First catch your corpus: methodological challenges in constructing a thematic corpus[1]

Alison Sealey

Chris Pak

Abstract

This paper describes the process by which we have constructed a corpus of heterogeneous texts about non-human animals. It aims to contribute both methodologically – in respect of the challenges of compiling a thematic corpus – and substantively – in relation to the identification of some features of discourse about animals. Having introduced the research project and its guiding questions, the article describes the principles of data selection and the procedures used in analysis. We highlight the methods we devised both to avoid the potential circularity associated with pre-determined search terms, and to overcome the limitations of a relatively small corpus containing a wide range of relevant vocabulary. We go on to report some initial findings on the most frequent animal naming terms and adjectives describing them, including a small case study of the adjectives 'live' and 'dead'. The article concludes by indicating the ways in which the iterative methods we have employed are open to further extension, and points to some methodological and substantive implications of this enterprise.

* * * * *

This paper reports on methodological issues raised by constructing a corpus of heterogeneous texts that are all, in some way, about non-human animals. Although some of the issues are specific to our particular theme, we hope to contribute to the generic enterprise of corpus construction by exploring the

[1] Acknowledgements. This research is funded by the Leverhulme Trust (RPG 2013-063). We are grateful to the other members of the project team - Guy Cook, Clyde Ancarno, Anda Drasovean, Emma McClaughlin – and, while we take responsibility for the article, we should also like to thank everyone who provided data, practical help and/or comments on previous drafts, including: Laurence Anthony, Vaclav Brezina, Nickie Charles, Elsevier, Mel Evans, Gaëtanelle Gilquin, Nick Groom, Susan Hunston, George Jacobs, Dan Johnson, the Mass Observation Project, Joe McNamara, Lee Oakley, Kate Pearce, Alice Roberts, Jasper Roe, Paul Thompson and the Centre for Corpus Research, University of Birmingham, and two anonymous referees.

challenges associated with building a corpus whose focus is thematic, in contrast with the better-established methods of building corpora to facilitate description of a specific linguistic variety or the discourse associated with specified genres, for example. The first section situates the construction of the corpus within the aims of the research project, and the second explains the parameters we set around potential data. In section three we summarise the iterative procedures we followed for identifying and selecting data for inclusion, before presenting some initial findings in section four. Here we provide an overview of the most frequent animal naming terms, going on to explore the adjectives in the corpus used to describe animals, followed by a small case study focusing on two of these. We conclude by drawing out the contribution of this approach to corpus construction for the analysis of discourse using corpus-assisted methods and for the enterprise of extending it to language about the non-human.

1. Background to the project: 'People', 'products', 'pests' and 'pets': the discursive representation of animals

The motivations for our project include a growing recognition (within and beyond academia) of the importance of exploring the boundaries and relationships between humans and animals. Developments of various kinds – biological, technological, political – as well as 'a new sense of our precariousness as a species in the face of ecological threats and climate change' (Rose, 2013: 4) are obliging social researchers to shift from largely exclusive concerns with the relations and interactions between people to include considerations of the other living organisms with which we share the planet. Research in this field is developing fast (e.g. journals such as *Society & Animals*; *Anthrozoös*), and includes debates among natural scientists and philosophers about the language used to talk and write about animals (for an overview see Sealey and Oakley, 2013), while applied linguists are beginning to acknowledge the posthuman (e.g. Pennycook, 2016) and discourse analysts are turning their attention to environmental issues in general (e.g. Alexander, 2009; Fill and Mühlhäusler, 2001; Harré et al., 1998), and the linguistic representation of animals in particular. Such studies are often motivated by concerns about the way animals are treated by humans (e.g. Stibbe, 2001, 2003, 2012, 2014; Dunayer, 2003; Glenn, 2004; Goatly, 2002, 2006; Kemmerer, 2006; Keulartz and van der Weele, 2008; Kheel, 1995); and analyses – often mainly of vocabulary – may highlight parallels with the representation of stigmatised social groups. Of the studies using corpus analysis to find patterns across large numbers of texts, most use a general corpus or the internet as the source of data (e.g. Gilquin and Jacobs, 2006; Gupta, 2006). Probably the most comprehensive work to date on issues such as those with which our project is concerned is Stibbe (2012), which takes a Critical Discourse Analysis approach to texts from a wide range of genres. As he observes, ‘the language used to describe animals in a Disney documentary is quite different from that of a slaughterhouse instruction manual,’ (2012: 7) and his book is divided accordingly into chapters organised around such thematic distinctions.

In deciding how to compile our own corpus, we wanted to ensure that we would be able not only to explore empirically claims found in existing commentaries, but also to identify any unanticipated trends or patterns – the ‘non-obvious meanings’ (Partington, 2008) which corpus-assisted discourse analysis can help to identify. For this reason, rather than using our intuition to predetermine sets of either animals or sub-topics (such as meat production, for example) about which to look for texts to include, we have instead taken an eclectic and inductive approach, as explained below. The overarching question guiding this strand of the project is ‘How are animals in the corpus represented by the language used?’ A more specific sub-question, which we consider towards the end of this paper, is: ‘What kinds of description are associated with different kinds of animal?’

2. The parameters of the corpus

2.1 A thematically organised corpus

Corpus type (e.g. general, learner, diachronic) is linked with analytic or applied objectives (linguistic description, pedagogy etc.), and our own project is discourse analytic (Partington et al., 2004; Conrad, 2002). Even within this kind of approach, compiling a corpus according to a shared topic or theme is not yet a well-established practice and could be considered controversial, or even ill advised (Sinclair, 2005), although there are exceptions. For example, Baker (2006: 26) cites the topic-based corpus compiled by Johnson et al. (2003) as one example of a ‘specialized corpus’, which he suggests is ‘perhaps the most important type of corpus (in terms of discourse analysis),’ and he and his colleagues have used newspapers as their corpus data to explore the topics of refugees, asylum seekers, and immigrants (e.g. Baker et al., 2008; Gabrielatos & Baker, 2008). Of course, the labels applied to the people so designated are themselves not neutral (Pace and Severance, 2016). Likewise, determining the vocabulary that denotes the topic of ‘animals’ is not unproblematic, as we explain below. On the other hand, embodied animals are less of a discursive ‘construct’ than the social categories by which human beings are classified (see Sealey, 2014). Nevertheless, the first challenge we faced was to identify linguistic items that we were confident denote (instances of) these entities.

2.2 The boundaries of our theme: what is an animal?

Engagement with the idea of 'animal' as a category can challenge common sense, rather as the findings from corpus analysis may challenge assumptions about language. Despite our intuitions that we know what animals are, the classification of living beings as 'animals' is not unproblematic (e.g. Dupré, 2001, 2002, 2012; Ingold, 1988; Margulis, 2007). Part of our quest is to discover what people include in their own concepts of 'animals', and to explore both how far the casual, practical taxonomies of the non-specialist and those of 'experts' overlap, as well as to identify any evidence there may be for linguistic categorisations influencing people's perceptions of, and behaviour towards, various kinds of creatures in different registers. We maintain that this focus on terms denoting the non-human has implications for corpus-assisted discourse analysis more broadly, in that the selection of search terms for any such project necessarily makes assumptions about relations between: individuals, the categories by which they are classified, and the linguistic labels for both.

