*3 Quantifying lexical usage: vocabulary pertaining to ecosystems and the environment

Kate Wild,[1] Andrew Church,[2]
Diana McCarthy and Jacquelin Burgess

Abstract

A recent development in corpus linguistics has been the integration of critical discourse methodologies, which allow in-depth contextual and qualitative analyses, with corpus linguistic methodologies, which allow broader quantitative analyses. Our study is a contribution to this approach. We present the methods used in a study of vocabulary pertaining to the environment, undertaken as part of the UK National Ecosystem Assessment. A clear and replicable methodology was developed and applied to three custom-built specialised web corpora and a reference web corpus; automatic analysis of collocations found using the Sketch Engine was complemented by manual analysis; and a small-scale replicability check was carried out to ensure that investigator divergence was minimal. We outline the approach and some of the key findings, and we also suggest areas for further refinement/investigation.

Keywords:

1. Introduction

In this paper, we present the methods and findings of a corpus analysis of environmental vocabulary. The research was commissioned by the UK National Ecosystem Assessment (UK NEA) – a major study designed to provide a comprehensive overview of the state of the UK's natural environment (UK NEA, 2011). One element of the UK NEA's work was to discover how ecosystems and the natural environment are discussed in the public sphere in Britain.[3] The analysis undertaken for the UK NEA was based

[1]

[2] School of Environment, Cockcroft Building, University of Brighton, Lewes Road, Brighton, BN2 4GJ, United Kingdom.
Correspondence to: Andrew Church, *e-mail*: A.Church@brighton.ac.uk

[3] Ecosystems are defined as complexes where plants, animals, microorganisms and the non-living environment interact as a functional unit (UK NEA, 2011). The UK NEA focussed on

on UKWaC, a web corpus of over 1.5 billion tokens of UK English in the public domain, and three purpose-built specialised web corpora of language relating to ecosystems, which was taken from academic websites, government websites, and newspapers, NGO websites and blogs. Our research aims were to discover key collocates of selected environmental terms, and to establish whether they tend to be used in positive, negative or neutral contexts. To achieve these aims, we developed a methodology which combined automatic and manual approaches, and integrated methods from the fields of corpus linguistics and critical discourse analysis. Although our study was small-scale, with a strict time-limit of three months, we argue that our methodological approach is a useful step towards social-scientific rigour and replicability between researchers in corpus-based critical discourse studies.

Under Section 2 we describe the theoretical and methodological background to our study, and under Section 3 we give a detailed description and evaluation of our methodology. While the focus of this paper is on methodology, we briefly describe our key findings under Section 4; and under Section 5 we present our conclusions and suggestions for future research.

2. Background

2.1 Studies of environmental language

Over the past few decades there has been a growing body of research into the language used to discuss environmental issues, and in the 1990s a new discipline, *ecolinguistics*, emerged. One early area of interest was grammatical agency: for example, Goatly (1996) and Schleppegrell (1997) examined the extent to which passive and nominalised forms are used in texts concerning the environment so as to avoid ascribing agency – and thus responsibility – to people, organisations or practices. This type of analysis is broadly aligned with critical discourse analysis (CDA), which has been the theoretical framework underlying much recent work on environmental language: for example, Kuha (2009) analyses statements about global warming in US newspapers and whether they present climate change and its causes as a certainty or not; Alexander's (2009) monograph analyses the contexts and linguistic features of several texts relating to the environment; and Carvalho and Burgess (2005) examine how the political orientations of British broadsheet newspapers resulted in different framings of climate change between 1985 and 2003. Other studies approach the matter from the viewpoint of corpus linguistics (CL) and

---

ecosystem services, which are the diverse benefits people obtain from ecosystems such as water, food, flood control, nutrient recycling and leisure opportunities.

examine selected lexical items: thus, Nerlich and Koteyko (2009) survey compounds with *carbon* in English-language blogs and newspapers, while Grundmann and Krishnamurthy (2010) compare references to climate change and global warming in English, French and German web corpora. Bevitori (2010) examines representations of climate change in newspapers, using a corpus-assisted discourse methodology.

Our study builds on such research by examining a wide range of terms that relate to the environment, and more specifically to ecosystems, rather than a selection of particular lexical or grammatical features, and by drawing from both CL and CDA methodologies. In the following sections, we give more details about these methods and approaches.

2.2 Methodological framework I: corpus linguistics (CL) and critical discourse analysis (CDA)

As noted above, several studies of environmental language have used a CDA approach, and this in keeping with CDA's aims of exploring power relations and inequality (in the case of environmental issues, these do not necessarily involve relations between people, but the relationship between people and the rest of nature). CDA is usually described as a research movement rather than a method (Baker *et al*., 2008: 273; and Fairclough *et al*., 2011), but it typically involves inter-disciplinarity, qualitative analysis and the extensive examination of 'social, political, historical and intertextual contexts, which go beyond analysis of the language within texts' (Baker *et al*., 2008: 273–4).

CDA has been subject to various criticisms, the most serious being an apparent lack of rigour and transparency in the selection of texts or linguistic features (Stubbs, 1997; and Widdowson, 2004: 166), and the ensuing argument that CDA practitioners can simply select the aspects of texts that agree with their own hypotheses or political agendas. One response to this is to complement CDA with a CL approach; this limits the bias of wholly manual analysis and also provides a framework against which to measure the distinctive features of a text. As Widdowson (2004: 115) points out: 'Clearly to say that a particular association is usual or unusual is to make a comparative statement: a norm is presupposed. And this, of course, is where corpus descriptions are of particular relevance, for they can provide a norm'. Several recent CDA studies have used CL techniques, although sometimes with a tendency towards corpora with insufficient data (see the survey of studies in Baker *et al*., 2008). For example, Alexander (2009) bases his CDA study of environmental discourse on very small corpora (ranging from about 800 to 5,000 words). Several chapters are devoted to an analysis of the BBC Reith Lectures (lectures on the topic 'Respect for the Earth' by various politicians,

scholars, activists and businesspeople), but these constitute fewer than 5,000 words, and this makes it difficult to find any meaningful patterns. For example, the analysis of *earth* was potentially fruitful, but for each lecture there were only two or three occurrences of *earth* and no collocate occurred more than once. In addition, there was no comparison with the collocates of *earth* in a reference corpus – in other words, there was no attempt to establish the norm (as per the Widdowson quote, supplied above).

Criticisms of CL should also be noted, in particular the argument that CL techniques divorce a text from its context. To some degree, CDA offers a means of addressing these problems of CL, and as Baker *et al.* (2008: 279) point out:

> These criticisms seem to stem from restricted conceptions of CL, and would apply more accurately to CL studies that limit themselves to the automatic analysis of corpora, and are of a descriptive rather than an interpretative nature. The examination of expanded concordances (or whole texts when needed) can help the analyst infer contextual elements in order to sufficiently recreate the context.

Our study aimed to adhere to this more 'interpretative' version of CL.

