mixing narrative description and simulated reported speech, references to characters and situations that are depicted as life-like, as well as featuring a number of at least partly evaluative reporting verbs, such as *opined* and *emended*. **Sample B**, on the other hand, contains no reported speech and reporting verbs, although it's clearly also narrative – albeit non-fictional –, with a relatively complex sentence structure, including numerous relative and adverbial clauses, and an overall high degree of formality. **Sample C**, in contrast, exhibits clear characteristics of (simulated) spoken language, much shorter and less complex syntax, even single-word 'sentences', with names, titles and informal terms of address (*old chap*) used when the characters are addressing/introducing each other, exclamations, contractions, and at least one hesitation marker (*Er*). And even though the language in the latter sample seems fairly natural, we can still easily see that it comes from a scripted text, partly because of the indication of speakers (which I've highlighted in bold-face), and partly due to the stage instructions included in square brackets.

As we haven't discussed any of the issues in processing such text samples yet, it may not be immediately obvious to you that these different types of register may potentially require different analysis approaches, depending on what our exact

aims in analysing them are. For instance, for **Sample A**, do we want to conceptually treat the reported speech as being of the same status as the descriptive parts, and do we thus want to analyse them together or separately? Or are we possibly just interested in how the author represents the direct speech of the characters in the novel, and would therefore want to extract only that? And if so, how would we best go about this?

Sample B is probably relatively straightforward to analyse in terms of perhaps a frequency analysis of the words, but what if we're also interested in particular aspects of syntax or lexis that may be responsible for its textual complexity or the perceived level of formality, respectively? And, last but not least, concerning **Sample C**, similarly to **Sample A**, which parts of the text would we be interested in here and how would we extract them? Are the stage instructions equally important to us as the direct speech exchanges between the characters? Or, if, for example, we're interested in the average number of words uttered by each character, how do we deal with hesitation markers? Do we treat them as words or 'non-words' simply to be deleted? As I've already tried to hint at in the beginning of this paragraph, the answers to these questions really depend on our research purpose(s), and can thus not be conclusively stated here.

Something else you may have noticed when looking at the samples I've provided above is that they're all from the early 20th century. As such, the language we encounter in them may sometimes appear overly formal (or even archaic) to some extent, compared to the perhaps more 'colloquial' language we're used to from the different registers these days. I've chosen extracts from these three particular texts and period for a number of reasons: a) their authors all died more than 70 years ago so the texts are in the public domain; in other words, there are no copyright issues, even when quoting longer passages; b) they are included in corpus compilations; and c) they not only illustrate register/genre differences but also how the conventions for these may change over time, as can be seen, for example, in the spelling of *to-day* in the final extract.

As pointed out above, another interesting aspect of these samples is that they exhibit particular formatting issues, which again may not be immediately apparent to you yet, but are due to somewhat bizarre typographical conventions. If you look closely at the samples, you can see that in **Sample A** there are double dashes marking the parenthetical counterpart (i.e. reference resolution) "their honey-moon" to the sentence-initial cataphoric pronoun "IT". What is in fact problematic to some extent for processing the text is that these dashes actually look like double hyphens, i.e. they're not surrounded by spaces on either side, as would be the normal convention. Now, many computer programs designed to count words will split the input text on spaces and punctuation. Unfortunately, though, this would leave us with some very strange 'words' (that superficially look like hyphenated compounds), *them—their* and *honey-moon—over*, in any resulting word-frequency list. This is obviously something we do not want and which introduces errors into any automatic analysis of the data. Something similar, albeit not to signal a parenthetical but instead some kind of pseudo-punctuation, happens

again for "Yes—or" a little further down in the text. We can already see, therefore, from this relatively short sample of text that a failure to deal with this feature could cause issues in a number of places throughout the text. The same problem occurs in the other two samples, only that there the dash doesn't actually consist of two separate characters, but one single *m-dash*.

A different problem occurs in the use of initial capitals in **Samples A** and **B**. As you can see, the words *it* and *the* appear in capital letters throughout, signalling the beginning of the chapter typographically. Again, as 'human consumers' of the text, this will not cause any processing problems, but for the computer, *the*, *The*, and *THE* are in fact three different 'words', or at least word forms. Thus, even single initial capitals at the beginning of sentences may become issues in identifying and counting words on the computer. We'll talk more about this type of issue in Section 4.4.1, where we'll explore ways of dealing with such features of the text in order to retain relatively 'clean' data.

1.1.3 Collecting and analysing data

When collecting our own data, we obviously need to consider *methodologies* that allow us to collect the right types and amount(s) of data to answer our particular *research questions*. This, however, isn't the only type of consideration necessary, but we also need to bear in mind *ethical issues* involved in the collection – such as asking people for permission to record them or to publish their recordings, etc. – and which type of *format* that data should be stored in so as to be most useful to us, and potentially also other researchers.