Numerically, the smallest animals constitute the largest proportion of the world's creatures: arthropods, a group that includes insects, arachnids and crustaceans, outnumber mammals by a ratio of about 312 to 1 (Basset et al., 2012), while micro-organisms are even more numerous, their diversity exceeding that of all other life-forms (Dupré, 2012: 165). Yet research into 'folk' categories (Atran and Medin, 2010; Berlin, 1973; Berlin et al., 1973) and the 'semantic primes' of natural kinds suggests that it is the more visible, larger categories of creature that people typically name when prompted to report on their awareness and experience of 'animals'. So while we have included in our corpus texts featuring spiders, snails and insects, we took the decision to exclude from our working definition of names for 'animals' words that denote any type of creature that is not normally visible to the naked eye.

2.3 Which linguistic variety?

We are investigating discourse about animals in 'contemporary British English' (although one of the PhD students funded by the project is conducting a contrastive study between English and Romanian discourse about five specific animals, and the other a diachronic study of news discourse about three wildlife species since the mid-eighteenth century). For the main corpus we take two decades as our time-frame, from 1995 – 2015. Preliminary work has identified differences in attitudes towards human-animal relationships between, for example, the USA and the UK, where norms - and indeed legislation - in this domain are different, and the language varies too. However, the identification of any text as being in 'British English' is somewhat problematic (c.f. critiques of the reification of languages and of 'native speakerism'), and some of the texts in which we are interested may be co-authored by several writers with varied linguistic backgrounds. So we include discourse originating wholly or mainly in the UK, insofar as its provenance can be established.

2.4 What kind of discourse?

As with many other studies of a similar kind, our potential 'universe' of discourse is huge and unquantifiable. We cannot hope to establish either the nature or the extent of all the texts that could potentially be included in our corpus, nor even to claim that it constitutes a representative sample of all the discourse about animals that exists in contemporary British English; (c.f. Leech, 1991: 27, on representativeness as 'largely as an act of faith'). To some extent, our decisions about what to include must be made on pragmatic, practical or opportunistic grounds. On the other hand, in line with the emphasis in corpus analysis on procedures that are replicable (e.g. Stubbs, 1996; 1997), we present the principles guiding our decisions. Although there continues to be debate in the literature about representativeness and sampling in corpus construction, there is general agreement that there is a potential circularity in pre-selecting the terms associated with a given theme. To avoid this, 'the design criteria … should be external to the texts, relating to the use of language in a recognized context that exists outside the realm of language analysis' (Adolphs, 2006: 21; Sinclair, 2005). In line with this approach, key parameters in data selection for our corpus are the orientations, interests and purposes expressed in discourse about animals from a range of genres. Again, we argue that the potential of this approach is not restricted to our topic. For example, the dominance of news as a source of corpus data about a wide range of issues in (critical) discourse analysis is understandable: not only are news texts increasingly readily available in digital form, but, as critical discourse analysts recognise, they are a powerful conveyor of dominant ideologies (e.g. van Dijk, 2013). However, if stances towards a theme are likely to vary across genres – as is the case here – then we believe it is advisable to be wary of restricting a corpus to news texts alone.

We identify in the title of our project some of the principal orientations which people may have towards animals, and our complete list, based on the literature (e.g. DeMello, 2012; Herzog, 2010; Ingold, 1988), is more extensive. Animals feature in human experience and discourse as: objects of observation, study or entertainment (in the 'wild', in laboratories, in zoos); companions; tools (for transport and/or work); commodities (for meat, other edible products, fur and clothes), competitors (with each other and with humans, in sport, as quarry in hunting, racing, fighting) and ‘out of place’ (‘pests’ / ‘vermin’). These are not mutually exclusive categories: creatures hunted for sport, such as ‘game’ birds or fish, may then be eaten; ‘pests’ or ‘vermin’ may be executed clinically (e.g. by fumigation) or hunted down in sporting rituals (e.g. foxes). Likewise, there are often no neat divisions between kinds of animal and orientations towards them: a dog may be treated as a pet and also used for guarding the home or acting as a blind person’s ‘sight’; a sign outside a farm, observed by a colleague, read ‘Rabbits for sale: pet or meat’; Herzog (2010) recounts the attentive care afforded laboratory rats and mice, in contrast to the casual way in which those that escape and become ‘pests’ are dispatched.

Part of our strategy, then, is to include a range of ‘interests’ in our data, in the sense of the beliefs, attitudes and values of those producing the discourse, given that people with differing views about our topic are obviously likely to represent the creatures they talk and write about in different ways – one person’s ‘laboratory experiment on a specimen’ is another person’s ‘torture of a sentient being’. We include a wide range of potential stances towards our topic, since part of our goal is to discover more subtle contrasts in the ways animals are represented than the obvious, semantically loaded ones such as these, given that even apparently ‘neutral’ discourse is saturated with evaluative stance (Hunston, 2007, 2010; Martin and White, 2005; Stubbs, 2001).

We have also included texts from a range of registers and genres (or ‘discourse types’, c.f. Partington 2010). Like ‘orientations’, ‘discourse genres’ don’t present themselves unproblematically in unambiguous categories, and they cut across our other parameters (as these also cut across each other). We have included texts with various functions (acknowledging that any text may perform more than one). At the most general level are the functions of ‘inform’, ‘instruct’, ‘persuade’ and ‘entertain’. To deal with the last one first, the representation of fictional animals in literature is a rich and interesting area of study, and could be explored in an extension of our project. However, since our focus is on actual creatures that exist independently of our discourse, and the representation of experiences of, perceptions of, and beliefs about real animals, one of our boundaries excludes fiction from this corpus. Other kinds of ‘entertainment’ genres include media broadcasts about animals, which also serve to educate or inform. The dissemination of research findings in academic journals is primarily informative, and press reports about animals are meant to inform too, but they may simultaneously take persuasive positions (for example on an issue such as badger culling). Commercial texts such as advertisements for animal products are inherently persuasive, though they may be instructive also, as well as being legally obliged to contain factual information, while the purpose of texts produced in support of campaigns is persuasive by definition. We have included examples of all of these in our corpus, recognising that these kinds of purposes are realised by different text types in different modalities. We acknowledge that this spreads coverage of each type relatively thinly, but believe that, for a topic which is under-researched, this potential disadvantage is outweighed by the opportunities offered for discovering patterns in different kinds of discourse that would not be found in a corpus of one genre alone, such as news, for example.

2.5 Where is discourse about our topic to be found?

In this section we explain how we dealt with the 'corpus-theoretical paradox' (Aarts and Bauer, 2008) which faces anyone compiling a corpus that involves selecting search terms in order to identify relevant texts. In our case the challenge is particularly daunting, because of the huge number of terms that can denote an 'animal'. Yet it is also worth noting that few topics investigated by discourse analysts present themselves with objective, concise lists of relevant terms. Mautner (2015: 157) claims that 'phenomena with varying and unpredictable lexical realizations' are much less suitable for corpus analysis than studies that 'crystallize … around discrete lexical items', and she illustrates the latter with findings about the patterning of the single words *unemployed* and *hardworking* in a corpus of news texts. However, the selection of even items such as these rests on prior knowledge about which words are likely to shed light on the topic of interest. Our approach to search term selection accords with that advocated by Biber (1993: 243), where '[t]he actual construction of a corpus … proceed[s] in cycles: the original design based on theoretical and pilot-study analyses, followed by collection of texts, followed by further empirical investigations of linguistic variation and revision of the design.'

