CHAPTER 3

Analysing Environmental Narratives Computationally

Ross S. Purves
Department of Geography; URPP Language and Space, University of Zurich, Switzerland

Olga Koblet
Department of Geography, University of Zurich, Switzerland

Benjamin Adams
Department of Computer Science and Software Engineering, University of Canterbury, Christchurch, New Zealand

This book's origins lie in a desire to showcase, through an interdisciplinary approach, the potential for computational methods in analysing text that describes the environment. Our argument is that these computational methods need not be complex, but rather that through a combination of well-designed research questions, appropriate text collections, a sensible choice of methods and careful interpretation, we can gain new and useful insights. In this chapter we work step-by-step through a set of basic building blocks to illustrate how environmental narratives can be computationally analysed. To make our arguments clear and concrete, we divide the material that follows into two parts. Each section starts with a general overview, introducing key concepts, literature

How to cite this book chapter:
Purves, Ross S., Olga Koblet, and Benjamin Adams (2022). "Analysing Environmental Narratives Computationally." In: *Unlocking Environmental Narratives: Towards Understanding Human Environment Interactions through Computational Text Analysis.* Ed. by Ross S. Purves, Olga Koblet, and Benjamin Adams. London: Ubiquity Press, pp. 43–84. DOI: https://doi.org/10.5334/bcs.c. License: CC-BY 4.0

and methods in a general sense. Accompanying this material are worked examples, where we show how the methods discussed can be used in practice to explore one particular form of environmental narrative – discussions of the role of the Forestry Commission in the UK since its formation in 1919.

3.1 Narrative Forms

We are all familiar with different ways of writing. Good recipes are well ordered, telling us exactly how to perform each step on the way to the perfect chocolate cake, and leaving no ingredients to our imagination. Political texts carefully argue, picking out evidence backing a particular position and applying thoughtfully chosen rhetoric, to persuade us of the merits of voting for a proposition. Travel guides emphasise and richly describe particular events or places, usually with the aim of informing our visit to the same location. There are many different ways of characterising such texts, focusing, for example, on narrative form or the genre of writing. These different forms matter, because understanding how they work is important to the ways in which we interpret their contents. Rather than delve deeply into literary theory here, we would like to introduce a few important concepts which can inform how we approach computationally analysing a text.

The first of these is the **hermeneutic circle** (Martin, 1972; Boell and Cecez-Kecmanovic, 2010). Simply put, this means that our understanding of a complete text is based on our reading of its individual parts, and our understanding of the parts is based on our reading of the whole. The key idea here is that to properly understand a text we need to both interpret it as a complete work *and* explore the individual parts making up the text, hence the interpretative circle involves a continuous back and forth. At its most reductionist level, the hermeneutic circle implies that we cannot know the meanings of individual words without seeing them in context. So, given a set of parts (words) {man, dog, bit, the} we do not know whether the man bit the dog, or the dog bit the man. To interpret the sentence (the whole) we need to understand the parts and know that "man" and "dog" are objects, and "bit" is an action. Interestingly, at least in English, we can discard "the" without any change to the meaning, which is conveyed by context (men and dogs can bite) and word order.[1] For those of us working computationally, extending these ideas to longer texts requires a realisation that interpreting a text at the word- and sentence-level is not enough to fully understand it, and equally importantly, to realise that external context (e.g., the political position of an author) might also be required to understand a text.

[1] The role of contextual factors in understanding the meaning of a text is also studied in pragmatics where not only linguistic context but also situational context (e.g., the writer's intention, who the audience is, what the social environment is, etc.) plays a role (Green, 1996).

This importance of context and its influence of our interpretation of a text is lucidly demonstrated by the short discussions of the "Glencoe Road" text in the introduction. Each interpretation was influenced not only by the disciplinary backgrounds of the discussants, but also by their geographical knowledge (e.g., whether they were personally familiar with the region) and their general knowledge of the UK, its politics and debates.

Central to our understanding and analysis of environmental narratives are the ways in which we extract and analyse references to space and place. Doing so naively, and simply treating locations as a reality in which a narrative takes place would be to ignore much of what we know about the use of metaphor and narrative in language. Thus, the space in which a literary text takes place is not only to be interpreted literally, but also metaphorically and culturally (Lotman, 1990; Lakoff and Johnson, 2008). For example, high places are generally associated with success and well-being, and low places with failure and depression.[2] These metaphors pervade written language, and are so effective that most of us do not notice them until we encounter their (unfamiliar) use in a non-native language. Dealing with such metaphorical uses of language is essential if we wish to computationally analyse text, as otherwise a computer has no way of distinguishing between, for example, fictive motion (we followed the ridge) and real motion (we followed the car) (Egorova, Tenbrink, and Purves, 2018). The semiotician, Yuri Lotman, pointed out a second, obvious, but also often neglected point. The space described in a (literary) text is an abstraction of (some) reality, not a copy. This abstraction is influenced by both explicit choices (e.g., those of a nature writer to emphasise sounds and sights in a landscape), but also less deliberate choices, influenced by, for example, culture, background and language.

Understanding, or at least acknowledging, narrative form is important for computational analysis, since it profoundly influences both the questions that it is reasonable to ask of a corpus and the results that a computational analysis can produce. These questions can take multiple forms, but it is important to consider them before commencing an analysis. We emphasise here the importance of formulating guiding questions before starting work and, where appropriate, **hypotheses**.

Hypothesis

One reading of the text on the "Glencoe Road" in Chapter 1, that provided by Graham Fairclough, focused on the political nature of the argument presented in the article. Fairclough notes the importance and value of individual voices, specifically that of the elite in

[2] The pervasive use of spatial metaphors in language (e.g., "rising up the ranks", "going through a rough patch") means that much spatial language in a text – even within texts that we would characterise as environmental narratives – might have no relationship whatsoever to environmental spaces.

the form of Sir John Stirling Maxwell. He identifies a metaphorical spatial disconnection between the well-connected and privileged elite and a region on its periphery. Temporally, he positions the article with respect to the importance of two movements - one connected with modernisation and progress through the motor car, and the other with the start of various movements connected with countryside conservation.

Fairclough's reading of the text contained a footnote which provides the starting point for the exemplary analysis carried out in this chapter. He notes that Maxwell was also "Chair of the Forestry Commission from 1929 to 1932, an organisation that some argue has had a less benevolent impact on the British landscape."

The Forestry Commission was formed in 1919, as a response to the depletion of Britain's forests during the First World War. Its original role was concerned with forestry as a means of producing timber, but over time this has changed to one more and more influenced by debates concerned with landscape beauty, biodiversity and recreational use of forests.

To explore the role of the Forestry Commission over time as recorded in text we turn to Hansard, edited transcripts of the proceedings of the British Parliament, which documents in great detail the activities of the House of Commons and the House of Lords. Inspired by Fairclough's footnote, we set out to explore how the Forestry Commission has been discussed in the British Parliament since its inception in 1919, with a focus on the perceived impact of the Forestry Commission on landscape. As a second, more contemporary source reflecting the views of those active in the countryside, we explored descriptions from Geograph[3], a crowdsourced collection of more than six million images and texts describing 1 km grid squares across Great Britain.

We hypothesise that discussion of the Forestry Commission in parliament will focus on the perceived negative impact of forestry, and the Commission's activities, on landscape. In Geograph we expect to see similar patterns, but with a more direct concentration on the visual impact of Forestry Commission activities in the landscape, since in this case our collection consists of texts describing photographs.

3.2 Building a Corpus

The first step in working with texts is compiling a large, systematic text collection, commonly known as a corpus (corpora in plural). The meaning of

[3] https://www.geograph.org.uk/

large and *systematic* depends on the research questions we want to answer with this corpus. Creating corpora can be traced back to at least the 1950s, when literary scholars started compiling systematic text collections of, for example, the complete works of one author or of a variety of authors covering the same **time period** (Pustejovsky and Stubbs, 2012). From the 1960s, influenced by the needs of corpus linguistics, corpora capturing usage of American and British English started to be created, in the form of, for example, the Brown Corpus and Lancaster-Oslo-Bergen corpora, respectively. Since both set out to reflect general usage of *language*, texts in these corpora include both written and transcribed spoken texts as well as different **genres** and **domains**. Genre describes general characteristics of texts, for example, broadcast news, as in the corpus of regional newspapers from the UK compiled by Lansdall-Welfare et al. (2017). Domain refers to main subject of the texts, which might influence word sense (e.g., *bank* in financial texts and *bank* in river renaturalisation texts) and the specific vocabulary used in a corpus (Augenstein, Derczynski, and Bontcheva, 2017). Examples of domain-specific corpora include a multilingual corpus of mountaineering texts called Text + Berg (Sennrich et al., 2009), the Nottingham Corpus of Geospatial language (Stock et al., 2013), or a corpus of reports covering 18 years of international climate negotiations (Venturini et al., 2014). Comparison of a domain-specific corpus (e.g., a corpus of travel reports) to a general one (e.g., the British National Corpus) can reveal which words or phrases are distinct or appear statistically significantly more or less often than in a domain-specific corpus (Kilgarriff, 2001) as we will see later in Section 3.4.1.

An additional important property of a corpus in an environmental context is its **spatial coverage**, that is to say the distribution of places described in a corpus. For example, the Corpus of Lake District Writing (CLDW) (Butler et al., 2017) covers the English Lake District, while the Palimpsest corpus (Alex et al., 2016) is a collection of fictional and historical texts related to Edinburgh, Scotland.

The massive growth in the availability of digitised texts has greatly increased the range and variety of resources from which corpora can be created. These include very large collections of digitised books, the web itself and more specific collections such as newspaper archives, collections of legal documents, scientific articles or some of the corpora described above. Examples of such collections include digitised books hosted by Google Books[4] or Project Gutenberg[5], crawls of the web such as the Common Crawl[6], historical newspaper archives as provided by the Chronicling America project[7] and archival records such as Hansard recording UK parliamentary debates[8]. However,

[4] https://books.google.com/

[5] https://www.gutenberg.org/

[6] https://commoncrawl.org/

[7] https://chroniclingamerica.loc.gov/

[8] https://api.parliament.uk/historic-hansard/index.html

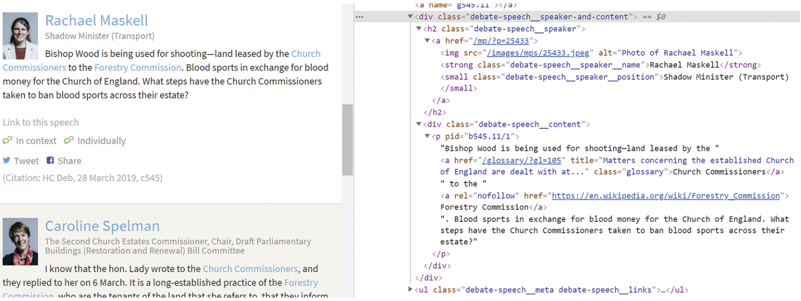
such resources are often too general to allow exploration of specific research questions, and the first step towards such an analysis is defining what properties the required corpus should have (e.g., language, spatial coverage, domain, etc.). Having defined such properties, we can extract potentially relevant documents by, amongst other possibilities, restricting ourselves to a particular genre (e.g., natural history writing in the Country Diaries of the Guardian[9]), using search terms relating to concepts in a given domain (e.g., 'glacier', 'ice', and 'mountain') or compiling lists of place names in the region of interest (e.g., 'Loch Lomond', 'Balloch', etc.) to extract documents referring to specific locations (Davies, 2013).