When using other people's existing data, there are usually issues in accessing data stored in their specific format(s) or converting the data to a format that is more suitable to one's own needs, as we've just seen above, such as removing unwanted types of information or transforming overly specific information into simpler forms of representation. In this textbook, we'll also look at some of the important aspects of collecting or adapting data to one's needs, as well as how to go about analysing and presenting them in various ways, once a suitable format has been established.

In order to be able to work with electronic data, we also need to become familiar with a variety of different programs, some written specifically for linguistic analysis, some for more general purposes of working with texts. One of the key features of this book is that the programs I'll recommend to you are almost exclusively obtainable free of charge, i.e. so-called *freeware*. This doesn't mean that there aren't other excellent programs out there that may do some of the tasks we want to perform even better, or in simpler or more powerful ways, but simply reflects the fact that there are already many useful free programs available, and also my own philosophy that we shouldn't need to spend substantial amounts of money just to enable us to do research. This is at least part of the reason why I make most of my own programs available to the research community in this way, apart from the fact that this makes my own research (results) more easily reproducible by

others, and therefore caters for the aims of satisfying *accountability* and academic honesty. For the sake of completeness, though, I'll generally try to at least refer to alternative commercial programs, but without discussing them in any detail.

Corpus linguistics, as a form of data analysis methodology, can of course be carried out on a number of different operating systems, so I'll also try to make recommendations as to which programs may be useful for the most commonly used ones, Windows, Mac OS X, and Linux. Because there are many different 'flavours' of Linux, though, with a variety of different windowing interfaces, I'll restrict my discussions to two of the most popular ones, KDE and Gnome. Unfortunately, I won't be able to provide detailed support on how to actually install the programs themselves, as this may sometimes involve relatively detailed information about your system that I cannot predict. Instead, however, I'll actually try to avoid/pre-empt such issues by recommending default programs that are probably already installed, provided that they do in fact fulfil all or at least most of our needs.

1.2 Outline of the Book

This book is organised into four sections. The first section (comprising Chapters 1 and 2) begins with a very brief introduction to the history and general design of corpora, simply to 'set the scene', rather than to provide an extensive coverage of the multitude of corpora that have been created for different purposes and possibly also made available for free or in the form of various types of interfaces. More extensive coverage on the subject, including more theoretical implications, is already provided in books like Kennedy (1998), Meyer (2002), or Lindquist (2009), so these texts can always be consulted for reference if necessary, and we can instead focus on more practical issues. For a more detailed interesting discussion of some of the different 'philosophical' approaches to corpus linguistics, you can consult McEnery and Hardie (2012).

The introductory section is followed by an overview of different methods to compile and prepare corpora from available online resources, such as *text archives* or the *WWW*. This section (spanning Chapters 3 and 4) should essentially provide the basis for you to start building your own corpora, but also introduces you to various issues related to handling language on the computer, including explanations of different file types you may encounter or want to use, as well as certain types of meta-information about texts.

Section 3 (Chapters 5 to 10) then deals with different approaches to corpus-based linguistic data analysis, ranging from basic searching (concordancing) via learning about more complex linguistic patterns, expressed in the form of regular expressions, to simple and extended word (frequency) list analyses. This part already contains information on how to tag your data morpho-syntactically, using freely available tagging resources, and how to make use of tagging in your analyses. The final section then takes the notion of adding linguistic information to

your data further, and illustrates how to enrich corpus data using basic forms of XML in order to cyclically improve your analyses or publish/visualise analysis results effectively.

As corpus linguistics is a methodology that allows us to develop insights into how language works by 'consulting' real-life data, it should be fairly obvious that we cannot learn how to do corpus research on a purely theoretical basis. Therefore, as far as possible, all sections of this book will be accompanied by practical exercises. Some of these will appear to be relatively straightforward, almost mechanical, ones where you simply get to follow a sequence of steps in order to learn how to use a specific function inside a program or web interface, while others are more explicitly designed to enable you to develop your own strategies for solving problems and testing hypotheses in linguistics. Please bear in mind, though, that for the former type of exercise, simply following the steps blindly without trying to understand why you're doing them will not allow you to learn properly. So, as far as possible, at each point you should try to understand what we're trying to achieve and how the particular program we're using only gives us a handle on producing the relevant data, but does not actually answer our research questions for us. In the same vein, it's also important to understand that once we actually have extracted some relevant data from a corpus, this is rarely ever the 'final product'. Such data generally either still needs to be interpreted, filtered, or evaluated as to its usefulness, if necessary by (re-)adjusting the search strategy or initial hypotheses and/or conclusions, or, if it's to be used for more practical purposes, such as in the creation of teaching materials or exercises, to be brought into an appropriate form.

As we move on and you learn more and more techniques, the exercises will also get more complex, sometimes assuming the size of small research projects, if carried out in full detail. As a matter of fact, as these exercises require and consolidate a lot of the knowledge gained in prior sections, they might well be suitable for small research projects to be set by teachers, and possibly even form the basis of BA theses or MA dissertations.