Recently, there has been a drive towards a more robust methodology for combining CL and CDA techniques. For example, the Refugees, Asylum Seekers and Immigrants Project (RASIM; see Baker and McEnery, 2005; Baker *et al.*, 2008; and Gabrielatos and Baker, 2008) proposed a novel integrated approach, whereby CDA-style contextual reading informed the building of a corpus and a CL analysis of frequencies and keywords, *etc.*, which in turn led to a CDA analysis of selected texts from the corpus, and so on (see the nine-stage model suggested by Baker *et al.*, 2008: 295). Another approach which integrates corpus and discourse methodologies is Corpus-Assisted Discourse Studies (CADS). As the name implies, CADS focusses on discourse but uses corpora to 'uncover, in the discourse type of investigation, the *non-obvious meaning*' that might elude a human reader (Partington, 2010: 88), but also employs a more varied approach to corpora than traditional corpus linguistics. (For a discussion of this approach, see Partington, 2010: 89–90.)

In our study, we aimed to contribute to such approaches in the following ways: by comparing specialised corpora with one another and with a large reference corpus; by developing a clear and replicable methodology; by carrying out a small replicability study to ensure consistency of findings irrespective of the individual analyst; and by combining automatic methods with manual methods at each stage, in order to draw from the best of both CL and CDA.

2.3 Methodological framework II: using the web as corpus

Our study was based on four corpora: the reference corpus, UK Web as Corpus (UKWaC), and three custom-built specialised corpora. Since all of these corpora were built from searching the web, we will comment, first, on the use of the web as a corpus.

There are various ways in which the web can be used as a corpus: the simplest is to use a search engine to retrieve results – for example, to check a standard spelling (Kilgarriff and Grefenstette, 2003), although the use of search engines in this way has been questioned, since the results obtained cannot be reproduced (Kilgarriff, 2007). In this paper, we use 'web as corpus' to refer to the use of the web as a source of data which a processor crawls to retrieve documents and automatically compile a corpus. UKWaC, a corpus containing 1.5 billion tokens, which was built from web domains ending in .uk, is an example of such a corpus. UKWaC was built in 2007 and, thus, contains a stable representation of material in UK websites at that time. Its construction is described in Ferraresi *et al*. (2008).

As Baroni and Ueyama (2006) argue, there are several advantages and disadvantages associated with such a corpus. The first advantage is that it is much quicker and cheaper to build than a manually constructed corpus; it can, therefore, be much larger and it can also, in principle, be made freely available.[4] Ferraresi *et al*., (2008: 1) note that UKWaC is 'among the largest resources of its kind, and the only web-derived, freely available English resource with linguistic annotation'. Secondly, it can contain genres that are not found in traditional written corpora, including, for example, blogs and discussion forums. Thirdly, more up-to-date versions can be created much more easily.

On the other hand, if the corpus is built quickly it may contain more 'noise' (though in the case of UKWaC, various clean-up processes were employed; see Ferraresi *et al*., 2008). The text-types included are limited to texts that have been posted on the web, which means that postings from individuals are made only by the computer literate. The corpus will usually contain less, or no, meta-data, so the researcher has to go to the source web page in order to find more information about a particular text. And, perhaps most importantly, the corpus-builder has less control over what goes into the text: as Hundt *et al*. (2007: 2–3) point out, 'we still know very little about the size of this "corpus", the text types it contains, the quality of the material included or the amount of repetitive "junk" that it "samples"'. However, questions of sampling and duplicate material can be addressed by the corpus

[4] Though note that there are issues regarding copyright (Sharoff, 2006).

builder. Ferraresi *et al.* (2008) explain their processes of selecting a range of seed words in order to ensure a breadth of source data, as well as their rigorous de-duplication processes. Indeed, in defence of the use of web as corpus, Baroni and Ueyama (2006: 32) quote a personal communication from Serge Sharoff:

> ...a well constructed Web corpus might provide a straightforward operational answer to the eternal question of what is a 'representative' corpus representative of: a Web corpus could be a corpus that samples, in the right proportions, the types of linguistic contents that an average user typically accesses online in a certain period of time.

Our reasons for selecting UKWaC were largely pragmatic: it is the largest and most up-to-date available corpus of UK English.[5] These advantages can be exemplified by comparing data from UKWaC (over 1.5 billion tokens, containing data from about 2007) with data from the British National Corpus (BNC; approximately 112 million tokens, containing data from about 1970 to 1993).[6] Table 1 shows the raw and normalised frequencies of selected environmental terms in each corpus. *Environment*, and to a lesser extent *ecosystem*, have sufficient data in both corpora to be subject to linguistic analysis. However, for infrequent words such as *geodiversity*, and phrases such as *natural capital*, the BNC is simply too small: one cannot identify usage from a handful of occurrences. Furthermore, as Pearce (2008: 6) notes, 'the BNC is becoming a historical corpus'. Thus, although we acknowledge that in some respects the BNC is a more varied and balanced corpus than UKWaC, UKWaC was the best source available to us in terms of size and the period it covers. However, we are mindful of potential problems. Since, as we noted above, there is less information about source material in a web-derived corpus, we were took care to examine context and to check sources at the stage of manually analysing concordance lines (see Section 3.3.3, below).

==Insert Table 1 about here==

2.4 Methodological framework III: the Sketch Engine

[5] However, a web corpus is harvested at a particular point in time and will, therefore, also age. While UKWaC was the most up-to-date corpus available at the time of our study, it is recognised that the material is older than the material we collected in spring, 2010.

[6] For more information about the make-up of the BNC, see: http://www.natcorp.ox.ac.uk/corpus/index.xml

The interface we used was the Sketch Engine (see Kilgarriff *et al*., 2004), a corpus query system which has been used in a variety of studies in lexicography, linguistics and corpus linguistics (see for example Atkins, 2010; Culpeper, 2009; Nastase *et al*., 2006; Pearce, 2008; and Pustejovsky *et al*., 2010). Our reason for choosing this interface was that it allows the user to build corpora and to interrogate these with several tools in addition to standard concordances: thesauri, word sketches and sketch differences. These tools are described under Section 3, below.

2.5 Summary

Our aim was to use tools and methodologies which would allow us to access large amounts of contemporary English; to carry out as much automatic analysis as possible in order both to save time and to capture information that might be missed by manual analysis; and to supplement automatic analysis with manual checks in order to capture context, stance and other features that are less easily identified by a computer program. In this section, we provide an overview of how we sought to achieve these aims: by using web-based corpora so as to retrieve large amounts of up-to-date material; by using the Sketch Engine in order to carry out extensive collocation analysis; and by using a combination of CDA and CL methodologies in order to integrate automatic and manual analysis.

3. Methodology and analysis

In this section, we outline the methodology for each stage of the study: the identification of words and phrases to be analysed; the building of specialised corpora; and the analysis. We will show that, at each stage, a combination of automatic and manual methods was employed to ensure comprehensiveness and inter-analyst reliability.