How can we work with sources such as those described above? Many are too large to simply copy and process in their totality locally. Often, the publishers of such data make them available through an **application programming interface (API)**. An API allows us to query a system with a defined request, and in return receive a response message. Such requests are usually returned as structured data in formats such as Extensible Markup Language (XML) or JavaScript Object Notation (JSON). These are typically hierarchical and allow data to be explored using attribute-value pairs. Thus, in the JSON fragment of a parliamentary debate from Hansard shown in Listing 3.1, "name" is an attribute with the value "Rachael Maskell". Since 'attribute-value' pairs are stored hierarchically, "name", "party" (political affiliation), and "constituency" (the geographic area represented by a Member of Parliament in the UK) are all accessible as child attributes of the parent attribute "speaker", in this case associated with "hdate", a date, presumably on which the question referred to by "body" was posed.

```json
{
   "body": "Bishop Wood is being used for shooting--land
       leased by the Church Commissioners to the Forestry
       Commission. Blood sports in exchange for blood money
        for the Church of England. What steps have the
       Church Commissioners taken to ban blood sports
       across their estate?",
   "hdate": "2019-03-28",
   "speaker": {
      "name": "Rachael Maskell",
      "party": "Labour/Co-operative",
      "constituency": "York Central"
   }
},
```

Listing 3.1: A JSON fragment from a UK parliamentary debate retrieved using the API provided by TheyWorkForYou (https://www.theyworkforyou.com/api/).

[9] https://www.theguardian.com/environment/series/country-diary



Rachael Maskell
Shadow Minister (Transport)

Bishop Wood is being used for shooting—land leased by the Church Commissioners to the Forestry Commission. Blood sports in exchange for blood money for the Church of England. What steps have the Church Commissioners taken to ban blood sports across their estate?

(Citation: HC Deb, 28 March 2019, c545)

Caroline Spelman
The Second Church Estates Commissioner, Chair, Draft Parliamentary Buildings (Restoration and Renewal) Bill Committee

I know that the hon. Lady wrote to the Church Commissioners, and they replied to her on 6 March. It is a long-established practice of the Forestry Commission, who are the tenants of the land that she refers to, that they inform

```
    <a name="g545.11"></a>
  ▼<div class="debate-speech__speaker-and-content"> == $0
    ▼<h2 class="debate-speech__speaker">
      ▼<a href="/mp/?p=25433">
          <img src="/images/mps/25433.jpeg" alt="Photo of Rachael Maskell">
          <strong class="debate-speech__speaker__name">Rachael Maskell</strong>
          <small class="debate-speech__speaker__position">Shadow Minister (Transport)
          </small>
        </a>
      </h2>
    ▼<div class="debate-speech__content">
      ▼<p pid="b545.11/1">
          "Bishop Wood is being used for shooting—land leased by the "
          <a href="/glossary/?gl=105" title="Matters concerning the established Church
          of England are dealt with at..." class="glossary">Church Commissioners</a>
          " to the "
          <a rel="nofollow" href="https://en.wikipedia.org/wiki/Forestry_Commission">
          Forestry Commission</a>
          ". Blood sports in exchange for blood money for the Church of England. What
          steps have the Church Commissioners taken to ban blood sports across their
          estate?"
        </p>
      </div>
    </div>
  ▶<ul class="debate-speech__meta debate-speech__links">…</ul>
```

Figure 3.1: Example of a webpage with its HTML structure.
Source: https://www.theyworkforyou.com/debates/?id=2019-03-28b.545.8#g545.11

Using an API and search terms to extract relevant documents from a given collection is the preferred approach where possible since it makes work easier to reproduce, and APIs typically also provide, firstly, metadata expanding on the semantics of attributes and, secondly, explicit terms and conditions with respect to the use of data (Section 3.3). However, many sources have been digitised and made accessible online without the implementation of APIs allowing direct querying of documents. For example, the content retrieved using an API query in Listing 3.1 is also accessible as a web page (shown on the left of Figure 3.1). The raw HTML used to render this web page is shown on the right of Figure 3.1. Inspecting this source, it is possible to identify classes we are particularly interested in, for example, the question ("body" above) is stored in the class "debate-speech__content" and the speaker ("name") in class "debate-speech__speaker__name". We can access these classes and extract the information they contain using a web scraper. Different programming languages have libraries specifically created for this task, for example, Scrapy[10] and Beautiful Soup[11] in Python, rvest[12] in R, and JSoup[13] in Java. This approach allows us to extract the content of a single web page or a family of web pages with the same or very similar structures. It is therefore well-suited to automatically extracting content from consistently structured web pages such as Wikipedia or content that follows a well-defined template, for example, reports produced by government agencies.[14] In doing so, it is important to consider any copyright and ethical considerations (Section 3.3).

[10] https://scrapy.org/
[11] https://pypi.org/project/beautifulsoup4/
[12] https://cran.r-project.org/web/packages/rvest/
[13] https://jsoup.org/
[14] Often websites like Wikipedia prefer that you download their data as packaged database dumps (see https://dumps.wikimedia.org), rather than via web scraping, which can put strain on their web servers and slow down the website for casual users. It is important to read the terms of use of any website that you are planning to scrape and avoid pages that are explicitly banned from automated scraping.

When we explore environmental narratives using online sources, we are often interested in building a corpus of documents describing a particular theme and/or region. If we aim to explore different domains, genres or perspectives, we may also want to analyse different sources, and scraping a single website is no longer sufficient. Here, we can take a similar approach to that described above using an API to search a specific collection or using a general web search API to search the web as a whole. The first step in such work is compiling a list of search terms (e.g., place names in a given region and word associated with a particular topic). Search engine APIs usually return uniform resource locators (URLs) rather than full document text in a first step, and the content of these URLs can be extracted using the scraping methods described above (with the important difference that the structure of individual web pages is likely to vary widely). Since irrelevant documents are also highly likely to be returned (e.g., hotel rooms named after places), filtering steps are also necessary when building a corpus using this approach. The BootCaT tool[15] makes collecting documents using search engine APIs possible without programming experience and although it is limited to 100 URLs, is very useful in exploring and testing queries and ideas.

Corpora

Since we wished to explore how the Forestry Commission was discussed in the UK Parliament, we first looked for online sources of debates. Hansard is the official written record of British parliamentary proceedings, is available and searchable online at https://hansard.parliament.uk/ and has been used to explore for example infrastructure in the British Empire (Guldi, 2019). To search the collection for discussions about the Forestry Commission programmatically, we took advantage of a third-party API implemented by the UK-based organisation mySociety (https://www.mysociety.org/). The API (https://www.theyworkforyou.com/) allows us to query for information on a variety of dimensions, including debates held by the upper (House of Lords) and lower chambers (House of Commons) of parliament. In an initial exploration we searched Hansard for all mentions of 'Forestry Commission', returning the transcripts of debate, their dates, speakers and whether the debate took place in the Commons or Lords. For the House of Commons, the API returned 1985 documents, contrasting sharply with the 190 returned from the House of Lords. We quickly established that while the documents from the Commons dated back to 1919

[15] https://bootcat.dipintra.it/

(the year the Forestry Commission was founded), those from the Lords started only in 1999. We therefore decided to concentrate on Commons debates. In a second filtering step, we identified a large number (408) of very short documents with no speaker assigned. Our final corpus therefore contained 1577 documents from Hansard debates recorded in the House of Commons dating back to 1919.

When building a corpus, identifying appropriate sources and exploring their properties, before starting analysis is important. Doing so requires a basic knowledge of the expected properties of documents related to the theme (here, we know when the Forestry Commission was founded) and collection (we would expect broadly similar temporal periods in documents returned from the two chambers of the UK parliament).

Geograph texts can be downloaded as a single corpus, and we queried the complete database for all descriptions mentioning the Forestry Commission, identifying 3114 such texts.

3.3 Copyright and Ethics

In an era where large volumes of text are readily available online, it is all too easy to gain the impression that, quite literally, anything goes. We can, as was discussed above (Section 3.2), develop crawlers to scrape data and build corpora based on any content that is visible on the web. However, when we build such corpora we need to be able to answer two, linked, but separate questions. Firstly, are we legally allowed to use these texts in the way that we plan to? And secondly, and equally importantly, are there ethical issues that should be considered before we commence our study? It is important to understand that legal and ethical standards change in space and time. For example, copyright laws vary according to legal jurisdictions and acceptable ethical practices change over time. In what follows, we give a non-exhaustive list of issues to consider when designing an experiment, and conclude with a checklist of questions to ask before starting work.

The increased importance of reproducibility in research has brought with it a recognition of the need to provide data and code together with scientific papers reporting on research results. This welcome development allows researchers to replicate existing results, and build upon them more easily than in the past. With respect to research on text, shared datasets allow the development of common baselines (e.g., with respect to identifying locations or characterising sentiment) based on published corpora and related annotations. Furthermore, given the complexity and challenges involved in building domain-specific corpora, reusing these for other research reduces duplication of effort and allows research to more directly and comparably build upon previous work.

However, before publishing a corpus, it is important to understand the notion of **copyright**. Simply put, copyright protects the creator of a work from its reproduction without their permission. Copyright laws vary widely geographically, but typically are of long duration, often extending 50 years or more beyond the death of the creator or author. Thus, books, newspaper articles, scientific papers and images are all usually protected by copyright which requires explicit permission for the reproduction and publication of material. Copyright holders may give permission for academic use, but simply by providing access to content, copyright holders do not relinquish their rights. In some countries, the notion of fair use allows limited quoting or reproduction of content in certain contexts, with, for example, quoting from a song or a book permitted as part of a review, reportage or even parody. In the UK for instance, academics and students carrying out non-commercial research are explicitly allowed to carry out text and data mining of sources to which they already have access through, for example, subscriptions[16].

For many works, information about copyright is explicitly provided. Thus, scientific publishers and newspapers publish copyright statements, and explain how works can be licensed for further use. Typically, such licensing is complex and may involve additional fees, based for example on the number of users accessing the content. One important development is the increase in the use of explicitly permissive licences, such as Creative Commons. Here, authors give permission for their work to be reused in different ways. The most open such license is Creative Commons Zero (https://creativecommons.org/share-your-work/public-domain/cc0/), which places a work in the public domain with no restrictions whatsoever. However, much more widespread are licences which require attribution of a work, restrict commercial reuse or prohibit derivative works.