Of course, you won't be left alone in figuring out the solutions to these exercises; both types will be solved at the end of each respective section, either in the form of detailed and precise explanations, or, whenever the results might be open to interpretation, by appropriate comments illustrating what you could/should be able to observe. For the more extensive exercises referred to in the previous paragraph, I'll often start you off with suitable explanations regarding the procedures to follow, and also hint at some potential issues that may arise, but will leave the completion up to you, to help you develop your awareness independently. Furthermore, as real corpus linguistics is not just about getting some 'impressive' numbers but should in fact allow you to gain real insights into different aspects of language, you should always try to relate your results to what you know from established theories and other methods used in linguistics, or even other related disciplines, such as for example sociology, psychology, etc., as far as they may be relevant to answering your research questions. This is also why the solutions to,

and discussions of, the exercises may often represent those parts of the book that cover some of the more theoretical aspects of corpus linguistics, aspects that you'll hopefully be able to master once you've acquired the more practical tools of the trade. Thus, even if you may think you've already found a perfect answer to an exercise, you should probably still spend some time reading carefully through each solution.

As this textbook is more practical in nature than other textbooks on corpus linguistics, at the end of almost all chapters, I've also added a section entitled 'Sources and Further Reading'. These sections essentially provide lists of references I've consulted and/or have found most useful and representative in illustrating the particular topic(s) discussed in the chapter. You can consult these references if you want to know more about theoretical or practical issues that I am unable to cover here, due to reasons of space. These sections may not necessarily contain highly up-to-date references, for the simple reason that, unfortunately, later writings may not always represent improvements over the more fundamental works produced in some of the areas covered. Once you understand more about corpus linguistics, though, you may want to consult the individual chapters in two of the recent handbooks, O'Keeffe & McCarthy (2010) and Lüdeling & Kytö (2008), so that you can evaluate the progress made over recent years yourself.

1.3 Conventions Used in this Book

In linguistics, there are many conventions that help us to distinguish between different levels of analysis and/or description, so as to better illustrate which different types of language phenomena we're dealing with at any given point in time. Throughout this book, I'm going to make use of many, if not most, of these conventions, so it's important to introduce them at this point. In addition to using these conventions as is done in linguistics, I may also use some of them to indicate special types of textual content relevant to the presentation of resources in this book, etc.

Double quotes ("…") indicate direct speech or short passages quoted from books.

Single quotes ('…') signal that an expression is being used in an unusual or unconventional way, that we're referring to the meaning of a word or construction on the semantic level, or to refer to menu items or sometimes button text in programs used. The latter may also be represented by a stylised button text, e.g.  .

Curly brackets ({…}) are used to represent information pertaining to the level of morphology.

Angle brackets (<…>) indicate that we're dealing with issues related to orthography or spelling. Alternatively, they're also used in certain types of linguistic annotation.

Forward slashes/square brackets generally indicate that we're discussing issues on the levels of phonology or phonetics. Within quoted material, they may also signal amendments to the original material made in order to fit it into the general sentence structure.

Italics are used to represent words or expressions, sometimes whole sentences, that illustrate language materials under discussion. In some cases, they may also be used to indicate emphasis/highlighting, especially if co-occurring with boldface.

Small caps are used to indicate lemmas, i.e. forms that allow us to conveniently refer to all instances of a verb, noun, etc.

`Monospaced font` indicates instructions/text to be typed into the computer, such as a search string or regular expression.

1.4 A Note for Teachers

The relatively low number of chapters may make this book appear deceptively short, and you might be wondering whether it would be suitable for a course that runs for a whole semester of up to 18 weeks; there's no need to worry, though, that you may necessarily have to supplement it with further materials, although this is of course possible.,

The sections and chapters of the book have been arranged to be thematically coherent, but, if you're planning to use it as a textbook in class, you'll frequently find that one chapter corresponds to more than one classroom unit. I'd therefore suggest that, while preparing specific topics, even – or especially – if you may already be an expert in the field, you at least try out the exercises carefully yourself, and then attempt to gauge how long it may take your students to carry them out. If your audience is already highly technically literate and has a strong background in linguistics, then obviously the exercises can be done much more quickly. If, on the other hand, your students are somewhat 'technophobic' or do not yet have a strong background in linguistics, you may either wish to spread the content over multiple units, or set at least some of the exercises as homework. In order to save time, you can also ask your course participants to perform certain preparatory tasks, such as downloading and installing different pieces of software, or registering for online resources, prior to coming to class.

1.5 Online Resources

This book also has an accompanying web page, where you'll be able to find some online exercises, links to my own software, updated information about programs or features discussed in the book, etc. The web address for this page is http://martinweisser.org/pract_cl/online_materials.html, and you'll probably want to bookmark this straight away, so that you'll be able to access it for future reference.

All my own software is provided under GPL 3 licence, so you can download and distribute it freely. The programs were originally designed to run on Windows, but can easily be used through Wine (https://www.winehq.org/) on Mac OS X or Linux. Additional information on how to do this can be found at http://martinweisser.org/ling_soft.html.