3.1 Developing a list of lexical items for analysis

The first stage of the study was to clarify the object of analysis – that is, what lexical items to examine. Unlike other studies of environmental language, our aim was to analyse a broad range of terms. We also wanted to employ automatic methods to help us to identify relevant terms, rather than rely solely on a pre-conceived list. For the UK NEA, one of the primary objectives of this analysis was to discover the range of meanings associated with the concept of

ecosystems circulating in the public sphere; as a result, we used the term *ecosystem* as a starting point. We used the thesaurus feature of the Sketch Engine, which generates a list of lemmas that occur with similar collocates, and in similar grammatical relations, to an input lemma. A thesaurus of *ecosystem* (see Table 2) retrieved related terms such as *habitat*, *flora* and *biodiversity*.[7]

==Insert Table 2 about here==

We then repeated the thesaurus process for each new term (that is, we retrieved a thesaurus of *habitat*, one of *flora*, and so on). We also added key collocates from word sketches, such as *pollute* and *destruction*. Finally, the list was checked and supplemented by members of the UK NEA research team, who wished to include in the analysis terms that were central to the analytical approach that was taken for the UK NEA.[8] The final list included 136 lexical items, as shown under Figure 1.

==Insert Figure 1 about here==

3.2 Building the corpora

In order to compare the usage of environmental terms across genres, we built three specialised corpora using WebBootCat. (For a description of this tool, see Baroni *et al*., 2006.) We used as seed words/phrases the 136 items listed under Section 3.1; WebBootCat generated queries involving three of these seeds at a time and automatically retrieved web pages which contained those seeds in the query. However, because several of the seeds have other non-relevant meanings (for example, a web page containing *loss*, *resource* and *management* might have nothing to do with environmental issues), we also stipulated a 'whitelist' of the less polysemous terms (*ecological*, *sustainable*, *conservation*, etc.), of which a webpage had to include at least three occurrences. This ensured that the specialised corpora contained only relevant, environment-related web pages. Finally, we specified domains and/or websites for each of the three specialised corpora, as shown under Table 3. These corpora are highly

[7] Table 2 replicates the information on the Sketch Engine website: on the website, clicking on a lemma opens a new window with concordance lines.

[8] While the input from the UK NEA team could be seen, to some extent, to manipulate the contents of the corpus, we would argue that this combination of automatic and manual processes provided a useful balance: the automatic process identified key terms that might have been missed by a human compiler, while the manual process identified relevant environmental terms which, although not captured by the Sketch Engine, were seen as central by experts in the environmental field.

comparable in that they were created at the same time (April, 2010) from the same set of seeds, and are of similar size.

Our original aim was to create separate corpora of news articles, blogs and NGO websites. However, because of problems with timeouts (especially with blogs), we found it difficult to create sufficiently large corpora – hence the decision to combine these genres into one 'public' corpus.[9] While these are in some respects distinct genres, a truly homogenous corpus is very difficult to achieve. For example, even a corpus containing only data from news websites would contain different text types depending on whether they are based on broadcast or print media sources, and on the character of the items: news report, editorial comment, feature, readers' comments, *etc*., each have different linguistic features (Carvalho, 2005). There are substantial areas of overlap in the public corpus; for example, NGO websites often contain blogs, while the readers' comments on newspaper articles are more akin to blogs than they are to the articles themselves.[10] The reason for developing these different corpora for comparison was that earlier qualitative social research had suggested that, whilst environmental policy makers were comfortable with the term *ecosystem*, it was not well understood by the public (Defra, 2007) and the different corpora would indicate if these differences persisted.

==Insert Table 3 about here==

3.3 Analysis

We now describe the stages of our analysis in detail: identifying collocates in word sketch; comparing collocates using sketch difference; manual analysis of concordance lines; and comparing frequencies.

3.3.1 Word sketches

We used word sketches to discover salient collocates of each lexical item in UKWaC and in the three specialised corpora. Word sketches are 'one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour' (Kilgarriff *et al*., 2004: 105). This can be exemplified by examining

[9] Timeouts occur if the search engine does not respond quickly enough and, therefore, the program automatically stops searching; this results in corpora with insufficient data.

[10] It is also possible that there is overlap between the corpora; for example, newspapers might reproduce material from government websites for dissemination. Furthermore, there may be some overlap between the material in the specialised corpora and in UKWaC.

part of the word sketch of *ecosystem* in UKWaC (Table 4).[11] Each column shows a particular grammatical relation in which *ecosystem* occurs (as object, subject, modifier, *etc*.). The lemmas in the column are organised by salience (although one can choose to organise them by frequency instead). The figures in the first column for each relation are raw frequencies; in the second column, salience (logdice), which is calculated as described by Rychlý (2008), is shown.

==Insert Table 4 about here==

The advantages of word sketches over traditional methods of identifying collocates are discussed in Pearce (2008: 4), who remarks that traditional methods tend to highlight particular parts of speech and unusual words. A word sketch, on the other hand, 'shows which grammatical roles a lemma prefers or avoids, and also displays its collocates in dozens of grammatical relations' (Pearce, 2008: 5).

Where there were interesting differences between the corpora, we explored the concordance lines in more detail. For example, Table 5 shows the ten most salient collocates of *nature*, where *nature* is a modifier, in the four corpora. There are notable similarities: in all four, *reserve* and *conservation* are the top two collocates. Some of the collocates in UKWaC are, not surprisingly, more general than those in the other corpora, (e.g., *nature lover* and *nature photography*), although all relate to *nature* in the sense 'the physical world and living things'.[12] One of the main differences is in the collocates *interest* and *value* (which are shaded in the table). These do not occur at all in the top ten in UKWaC; they are very frequent and salient in the government corpus, and much less so in the academic and public corpora.

==Insert Table 5 about here==

If we examine the concordance lines where *nature* collocates with *value* in the government corpus, we find that they often occur in planning proposals. There are frequent references to *nature conservation value*, and to *recognising* and *identifying* sites of nature value. There are also references to areas which have limited nature value, and which might, thus, be used for development (though only in hypothetical situations); for example, one website suggests that

[11] Table 4 replicates the information presented on the Sketch Engine website, with the headings 'raw frequency' and 'salience score' added for clarity. On the website, a click on a frequency figure opens a new window showing concordance lines.

[12] In other grammatical relations, UKWaC contains a lot of collocates relating to non-environment senses, (e.g., *human* (*human nature*) and *problem* (*the nature of the problem*)). We did not spend time examining these non-relevant senses.

'developments of this kind should only take place on **land of lower nature conservation value**' (government corpus).[13] We return to the question of the 'value' of nature in the discussion of our findings under Section 4.

3.3.2 Sketch differences

For selected pairs, we also carried out sketch differences. The sketch difference tool compares the collocates of two lemmas: it shows shared collocates, as well as highlighting those collocates which, according to thresholds in Sketch Engine, are more salient for one lemma than the other.[14] For example, Table 6 shows part of a sketch difference containing words modified by *rural* and *urban* in UKWaC (the full page contains five other grammatical relations).[15] This tool is a useful way of highlighting differences between related terms: *rural* has more collocates relating to beauty and peace (*idyll* and *retreat*), while *urban* is more frequently used in contexts of change (*renaissance*, *renewal* and *regeneration*). However, it also highlights similarities: both *rural* and *urban* occur with terms relating to poverty (*poor* and *poverty*).