In general, collections of unstructured text used for analysis of environmental narratives can be categorised with respect to licensing in four broad categories:

- Licence allows redistribution and adaption under no, or some conditions (e.g., Wikipedia and Geograph) allowing corpora to be created directly incorporating such material. Where licences vary, then care must be taken in merging materials.
- Curated corpora created by third parties and shared under clear licence conditions (e.g., Text+Berg where licence has been negotiated with copyright owners or the Corpus of Lake District Writing which consists of historical, out-of-copyright documents).
- Licensed corpora (e.g., GeoCLEF) of newspaper articles, where redistribution is subject to restrictions and permission from the licensee or copyright holder.
- Scraped corpora of content from the web, blogs and so on where fair use may be implicitly assumed but copyright is unclear.

[16] https://www.gov.uk/guidance/exceptions-to-copyright

In practice, it appears that many researchers working with text build corpora with limited regard to the situation with respect to licensing. For example, the Geograph project consists of millions of images and accompanying textual descriptions located all over the British Isles. Individual contributions are licensed with a Creative Commons BY-SA licence (https://creativecommons.org/licenses/by-sa/2.0/). This licence allows copying, redistribution and transformation of the content for any use, even commercially, under two conditions. Firstly, appropriate credit to the author must be given. Secondly, any new material developed based on the source must be distributed under the same licence conditions. We identified 32 research papers which had used these Geograph data, from many different research groups. Of these, only 17 attributed the data as required in the licence, and even less made their results available under an equivalent licence.

This lack of regard for clearly communicated licensing conditions for the reuse, adaption and distribution of text brings us to the question of ethics. Ethics have traditionally been policed academically by institutional review boards, whose domain has gradually extended from medical research, through social sciences to data analysis. Geographically, the degree of ethical scrutiny of research with respect to data and methods has varied, leading to different notions of acceptable ethical practice. Zimmer (2018) set out an ethical framework in the context of big data, exploring ethics in the context of particularly extreme example, where researchers scraped the content of an online dating website, arguing that the data were in any case public. Zimmer introduces some generally accepted principles of research ethics, including minimising harm, gaining informed consent and maintaining privacy and confidentiality. To these, we add an additional idea, transparency.

When using text to explore environmental narratives, we can aggregate individual texts to allow a macroanalysis, and zoom into particular details to perform microreading. These scales are important, as they also have ethical implications relating to the three principles identified by Zimmer. Macroanalysis typically obfuscates individual contributions, reducing our ability to identify individuals. Microreading, by contrast, emphasises context when reading a text, and zooms in to individual statements.

Minimising harm implies that participants, or in our case those who create content, are not subject to any harm through being involved in research. Analysing text, at first glance, appears to be a wholly innocuous activity. However, for example, by identifying illegal or controversial opinions in text, we can potentially expose authors to harm through, for instance, legal sanctions or unwanted online pressure. Using text to explore the properties of landscape may identify regions which are worthy of protection, and thus contribute to overall public good. But if text analysis can be used to identify such regions, then conversely it also has the potential to highlight regions where protection is no longer appropriate, at least according to our sources. Removing protection from an area may have long-term negative consequences, such as a reduction in tourist visits and income. Put simply, if we analyse environmental narratives with an expectation that the information derived can be used in decision-making, then thought should be given to the potential consequences of these decisions.

If individual texts are analysed and presented for microreading, then traditional ethics would require **informed consent**, where participants are informed in advance of the benefits and risks of participation, the aims of the research, and are given the opportunity to withdraw at any time. Text analysis rarely involves informed consent, since analysis is carried out on content produced for other purposes and often without the knowledge of the creators. Thought should be given to how this can be done ethically, for instance, by linking to rather than copying content, such that where creators delete it, it is no longer used in analysis or presentation of results. In particular cases – for instance, if mapping potentially controversial statements – it may be desirable to ask contributors for informed consent before analysis and publication.

Privacy and confidentiality are not only ethical issues but also legal ones. In Europe, the General Data Protection Regulation (https://gdpr-info.eu/) sets out clear rules with respect to the processing of personal information. Ethically though, irrespective of the legal situation, we need to consider the rights of individuals to privacy by, for example, not seeking to use other data to identify individuals and giving careful consideration to whether any personal information (e.g., age or sexual orientation) should be collected without specific, informed consent.

However, these principles bring with them other challenges. If we are to maintain confidentiality, that in turn implies not attributing material to its authors – which, of course, directly contradicts the need for attribution set out above. We therefore propose an additional ethical idea, which researchers should consider, **transparency**.

Transparency means that when we build a corpus, we make clear how we did so, what licences the content had, and link to the original materials rather than storing copies. It also implies that the creators of content have access to, and can comment on the results of any research, thus building and maintaining a dialogue about the use of text in research. Transparency allows individuals whose content has been analysed the opportunity to criticise our interpretation of their material and, potentially, to 'set the record' straight or ask for their material to be removed.

For the researcher starting out with text we make the following suggestions with respect to copyright and ethics:

- Identify a range of candidate text collections for the research question under investigation.
- Research, and document, the copyright conditions under which chosen text collections are published. Consider whether fair use is applicable.

- Note any conditions under which data can be used (e.g., attribution, non-commercial, share-alike).
- Ascertain whether research requires institutional ethics review.
- If analysis only involves macroreading, ensure that results are shared appropriately and discussion is possible.
- Where microreading of corpus is important, consider how contributions can be withdrawn, harm minimised, privacy and confidentiality maintained, and a transparent dialogue enabled.

Licensing of our collections

Hansard is the public record of the UK parliament, and it is published under an Open Parliament Licence[17], a very open licence, which allows commercial and non-commercial adaptation and exploitation subject to attribution. The TheyWorkForYou project stores these data in a more structured way, (https://www.theyworkforyou.com/) allowing querying using their API. Since Hansard is a public record, and the individuals we can identify are elected representatives, there are, we judge, no ethical issues in the use of these data. However, it is worth noting that in exploring historical archives we may uncover utterances which are no longer considered acceptable, and it is important to report on such material in context.

For Geograph, the data are published under a CC BY-SA licence. This in turn requires that firstly, we acknowledge individual authors of contributions we use and quote from in our analysis. Where we analyse the corpus as a whole (e.g., looking at the frequencies of individual words) we should acknowledge the creators of the corpus in a general sense. Secondly, this licence specifies that we should allow others to use our results under the same licensing terms (so-called share alike).

3.4 Corpus Linguistics and Natural Language Processing

Computationally analysing text can take a number of forms and is referred to in different fields as **corpus linguistics**, **natural language processing (NLP)**, or simply **text analysis** (Manning and Schutze, 1999; McEnery and Wilson, 2003). In this book we focus specifically on the processing of written language (text). Some areas of research on text aspire to what is referred to as general artificial

[17] https://www.parliament.uk/site-information/copyright-parliament/open-parliament-licence/

intelligence, investigating how computers can learn to "understand" language in the same way that people do. Although the development of such computational systems could in theory give us great insight into how people think about their environments, and great progress is currently being made, we are still far from building them (Suissa, Elmalech, and Zhitomirsky-Geffet, 2022). Natural language is often ambiguous and humans, as we have discussed above, bring a wealth of background knowledge when making sense of it. When we analyse text data we must, therefore, make many simplifying assumptions.

From a practical perspective, NLP methods today provide tools to statistically analyse written text to uncover patterns in language use, topics, sentiment and different perspectives among other things. Because these methods are computational they can be used on much larger amounts of text than humans are able to read in a limited time frame, and this is where their main power lies in helping us to understand and analyse environmental narratives. However, we must also remember the limits of the models to fully capture the many nuances of natural language that a typical human reader can easily grasp.

NLP is an extremely active area of research with thousands of new articles published each year, which push the envelope on state of the art results for various shared tasks. Increasingly these methods rely on complex deep learning models that require massive amounts of data, computational resources and energy to develop. Our goal in this section is not to provide a comprehensive summary of the more advanced techniques that are currently being developed, but rather to give an introduction to a selected suite of powerful and established methods that are well-suited for environmental narrative analysis. As we will see in the case studies described in the second part of this book, simpler, established methods can be quite effective when used appropriately and in practice are often much easier to apply.

Typically, there are two stages to any environmental narrative analysis that uses NLP. The first stage is **pre-processing** and primarily involves applying methods that translate the raw text in a corpus into a form amenable to computational analysis. Many pre-processing steps involve working with text to divide it into meaningful chunks that are obvious to a human reader. These might include identifying words, sentences, paragraphs or utterances from individual speakers. Normalisation tries to match words or sequences of words onto a single canonical form (e.g., working out that '12 pm' and 'midday', U.S. and USA, or 'Zurich' and 'Zürich' all convey the same meaning). Stemming (and a closely related technique, lemmatisation) reduces inflected words to root forms with the same aim (e.g., the stems of 'snowing' and 'snowed' are 'snow').

This stage also involves **encoding** the language in the texts as **features** as well as creating new features from the raw data. There are multiple levels of structure that humans use to make sense of natural language and thus there are multiple levels at which we might encode language as features. These levels span from individual (potentially normalised and/or stemmed) words and n-grams (sequences of words), parts of speech (e.g., adjectives, nouns) and other aspects of syntax and grammar all the way to semantics and discourse. For environmental texts we might also emphasise certain features related to the domain, for example, using lists of nouns describing natural features such as 'hill', 'mountain', 'river' and so on.

A second stage often involves analysis of features to answer questions about the language in the corpus. Two important categories of NLP analysis are classification tasks and sequencing tasks. Applications of classification-based analysis in NLP include tasks such as topic classification, document similarity, sentiment analysis and stance detection. Sequencing-based applications include building translation or summarisation systems as well as interactive systems that generate answers to queries. Although both categories have potential utility in environmental narrative research, we focus primarily on classification here as it is much more straightforward to implement with existing tools, and has great utility in understanding environmental narratives.

3.4.1 Pre-processing and encoding natural language as features

A corpus of natural language text is, at its core, simply a collection of ordered words, sometimes organised into discrete documents (possibly along with metadata about those documents, such as the author or labels). The words and occasionally the individual characters that make up a document are the basic elements used to analyse text.

In some models, the order of the words in a document is not considered. These are called **bag-of-words (BOW)** models, and are predicated on the idea that the frequency of words in a document is enough of a statistical signal for us to discover meaningful information about the corpus' content. In other words, the sequence that the words occur in, which provides humans information about grammar and much of the meaning of the text, is ignored. While on the face of it a BOW model might appear overly simplistic, it can be surprisingly effective when we are dealing with large corpora.