If discourse about a topic is known to feature in identifiable types of source, then those sources are useful in the initial phase of collecting texts for the corpus. In our case, television programmes about the lives of animals are one obvious source. Along with these we have access to an extant corpus of texts about organic and non-organic food (Cook, 2007), from which we have extracted the texts relating to animal products. A further source of this kind is organisations promoting particular positions in relation to animals, such as charities and campaign groups. We have aimed to use text-external criteria to guide the selection: sites listing UK-based animal pressure groups were consulted to identify the more influential ones, and publicly available texts from the websites of these organisations are included in the corpus. Legislation that is explicitly about animals is similarly identifiable in a fairly direct way, and was selected using Lexis Nexis. These kinds of texts constitute one section of our data.

A second source of data is elicited from people who are asked to reflect on their responses to our topic. We have been given access to two data-sets collected by a sociologist who studies the role of animals in people's lives. One is 103 written responses to a 'directive' in 2009 on 'Animals and humans' by the Mass Observation Project. (For more detailed information, see Sealey and Charles, 2013.) The other comprises transcripts of interviews by Charles with guardians/keepers of companion animals. In other, qualitative strands of our project, we have conducted 17 metadiscursive interviews with producers of texts such as those included in the first type of data outlined above, as well as 9 pairs of reflective focus-group sessions with readers and viewers from a range of backgrounds and with varying kinds of interests in our topic. The aim here is to complement the corpus analysis (what is said/written) with evidence from producers and audiences for such texts, thus attending to the 'triangle of communication' (Cook, 2004). Transcripts of this elicited data are included in our corpus.

The third types of texts for our corpus are less readily identified. Facing similar challenges to ours, Gilquin and Jacobs (2006) used the BNC as data, and various Internet sources[2] as a basis for their list of search terms comprising 914 words for animals, including: '(a) general and common nouns (e.g. *bird, cat, dog, horse*); (b) nouns for males, females, and offspring (e.g. *calf, mare, stag*); (c) some specialized nouns (e.g. *drosophila, ostracod, whydah*); and (d) the main breeds of cats, dogs, and horses (e.g. *angora, collie, shire*).' These authors have been kind enough to share the resulting list with us, and we have incorporated it into the process by which we have filtered the search terms used to identify candidate texts for inclusion in the corpus, as explained below.

[Insert Table 1 about here]

3. Piloting and selecting

[2] http://dir.yahoo.com/Science/Biology/Zoology/Animals__Insects__and_Pets/Complete_List_of_Animals_by_Name/Complete_Listing/, http://www.mcwdn.org/Animals/AnimalsIndex.html, http://www.greenapple.com/~jorp/amzanim/aninfct.htm, http://www.geocities.com/RainForest/4076/indexlist.html, http://www.enchantedlearning.com/subjects/animals/Animalbabies.shtml, http://dir.yahoo.com/Science/Biology/Zoology/Animals__Insects__and_Pets/Mammals/Cats/Breeds/, http://dir.yahoo.com/Science/Biology/Zoology/Animals__Insects__and_Pets/Mammals/Dogs/Breeds/All_Breeds/, http://dir.yahoo.com/Science/Biology/Zoology/Animals__Insects__and_Pets/Mammals/Horses/Breeds/

To summarise, we have explained the challenges we faced in identifying appropriate data for our corpus of 'discourse about animals', along with the principles we have adopted to meet them. We have set parameters around: linguistic variety, time period, genres, stance and interests of text authors, and categories of animals, including the ways in which people are likely to interact with them, and we have been able to take advantage of previous studies (Cook et al., 2003; Cook, 2007), some student projects about this theme, the support of co-operative colleagues and organisations, and pilot analyses of sub-sets of the data already collected (Sealey and Charles, 2013; Sealey and Oakley, 2013, 2014).

For incorporation into the corpus, each text was processed so that all the data is consistent in layout and searchable in identical ways using corpus tools. Pdfs were converted into plain text files using *AntFileConverter* (Anthony, 2015a). Tags for headings, paragraph breaks etc. are used consistently with simple XML markup, and each text has been given a unique identifier. Core attributes are recorded both in a metadata spreadsheet and in-text headers, for maximum flexibility with different corpus tools.

3.1 The composition of the initial corpus and 'Master List' of potential search terms

The second cycle of this approach to corpus construction uses the existing mini-corpus (in this case around a quarter of a million words) to generate a list of search terms to use with more general discourse types. That is, from this corpus of texts that are self-evidently and/or elicited to be about our topic, we used *AntConc* (Anthony, 2014) to derive a complete word list, manually identifying words within it that denote animals, which generated a list of 419 items. This was a labour-intensive process, necessitating not only the identification of words denoting animals but also the checking of unfamiliar terms (using the *Encyclopedia of Life*[3] in the first instance, or a Google search if no hits were returned, often due to mis- or alternative spellings). In addition to variant spellings, other anomalies and inconsistencies had to be resolved (e.g. automated searches for 'pig' will by default not distinguish hits on 'guinea pig'; 'humpback whales' are referred to variously as 'humpback' and 'humpbacks' in the plural).

The eventual composite list of candidate 'words for animals' we refer to as our 'Master List', which was subject to continuing modification in light of the iterative processes outlined here. We went on to compare the words identified by the

[3] http://eol.org

processes described above with the 914 terms used by Gilquin and Jacobs in their 2006 study (see Section 2.5); we included many of these terms in our Master List, but excluded those that denote organisms not visible to the naked eye.

3.2 Using and refining the Master List

Refinement and consolidation of this Master List was a rather paradoxical process: the list was extended 'vertically' as new terms occurred in texts that we added to our collection, and at the same time the list of items to explore in depth was shortened, as we paid attention to the 'horizontal' dimension – that is, whether items occur across all or most genres and text types. Therefore, we calculated which terms to focus on in detail by identifying within each sub-corpus:

- where in a ranked list each naming term for an animal occurs, when all the types in any given text are ranked by frequency
- the normalised frequency of each term: each naming term for an animal that occurs in one of these sub-corpora is ranked for its frequency per 1000 words, to take account of the varying sizes of the sub-corpora
- the percentile rank in the frequency list in each case (i.e. the ranking of a term within a wordlist converted into a percentage).[4]

As we added to the range of material from texts that are self-evidently about our topic, it became more apparent which animals are named most consistently across different genres of discourse (see Section 4.1)

3.3 Using the Master List to identify new texts

Once complete, the Master List comprised a set of terms that could be used for identifying texts whose topic is animals, but which originate from general sources that are potentially about a wide range of subjects. For example, we

[4] To explain this further: to identify which terms occur more or less frequently across the different genres we have collected, simple ranking comparisons are inadequate. For example, a corpus containing 1000 types and one containing 100 types may each contain a particular term that is ranked in fiftieth position in their respective frequency lists. However, being ranked at 50 out of 1000 (and thus in the highest 5% of words used) represents a different significance of the term within the corpus than being ranked fiftieth out of 100 (and thus in only the highest 50% of words used). Therefore, the ranking of each term has been converted into a percentile figure using this formula in *Excel*: '((Types - Ranking) / Type)*100'.

received access to a corpus of 10,000 articles sampled from 50 academic journals (mainly scientific) published by Elsevier in the period 2001–2010. Some, but by no means all, feature animals, so we used the words denoting animals from the initial Master List to generate a single string (…pigeon OR sheep OR puppy OR squirrel OR bear OR hamster OR pony OR collie OR insect OR cow OR budgie OR frog OR hen …), identifying 2000+ articles as potentially relevant. The titles and abstracts suggested by this process were then manually checked, and false hits eliminated.