==Insert Table 6 about here==

Pearce (2008: 21) suggests that when using sketch difference, 'the analyst is in danger of exaggerating the differences and overlooking similarities'. However, this is perhaps a tendency of corpus/discourse studies in general, rather than a fault of the sketch difference tool itself, which does display shared collocates. Taylor (2011) shows that corpus-assisted discourse studies have had a tendency to focus on differences and neglect similarities; she notes the importance of searching for similarities in her study of 'boy/s' and 'girl/s'. We took care to identify shared collocates as well as different ones.

3.3.3 Analysis of concordance lines

[13] See: http://www.aberdeenshire.gov.uk/planning/inquiry/PrecogRSPB%20Scotland%20Summary.pdf

[14] In the newest version of the Sketch Engine, it is now possible to compare lemmas across sub-corpora as well, and, indeed, to compare corpora processed with the same part-of-speech tagger by first combining these to form a 'super-corpus'.

[15] Table 6 is a simplified version of the original. Sketch differences on the Sketch Engine website are colour coded, with strong collocates of the first lemma in different shades of green depending on salience score, strong collocates of the second lemma in different shades of red, and equally strong collocates of both lemmas in white. As with word sketches and thesauri, one can click on a frequency figure to open a window of concordance lines.

Automatic methods (words sketches and sketch differences) were consistently reinforced by the manual analysis of expanded context. In addition, for all the lexical items in the study, a random sample of 100 citations in UKWaC was analysed. For phrases, which could not be examined using word sketch, we also analysed random samples of fifty citations in each of the specialised corpora. This stage of the study allowed us to identify uses and connotations of environmental vocabulary that might not be revealed through collocation analysis alone; in particular, we were able to detect whether particular lexical items tend to be used positively, negatively or neutrally in the different corpora.

As an example of this method, we present our manual analysis of *national park*. Of a sample of 100 citations in UKWaC, all are either neutral or positive, with frequent references to the beauty of particular national parks in a number of different countries and to their value and functions; for example:

> Today, the Northumberland **National Park** acts as a welcome 'lung' for Tyneside's cyclists.
>
> (UKWaC)[16]

> Banff is a small, attractive, frontier town famous for its railway history and wildlife. It is a genuinely unspoilt resort in the middle of the **National Park**.
>
> (UKWaC)[17]

In the government corpus, there are neutral references to national park boundaries and borders, and also positive references to the value of national parks:

> The majority of the **National Park** is of outstanding national, European and international value for its nature conservation interest.
>
> (government corpus)[18]

In the academic corpus, citations are again neutral or positive, with references to the need to protect or preserve national parks; for example:

> UNESCO ... requested that alternative mining sites be sought that would pose less of a threat to the integrity of Jasper **National Park**.
>
> (academic corpus)[19]

---

[16] See: danorth.fsnet.co.uk
[17] See: http://www.skiworld.ltd.uk/resort/Canada/Banff_and_Lake_Louise/27
[18] See: http://www.newforestnpa.gov.uk/text/biodiversity_topic_paper.pdf
[19] See: http://www.geog.leeds.ac.uk/courses/level3/geog3320/studentwork/groupe/report.html

By contrast, the public corpus contains three citations in a sample of fifty – two from *The Guardian* and one from the BBC news website – which discuss problems with the idea of national parks (rather than specific National Parks); for example:

> Climate change, invasive species and diseases do not stop at the borders of **national parks**.
>
> (public corpus)[20]

This contributes to our overall finding that, in the public corpus, there is more of an emphasis on problematic aspects of environmental issues than in the other corpora.

Furthermore, analysing concordance lines yielded much richer findings than would have been possible with collocate analysis alone. In particular, manual analysis was necessary to identify the broader contexts of issues concerning particular lexical items.

3.3.4 Comparing frequencies

We also compared frequencies of the lexical items across the three specialised corpora, to check for any notable differences (and similarities). We did not put too much weight on the findings of this stage, as it was possible that the frequencies were affected by WebBootCat's random generation of queries from the seeds (for example, a particular seed may have been included in more queries in one corpus than in another)[21]. However, some broad tendencies may be noted. Table 7 shows the frequencies of selected lexical items in UKWaC and in the three specialised corpora (the frequencies in UKWaC are simply given for reference, since these are inevitably lower, per 10,000 tokens, than in the specialised corpora).

==Insert Table 7 about here==

The frequencies suggest that the public corpus has more of a focus on environmental problems (e.g., *climate change*, *global warming*, *destroy* and *extinct*), while the government corpus has higher frequencies for more positive or neutral terms (e.g., *biodiversity*, *habitat* and *green space*). We also paid

[20] See: http://news.bbc.co.uk/1/hi/sci/tech/7506109.stm

[21] This is a problem of current web crawling technology in general, as it depends on the various parameters of the search engines. It is an issue that needs to be recognised, but it does not negate the benefits of being able to build bespoke specialised corpora.

attention to similarities: the frequencies for *environment* and *environmentally* are much the same across the corpora, although *environmental* is more frequent in the academic corpus (largely in the phrase 'environmental change'), which may, perhaps, be because of the greater lexical density of academic prose. It is also notable that *sustainable* and *sustainability* are most frequent in the academic corpus, but the antonym *unsustainable* is most frequent in the public: this, again, suggests the public focus on the negative aspects of environmental issues. However, this may be due to the inclusion of newspapers in the public corpus, which tend to focus on negative news stories. In future analyses, we would hope to separate newspapers from other parts of the public corpus.

3.4 Replicability

It was noted under Section 2.2 that one of the criticisms of CDA is its tendency towards subjectivity. Using corpus evidence is one way of addressing this issue; another is using triangulation methods to check replicability of findings. Marchi and Taylor (2009) discuss various methods of triangulation that might be employed: methodological triangulation (approaching the same data with different methods); data triangulation (carrying out the same study on different sets of data); theoretical triangulation (approaching the same data from differing theoretical stances); and investigator triangulation (two or more researchers analysing the same set of data with the same research question). They focussed on investigator triangulation, and carried out a study in which both co-authors analysed the same set of data, and checked for converging and dissonant results. A similar study was carried out by Baker (2011) where five researchers analysed the same set of data using a combination of CL/CDA techniques.

As in many previous studies, we did not carry out a full-scale replicability study, but we did conduct a small 'sanity check' which relates to previous work on investigator triangulation. Two researchers (the first and third authors of this paper) worked on the methodology together, but due to time constraints needed to share the analysis. In order to ensure that this would not bias the results, three of the lexical items (*access*, *allotment* and *global warming*) were analysed and findings were compared.