Examining the count or **frequency** of a **term** (or token) is the simplest kind of BOW analysis we can perform. We use term to refer to both individual words as well as n-grams, sequences of adjacent words. For example, 'adjacent words' is a 2-gram, also sometimes referred to as a bigram. However, simple frequency measures do not alone tell us how important a term is in a given document, since we first have to control for how common these words are in language in general, and in a corpus in particular. Some words (e.g., 'the', 'in', 'of') occur often in language in general and high frequency is an artefact of general language use. The simplest approach to dealing with this issue is the removal of these so-called **stop words** using lists of common terms in a language to retain only words thought more likely to contain meaning.

Corpus linguistics

Table 3.1 shows some basic properties of our corpus. Note how the number of tokens decreases based on what we treat as tokens. The lower value (not including punctuation) is more representative for most of the methods which will be applied here, since these are based on the BOW model described above, ignoring punctuation. Note also the mean (810) and median number of tokens per document (462). These numbers are especially interesting if we compare them to other corpora or, for example, if we compare different time periods within the same corpus. Information about the size of the corpus and its language should be always included in its description.

Table 3.2 illustrates, after normalising to lower case, token counts for the 20 most frequent words in our corpus. The words in the all tokens list convey no semantics with respect to the topic of the debates and include articles ('a', 'the'), prepositions (e.g., 'of', 'to', 'in'), conjunctions ('and', 'that'), verbs ('be', 'have'), and pronouns ('i', 'we'). Such words are typically included in stop word lists and removing these results in a revised frequency list. This list contains many words related to the general business of parliament. Some of these are obvious, e.g., 'government', 'people', 'minister', 'house'. Others require more knowledge of the language used in parliamentary debates, for example, members are referred to as "the right honourable member" or "my honourable friend" in speeches, and all of these words (or abbreviations thereof) appear as frequent tokens (e.g., 'hon', 'member', 'right', and 'friend'). This second list, we hypothesise, thus tells us something about the nature of parliamentary debates in general, but little or nothing about those discussing the Forestry Commission (apart from the obvious appearance of our search terms 'forestry' and 'commission' and, possibly, 'scotland', reflecting that country's much more forested nature).

Count	Total corpus	Mean per document	Median per document
All tokens	1450791	920	533
Tokens without punctuation	1277661	810	462
Sentences	48136	31	17

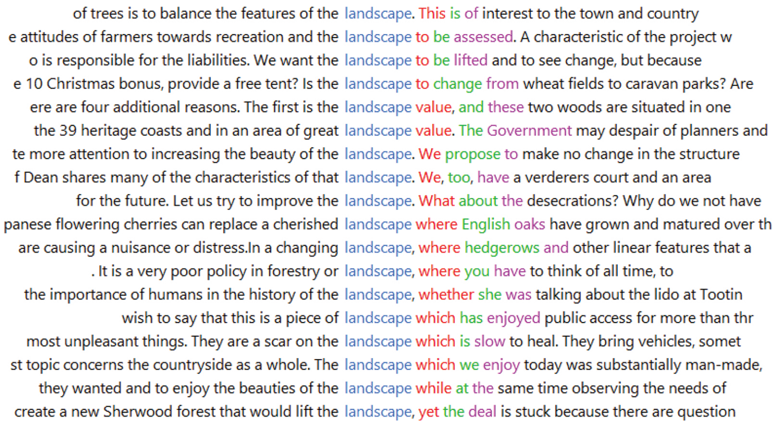
Table 3.1: Basic counts for corpus.

Rank	All tokens	Count	All tokens No stop words	Count
1	the	98,580	hon	5299
2	of	49,039	government	4062
3	to	41,450	commission	3570
4	and	32,112	one	3512
5	in	30,294	forestry	3405
6	that	28,290	land	2843
7	a	22,349	member	2436
8	is	21,433	people	2397
9	i	17,995	right	2303
10	for	15,006	many	2205
11	it	14,249	bill	2160
12	be	13,168	minister	2101
13	have	11,097	house	2038
14	not	10,155	new	1985
15	we	9927	friend	1971
16	are	9604	made	1871
17	on	9372	time	1845
18	which	9023	Scotland	1811
19	this	8667	great	1784
20	as	8279	years	1774

Table 3.2: 20 most frequent terms in corpus before and after filtering for stop words.

We can also look at term usage across a corpus by weighting the importance/relevance of a term for a document in comparison to its use in a corpus overall. One popular method is **term frequency-inverse document frequency (TF-IDF)**. TF-IDF is the product of the term frequency in a given document with the logarithm of the inverse fraction of the total number of documents in which the term occurs. It gives words common in all documents across the corpus lower scores than those which are common in a small sub-set of documents from the corpus. TF-IDF is a simple but very effective way of ranking the importance of terms in documents, where the corpus overall is used to calculate IDF, and in corpora, where another corpus (e.g., the British National Corpus described above) is used to estimate "normal" use of a term.

Having identified potentially interesting words, we can explore individual words qualitatively and quantitatively using a variety of methods. Perhaps the simplest is to use the idea of **concordance** to explore a word in context. Here,



```
                of trees is to balance the features of the landscape. This is of interest to the town and country
    e attitudes of farmers towards recreation and the landscape to be assessed. A characteristic of the project w
      o is responsible for the liabilities. We want the landscape to be lifted and to see change, but because
   e 10 Christmas bonus, provide a free tent? Is the landscape to change from wheat fields to caravan parks? Are
       ere are four additional reasons. The first is the landscape value, and these two woods are situated in one
        the 39 heritage coasts and in an area of great landscape value. The Government may despair of planners and
    te more attention to increasing the beauty of the landscape. We propose to make no change in the structure
    f Dean shares many of the characteristics of that landscape. We, too, have a verderers court and an area
                 for the future. Let us try to improve the landscape. What about the desecrations? Why do we not have
 panese flowering cherries can replace a cherished landscape where English oaks have grown and matured over th
      are causing a nuisance or distress.In a changing landscape, where hedgerows and other linear features that a
                  . It is a very poor policy in forestry or landscape, where you have to think of all time, to
       the importance of humans in the history of the landscape, whether she was talking about the lido at Tootin
                       wish to say that this is a piece of landscape which has enjoyed public access for more than thr
      most unpleasant things. They are a scar on the landscape which is slow to heal. They bring vehicles, somet
    st topic concerns the countryside as a whole. The landscape which we enjoy today was substantially man-made,
        they wanted and to enjoy the beauties of the landscape while at the same time observing the needs of
    create a new Sherwood forest that would lift the landscape, yet the deal is stuck because there are question
```

Figure 3.2: Example concordances in the Hansard corpus.

the corpus as a whole is searched for all instances of a potentially interesting word, and these are then visualised in within the sentence or passage of text in which they occur.

Concordance

In Figure 3.2, example concordances with 'landscape' in our Hansard corpus are shown, highlighting the three words occurring after landscape. The sentences as a whole quickly reveal different ways in which landscape is discussed, for example with respect to its components, such as 'English oaks', 'hedgerows', 'woods', 'wheat fields'; change and value, such as 'change', 'desecration', 'lifted'; and access and recreation, such as 'public access', 'carvan parks'.

The notion of **co-occurrence** takes this one step further by looking at individual words which are found within a given distance of a search term or "node". For example, in the previous sentence the words 'a:2', 'given:1', and 'of:1' co-occur within a two word window of 'distance'. If we then remove stopwords we are left only with 'given'. Note that the order of steps is important here – if we first remove stopwords and then identify co-occurrences within a two-word window we find 'within:1', 'given:1', 'search:1', and 'term:1'. The influence of such seemingly minor choices can be very important, and reporting these choices is crucial if research is to be reproducible.

In a large corpus we can explore such co-occurrences in detail, and in particular look for meaningful combinations of words, termed **collocates**. We can find collocates in a corpus by looking for statistically significant co-occurrences. Statistical significance implies that the two words co-occur together more than we would expect by (random) chance given the overall frequencies of words in a corpus. We can use a wide range of measures to calculate statistical significance, all of which can be interpreted in slightly different ways. For example 'mutual information' favours exclusive and infrequent co-occurrence, while the 'T-Score' favours non-exclusive and frequent co-occurring terms (Brezina, 2018). Significant co-occurring terms can be ranked using simple frequency, or measures such as mutual information and T-scores. In analysing environmental narratives we are interested in not just statistically significant collocation, but also those which convey meaning in a specific context. Such phrases are recognisable in language, familiar to native speakers and seemingly logical substitutions sound clumsy or wrong. For example, though 'white mountains' and 'snowy mountains' contain similar information about the visual appearance of snow-covered distant mountains, the latter is a much more natural construction.

These approaches can be adapted by attaching more semantics to the corpus. Perhaps most simply, words can be merged to group those with the same meaning using normalisation, stemming and lemmatisation as described above. Normalisation approaches might include reducing all words to lowercase, creating canonical forms of words including diacritics, merging singular and plural forms of words and resolving different spellings to a single canonical form (e.g., 'colour' vs. 'color'). All of these methods can be further refined by also including information about parts of speech through **part-of-speech (POS) tagging**. A POS tagger assigns each word in a document a tag, such as verb, noun, adjective, preposition and so on. A related family of tools, dependency parsers analyse the grammatical structure of sentences and transform them to so-called dependency trees.

Dependency parsing

We used the Python library spaCy[18] to process the following fictive sentence: 'Beautiful, peaceful landscape in Bacton Woods'. The dependency tree shown in Figure 3.3 includes dependency labels (amod - adjectival modifier, prep - prepositional modifier, pobj - object of preposition, and compound) and part of speech tages for each words (ADJ - adjective, NOUN, ADP - adposition, PROPN - proper noun).

The compound place name 'Bacton Woods' is correctly recognised as such and the words it consists of are correctly labelled as proper nouns. Both 'beautiful landscape' and 'peaceful landscape' are identified as adjectival modifiers. Such compounds can be useful as features for machine learning as will be described in more detail in Section 3.4.2.

[18] https://spacy.io/api/annotation#dependency-parsing

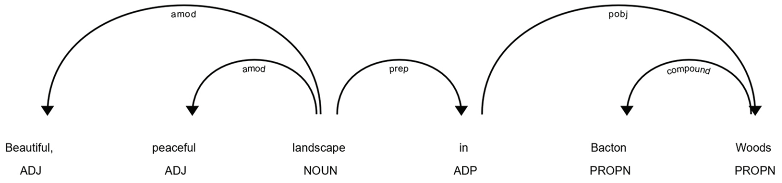


Figure 3.3: Sample dependency tree.