The most heterogeneous data source on which we drew was newspapers. Readers may wish to consider how far the challenges we encountered in identifying the relevance of texts returned by our search terms are specific to our topic. One of these challenges is the return of texts including only metaphorical uses of a search term. There is a rich seam of animal-related idioms and metaphors running throughout the language, and dozens of variations of the HUMAN IS ANIMAL metaphor (see Goatly 2006), so we needed to process returns of references to 'moles' denoting 'insiders who spy', for example, of 'lamb' that turned out to be in the phrase 'mutton dressed as lamb', and multiple instances of the idiom 'dog eat dog'. While metaphor is a significant issue for computational management of linguistic data generally (e.g. Association for Computational Linguistics, 2013), the extent of the porosity in the boundaries between discourse about humans and other animals is likely to be a more substantive finding from our research (e.g. Sealey, 2016). This is evident not only in metaphorical uses of animal naming terms, but also in the range of proper nouns that are identical to words for animals (e.g. individuals and organisations such as 'Peter Bird', 'Fox News', 'Badgers Healthcare Providers'), while we also encountered organisations named the 'Foxhounds' or 'Hounds', where the term was used to denote both hunters and dogs.

In light of this, we could have restricted our search within newspapers to events in which animals are central to a story (such as the horse-meat 'scandal' or the badger cull), selecting perhaps only texts with multiple mentions of specific animal naming terms. However, this approach would overlook the more incidental ways in which animals are referred to in news texts, which is contrary to the principle of corpus construction that we have established, of not presupposing what is likely to be present in the discourse. Given that our topic involves so many potential search terms, we could not automate the elimination of 'noise' in the way some other researchers have done (Kantner et al., 2011). Therefore, we devised our own procedures to identify a wider range of news discourse about animals while eliminating irrelevant texts.

Ten newspapers were selected with reference to a list of those with high circulation,[5] compiled from the ABC and reflecting the most up-to-date reports on circulation figures; they are: *The Sun, Daily Mail, Daily Mirror, Daily Telegraph, Daily Express, Daily Star, The Times, The Guardian, The Independent, Evening Standard*. For each year in our time-span, six evenly spaced months were chosen, and four numbers were randomly generated for each month of these years. All articles for the first of the randomly generated dates for each newspaper were identified. If this date fell on a day with no articles published, the second of the four randomly generated numbers was selected, and, if that was not suitable, the third was used, and so on. The returned articles were searched using the string of terms in the Master List, and all the articles that were returned were downloaded. A second random number was generated to select a single article from each of the days. If the article was irrelevant – either because its reference to animals was metaphorical, or because there was only a single, passing reference to an animal – another randomly generated number was used until a relevant hit was obtained and these articles were saved - a time-consuming process. This procedure resulted in a news sub-corpus comprising 1023 texts (466340 words).

4. Initial findings

The initial findings presented in this section are of substantive interest for our research into contemporary British English discourse about animals. But they are also of methodological interest, in that they illustrate the results of corpus construction guided by the iterative, largely inductive (and partially opportunistic) approach described above, as opposed to one based on the *a priori* selection of a unitary data source or search terms based only on intuition.

4.1 Most frequent naming terms

Our Master List has, as indicated above, foregrounded the most frequent words for animals found across a wide range of genres. The total number of animal terms stands at approximately 2600, distributed across sub-corpora as indicated in Table 2. Table 3 lists the 20 items that occur across at least nine of the sub-corpora, with the percentile rankings for each.

[Insert Table 2 and Table 3 about here]

[5] http://en.wikipedia.org/wiki/List_of_newspapers_in_the_United_Kingdom

The animals that are referred to consistently frequently in this heterogeneous corpus, as identified in this list, may be grouped by various properties. Two of the naming terms are superordinates (*fish, bird(s)*). While all the others are naming terms for mammals, it is noticeable that *mammal(s)* is much less frequent, consistent with the notion that it acts as the default, implicit superordinate; moreover, the naming terms for mammals denote those that feature prominently in human experience. That is, the terms are for 'domestic' animals, although this is itself a fairly imprecise term, and one that derives from human priorities and practices: a 'domestic' animal is defined by the Cambridge Dictionary as one 'that is not wild and is kept as a pet' (which includes *dog(s)* and *cats*), 'or to produce food' (*pig(s), cow(s), cattle, sheep*). Apart from *fox*, the other three species named could potentially be food, but there are cultural connotations (stigma, taboo or snobbery) to the consumption of the flesh of these animals that do not apply to the others. Moreover, rabbits are an archetypal mixed category ('pest', 'pet', 'food'), while horses are used for labour and sport. In this sense, *fox* is anomalous, in that it names a species about which our corpus suggests there is social concern, but of a different sort from the other most commonly named kinds of animals.

4.2 Animals described: illustrative examples

Readers will no doubt have realised that with so many search terms relevant to our theme, and such a relatively small corpus, the available data for identifying patterns around any particular animal naming term is limited. We conclude this section with a description of how we facilitated analysis by tagging all the animal naming terms, to two levels of delicacy. Each occurrence of an animal naming term from the Master List was tagged with a symbol ('¬'). In addition, a further code was devised according to common, 'folk' categories for classes of animals: 'amphibian', 'bird', 'fish', 'insect' (in which we include arachnids, for pragmatic reasons), 'mammal', 'mollusc' (gastropods and cephalopods) and 'reptile'. Because the word 'worm(s)' appears frequently in the corpus, denoting creatures of various types, it was assigned a separate code. Three other categories were also used: 'ambiguous' was assigned to terms that are used for two or more animals of different types (often scientific terms such as 'Japonicus' and 'Obesus'); the 'other' category includes animals that could not be assigned to the other categories, such as crustaceans. It was at this stage that, despite not having used words for animal products as search terms, we recognised the usefulness of tagging the category of 'products', for words denoting animal-derived substances, such as meat, eggs and sperm. An R script was used to cross-reference the Master

List with each file in the corpus, inserting the class code before the animal term, and the symbol '¬' following each animal naming term.

These operations allow us to identify the frequency of co-occurrences of specific words with all animal terms, with specific classes of animal, or with specific animals. For example, we analysed the frequency of adjectives to the immediate left of the tagged animal naming terms by searching a POS tagged corpus for the animal symbol and any preceding adjectives. In order to ensure that the frequency of occurrences of a particular adjective and the animal symbol was comparable across the variably sized sub-corpora, log-dice values were calculated, following Rychlý (2008). First, an aggregate of the frequency of the specific adjective and the frequency of the animal symbol within one sub-corpus was divided by the frequency of co-occurrences of the two terms (within the same sub-corpus), multiplied by 2. A log-dice value was then generated by adding 14 to the binary logarithm (log2) of the result. An average log-value for each adjective was then calculated to derive a single log-value for the particular adjective within one sub-corpus, and across all sub-corpora. Using this method, we were able to identify the most frequently occurring adjectives preceding an animal naming term for each sub-corpus, and for the corpus as a whole.