==Insert Table 8 about here==

As an example of our results, Table 8 shows the features of *allotment* identified by Researcher 1 and Researcher 2. There are some differences: the collocates identified did not overlap exactly, and were grouped differently by the two researchers; also, the point about the academic corpus's focus on lack

of allotments was identified by Researcher 2 but not Researcher 1. However, the overall results are encouraging: although the collocates identified were not identical, the general feel of these is the same; and several other points were replicated exactly. For instance, both researchers identified the same citation as an example of the way that negative collocates of *allotment* are used in a positive context:

> … those millions of us who live in town are particularly fortunate because our "local countryside" is made up of parks and cemeteries, railway sidings, waste tips, overgrown quarries, **abandoned allotments**, neglected gardens. In other words, a wonderful mosaic of wildlife habitat.
>
> (public corpus)[22]

This citation led us to explore the issue of the subversion of semantic prosodies (see Hunston, 2007; and Louw, 1993). While words like *abandon* and *neglect* have negative semantic prosodies, they are used here to subvert expectations: it is only by being ignored by humans that wildlife can thrive.

While this was a small replicability check, it provided encouraging indications that, given the same methodology and same data, our findings could be reproduced. However, there is likely to be some element of researcher divergence in a study such as this, which involves subjective interpretation of the data. In future work on these three corpora, we would extend this with a larger set of lexical items, fresh analysts who had not been involved in devising the methodology, different analysts using different tools, and, if possible, another set of corpora created under the same conditions to check for data triangulation. We are also interested in intra-analyst replicability – that is, getting an analyst to repeat the procedure six months later.

4. Findings

Although the focus of this paper is on the methodology behind the study, we will briefly summarise some of our key findings.

- **The conceptualisation of nature**. Raymond Williams's entry on *nature* in his seminal study *Keywords* (1983 [1976]: 221–4) argued that:

  > Nature is perhaps the most complex word in the language [...] Nature has meant the 'countryside', the 'unspoiled places',

[22] See: http://www.telegraph.co.uk/gardening/4792215/Just-wild-about-suburbia.html

> plants and creatures other than man. The use is especially current in contrasts between town and country: nature is what man has not made, though if he made it long enough ago – a hedgerow or a desert – it will usually be included as natural.

The multiple conceptions of *nature* emerged strongly in our findings. On the one hand, nature is typically presented as distinct from humankind (in phrases such as 'both human needs and nature...' and 'people experiencing nature...'). On the other hand, there are occasional citations – mainly in the academic corpus – that present humans as part of nature; for example, 'the relationship between humans and the rest of nature...' Nature is often presented as a commodity, especially in the government corpus, as shown in the collocation of *nature* with *interest* and *value* (see Section 3.3.1). The same applies to other environmental vocabulary, as can be seen in citations such as 'investment in key environmental assets'. There are also explicit measurements of its monetary value, as in: 'Within the project area, ecosystem services may be c. $1.5 billion y-1, or $105 person-1 y-1.'[23] The UK NEA (2011) itself was also designed to produce economic values of changes in approaches to ecosystem management. In the public corpus, nature is not usually measured explicitly in monetary terms: indeed, this conceptualisation is sometimes questioned. However, environmental terms are frequently used to promote products and services, especially on tourism websites.

- **Attitudes towards nature**. There are mixed attitudes towards nature, especially when it is conceptualised as distinct from humans. There are positive framings of 'untouched', 'pristine' nature, but also negative portrayals of nature which is not controlled, in phrases such as 'overgrown vegetation' and 'unsightly wilderness'. However, we also found evidence of seemingly negative terms such as *overgrown* and *abandoned* being used in positive contexts to challenge readers' assumptions, (i.e., to present in a positive light the idea of nature without human interference). This strategy was mentioned in the discussion of *allotment* under Section 3.4; another example is a description of a part of Cornwall: 'Here, the gorse survives, and the prickliness of scavengers and borderline farmers … ultimately defends the land against those who would declare it valuable for tourism only' (UKWaC).[24] Other attitudes relate to fear and anxiety: although the

[23] See: http://template.bio.warwick.ac.uk/staff/aprice/bibliography.htm
[24] See: http://www.artspacegallery.co.uk/MainImagePages/Artists/Paintings/Atkins/Feaver_article.htm

countryside and green spaces are generally perceived positively, there are also indications of fear associated with isolated areas: for example, green spaces and urban parks are sometimes associated with crime.

- **Differences between the specialised corpora**. Some differences between the specialised corpora have already been noted. For example, in the public corpus there is more of a focus on environmental problems, whereas in the government corpus, there is more of a focus on the use of ecosystems and their value to the public. The corpora also differ in terms of stance: in particular, whilst there is plenty of evidence of pro-environmental attitudes in all corpora there is also frequent evidence of scepticism in the public corpus, (for example, the questioning of scientific and expert knowledge, and cynical attitudes towards self-interested uses of environmental issues). Two uses of *green* in the public corpus exemplify the latter: a blog comments on the promotional use of the word, arguing that 'We won't consume ourselves to freedom by tacking "green" onto every enterprise like a postscript',[25] while a newspaper article presents politicians' attention to environmental issues in a rather disparaging light: 'The would-be prime ministers have also touted their green credentials....'[26] However, the negative and sceptical attitudes in the public corpus may be attributable to the inclusion of newspapers in this corpus, which tend to focus on 'bad news'. It would be of interest, in future studies, to analyse blogs and other material separately in order to discover the prevailing attitudes in other types of public discourse.

- **Lexical items not widely used or understood**. There was evidence that several of the lexical items under analysis (including *biome*, *biotope* and *ecology*) are not widely used, in that they are very infrequent in the corpora, or that they tend to appear only in book titles or other specific contexts. Others appear to be new or not yet widely understood, in that they are often presented in inverted commas (e.g., 'the US consumes "natural capital" at about the average rate')[27] or with an accompanying explanation (e.g., 'Biodiversity is a term which simply means "the variety of life"').[28]

---

[25] See: http://www.greenpeace.org.uk/blog/about/deep-green-going-deeper

[26] See: http://www.guardian.co.uk/environment/2006/apr/20/business.greenpolitics. A concordance of *tout* in UKWaC shows that, in the sense 'persuade people of the merits of something', it has a negative prosody: things that are *touted* are usually not as good as the presentation suggests.

[27] See: http://www.workface-limited.co.uk/html/newscientist_20060427.html

[28] See: http://www.offwell.free-online.co.uk/woodland_manage/conserva_manage.htm

5. Conclusions

The methodology used in our study allowed us to survey a wide range of environmental lexis from a variety of angles: we examined collocates, connotations and evaluative tendencies; and we compared usage across genres. We hope to have contributed to the growing field of corpus-based critical discourse studies in three key ways:

(*a*) We built specialised corpora from the web using the same set of seeds, thus allowing systematic comparison across genres and with a reference corpus;
(*b*) We combined automatic and manual methods at all stages (the creation of the list of lexical items to be analysed, the building of specialised corpora from the web and the analysis). In this way, the comprehensiveness and breadth allowed by automatic methods was complemented by the in-depth focus achieved by manual analysis; and,
(*c*) Before carrying out the study, we defined a clear methodology that could be reproduced by other researchers with the same data. Our small-scale check of replicability produced encouraging results, in that there was substantial overlap between the findings of the two researchers involved.