Comparison with a general corpus of English

One approach to finding more semantically interesting terms in the frequency lists discussed above is to compare frequent adjectives in our debates to frequencies in a general corpus of English (the British National Corpus [BNC]), which contains both written and spoken English). In Figure 3.4 we plot the ranks of the 20 most

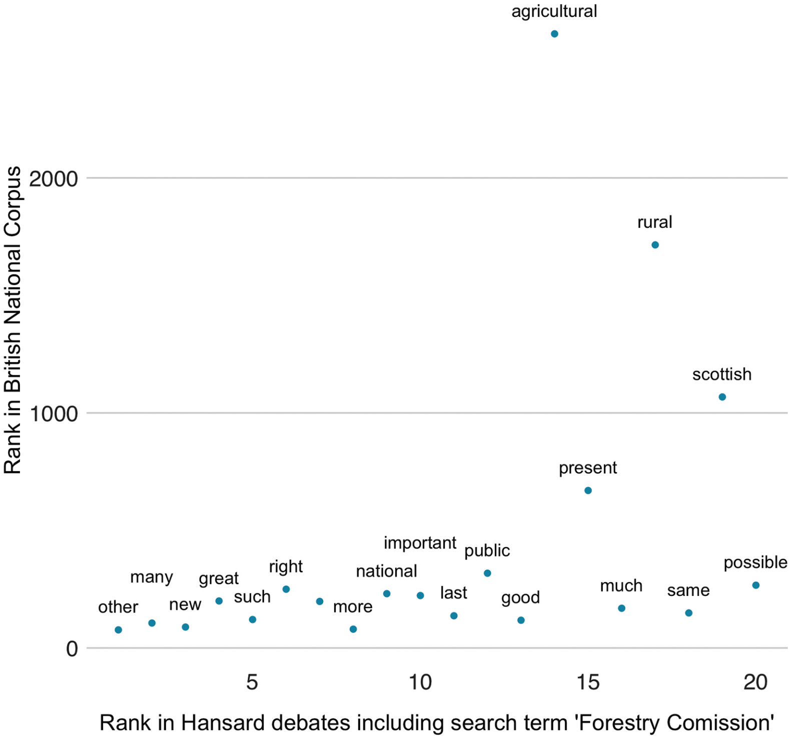


Figure 3.4: Adjectives ranked in BNC and the Hansard corpus.

frequent adjectives in our Hansard corpus against their corresponding ranks in the BNC. Many of these adjectives are relatively common in the BNC, with 16 of the 20 in the most 500 frequent words found in the BNC. Three however, appear to be much rarer in general text, 'scottish', 'rural', and 'agriculture', all of which have ranks higher than 1000. These frequencies suggest that agriculture and rural landscapes are often discussed in relation to the Forestry Commission, and that Scotland often appears to be referred to in this context.

These results are obvious, but they illustrate how we can preprocess our corpus to gain first insights into important topics. The choices we make along the way (e.g., in normalising text, using POS taggers and stemming word to their roots) all influence results in both predictable ways (e.g., stemming "year" and "years" will aggregate all counts referring to time in this way) and less obvious ways. For instance, roughly 50% of the instances of forestry were classified as nouns and 50% as adjectives.

Co-occurrence and TF-IDF

Another way to finding semantically interesting terms is to use TF-IDF. However, first, since our initial hypothesis was that discussions of the Forestry Commission would focus on negative impacts on landscape, we decided to explore co-occurrence with landscape. As explained above, we had to choose some parameters for our search window in the corpus. We decided to retrieve all instances where a word co-occurred more than three times with landscape, in a search window spanning three words before and after our node (landscape). We ranked co-occurring terms by both raw frequency and mutual information (recall this measure favours exclusive, infrequent co-occurrence) (Table 3.3).

The raw frequencies reveal that landscape is often referred to in terms of beauty, and by looking at ranks based on mutual information we see that this relationship is indeed specific to landscape and is captured in related words such as 'beauties' and 'beautiful'. Furthermore, the importance of change and preservation is captured by terms such as 'change', 'lift', 'enhance', and 'unique'. These results suggest that landscape does indeed appear to be discussed in the context of impacts.

TF-IDF can reveal the relative importance of these words and provide some hints as to context. For example, the very first mention of 'landscape' in our texts in 1925 (1925-03-30a.992.4.txt)

Co-occurring term ranked by frequency (count)	Co-occurring term ranked by mutual information (count)
beauty (13)	lift (3)
change (9)	coverage (4)
rural (7)	beauties (3)
forestry (6)	accustomed (3)
countryside (6)	enhance (4)
features (6)	beauty (13)
areas (5)	features (6)
farmers (5)	unique (3)
coverage (4)	beautiful (3)
hon (4)	wildlife (3)

Table 3.3: Ten most frequent terms co-occurring with landscape ranked by frequency and mutual information.

> seems to be in the context of roads with the following terms highest ranked according to TF-IDF: [‘roads’, ‘road’, ‘fund’, ‘tolls’, ‘unclassified’, ‘ought’, ‘gbp’, ‘blind’, ‘corners’, ‘42000000’]. A later text, reporting a debate in 1977 (1977-05-05a.705.4.txt), seems to focus on recreation, since its highest ranked TF-IDF terms are: [‘recreation’, ‘sport’, ‘recreational’, ‘facilities’, ‘education’, ‘sports’, ‘schools’, ‘regional’, ‘enthusiasm’, ‘fringe’].
>
> However, to understand the whole picture we need to move from a bag of words representation, or “macroanalysis”, and start to explore in more detail ways in which landscape has been discussed in parliamentary debates over time through ‘microreading’.

Reading of individual texts, often termed in the context of computational analysis **microreading**, is an obvious, and important, but sometimes neglected task (Jockers, 2013)[19]. It involves, as discussed in the introduction, reading and interpreting individual passages or texts that have been identified as potentially of interest computationally, and is closely related to the notion of hermeneutic analysis introduced earlier. In microreading, we use our knowledge of the corpus and associated context to interpret and refine our understanding of text.

[19] Macroanalysis and microreading are also often termed distant and close reading (Moretti, 2013). We prefer the macroanalysis/ microreading distinction here, since it emphasises that computational methods often do not involve any reading in a traditional sense, but rather a macroanalysis of, for example word frequencies.

Microreading

If we zoom into the texts we identified above, then we quickly find that the speaker in 1925, Lieut-Colonel Wilfrid Ashley of the Conservative Party, was, as suggested in the TF-IDF analysis, concerned with the link between roads, the landscape, and trees. Only by reading the text though do we see how his speech mirrors the original text analysed by Graham Fairclough, discussing roads in terms of both progress and aesthetics, and proposing trees as a way of mitigating their ugliness.

'The origin of this Bill is rather quaint. The year before last, when I had the honour of being Parliamentary Secretary to the Ministry of Transport, it was brought home to me very forcibly that these new roads which had been constructed, and were being constructed in the vicinity of the Metropolis, however excellent they might be from the transportation point of view and however useful from the national aspect are extremely ugly. I think the House will agree with me that a great wide stretch of road surface, in most parts bounded by concrete posts bound together by iron wires, is not a very graceful or grateful addition to the landscape. So I went into the matter rather fully, and came to the conclusion that, if proper trees be planted alongside some of these great roads, it would, at any rate, in a few years take off the bareness of the aspect and replace many trees which had had to be cut down when these new roads were made.'

The second text, a speech by Mr Kenneth Marks (Labour Party) also demonstrates nicely that TF-IDF correctly identified the importance of recreation. However, contrary to our original expectations, the Forestry Commission is here being discussed in a positive sense with respect to its impact on landscape.

'The commission has been helping with the reclamation of derelict land in river valleys, tree planting, general landscape improvement works and provision for informal recreation facilities, such as picnic areas, footpath and bridleway systems, and information facilities. The commission has also engaged the Civic Trust for the North-West to carry out an experimental project to promote recreational use, to encourage conservation and to stimulate local interest in the Tame Valley.'

Reading both of these texts also reveals something of the formal and rather complex nature of parliamentary language, which has also changed over time, a further issue for consideration in their analysis.

Of particular importance to environmental narratives are a family of methods collectively known as **named entity recognition (NER)**. These focus on identifying and classifying proper nouns such as the names of people, organisations and places. An important task in NER, and dealing with semantics more generally, is disambiguation. Lexical ambiguity refers to the phenomenon where a word has multiple possible meanings. Without additional context, we cannot resolve lexical ambiguity. For example, the word 'duck' can mean to crouch down or refer to a water dwelling bird. Given a POS tagger, and an associated sentence, these two meanings can be disambiguated since one sense of the word is a verb and the other a noun. A specific form of lexical ambiguity is referent ambiguity, where the same name is used to refer to multiple places (an extreme example is that there are more populated places named Springfield in the United States than there are US states). Assigning semantics to words is prone to errors which are often related to ambiguity.

Having identified named entities and dealt with referent ambiguity, they can be related to a unique instance (a specific person or place). By using external knowledge bases, such as place name gazetteers (Hill, 2009) containing information about these instances, we can add additional semantics to a text such as place types (e.g., village, forest, mountain) and other rich metadata (e.g., coordinates or bounding boxes) and link information to additional sources.

The quality of the tools used to perform these tasks varies greatly. For example, POS tagging in English is generally reliable. Dependency parsing within a sentence is highly effective, but linking entities across a narrative remains a difficult problem. NER is a vibrant research area, where much progress has been made, but often with a focus on particular classes of entity (such as organisations) and text genres (such as news reporting) and performance is often poor when methods are transferred to new genres or entity classes.

Named Entity Recognition

To demonstrate the potential and problems of an out-of-the-box solution for NER we processed two Geograph descriptions using the Python library spaCy[20]. The result of the first example is fully correct, 'Forestry Commission' is recognised as an organisation [ORG] and 'Balgownie' as a geopolitical entity or simply a location [GPE]. In the second example, we also see that many entities are labelled correctly, 'early January' is recognised as a date [DATE], 'North Norfolk District Council' as an organisation [ORG]. However, we also see that the locations were not recognised as such, 'Bacton Woods' and 'Witton Woods' are both wrongly labelled (as organisation [ORG] and person [PERSON] respectively). spaCy is

[20] https://spacy.io/usage/linguistic-features#named-entities

trained on the corpus OneNotes Release 5.0[21], a collection of news, weblogs and transcribed telephone conversations. This example shows that this out of the box solution does not work perfectly on some of our texts, and that either use of additional rules or retraining the algorithm on annotated texts would be necessary.

Example 1:

Thinning out at Forestry Commission **ORG** mixed woodland at Balgownie **GPE** [22].

Example 2:

... are the prevalent woodland colours in early January **DATE** . Bacton Woods **PERSON** , also known as Witton Woods **ORG** , covers 113 **CARDINAL** hectares; the woodland is owned by the Forestry Commission **ORG** and partly managed by North Norfolk District Council **ORG** , who together form the Bacton Woods Countryside Partnership Project **ORG** [23].