As an illustration of our approach to the analysis of our corpus, we outline in this section the inductive approach with which we are addressing the sub-question, 'what kinds of description are associated with different kinds of animal?'. Using the process described above, we generated a list of the adjectives (as classified by *TagAnt*, Anthony 2015b) that collocate with an animal naming term; we report here on those occurring most frequently in attributive position immediately before the noun (i.e. in L1 position with the animal-name-tag as node)[6]. There are some false hits when the animal term is itself a modifier of a subsequent noun (e.g. one of the occurrences of the frequent phrase 'good dog' is in the string 'good dog shampoo'); nevertheless, the process does allow us to identify some general trends.

Among the most frequent adjectives used to modify animal naming terms in our corpus are several that also feature highly in a general corpus such as the BNC[7]. In the list in Leech et al. (2001) of the top 50 adjectives in the BNC, *other* is ranked highest, and in our list it is second; *new* is in third position in both lists,

[6] Because the journal sub-corpus is much larger and more specialised than the other sub-corpora, we have excluded it from this analysis.

[7] The comparison is not exact of course as we are focusing on attributive adjectives for animal nouns only.

and *old* is fourth in both; *young* is 6th in our list and 14th in the BNC list; *black* is 7th / 40th; *small* 10th / 7th; *little* 18th / 21st; *different* 19th / 8th; *white* 21st / 48th; *large* 24th / 9th; *British* 28th / 16th; *big* 30th / 20th; *local* 31st / 10th; *only* 43rd / 37th; *certain* is ranked 45th in both lists; *whole* is 49th / 47th. Adjectives that rank high in our list but not in the BNC list include: *wild*, in top position, and *pet* in 5th. There are also colour terms for describing animals that are not high in the BNC's more general list: *red* (17th) *grey* (25th) *blue* (44th).

Other researchers from various disciplines (see citations above) have identified anthropocentric attitudes as a prevalent feature of contemporary treatment of animals. We could gloss this as the assumption of the centrality of human beings' concerns, perceptions and values. For the discourse analyst, the notion of 'stance' is useful in this context. Analysts using a range of approaches have demonstrated how both affective and epistemic stance are integral to the expression of propositions, while cognitive linguists (e.g. Langacker, 1991) draw attention to the influence of our biological, cognitive endowment on the ways in which we construe our experience through language. Langacker also notes 'linguistic evidence [for] the greater intrinsic prominence of entities that are … human vs. non-human' (1993: 449).

Analysis of stance in discourse typically focuses on items such as stance adverbs and modal expressions, while the adjectives receiving attention tend to be semantically evaluative (e.g. *possible, amazing, easy,* Hunston, 2007 citing Charles, 2004). However, even common English adjectives such as *big, small, little* and *large*, as well as colour terms, denote a property of some entity as it is perceived in relation either to an observer or to some other entity. The fact that the observer, and/or the evaluator of relative size, is almost invariably human almost – but not quite – goes without saying. Our species' biological endowment leads us to prioritise in language the characteristics that we perceive most readily, whereas the perceptive capacities of many other species are very different, and some would probably be impossible to encode in human language. In identifying the adjectives in our own corpus that most frequently modify animal naming terms, we are reminded of this observation by Leech et al. (2001: 287) about the relative frequencies in the BNC of adjectives for regions and nations:

> This list makes the obvious, if unpalatable, point that a British corpus reflects the assumption that Britain stands at the centre of the known universe and that the importance of a region or nation diminishes roughly in proportion to its 'remoteness' from Britain.

Analogously, we might suggest that our list of adjectives frequently used to describe animals – like the list of the most frequent naming terms – reflects the assumption that what is important about them is defined by human concerns. In addition to descriptive adjectives that point to the properties of animals perceived visually by humans, two other relatively frequent modifiers of animal naming terms in our corpus are *live* and *dead*, at 9th and 12th positions respectively. These are both much higher than in the BNC list, where *dead* is ranked 99th and *live* 454th, suggesting that these qualities are particularly salient in discourse about animals. The choice of 'live' as an attributive adjective to describe an organic entity is a marked one. Non-organic entities such as 'broadcast' or 'performance', for example, may be classified as 'live' as opposed to 'pre-recorded', but human beings are implicitly presumed to be 'live', a quality which is thus not usually worthy of mention. Not so with non-human animals. We therefore conclude this indicative analysis of one aspect of our corpus with a closer exploration of these adjectives in context.

4.3 A small case study: 'live' and 'dead' animals

The absolute frequency of {*dead* + [animal naming term]} is greater than that of {*live* + [animal naming term]}, but the former ranks higher when the different sizes of sub-corpora are accounted for. The two strings are not evenly distributed among our sub-corpora, with these descriptions of animals predominating in campaigning literature and newspapers.

[Figure 1 about here]

Live premodifies 28 different animal naming terms (24 if plurals are not counted separately[8]), and *dead* 59 (48). There is some overlap: terms (singular or plural) which are premodified by both adjectives include: *animal, badger, bird, cow, fish, fox, rabbit, shark, sheep,* and *whale*. Among animals described in our corpus as 'live' but not 'dead' are: *chicks, cockerels, lambs,* and *pigeon(s)*, while among those premodified by 'dead' but not by 'live' are: *beetles, buzzards, crows, mice, salmon* and *turkeys*. The fact that speakers and writers draw attention to the live/dead status of animals points to ways they are experienced, perceived and valued in human society. This is illustrated by examples of concordance lines, which are grouped thematically below.

[8] Though note that plurals of animal naming terms are not always marked (e.g. *sheep, fish*) and some function as mass nouns – e.g. *lamb* (see discussion in Sealey & Charles 2013)

We classify one such theme as broadly ‘ecological’, including accounts in broadcast documentaries of how a dead animal represents food for others, while the fact that an animal may fall prey to another before dying naturally is also commented on:

vultures are quick to spot any opportunity. A **dead** yak has drawn a crowd	Broadcast (*Wild China*)
A **dead** tuna has attracted a deep sea conga eel and a six gill shark	Broadcast (*Earth Story*)
So this is a great bonanza for them [great white sharks]**-** the body of a **dead** whale. The carcass will draw in every great white for miles around	Broadcast (*Africa*)
the sturgeon is purpose built for hoovering up salmon eggs, small fish, crayfish and even the carcasses of **dead** salmon from the gravel river bed	Newspaper (*The Daily Express*)
These scavengers [turkey vultures] are quite prepared to attack **live** chicks	Broadcast (*Penguins: Spy in the Huddle*)
One cat brought in **live** fish from someone's pond and a **dead** guinea-pig	Mass Observation

In all these examples, despite the fact that the topic is behaviour among non-human animals, a human-like stance is inevitable. There is an anthropomorphic flavour to ‘drawn a crowd’, ‘great bonanza’, ‘quite prepared’, ‘purpose built’, and ‘hoovering up’. Although ‘scavenger’ may be intended as a neutral description (one that feeds on decaying matter), the connotations with behaviour that humans find disgusting are unavoidable.