The study could be extended in the future. In particular, it would be useful to apply the methodology to a larger sample of data to check whether the results can be generalised, and to make more extensive replicability checks on more data and with more researchers. The data sample could also be further refined with, for example, separate corpora for NGOs, newspapers and blogs, or with comparable corpora from different years to allow for diachronic analysis. Another avenue of research would be the use of automatic sentiment analysis, whereby a computer system automatically applies tags reflecting subjectivity and opinion to documents, sentences or even phrases (Pang and Lee, 2008). While state-of-the-art automatic analysis of evaluative language is somewhat shallow, with low accuracy compared to a human, this could enable us to find larger samples of potentially evaluative text for a broader analysis.

As Baker *et al*. (2008: 297) argue, 'theoretical and methodological cross-pollination [...] seem to benefit both CDA and CL'. Our study is an example of how such a combination of approaches could be applied, and also points to areas where further refinements to methodology might be possible. In addition, it reveals the importance of quantitative and qualitative lexical

analysis for understanding discourse over the environment and ecosystems in contemporary society.

References

Alexander, R.J. 2009. Framing Discourses on the Environment: A Critical Discourse Approach. New York: Routledge.

Atkins, B.T.S. 2010. 'The DANTE database: its contribution to English lexical research and in particular to complementing the FrameNet data' in G.-M. de Schryver (ed.) A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks, pp. 267–97. Kampala: Menha Publishers.

Baker, P. 2011. 'Discourse, news representations and corpus linguistics'. Plenary paper presented at Corpus Linguistics 2011, Birmingham, 20–22 July.

Baker, P. and T. McEnery. 2005. 'A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts', Language and Politics 4 (2), pp. 197–226.

Baker, P., C. Gabrielatos, M. Khosravinik, M. Krzyzanowski, T. McEnery and R. Wodak. 2008. 'A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press', Discourse and Society 19 (3), pp. 273–306.

Baroni, M. and M. Ueyama. 2006. 'Building general- and special-purpose corpora by Web crawling' in Proceedings of the Thirteenth NIJL International Symposium, Language Corpora: Their Compilation and Application, pp. 31–40. Tokyo.

Baroni, M., A. Kilgarriff, J. Pomikalek and P. Rychly. 2006. 'WebBootCaT: a web tool for instant corpora', Proceedings of Euralex. Torino, Italy.

Bevitori, C. 2010. Representations of Climate Change. News and opinion discourse in UK and US quality press: a Corpus-assisted Discourse Study. Bologna: Bologna University Press.

Carvalho, A. 2005. 'Representing the politics of the greenhouse effect', Critical Discourse Studies 2 (1), pp. 1–29.

Carvalho, A. and J. Burgess. 2005. 'Cultural circuits of climate change in UK broadsheet newspapers, 1985–2003', Risk Analysis 25 (6), pp. 1,457–69.

Culpeper, J. 2009. 'The metalanguage of impoliteness: using Sketch Engine to explore the Oxford English Corpus' in P. Baker (ed.) Contemporary Corpus Linguistics, pp. 64–86. London: Continuum.

Defra (Department for Environment, Food and Rural Affairs). 2007. Public Understanding of the Concepts and Language around Ecosystem Services and the Natural Environment. Department for Environment, Food and Rural Affairs, London.

Fairclough, N., J. Mulderrig and R. Wodak. 2011. 'Critical discourse analysis' in T.A. van Dijk (ed.) Discourse Studies: A Multidisciplinary Introduction, pp. 357–78. London: Sage.

Ferraresi, A., E. Zanchetta, M. Baroni and S. Bernardini. 2008. 'Introducing and evaluating UKWaC, a very large web-derived corpus of English' in *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. Marrakech, Morocco. Available online, at: http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf

Gabrielatos, C. and P. Baker. 2008. 'Fleeing, sneaking, flooding: a corpus analysis of discursive constructions of refugees and asylum seekers in the UK Press 1996–2005)', Journal of English Linguistics 36 (1), pp. 5–38.

Goatly, A. 1996. 'Green grammar and grammatical metaphor, or language and myth of power, or metaphors we die by', Journal of Pragmatics 25 (4), pp. 537–60. Reprinted in A. Fill and P. Mühlhäusler (eds, 2001) The Ecolinguistics Reader: Language, Ecology and Environment, pp. 203–25. London and New York: Continuum.

Grundmann, R. and R. Krishnamurthy. 2010. 'The discourse of climate change: a corpus-based approach', Critical Approaches to Discourse Analysis across Disciplines 4 (2), pp. 125–46.

Hundt, M., N. Nesselhauf and C. Biewer. 2007. 'Corpus linguistics and the web' in M. Hundt, N. Nesselhauf and C. Biewer (eds) Corpus Linguistics and the Web, pp. 1–6. Amsterdam and New York: Rodopi.

Hunston, S. 2007. 'Semantic prosody revisited', International Journal of Corpus Linguistics 12 (2), pp. 249–68.

Kilgarriff, A. 2007. 'Googleology is bad science', Computational Linguistics 33 (1), pp. 147–51.

Kilgarriff, A. and G. Grefenstette. 2003. 'Introduction to the special issue on web as corpus', Computational Linguistics 29 (3). Available online, at: http://www.kilgarriff.co.uk/Publications/2003-KilgGrefenstette-WAC Intro.pdf

Kilgarriff, A., P. Rychlý, P. Smrz and D. Tugwell. 2004. 'The Sketch Engine', *Proceedings of Euralex*, pp. 105–16. Lorient, France. (Reprinted in P. Hanks (ed.). 2007. Lexicology: Critical Concepts in Linguistics. London: Routledge.)

Kuha, M. 2009. 'Uncertainty about causes and effects of global warming in U.S. news coverage before and after Bali', Language and Ecology 2 (4). http://www.ecoling.net/journal.html

Louw, B. 1993. 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies' in M. Baker, G. Francis and E. Tognini-Bonelli (eds) Text and Technology: In Honour of John Sinclair, pp. 157–77. Philadelphia and Amsterdam: John Benjamins.

Marchi, A. and C. Taylor. 2009. 'If on a winter's night two researchers… A challenge to assumptions of soundness of interpretation', Critical Approaches to Discourse Analysis across Disciplines 3 (1), pp. 1–20.

Nastase, V., S. Jelber, M. Sokolova and S. Szpakowicz. 2006. 'Learning noun-modifier semantic relations with corpus-based and WordNet-based features' in Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006), pp. 781–7.

Nerlich, B. and N. Koteyko. 2009. 'Compounds, creativity and complexity in climate change communication: the case of "carbon indulgences"', Global Environmental Change 19, pp. 345–53.