Irrespective of the tool being used, manual annotation remains the gold standard for adding semantics to a corpus. Annotation is a time consuming but very valuable way of analysing a corpus. It can be used directly as an analytical tool, to validate results produced computationally, or to create training data used in machine learning approaches. Regardless of the application, annotation requires a set of rules defining categories for annotation and the rules used to identify them, some form of replication by multiple annotators (typically reported in the form of inter-annotator agreement after annotation of the same or overlapping corpora by independent annotators using the same rules), and a way of storing annotated texts for future use. Since annotation is a very common activity, many community standards already exist, such as the Text Encoding Initiative (TEI) guidelines[24]. Annotation tools aim to make the annotation process simpler, and are increasingly moving to online tools such as Inception (Klie et al., 2018) and Recogito (Simon et al., 2017), developed with a focus on annotation in the Spatial Humantities and compatible with common formats including TEI[25]. Annotation is closely related to "microreading" since both involve a detailed reading of the text, with the main difference relating

[21] https://catalog.ldc.upenn.edu/LDC2013T19

[22] Paul McIlroy, https://www.geograph.org.uk/photo/285266

[23] Evelyn Simak, https://www.geograph.org.uk/photo/650293

[24] https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html

[25] https://recogito.pelagios.org/

to purpose – microreading is often concerned with a qualitative interpretation, while annotation results are used as training, test and evaluation data in quantitative work. In practice, many projects develop bespoke annotation schemes specific to the task at hand, and Pustejovsky and Stubbs (2013) give a useful overview of a potential pipeline which they call the model-annotate-model-annotate cycle, emphasising the importance of iteratively modelling (i.e., specifying the concepts that should be annotated) and actually annotating data.

Text Encoding Initiative

The annotation can be done on different hierarchical levels, for example, we can annotate ‘early January’ simply as <date> or we can add elements ‘notBefore=“–01-01” notAfter=“–01-10”’ indicating when the date (or in this case time period) actually is. Similarly, the tag <placeName> can contain more detailed information, for example, it can be divided on <settlement> and <region> (Listing 3.2).