A sub-set of this ‘ecological’ theme is where dead animals are potentially an indication of a problem – especially for humans. Most of these examples represent human health concerns, e.g.:

Health authorities ordered every villager to be vaccinated as soon as the **dead** birds sparked the alert	Newspaper (*The Sun*)
fears that bird flu is heading to Britain increased yesterday as dozens of **dead** turkeys were found on a farm	Newspaper (*The Daily Mirror*)
Oxygen levels in the Thames were reduced to virtually nil along a stretch from Kew, Brentford and Isleworth, with **dead** bream and roach piled up on the banks and floating belly up in the water	Newspaper (*The Guardian*)
Urgent analysis of the **dead** birds is being carried out by scientists from British Nuclear Fuels	Newspaper (*The Independent*)

Since the alert began at the weekend - when 63 **dead** swans and wild fowl were found on marshes near the village -, 100 domestic birds have been culled	Newspaper (*The Sun*)

In addition, some dead animals are commented on in ways that reveal social norms, e.g.:

Dead cat dumped on doorstep A HORRIFIED couple found a kitten's severed HEAD on their doorstep	Newspaper (*The Sun*)
It is suspected some of the terriers and bulldogs were being used in illegal fighting and badger-baiting. Cats and **dead** rats were found in some of the outbuildings, suggesting owners were attempting to blood young terrier pups	Newspaper (*The Daily Mail*)
Providing alternative food, such as **dead** mice, rats or chicks, stopped harriers taking any grouse	Newspaper (*The Daily Express*)

These examples illustrate how proscriptions against certain kinds of human behaviour towards animals sit alongside the acceptance of others, such as the assumption that grouse should be preserved from birds of prey so as to be available for humans to shoot and kill.

Another major theme concerns humans as consumers of animals, and the 'pivotal transitional stage' when 'livestock become deadstock' (Wilkie, 2010: 16). Wilkie explores in detail how farmed animals are differentially valued when traded before and after being killed, while the discursive representation of animals as 'stock' has been noted by Stibbe (2006: 66), who examines the way salmon are routinely represented 'in economic terms … as a *commodity* … equated with *grain* and *timber*'. Examples of this theme in our corpus include:

It's like the word, the way we use 'beef' to describe **dead** cow and it's the way we use 'pork' to describe pig meat and stuff like that	Focus Group (18-23 year olds)
paying for leather and sheepskin adds substantially to the slaughterhouse value of the **dead** animal and financially supports the meat industry	Campaign (*Animal Equality*)
Some prices for **live** lambs being sold at markets are the best for nearly three years	Newspaper (*The Guardian*)

And finally in this section, we note how ‘live’ serves to emphasise the objections of some speakers and writers to ways of treating animals that are accepted by others.

It's a **live** dog, it's just sitting there waiting for the thing to come down on its head and I don't know where the image was taken but just in a part of the world where the animal is just no value at all	Interview (CEO of the Badger Trust)
If you wouldn’t visit an animal circus, then you shouldn’t visit a **live** reindeer parade	Campaign literature (Compassion in World Farming)
A **live** bird is placed in one compartment of the trap, to act as a decoy for other birds	Campaign literature (Animal Aid)
Support Kent Action Against Live Exports (KAALE) by attending their demos at Dover Port when **live** animal exports take place	Campaign literature (Compassion in World Farming)
Shooting of free-running badgers has never been carried out before and is likely to be extremely difficult for the marksmen involved, especially at night, resulting in injured **live** badgers	Campaign literature (Badger Trust)

We cited above the point made by Mautner about the fit between corpus-assisted discourse analysis and ‘discrete lexical items’ (2015: 157). In the case of our topic – and perhaps similarly under-researched topics – it is the iterative approach to corpus construction and analysis described above that has helped us to recognise how, in context, the apparently factual descriptors, *live* and *dead*, play a role in the discursive representation of human orientations – practical, epistemic, attitudinal – towards animals.

5. Conclusion

Our specialised corpus, compiled around the theme of animals, has some limitations. Its relatively small size and heterogeneity preclude certain kinds of analysis. We have mitigated this to some extent by tagging the extensive set of naming terms to facilitate searching for patterns that would not otherwise have been apparent. We are also able to compare findings from this corpus with reference corpora, a process that reveals, among other things, bidirectional influences of the metaphorical uses of animal terms. That is, as with the term ‘scavenger’ discussed above, socio-cultural norms are implicit in descriptions both of animals in human terms and vice versa. The iterative process of data selection, processing and analysis continues. In light of the findings about the most frequently mentioned animals across the genres in the corpus, we can select search terms to collect additional, animal-specific corpora, with the benefit of indicators derived from our corpus of genuine ‘hits’ as opposed to ‘noise’. The elicited, metadiscursive interviews and focus groups indicate intuitions and expectations which can be further investigated from within our corpus, and, in light of patterns revealed there we can again consult larger reference corpora for confirmatory or conflicting evidence.

The aim of this article has been to set out the considerations underlying the procedures we have used to compile a thematic corpus, including specifically the steps we have taken to avoid the potential circularity in such an enterprise. We hope these will be relevant to other researchers working with corpus-assisted discourse analysis. While some very particular challenges about categories, classification and taxonomies are raised by our theme, we suggest that these aspects should always be questioned by researchers using corpus-assisted methods to explore the discourse associated with social issues.

Meanwhile, increasing attention is being paid to the specific theme of our corpus, and to 'the influence of language on the life-sustaining relationships of humans with each other, with other organisms and with the natural environment' (Ecolinguistics Association, n.d.). We think that this theme invites some far-reaching debates about our interconnected world, and hope that our analysis of this corpus has a contribution to make.

References

Aarts, B., Bauer, M., 2008. Corpus construction: A principle for qualitative data collection, in: Atkinson, P., Bauer, M.W., Gaskell, G. (Eds.), *Qualitative Researching with Text, Image and Sound: a practical handbook for social research*. Sage, London, pp. 19 - 37.

Adolphs, S., 2006. *Introducing electronic text analysis : a practical guide for language and literary studies*. Routledge, London; New York.

Alexander, R.J., 2009. *Framing Discourse on the Environment: a critical discourse approach*. Routledge, New York and London.

Anthony, L., 2014. *AntConc*, 3.4.3 ed, Tokyo, Japan.

Anthony, L., 2015a. *AntFileConverter*, 1.2.0 ed. Waseda University, Tokyo, Japan.

Anthony, L., 2015b. *TagAnt*, 1.2.0 ed. Waseda University, Tokyo, Japan.

Association for Computational Linguistics, 2013. *The First Workshop on Metaphor in NLP*, in: Association for Computational Linguistics (Ed.), *The First Workshop on Metaphor in NLP*, Atlanta, GA, USA.

Atran, S., Medin, D., 2010. *The Native Mind and the Cultural Construction of Nature*. MIT Press.

Baker, P., 2006. *Using Corpora in Discourse Analysis*. Continuum, London.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T., Wodak, R., 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, 273 - 306.