Pang, B. and L. Lee. 2008. 'Opinion mining and sentiment analysis', Foundations and Trends in Information Retrieval 2 (1–2), pp. 1–135.

Partington, A. 2010. 'Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project', Corpora 5 (2), pp. 83–108.

Pearce, M. 2008. 'Investigating the collocational behaviour of MAN and WOMAN in the BNC using Sketch Engine', Corpora 3 (1), pp. 1–29.

Pustejovsky, J., A. Rumshisky, A. Plotnick, E. Jezek, O. Batiukova and V. Quochi. 2010. 'SemEval-2010 Task 7: argument selection and coercion', Association of Computational Linguistics, pp. 27–31.

Rychlý, P. 2008. 'A lexicographer-friendly association score' in P. Sojka and A. Horák (eds) Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008, pp. 6–9. Brno: Masaryk University.

Schleppegrell, M.J. 1997. 'Agency in environmental education', Linguistics and Education 9, pp. 49–67.

Sharoff, S. 2006. ‘Open-source corpora: using the net to fish for linguistic data’, International Journal of Corpus Linguistics 11 (4), pp. 435–62.

Stubbs, M. 1997. ‘Whorf’s children: critical comments on critical discourse analysis (CDA)’ in A. Ryan and A. Ray (eds) Evolving Models of Language Clevedon: Multilingual Matters. Reproduced online, at: http://www.uni-trier.de/fileadmin/fb2/ANG/Linguistik/Stubbs/stubbs-1997-whorfs-children.pdf

Taylor, C. 2011. ‘Searching for similarity: the representation of boy/s and girl/s in the UK press in 1993, 2005, 2010’ Paper presented at Corpus Linguistics 2011, Birmingham, 20–22 July.

UK NEA. 2011. The United Kingdom National Ecosystem Assessment: Synthesis of the Key Findings. Cambridge: UNEP-WCMC.

Widdowson, H.G. 2004. Text, Context, Pretext: Critical Issues in Discourse Analysis. Oxford: Blackwell.

Williams, R. 1983 [1976]. Keywords: A Vocabulary of Culture and Society. Revised edition. Oxford: Oxford University Press.

Table 1: Raw and normalised frequencies of selected environmental terms in the BNC and UKWaC

	BNC		UKWaC	
	Raw freq.	Per million tokens	Raw freq.	Per million tokens
ecosystem	283	2.52	10,406	6.65
ecosystem approach	4	0.04	208	0.13
ecosystem services	0	0.00	186	0.12
environment	14,361	128.02	340,729	217.68
geodiversity	0	0.00	163	0.10
natural capital	1	0.01	183	0.12
right-to-roam / right to roam	7	0.06	381	0.24
urban park	3	0.03	408	0.26

Table 2: Part of a thesaurus of *ecosystem* in UKWaC

Lemma	Salience score	Raw frequency
HABITAT	0.329	39,127
BIODIVERSITY	0.306	16,729
FLORA	0.301	7,575
ECOLOGY	0.279	10,747
VEGETATION	0.275	15,477
WETLAND	0.274	7,657
FLORA	0.231	6,160
WILDLIFE	0.227	42,206
RAINFOREST	0.217	6,233
GRASSLAND	0.215	10,076
FOREST	0.212	61,481
LANDSCAPE	0.200	73,990
FISHERY	0.199	11,521
WOODLAND	0.197	35,083
REEF	0.194	11,601
ORGANISM	0.189	22,425
CLIMATE	0.186	84,078
ENVIRONMENT	0.167	271,267

Figure 1: List of lexical items included in the study. Unless part of speech is specified, all parts of speech were examined (for example, *access* as both noun and verb)

access, *agricultural*, *agriculture*, *allotment*, *amateur*, *anxiety*, *aquatic*, *attachment*, *beautiful*, *beauty*, *biodiversity*, *biogeography*, *biological*, *biology*, *biome*, *biosphere*, *biotope*, *change*, *climate*, *climate change*, *coastal*, *commons*, *conservation*, *conservation group*, *conserve* (verb), *countryside*, *cultural*, *cultural diversity*, *cultural heritage*, *culture*, *damage*, *deforestation*, *desert* (noun), *destroy*, *destruction*, *diversity*, *dynamics*, *eco-*, *ecological*, *ecology*, *economic*, *economy*, *ecosystem*, *ecosystem approach*, *ecosystem services*, *environment*, *environmental*, *environmentally*, *expert*, *expert knowledge*, *extinction*, *farmland*, *fauna*, *fear*, *fishery*, *flora*, *forest*, *freedom*, *freshwater*, *garden*, *geodiversity*, *geology*, *global warming*, *grassland*, *green*, *green space*/*greenspace*, *greenhouse effect*, *habitat*, *heathland*, *hedgerow*, *heritage*, *independence*, *indigenous*, *interaction*, *invasive*, *landscape*, *landscape*(*d*) *garden*, *land-use change*, *loss*, *management*, *man-made*, *marine* (adjective), *national park*, *National Trust*, *native*, *natural*, *natural capital*, *natural history*, *nature*, *ocean*, *open space*, *organism*, *peace*, *peaceful*, *peacefulness*, *plant*, *politics*, *pollute*, *pollution*, *popular culture*, *population*, *professional*, *public access*, *rainforest*, *recreation*, *reef*, *reserve* (noun), *resource*, *right-to-roam*/*right to roam*, *Royal Society for the Protection of Birds* (*RSPB*), *rural*, *savanna*(*h*), *science*, *semi-natural*, *sense of place*, *soil*, *solitude*, *specialist*, *species*, *spiritual*, *sustainability*, *sustainable*, *terrestrial*, *topography*, *unsustainable*, *urban*, *urban park*, *value*, *vegetation*, *voluntary*, *volunteer*, *wetland*, *wilderness*, *wildlife*, *Wildlife Trust*, *woodland*

Corpus name	Domains/websites	Tokens	Typical text types
Academic	domains ending in .ac.uk	1,616,891	journal articles in university repositories, working papers, lecture handouts, course outlines
Government	domains ending in .gov.uk	1,691,559	reports of research projects, guidelines, planning and development proposals, public information documents
Public	.bbc.co.uk; .telegraph.co.uk; .timesonline.co.uk; .guardian.co.uk; .thesun.co.uk; .dailyrecord.co.uk; .blog.co.uk; .foe.co.uk; .rspb.org.uk; .woodlandtrust.org.uk; .greenpeace.org.uk; .wwf.org.uk; .nationalgeographic.co.uk; .nationaltrust.org.uk; .theecologist.org	1,326,849	news articles (including readers' responses), blogs, informative documents