Listing 3.2: TEI example of a Geograph description contributed by Evelyn Simak.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
<teiHeader>
<!---...-->
</teiHeader>
<text>
<body>
<l>
... are the prevalent woodland colours in
<date notBefore="--01-01" notAfter="--01-10">
early January</date>.
<placeName>Bacton Woods</placeName>,
also known as <placeName>Witton Woods</placeName>,
covers 113 hectares; the woodland is owned by
<orgName>the Forestry Commission</orgName> and
partly managed by <orgName>North Norfolk District
Council</orgName>, who together form the
<orgName>Bacton Woods Countryside Partnership
Project</orgName>.
</l>
</body>
</text>
</TEI>
```

3.4.2 Classification

Often we are interested in analysing texts grouped by a common topic, gender of the writer, time period and so on. To do so, we can either perform **unsupervised classification** by grouping texts based on their statistical similarities and adding the labels to the emerging classes, or **supervised classification**, where classes are defined in advance.

For **supervised classification**, training data has to be either already available or created prior to the classification, often through the process of annotation, sketched above. For classification tasks, a typical annotation workflow includes the following steps:

- Identification of desired classes.
- Creation of the set of clear rules allowing independent annotators to annotate texts consistently. Commonly, a small random sample of the data is selected to refine the rules and give examples.
- Independent annotation and calculation of inter-annotator agreement. For this task around 10% of the randomly selected texts are suitable. Inter-annotator agreement is calculated using a statistical measure, typically Cohen's or Fleiss' Kappa, depending on the number of annotators (Landis and Koch, 1977; Pustejovsky and Stubbs, 2012).
- If inter-annotator agreement is acceptable for type of texts (e.g., lower Kappa is acceptable for complex historical texts) one annotator can proceed with the rest of the annotation. Otherwise, the rules should be refined, another random 10% selected and annotated until inter-annotator agreement reaches the desired value.

Annotation

Since our initial hypothesis was that the Forestry Commission's actions are perceived as leading to negative impact on landscapes, we set out to classify all the Hansard texts containing the word 'landscape' into three classes: 'negative impact', 'positive impact' and 'neutral'. We expected the Hansard texts to contain a limited number of clearly opinionated texts towards the Forestry Commission. Therefore, we added another collection of texts – Geograph[26] – to enrich our corpus with texts more likely to contain positive or negative sentiments.

[26] https://www.geograph.org.uk/

To accomplish this task we, firstly, created a set of the following rules:

- 'Negative impact' included descriptions of current and former negative consequences of the Forestry Commission strategies, infrastructure or negative references to Forestry Commission buying of land practices. Examples: *The horrible trees on the left are privately managed and the horrible trees on the right are in a Forestry Commission holding*[27].; *This is the sort of planting which got Forestry Commission woodland such a bad name*[28].
- 'Positive impact' included texts describing positive influence on landscape, such as actions towards revival of native wood, positive effect on biodiversity and creation of infrastructure for recreation. Examples: *The Forestry Commission are encouraging the regrowth of natural woodland species in the Knapdale Forest*[29].; *This area of heathland and bog would be inaccessible to walkers without footbridges like this one, constructed by Forestry Commission engineers*[30].
- 'Neutral' descriptions include factual statements or describe effects on landscape without positive/negative judgement. Examples: *A Forestry Commission house in Penninghame Forest*[31].; *The forest had been replaced by spruce plantations here by the Forestry Commission. Policies have changed and this area is likely to revert to oak in the future, now that the spruce has been removed*[32].

Two annotators then annotated 15 Hansard texts according to the rules described above. Eleven of 15 descriptions (73%) were identically annotated by both annotators; however, Cohen's Kappa of 0.58 is moderate (Table 3.4), therefore, a further random 15 texts were selected, for which the annotators reached the substantial agreement of 0.78.

Cohen's Kappa is calculated based on a confusion matrix (Table 3.5), where the results of one annotator are

[27] *Richard Webb, https://www.geograph.org.uk/photo/166532*
[28] *Barbara Cook, https://www.geograph.org.uk/photo/104209*
[29] *Patrick Mackie, https://www.geograph.org.uk/photo/245251*
[30] *Jim Champion, https://www.geograph.org.uk/photo/92418*
[31] *Oliver Dixon, https://www.geograph.org.uk/photo/172754*
[32] *Richard Webb, https://www.geograph.org.uk/photo/187987*

Kappa statistic	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

Table 3.4: Relation between Kappa statistics and strength of agreement as proposed by Landis and Koch (1977).

	Negative	Neutral	Positive	Row totals
Negative	3	0	0	3
Neutral	2	6	2	10
Positive	0	0	2	2
Column totals	5	6	4	15

Table 3.5: Confusion matrix of the first annotation of the Forestry Commission text into three classes: negative, neutral, positive.

written horizontally, and the other vertically. Then, it is calculated according to the following formula:

$$k = \frac{\sum_a - \sum_{ef}}{n - \sum_{ef}},$$

where $\sum_a$ is sum of the agreements (the diagonal), n – total number of texts, and $ef = \frac{row\_total*column\_total}{overall\_total}$ – expected frequency per class.

Sum of the agreements:

$$\sum_a = 11$$

Total number of texts:

$$n = 15$$

Expected frequencies:

$$ef_{negative} = \frac{3*5}{15}, ef_{neutral} = \frac{10*6}{15}, ef_{positive} = \frac{2*4}{15}$$
$$\sum_{ef} = 5.53$$

Cohen's Kappa:

$$k = \frac{\sum_a - \sum_{ef}}{n - \sum_{ef}} = \frac{11 - 5.53}{15 - 5.53} = 0.58$$

Through the process of manual annotation the following distribution of texts according to classes emerged:

- negative 9
- neutral 24
- positive 13

Following the same rules, a single annotator annotated all Geograph texts containing 'Forestry Commission', which after filtering for identical descriptions contributed by the same author resulted in 3014 texts. The majority of the Geograph texts were also neutral:

- negative 105
- neutral 2687
- positive 322

However, it is important to note that many of the texts are not neutral in the traditional sentiment analysis sense. For example, the following description clearly shows negative sentiments towards Ordnance Survey mapping decisions, but does not provide any information about acceptance of the Forestry Commission actions: *A purple mess cluttering up the map states that this is Forestry Commission land*[33].

Having annotated data, for example into binary (e.g., positive, negative), nominal (e.g., forest, meadow, urban, lake) or ordinal (e.g., sentiment ranging from very negative through neutral to positive) classes, then it is possible to fit statistical models to text features or train machine learning models using text features. The most common representation of a text is through the so-called feature vectors (see Section 3.4.1). The simplest feature we can use in text processing is a vector containing zeros and ones representing absence and presence of the n most frequent unigrams in the whole corpus.

For example, if the five most frequent unigrams in a corpus, after stop word removal are 'timber', 'recreation', 'tree', 'beach' and 'sea', then a text mentioning only 'timber' will be represented as the vector $[1, 0, 0, 0, 0]$, and texts mentioning only 'recreation' and 'tree' will be represented as $[0, 1, 1, 0, 0]$. Other features could include (typically normalised) frequency of unigrams, frequency of other n-grams, number of words belonging to the same POS (e.g., adjectives),

[33] https://www.geograph.org.uk/photo/1826512

or frequency of defined syntax dependencies (e.g., adjectival modifiers). Thus, feature vector representation of the following sentence: *Beautiful, peaceful landscape in Bacton Woods* based on frequency of nouns, adjectives and total number of words is $[n\_nouns, n\_adjectives, n\_words]$ or $[3, 2, 6]$. Other common feature types include presence or absence of words from lists of relevant terms (e.g., sentiment lexicons containing terms commonly associated with positive or negative sentiment) or more complex compound features, for example capturing the similarity of texts.

Having encoded texts into features, we can apply a variety of statistical and machine learning methods to predict how other documents should be classified, including general linear models, random forests, naïve Bayes, support vector machines and neural networks. In practice, some classifiers work better than others for text data—and some work better on smaller datasets or can be trained more quickly, while others are more effective on very large datasets. Naïve Bayes is a relatively simple probabilistic model that assumes that the words used in a document are statistically independent of one another. This assumption of independence means that naïve Bayes is prone to error but it also can discover words that are important indicators of a category even in quite small data sets. It is also very fast to train.

Overfitting is an important problem in machine learning, where good performance is possible on a training data because the model slavishly fits to individual data points, and thus does not generalise well when presented with an unseen set of feature values. An example of a classifier not prone to overfitting is the random forest classifier, since it creates random sub-sets of the features and builds smaller trees using these sub-sets. In contrast, more sophisticated classifiers using, for example, neural networks, are highly effective for large corpora, but they are prone to overfitting the training data when working with smaller data sets, and they require extensive computational resources to train.

Evaluation of such models can be carried out in a number of ways. Very common are the calculation of **precision**, **recall** and **F1** scores. Precision is the proportion of correctly classified texts. For example, in a corpus of 20 texts, if 10 texts were classified as positive but only eight were annotated as positive, then the precision would be 8/10 (0.8). Recall is a measure for the completeness of a result, and is the proportion of texts belonging to a class which we return. Thus, if a total of 16 texts were annotated as positive in our example, then the recall would be 8/16 (0.5). The F1-score is the harmonic mean of precision and recall.[34] In our case, F1 would therefore be 0.62. F1 is particularly useful in comparing performance of classifiers with different feature sets, but less illuminating in isolation. Depending on the task at hand we may choose to optimise for precision, aiming to have a classifier which makes as few mistakes as possible, or recall, returning as many relevant examples as possible.

[34] Calculated as $F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$.

In closing this section, it is worth remembering one final point. In recent years awareness has greatly increased of the environmental impacts of all aspects of human behaviour, and computational methods are no different. Considering the potential environmental impacts of, for example, training machine learning approaches to classification is an important consideration (Bender et al., 2021).

Classification

The Countryside Act 1968 extended the powers of the National Parks Commission, renaming it the Countryside Commission, and extending its conservation and recreational remits. It also gave the Forestry Commission the explicit power to enhance access to forests for enjoyment and recreation. Since all of our texts are attributed with a date, and 1968 falls almost exactly midway through the time period captured in our Hansard corpus, we hypothesised that changes in the language used with respect to debates might allow us to classify texts according to their data of publication. Since data of publication was given as metadata, we could use this date directly to group debates in two classes 'before' and 'after' 1968.

To do so we randomly divided all our texts on two halves of training and test data. Training data consisted of 416 texts before and 373 after the year 1968, while test data contained 421 texts before and 367 after 1968. We then trained a random forest[35] on our training data based on a range of features to classify texts and evaluated model performance on our test data set.

Using only the 300 most frequent unigrams, our model has an F1 of 0.808, already a relatively good performance, suggesting that language does change between these dates. After filtering out very short descriptions, containing less than 20 unigrams, F1 increased very slightly to 0.810. Using bigrams (e.g., 'climate change') and length of descriptions as features did not improve the prediction ability of the model. We had hypothesised that these features might be effective in capturing, on the one hand, new issues such as climate change, and on the other more or less controversial topics (as opposed to simple descriptions). We suspect that the total number of texts was too small for these to improve model performance in this case.

[35] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

An additional useful property of random forests is that a measure of the relative importance of each feature on the prediction is returned. Since features here are simply vectors of unigrams, we can explore which unigrams are most likely to allow us to classify texts temporally. The 25 most useful unigrams were: *environment, asked, constituency, present, ask, quite, change, end, like, countryside, management, work, friend, regard, men, committee, timber, acres, kind, important, public, secretary, government, agriculture, land*. Some of these (e.g., 'environment', 'timber', 'change', 'countryside', 'public') may reflect changes in the topics being discussed with respect to forests, and potentially a move towards recreation and conservation and away from timber production.

About 70 descriptions in each class were classified wrongly. Many of these texts belong to the years around 1968. It is clear that themes of debates do not change sharply in 1968. However, the greatest number of wrongly classified texts (eight) was in 1978. A microreading analysis of these texts shows that many wrongly classified texts in 1978 mention the Forestry Commission in passing as one of examples of organisations as in the text below (1978-03-21a.1462.1.txt):

'Those powers at present are exercised not by Ministers of the Crown but by bodies such as the Forestry Commission and the Housing Corporation as specified in Schedule 7'.

Other texts refer to issues returned to throughout the history of the Forestry Commission as in the example below (1978-07-20a.904.2.txt):

'The total area of existing Forestry Commission forest in Wales is almost one-third of that in Scotland, but if we make allowance for Scotland's greater size we find that the proportion of land afforested is about the same'.

Using the simple features we selected for illustration here, we could not accurately predict the time of writing of such texts. In a typical iterative process, we might add additional features based on this microreading (e.g., presence of names of government organisations) to our feature vectors. In doing so however, it is important to retain 'unseen' data on which we test a final model. In this example we do not demonstrate an exhaustive list of such features, nor do we go beyond simple unigrams to more advanced features such as TF-IDF scores for terms, since our aim was to illustrate that a classifier can distinguish between two periods using simple features.

Regression

Since the metadata associated with debates is given on an interval scale as a year, we can also treat this problem as one of regression, and attempt to predict the year of a debate. Random forests can be used for both classification and regression[36], and making this change is straightforward computationally if the data provided furnish the necessary classes.

Using the same features as for our binary classifier (300 most frequent unigrams and filtering short texts with less than 20 tokens), we could train a random forest regression model with an r^2 of 0.413. This implies that about 40% of the variation in date attributed to a debate can be predicted by the choice of words alone, independent of their detailed context. Delving more deeply into the results, we note that only 168 descriptions (ca. 21%) are assigned a date with an error of more than 20 years. Once again, microreading, is an important way of exploring our corpus. For example, two contributions, from 2019 (2019-03-28b.545.11) and 1942 (1942-07-28a.330.7) were predicted with large errors of 1932 and 1990, respectively.

'Bishop Wood is being used for shooting—land leased by the Church Commissioners to the Forestry Commission. Blood sports in exchange for blood money for the Church of England. What steps have the Church Commissioners taken to ban blood sports across their estate?' (2019-03-28b.545.11)

'Sixty-three per cent. of the officials employed in Forestry Commission plantations in Wales are Welsh. Consequently there is not a preponderating number of English. Welsh officials are also employed in England, such interchanges being both desirable and necessary in the interests of the Forest Service as a whole.' (1942-07-28a.330.7)

Both of these texts are short, and without additional contextual information we suggest difficult or impossible for a human to date in a meaningful way. With additional information, the former text, discussing as it does 'blood sports' and related to the controversial ban and discussion around hunting in the UK in the early 21st century can easily be dated, but once again the features used in our model are not capable of identifying such changes. Rather we suggest, that such outliers can provide informative ways of zooming in and out from our corpus and identifying emerging themes of potential interest.