Basset, Y., Cizek, L., Cuénoud, P., Didham, R.K., Guilhaumon, F., Missa, O., Novotny, V., Ødegaard, F., Roslin, T., Schmidl, J., Tishechkin, A.K., Winchester, N.W., Roubik, D.W., Aberlenc, H.P., Bail, J., Barrios, H., Bridle, J.R., Castaño-Meneses, G., Corbara, B., Curletti, G., Duarte da Rocha, W., De Bakker, D., Delabie, J.H.C., Dejean, A., Fagan, L.L., Floren, A., Kitching, R.L., Medianero, E., Miller, S.E., Gama de Oliveira, E., Orivel, J., Pollet, M., Rapp, M., Ribeiro, S.P., Roisin, Y., Schmidt, J.B., Sørensen, L., Leponce, M., 2012. Arthropod diversity in a rainforest. *Science* 338, 1481-1484.

Berlin, B., 1973. Folk systematics in relation to biological classification and nomenclature. *Annual Review of Ecology and Systematics* 4, 259 - 271.

Berlin, B., Breedlove, D.E., Raven, P.H., 1973. General principles of classification and nomenclature in folk biology. *American Anthropologist* 75, 214-242.

Biber, D., 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8, 243-257.

Conrad, S., 2002. 4. Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics* 22, 75-95.

Cook, G., 2004. *Genetically Modified Language*. Routledge, London.

Cook, G., 2007. *The discourse of organic food promotion: language, intentions, and effects. ESRC End of Award Report, RES-000-22-1626*. ESRC, Swindon.

Cook, G., Reed, M., Twiner, A., 2003. “But it's all true!”: commercialism and commitment in the discourse of organic food promotion. *Text & Talk - An Interdisciplinary Journal of Language, Discourse Communication Studies* 29, 151–173.

DeMello, M., 2012. *Animals and Society: an introduction to human-animal studies*. Columbia University Press, New York.

Dunayer, J., 2003. English and speciesism. *English Today* 19, 61-62.

Dupré, J., 2001. In defence of classification. *Studies in History and Philosophy of Biological and Biomedical Sciences* 32, 203–219.

Dupré, J., 2002. *Humans and Other Animals*. Clarendon Press, Oxford.

Dupré, J., 2012. *Processes of Life*. Oxford University Press, Oxford.

Ecolinguistics Association, n.d. The Ecolinguistics Association: About. http://www.ecoling.net. Last accessed 1st May 2016.

Fill, A., Mühlhäusler, P., 2001. *The Ecolinguistics Reader: language, ecology and environment*. Continuum, London.

Gabrielatos, C., Baker, P., 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, 5 - 38.

Gilquin, G., Jacobs, G.M., 2006. Elephants who marry mice are very unusual: the use of the relative pronoun who with nonhuman animals. *Society & Animals* 14, 79 - 105.

Glenn, C.B., 2004. Constructing consumables and consent: a critical analysis of factory farm industry discourse. *Journal of Communication Inquiry* 28, 63-81.

Goatly, A., 2002. The representation of nature on the BBC World Service. *Text* 22, 1-27.

Goatly, A., 2006. Humans, animals, and metaphors. *Society and Animals* 14, 15 - 37.

Gupta, A.F., 2006. Foxes, hounds, and horses: who or which? *Society & Animals* 14, 107 - 128.

Harré, R., Brockmeier, J., Mühlhäusler, P., 1998. *Greenspeak: a study of environmental discourse*. Sage, London.

Herzog, H., 2010. *Some We Love, Some We hate, Some We Eat: why it's so hard to think straight about animals*. Harper Perennial, New York.

Hunston, S., 2007. Using a corpus to investigate stance quantitatively and qualitatively, in: Englebretson, R. (Ed.), *Stancetaking in Discourse: subjectivity, evaluation, interaction*. John Benjamins, Amsterdam, pp. 27 - 48.

Hunston, S., 2010. *Corpus Approaches to Evaluation: phraseology and evaluative language*. Taylor & Francis.

Ingold, T., 1988. *What is an Animal?* Unwin Hyman, London.

Kantner, C., Kutter, A., Hildebrandt, A., Püttcher, M., 2011. How to get rid of the noise in the corpus: cleaning large samples of digital newspaper texts. *International Relations Online Working Paper Series*.

Kemmerer, L.A., 2006. Verbal activism: 'Anymal'. *Society & Animals* 14, 9 - 14.

Keulartz, J., van der Weele, C., 2008. Framing and reframing in invasion biology. *Configurations* 16, 93 - 115.

Kheel, M., 1995. License to kill: an ecofeminist critique of hunters' discourse, in: Adams, C.J., Donovan, J. (Eds.), *Animals and Women: feminist theoretical explorations*. Duke University Press, Durham, NC, pp. 85-125.

Langacker, R.W., 1991. *Foundations of Cognitive Grammar*. Stanford University Press, Stanford, CA.

Langacker, R.W., 1993. *Universals of construal, Annual Meeting of the Berkeley Linguistics Society*, pp. 447-463.

Leech, G., 1991. The state of the art in corpus linguistics, in: Aijmer, K., Altenberg, B. (Eds.), *English Corpus Linguistics*. Longman, Harlow, pp. 8 - 29.

Leech, G., Rayson, P., Wilson, A., 2001. *Word Frequencies in Written and Spoken English, based on the British National Corpus*. Longman, London.

Margulis, L., 2007. Power to the protoctists, in: Sagan, D. (Ed.), *Dazzle Gradually: reflections on the nature of nature*. Chelsea Green Publishing.

Martin, J.R., White, P., 2005. *The Language of Evaluation: appraisal in English*. Palgrave Macmillan, London.

Mautner, G., 2015. Checks and balances: how corpus linguistics can contribute to CDA, in: Wodak, R., Meyer, M. (Eds.), *Methods of Critical Discourse Studies*. Sage, London, pp. 154 - 179.

Pace, P., Severance, K., 2016. Migration terminology matters. *Forced Migration Review* 51, 69 - 70.

Partington, A., 2008. The armchair and the machine: corpus-assisted discourse research, in: Taylor Torsello, C., Ackerley, K., Castello, E. (Eds.), *Corpora for University Language Teachers*. Peter Lang, Bern, pp. 95-118.

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

Partington, A., Morley, J., Haarman, L., 2004. *Corpora and Discourse*. Peter Lang, Berlin.

Pennycook, A., 2016. Posthumanist applied linguistics. *Applied Linguistics* amw 016, online.

Rose, N., 2013. The human sciences in a biological age. *Theory, Culture & Society* 30, 3–34.

Rychlý, P., 2008. *A lexicographer-friendly association score*, in: Sojka, P., Horák, A. (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN)*. Brno: Masaryk University, pp. 6 - 9.

Sealey, A., 2014. Cats and categories—reply to Teubert. *Language and Dialogue* 4, 299-321.

Sealey, A., 2016. *What do we talk about when we talk about animals? , The Animal Challenge to the Social Sciences*, University of Leicester.

Sealey, A., Charles, N., 2013. “What do animals mean to you?”: naming and relating to nonhuman animals. *Anthrozoös* 26, 485-503.

Sealey, A., Oakley, L., 2013. Anthropomorphic grammar? Some linguistic patterns in the wildlife documentary series *Life*. *Text & Talk* 33, 399–420.

Sealey, A., Oakley, L., 2014. Why did the Canada goose cross the sea? Accounting for the behaviour of wildlife in the documentary series *Life*. *International Journal of Applied Linguistics* 24, 19 - 37.