Table 3: Specialised corpora

	Raw frequency	Salience score		Raw frequency	Salience score		Raw frequency	Salience score		Raw frequency	Salience score
Object of	1,633	**1.4**	Subject of	1,157	**1.7**	Modifier	5,350	**2**	Modifies	1,956	**0.7**
degrade	25	6.69	*function*	25	4.99	*aquatic*	161	8.92	*functioning*	43	6.63
conserve	22	5.65	*model*	28	4.42	*terrestrial*	160	8.51	*resilience*	14	5.7
disrupt	16	4.96	*evolve*	11	3.11	*marine*	381	8.35	*dynamics*	47	5.33
function	20	4.64	*consist*	10	2.15	*fragile*	129	7.88	*ecology*	20	5.06
damage	36	4.47	*suffer*	12	1.47	*freshwater*	76	7.67	*degradation*	10	4.41
threaten	39	4.47	*process*	5	1.46	*semi-natural*	26	6.89	*function*	117	3.45
harm	11	4.45	*approach*	6	1.34	*mangrove*	28	6.81	*biodiversity*	8	3.39
reconstruct	5	4.12	*depend*	6	1.2	*planktonic*	19	6.76	*stability*	18	3.38
restore	30	4.04	*respond*	5	1.19	*wetland*	46	6.72	*integrity*	12	3.23
impact	7	4.01	*result*	5	0.98	*pelagic*	18	6.5	*approach*	141	3.15
invade	6	3.9	*comprise*	5	0.94	*arctic*	19	6.47	*habitat*	18	3.03
preserve	23	3.72	*occur*	7	0.71	*tropical*	64	6.4	*restoration*	9	2.83

Table 4: Part of a word sketch of *ecosystem* in UKWaC

UKWaC			Academic corpus			Government corpus			Public corpus		
	Raw frequency	Salience score		Raw frequency	Salience score		Raw frequency	Salience score		Raw frequency	Salience score
reserve	4,924	10.8	*conservation*	217	10.5	*conservation*	808	12	*reserve*	210	12.2
conservation	2,102	9.46	*reserve*	43	10.1	*reserve*	305	11.8	*conservation*	215	11.2
trail	1,010	8.83	*conservationist*	4	8.02	*interest*	119	10	*legislation*	13	8.72
lover	481	8.02	*tourism*	6	7.6	*value*	104	9.66	*trail*	5	7.93
walk	272	5.98	*interest*	8	7.13	*trail*	31	9.08	*importance*	7	7.73
designation	50	5.63	*agency*	3	6.57	*importance*	33	8.33	*interest*	9	7.64
conservationist	34	5.59	*value*	7	6.19	*designation*	17	7.86	*body*	8	7.55
photography	78	5.17	*area*	5	4.67	*space*	30	7.7	*value*	10	7.5
extent	113	5.09	*management*	5	4.65	*site*	34	6.77	*adviser*	3	7.13
conservancy	20	4.99	*resource*	3	4.48	*conservancy*	5	6.56	*objective*	6	7.06

Table 5: Ten most salient collocates of *nature* (where *nature* is modifier) in UKWaC and in the academic, government and public corpora. (Salience scores, logdice, are calculated as described by Rychlý (2008))

Table 6: Part of a sketch difference of *rural* and *urban* in UKWaC

	Raw frequency (rural)	Raw frequency (urban)	Salience score (rural)	Salience score (urban)
Stronger collocates of *urban*				
sprawl	0	492	0	8.2
legend	0	389	0	7.1
myth	11	363	1.7	7.0
renaissance	20	483	3.2	8.1
renewal	15	347	2.4	7.1
fringe	44	250	4.2	7.0
regeneration	250	1,480	6.1	8.9
Collocates of both *rural* and *urban*				
environment	679	1,947	5.4	6.9
dweller	156	213	6.1	6.9
landscape	631	698	6.8	7.1
poor	418	291	7.5	7.3
population	895	571	6.4	5.8
area	15,903	9,089	8.2	7.4
poverty	383	188	6.2	5.3
settlement	512	248	6.8	5.9
setting	992	442	7.1	6.0
district	505	185	6.8	5.5
village	971	299	6.8	5.1
Stronger collocates of *rural*				
location	1,241	227	6.9	4.5
community	5,675	622	7.7	4.5
hinterland	159	11	6.2	2.7
retreat	229	15	6.4	2.7
livelihood	246	10	6.6	2.3
economy	2,146	99	8.0	3.7
idyll	243	0	6.9	0

Table 7: Frequencies of selected items in UKWaC and in the academic, government and public corpora. For each lemma, the highest normalised frequency of the three specialised corpora is underlined, except in cases where there is little difference between the three

Lexical item	UKWaC		Academic		Government		Public	
	Raw freq.	Freq. per 10,000 tokens	Raw freq.	Freq. per 10,000 tokens	Raw freq.	Freq. per 10,000 tokens	Raw freq.	Freq. per 10,000 tokens
biodiversity	25,858	**0.165**	3,490	**21.585**	6,129	**36.233**	2,042	**15.390**
biotope	332	**0.002**	51	**0.315**	8	**0.047**	7	**0.053**
climate change	36,531	**0.233**	1,657	**10.248**	1,313	**7.762**	2,451	**18.472**
destroy	78,214	**0.500**	99	**0.612**	129	**0.763**	268	**2.020**
ecosystem	10,406	**0.066**	3,108	**19.222**	1,061	**6.272**	997	**7.514**
environment	340,729	**2.177**	3,258	**20.150**	3,491	**20.638**	2,601	**19.603**
environmental	191,775	**1.225**	6,280	**38.840**	3,194	**18.882**	2,349	**17.704**
environmentally	13,142	**0.084**	175	**1.082**	269	**1.590**	177	**1.334**
extinct	4,935	**0.032**	48	**0.297**	69	**0.408**	186	**1.402**
global warming	11,708	**0.075**	166	**1.027**	64	**0.378**	350	**2.638**
green space /greenspace	5,147	**0.033**	96	**0.594**	1,411	**8.341**	142	**1.070**
habitat	41,961	**0.268**	2,567	**15.876**	5,315	**31.421**	2,513	**18.940**
sustainability	28,390	**0.181**	1,035	**6.401**	597	**3.529**	384	**2.894**
sustainable	85,021	**0.543**	2,357	**14.577**	1,557	**9.205**	1,574	**11.863**
unsustainable	3,544	**0.023**	52	**0.322**	39	**0.231**	113	**0.852**

Feature	Researcher 1	Researcher 2
Collocates in UKWaC	neutral: *gardening*, *holder*, *plot*, *gardener* negative: *derelict*, *disused*, *unused*, *overgrown*	relating to gardens: *gardening*, *gardener*, *garden* relating to ownership: *smallholding*, *rent* negative: *derelict*, *disused*, *overgrown*
Positive/negative evaluation	neutral or positive, relating to their uses and benefits	neutral or positive, relating to their uses and benefits
Frequencies in specialised corpora	*allotment* is most frequent in the government corpus, and rare in the academic corpus	*allotment* is most frequent in the government corpus, and rare in the academic corpus
Other differences between specialised corpora	–	academic corpus: focus on the lack of allotments
Other findings:	positive use of negative collocates	positive use of negative collocates

Table 8: Comparison of two researchers' analyses of *allotment*