[36] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

In contrast to supervised classification methods which require annotated training data, **unsupervised methods** require no training data. Such methods are often used to explore corpora, and can provide powerful and straightforward ways of identifying common threads of discourse within a corpus. Perhaps the most well known such family of methods is topic modeling (Blei, 2012). One very commonly used form of topic modeling is **latent Dirichlet allocation (LDA)** (Blei, Ng, and Jordan, 2003). The basic notion, if not the mathematics underlying the approach, is relatively straightforward. Imagine a corpus of documents derived from a newspaper, where each article is stored as a document. Different newspapers publish different genres of articles, ranging, for example, from sports reporting through editorials and travel reporting to celebrity gossip, political reporting, local news and foreign affairs. A given article might though combine aspects of these genres, for example, a story reporting on Brexit negotiations combines both political reporting and foreign affairs. LDA, given the raw text of articles, attempts to do two things:

1. Identify a set of *n* topics which best differentiate individual documents based on the bag of words model;
2. Assign to every term in the probability that it belongs to a given topic.

The set of topics generated are based on the co-occurrence of terms in documents, and are often claimed to be easily interpretable (Chang et al., 2009). A new, unseen, document can then be associated with one or more topics, based on the terms making it up and their probability of belonging to individual topics. Topic modeling can therefore be used in three distinctive ways. Firstly, topic modeling can be used to explore a corpus. By generating a set of topics, examining the terms making up a topic and assigning labels to topics it is possible to in principle identify different forms of discourse. Importantly, the explorative process is sensitive to a range of input parameters, including crucially the number of topics and to the ways in which the corpus is pre-processed. Secondly, topic modelling can be used predictively, analogously to the supervised methods described above. For example, performing topic modelling and identifying three classes of readers' letters: those supportive of a government, those critical of a government and those discussing other matters. Given a new letter, we could then identify to which, if any, topic it best belonged. Finally, it is possible to use topic modelling to find semantically similar documents. For example, given a document that contains a specific mixture of topics, we can find other documents that share that particular mixture, or do so while also adding in another additional topic.

Topic modelling relies on the distributional hypothesis – neatly summed up by the linguist Firth in 1957 as 'You shall know a word by the company it keeps!'. The critical reader will, we hope, note that this also implies some dangers inherent in topic modelling. Topic modelling relies entirely on a bag of words model, and as we have seen the language used in different domains can vary considerably. Thus, language might not vary only according to the nature of a debate, but according to the domain or genre of writing, or indeed according to the backgrounds of the authors. Given, for example, a corpus of nature descriptions written by school children and adults, we might expect the age of the authors to be more decisive in determining topics than the content of the descriptions themselves.

Topic Modelling

In the classification step we annotated Geograph descriptions into three classes: negative, neutral and positive. The decision about the nature of the classes was made beforehand according to our hypothesis. However, these descriptions cover a variety of other topics and can be classified in many different ways. To explore these possibilities, we used a Python implementation of LDA[37]. One of the important decisions in topic modelling is the number of classes. There are methods to approach this problem quantitatively, but we simply experimented with 20 topics, and below are three examples.

Topic 1: Cycling/walking
Most probable words: *park, walkers, road, entrance, bit, car, narrow, route, signs, cycle, woodland, cycling, forest, heads, popular, lodge, farm, walking, land, village*

Example descriptions:
Looking east towards the Royal Oak pub, the "centrepiece" of Fritham *village*. The place is always busy with visitors on weekends and during the holidays. There are several Forestry Commission *car parks* which provide convenient access to the surrounding *Forest* (on foot, bike or horse)[38].

The tracks in the Forestry Commission *land* in the New *Forest* are very *popular* with *cyclists* and *walkers*[39].

Topic 2: Second World War airfields
Most probable words: *operated, road, mor, monadh, woodland, forest, wind, car, park, part, WWII, hill, following, used, airfield, loch, warning, timber, track, area*

[37] https://radimrehurek.com/gensim/
[38] Jim Champion, https://www.geograph.org.uk/photo/69306
[39] Nigel Mykura, https://www.geograph.org.uk/photo/6361160

Example descriptions:
Used by the Forestry Commission for storing *timber* & compost; a concrete approach *road* probably shows its origins as *WWII airfield* use[40].

The Forestry Commission *car park* at Janesmoor Pond is on the site of a Second World War *airfield* – odd bits of brick and concrete remain here and there. The gravel surface of the *car park* overlies the former service roads on the *airfield*. It is a very spacious *car park*, by New *Forest* standards, capable of accommodating large vehicles with horse trailers[41].

Topic 3: Access
Most probable words: *wood, road, country, access, park, also, route, woodland, public, forest, area, part, carved, track, footpath, car, conifer, mostly, plantation, accessible*

Example descriptions:
Although Forestry Commission, this *wood* is not mapped as *public access*. Presumably only leasehold – a confusing distinction to the *public*. However, a short distance along this *track* it is joined by a public *footpath* which starts at a different point on Fisher Lane[42].

Part of the Callan's Lane *Wood* Forestry Commission *public access* scheme, this *wood* forms the central *part* of the mixed *woodland*. This grass *track* leads to open farmland, with mixed broadleaved trees on the left and *conifers* on the right[43].

3.5 Where to Next?

Our aim in this chapter was to introduce a methodological tool box for undertaking the computational analysis of environmental narratives. This tool box contains not only concrete tools, such as those for part of speech tagging or NER, but also requires that we think about the questions we (can or should) ask of texts, narrative forms used in text, ways of sourcing or building corpora and often forgotten issues of copyright and ethics with respect to our sources.

As an example we set out to explore ways in which the Forestry Commission was discussed in the UK over the last 100 years in two contrasting corpora: speeches from the House of Commons and a collection of crowdsourced image descriptions. We hypothesised that the perceived negative impacts of the

[40] Mike Faherty, https://www.geograph.org.uk/photo/3884578
[41] Jim Champion, https://www.geograph.org.uk/photo/62968
[42] Robin Webster, https://www.geograph.org.uk/photo/335443
[43] Kate Jewell, https://www.geograph.org.uk/photo/405217

Forestry Commission on landscape would be visible in parliamentary debates, and that the visual impacts of forestry would be emphasised in the image descriptions. The methods we used demonstrate that landscape, and the impact of trees and forestry on it, were commonly discussed in our corpus. However, extensively annotating texts revealed that in practice many were neutral, and more texts in both collections were positive than negative. Automatic analysis of the Hansard corpus showed that language use does allow us to differentiate between texts, and indicated that the nature of issues discussed has changed over time, with something of a move towards recreation. Interestingly, the importance of recreation and access also emerged in the topic modelling of the Geograph collection, where contrary to our expectations not only what could be seen was discussed, but also the influence of forestry on access (thus implying that what the photographers do in a location was important) as well as historical land use (the Second World War airfields, perhaps reflecting the interests of the contributors).

Perhaps the most obvious result of our exploration is the importance of context and a constant interplay between source texts and computational analysis in our interpretation. This observation closes the circle of this chapter, returning us to the importance of the hermeneutic circle and emphasising the importance not of the tools we use to read, but rather the ways in which we combine these tools to gain knowledge.

3.6 Suggested Readings

Literary theory and spatial language

Bleicher, Josef (2017). *Contemporary Hermeneutics: Hermeneutics as Method, Philosophy and Critique*. Vol. 2. Routledge.

This book by Bleicher is an excellent introduction to the theory of hermeneutics for literary criticism.

Lakoff, George, and Mark Johnson (2008). *Metaphors we live by*. University of Chicago Press.

In this influential book Lakoff and Johnson show the deep role of metaphors, including spatial metaphors, in how people communicate through language.

Lotman, Yuri M (1990). *Universe of the Mind: A Semiotic Theory of Culture*. London: IB Taurus.

Lotman provides a view of narrative analysis from a cultural semiotics perspective with his approach to the topic of literary space being of particular relevance to environmental narratives.

Digital literary analysis

Moretti, Franco. Distant reading. Verso Books, 2013.

Jockers, Matthew L (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.

Both of these books explore how statistical and computational methods can be used to perform literary analysis on narrative texts. They make the case for 'reading' narratives in aggregate to understand the sociology of literature in new ways.

Corpus linguistics and construction

McEnery, Tony, and Andrew Hardie (2011). *Corpus linguistics: Method, Theory and Practice*. Cambridge University Press.

Kennedy, Graeme (2014). *An Introduction to Corpus Linguistics*. Routledge.

These are comprehensive textbooks that describe the use of corpus data to study language. Topics include both the construction of corpora and issues of ethics as well as methods of analysis.

Natural language processing

Manning, Christopher, and Hinrich Schutze (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Although the first edition of this book came out over 20 years ago, it remains an influential text and introduces the key statistical foundations for modern natural language processing.

Bird, Steven, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

This book provides an excellent introduction to practical tools for doing natural language processing using the Python programming language. All the computational methods described earlier in this chapter are represented with examples.

Blei, David M (2012). Probabilistic topic models. *Communications of the ACM* 55, no. 4: 77-84.

In this summary article, Blei introduces the family of probabilistic topic models with clear examples of their use.

References

Alex, Beatrice, Claire Grover, Jon Oberlander, Tara Thomson, Miranda Anderson, James Loxley, Uta Hinrichs, and Ke Zhou (2016). "Palimpsest: Improving assisted curation of loco-specific literature". In: *Digital Scholarship in the Humanities* 32.November 2016. ISSN: 2055-7671. DOI: https://doi.org/10.1093/llc/fqw050.

Augenstein, Isabelle, Leon Derczynski, and Kalina Bontcheva (2017). "Generalisation in named entity recognition: A quantitative analysis". In: *Computer Speech and Language* 44, pp. 61–83. ISSN: 10958363. DOI: https://doi.org/10.1016/j.csl.2017.01.012.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the dangers of stochastic parrots: Can language models be too big?" In: FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922.

Blei, David M (2012). "Probabilistic topic models". In: *Communications of the ACM* 55.4, pp. 77–84. DOI: 10.1145/2133806.2133826.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of Machine Learning Research* 3.Jan, pp. 993–1022. DOI: 10.5555/944919.944937.

Boell, Sebastian K and Dubravka Cecez-Kecmanovic (2010). "Literature reviews and the hermeneutic circle". In: *Australian Academic & Research Libraries* 41.2, pp. 129–144. DOI: 10.1080/00048623.2010.10721450.

Brezina, Vaclav (2018). *Statistics in corpus linguistics: A practical guide*. Cambridge: Cambridge University Press. DOI: 10.1017/9781316410899.

Butler, James O, Christopher E Donaldson, Joanna E Taylor, and Ian N Gregory (2017). "Alts, Abbreviations, and AKAs: Historical onomastic variation and automated named entity recognition". In: *Journal of Maps and Geography Libraries* 13.1, pp. 58–81. DOI: 10.1080/15420353.2017.1307304.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei (2009). "Reading tea leaves: How humans interpret topic models". In: *Advances in neural information processing systems*, pp. 288–296.

Davies, Clare (2013). "Reading geography between the lines: Extracting local place knowledge from text". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8116 LNCS, pp. 320–337. ISSN: 03029743. DOI: 10.1007/978-3-319-01790-7-18.

Egorova, Ekaterina, Thora Tenbrink, and Ross S. Purves (2018). "Fictive motion in the context of mountaineering". In: *Spatial Cognition and Computation* 18.4, pp. 259–284. ISSN: 15427633. DOI: 10.1080/13875868.2018.1431646.

Green, Georgia M (1996). *Pragmatics and natural language understanding*. London: Psychology Press.

Guldi, Jo (2019). "Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change". In: *Technology and Culture* 60.1, pp. 1–33. DOI: 10.1353/tech.2019.0000.

Hill, Linda L (2009). *Georeferencing: The geographic associations of information*. Cambridge: MIT Press.

Jockers, Matthew L (2013). *Macroanalysis: Digital methods and literary history*. Champaign: University of Illinois Press. DOI: 10 . 5406 / illinois / 9780252037528.001.0001.

Kilgarriff, Adam (2001). "Comparing corpora". In: *International Journal of Corpus Linguistics* 6.1, pp. 97–133. ISSN: 1384-6655. DOI: 10.1075/ijcl.6.1.05kil.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych (2018). "The INCEpTION platform: Machine-Assisted and knowledge-oriented interactive annotation". en. In: *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018). Santa Fe, USA: Association for Computational Linguistics, pp. 5–9. URL: http://tubiblio.ulb.tu-darmstadt.de/106270/.

Lakoff, George and Mark Johnson (2008). *Metaphors we live by*. Chicago: University of Chicago press.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data". In: *Biometrics* 33.1, pp. 159–174. ISSN: 0006341X. DOI: 10.2307/2529310.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, Nello Cristianini, Amy Gregor, Boon Low, Toby Atkin-Wright, Malcolm Dobson, and Richard Callison (2017). "Content analysis of 150 years of British periodicals". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.4, E457–E465. ISSN: 10916490. DOI: 10.1073/pnas.1606380114.

Lotman, Yuri M (1990). *Universe of the mind: A semiotic theory of culture*. London: IB Taurus.

Manning, Christopher D. and Hinrich Schutze (1999). *Foundations of statistical natural language processing*. Cambridge: MIT Press. ISBN: 0262133601. DOI: 10.1145/601858.601867.

Martin, Wallace (1972). "The hermeneutic circle and the art of interpretation". In: *Comparative Literature* 24.2, p. 97. ISSN: 00104124. DOI: 10.2307/1769963.

McEnery, Tony and Andrew Wilson (2003). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Moretti, Franco (2013). *Distant reading*. London: Verso Books.

Pustejovsky, James and Amber Stubbs (2012). *Natural language annotation for machine learning*. Vol. 1. Sebastopol: O'reilly. ISBN: 978-1-449-30666-3.

— (2013). *Natural language annotation for machine learning: A guide to corpus-building for applications*. Sebastopol: O'Reilly Media, Inc.

Sennrich, Rico, Gerold Schneider, Martin Volk, and Martin Warin (2009). "A new hybrid dependency parser for German". In: *Von der Form zur Bedeutung: Texte automatisch verarbeiten/From form to meaning: Processing texts automatically. Proceedings of the Biennial GSCL Conference 2009*, pp. 115–124.

Simon, Rainer, Elton Barker, Leif Isaksen, and Pau de Soto Cañamares (2017). "Linked data annotation without the pointy brackets: Introducing Recogito 2". In: *Journal of Map & Geography Libraries* 13.1, pp. 111–132. DOI: 10.1080/15420353.2017.1307303.

Stock, Kristin, Robert C. Pasley, Zoe Gardner, Paul Brindley, Jeremy Morley, and Claudia Cialone (2013). "Creating a corpus of geospatial natural language". In: *Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8116 LNCS, pp. 279–298. ISSN: 03029743. DOI: 10.1007/978-3-319-01790-7-16.

Suissa, Omri, Avshalom Elmalech, and Maayan Zhitomirsky-Geffet (2022). "Text analysis using deep neural networks in digital humanities and information science". In: *Journal of the Association for Information Science and Technology* 73.2, pp. 268–287. DOI: 10.1002/asi.24544.

Venturini, Tommaso, Nicolas Baya Laffite, Jean-Philippe Cointet, Ian Gray, Vinciane Zabban, and Kari De Pryck (2014). "Three maps and three misunderstandings: A digital mapping of climate diplomacy". In: *Big Data & Society* 1.2, p. 2053951714543804. ISSN: 2053-9517. DOI: 10 . 1177 / 2053951714543804.

Zimmer, Michael (2018). "Addressing conceptual gaps in big data research ethics: An application of contextual integrity". In: *Social Media+ Society* 4.2, p. 2056305118768300. DOI: 10.1177/2056305118768300.