Sinclair, J.M., 2005. Corpus and text - basic principles, in: Wynne, M. (Ed.), *Developing Linguistic Corpora - a guide to good practice*. Oxbrow Books, Oxford, pp. 1 - 16.

Stibbe, A., 2001. Language, power and the social construction of animals. *Society and Animals* 9, 145-161.

Stibbe, A., 2003. As charming as a pig: the discursive construction of the relationship between pigs and humans. *Society and Animals* 11, 375-392.

Stibbe, A., 2006. Deep ecology and language: the curtailed journey of the Atlantic salmon. *Society & Animals* 14, 61 - 77.

Stibbe, A., 2012. *Animals Erased: discourse, ecology, and reconnection with the natural world*. Wesleyan University Press, Middletown, CT.

Stibbe, A., 2014. Ecolinguistics and erasure: restoring the natural world to consciousness, in: Hart, C., Cap, P. (Eds.), *Contemporary Critical Discourse Studies*. Bloomsbury Academic, London.

Stubbs, M., 1996. *Text and Corpus Analysis*. Blackwell, Oxford.

Stubbs, M., 1997. Whorf's children: critical comments on critical discourse analysis (CDA), in: Ryan, A., Wray, A. (Eds.), *Evolving models of language*. British Association of Applied Linguistics / Multilingual Matters, Clevedon, pp. 100 - 116.

Stubbs, M., 2001. *Words and Phrases: corpus studies of lexical semantics*. Blackwell, Oxford.

van Dijk, T.A., 2013. *News as Discourse*. Routledge, Abingdon.

Wilkie, R., 2010. *Livestock/Deadstock: working with farm animals from birth to slaughter*. Temple University Press, Philadelphia.

Tables and Figures

Sub-Corpus	No of Files	No. of Types	No. of Tokens
Broadcasts	83	19835	614378
Campaign literature	470	16488	306680
Legislation	843	10201	627127
Food websites	258	7503	87118
Journals	1609	93567	5698531
News	1023	28777	466340
Contributions to the Mass Observation Project	103	9931	174938
Focus groups	19	8277	229059
Interviews with text producers	17	8068	157664
Interviews with guardians/keepers of dogs	19	8698	309719
Total	**4444**	**211345**	**8671554**

Table 1. The composition of the corpus

Number of Sub-Corpora	No. of Terms	Cumulative Total No. of Terms
10	36	36
9	39	75
8	49	124
7	80	204
6	102	306
5	113	419
4	146	565
3	201	766
2	272	1038
1	1167	2205
0	421	2626

The figures in the column ‘No. of Terms’ show the number of items added as the threshold number of sub-corpora in which items appear is reduced. The other figure in this column is the cumulative total. I.e., when the threshold is reduced from 10 to 9, for example, 39 new items are included, in addition to the 36 already listed.

The 421 animal terms in the ‘0’ row are those that appeared in Gilquin and Jacobs’ list of 914 animal terms, but which did not appear in any of our sub-corpora.

Table 2. The distribution of terms for animals occurring across the 10 sub-corpora

Term	Broadcasts	Campaign	Legislation	Food Websites	Journals	Mass Observation	News	Focus Groups	Interviews	Dog Keeper Interviews	Average of Percentile Rankings
fish	99.35	98.65	98.92	98.07	99.91	97.82	99.71	97.73	94.58	95.29	98.00
birds	98.89	99.51	89.71	95.47	99.53	98.68	99.53	95.32	93.64	82.06	95.23
dogs	94.75	99.61	96.80	63.71	99.85	99.32	99.64	98.27	97.32	99.18	94.85
pigs	94.72	96.53	83.14	98.75	99.81	95.78	98.82	98.30	97.78	83.96	94.76
sheep	97.28	97.56	89.99	94.86	99.80	94.85	98.87	96.99	88.58	84.92	94.37
bird	97.46	99.31	91.68	93.58	98.70	97.43	99.41	92.71	90.77	79.25	94.03
pig	97.24	92.44	62.96	96.76	99.52	92.53	98.56	97.60	97.69	79.95	91.53
dog	90.81	99.70	97.55	30.16	99.63	99.62	99.68	98.71	95.72	99.71	91.13
cows	94.22	97.65	69.34	97.21	99.80	92.96	97.15	96.69	92.53	69.45	90.70
horses	87.15	99.52	91.72	50.11	99.54	98.00	99.33	94.37	90.31	92.25	90.23
cattle	87.30	96.49	89.83	94.35	99.81	89.99	98.94	95.18	91.10	36.70	87.97
horse	88.09	99.30	94.67	23.46	99.21	98.10	99.59	95.26	80.52	93.35	87.16
cats	94.35	99.08	83.96	0.00	99.59	99.38	98.87	96.18	95.65	97.56	86.46
chicken	80.94	95.86	22.83	98.88	98.24	95.58	97.93	94.30	89.08	86.28	85.99
fox	90.80	99.07	67.55	26.46	97.66	97.01	98.67	97.05	97.88	85.35	85.75
rabbit	75.88	97.05	72.57	45.80	98.94	97.95	96.99	91.33	87.88	92.41	85.68
cat	91.96	99.23	82.98	0.00	99.12	99.53	98.99	96.40	85.31	97.77	85.13
deer	96.27	98.27	94.59	85.81	98.39	88.85	97.71	94.27	93.55	0.00	84.77

cow	96.87	92.13	21.18	85.87	99.39	79.16	97.67	96.74	92.09	69.46	83.06
rabbits	73.96	97.94	79.37	12.82	98.43	97.21	94.84	93.08	91.68	90.93	83.03

Table 3. The 20 most frequent animal naming terms by percentile rank (i.e. the ranking of a term within a wordlist converted into a percentage) in each sub-corpus and in the corpus as a whole

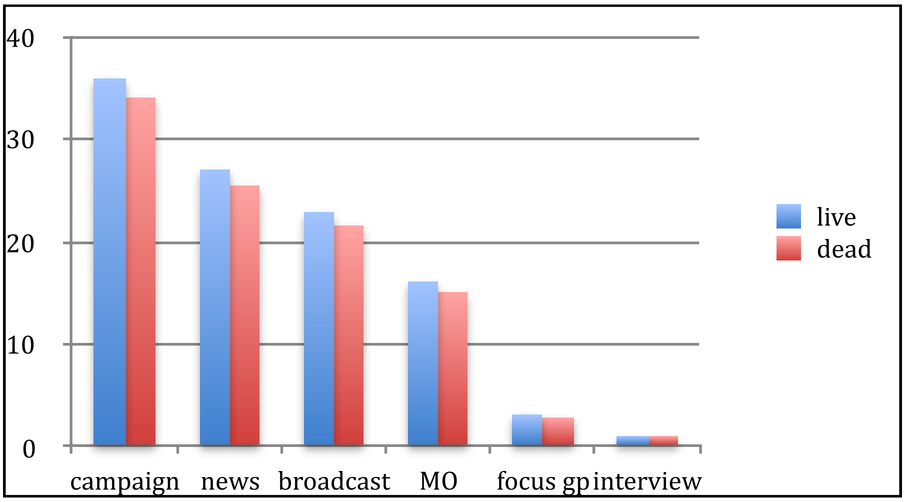

Figure 1. Sources of the adjectives *live* and *dead* immediately preceding animal naming terms by percentage from each of the sub-corpora (excluding the journal sub-corpus)