‘What is this corpus about?’: using topic modelling to explore a specialised corpus

Akira Murakami,[1] Paul Thompson,[1]
Susan Hunston[1] and Dominik Vajn[2]

Abstract

This paper introduces topic modelling, a machine learning technique that automatically identifies ‘topics’ in a given corpus. The paper illustrates its use in the exploration of a corpus of academic English. It first offers the intuitive explanation of the underlying mechanism of topic modelling and describes the procedure for building a model, including the decisions involved in the model-building process. The paper then explores the model. A topic in topic models is characterised by a set of co-occurring words, and we will demonstrate that such topics bring us rich insights into the nature of a corpus. As exemplary tasks, this paper identifies the prominent topics in different parts of papers, investigates the chronological change of a journal, and reveals different types of papers in the journal. The paper further compares topic modelling to two more traditional techniques in corpus linguistics, semantic annotation and keywords analysis, and highlights the strengths of topic modelling. We believe that topic modelling is particularly useful in the initial exploration of a corpus.

Keywords:

1. Introduction: exploratory techniques in corpus linguistics

One of the methodological challenges in corpus linguistics is how to approach a specialised corpus to discover what might be said of significance about it, but to do so with as few constraining preconceptions as possible. Important advances in the field include:

- Identifying what is distinctive about the corpus in question, by comparing word frequency with that in a more general corpus (e.g., using the Keywords function in WordSmith Tools; Scott, 1996);

[1] Department of English Language and Applied Linguistics, University of Birmingham, United Kingdom.
[2] Livesey House, University of Central Lancashire, Preston, PR1 2HE.
Correspondence to: Akira Murakami, *e-mail*: a.murakami39@gmail.com

- Characterising the semantic fields of the corpus in question using semantic annotation (e.g., Wmatrix; Rayson, 2008); and,
- Establishing frequently occurring phraseologies, variously defined as *n*-grams, phrase frames (Fletcher, 2007) or concgrams (Cheng *et al.*, 2009).

The advantage of all these methods is that they recontextualise the information the researcher deals with, by focussing on what is most frequent and/or what is most distinctive about the specialised corpus that is subject to investigation. In this paper, we shall argue that this might also be reductive. We investigate an alternative approach to word co-occurrence called 'topic modelling' and demonstrate how it may be used as a starting point for the investigation of a specialised corpus. We propose that corpus linguists may wish to adopt this method in initial scoping studies of a target corpus.

As its name suggests, 'topic modelling' might be conceptualised as a way of describing what a text is about. It is an alternative to keywords and semantic tagging, both of which could be said successfully to identify the 'aboutness' of a corpus. Unlike keywords, topic modelling operates on a single corpus, and does not depend for its operation on identifying what is most different about two corpora. Unlike supervised semantic annotation, the categories identified by topic modelling emerge from the methodology and the corpus rather than being predetermined.

In fact, the term 'topic modelling' is something of a misnomer. As described in detail below, the technique identifies lists of words which have a high probability of co-occurrence within a 'span' that is set by the researcher, but that is typically of hundreds or thousands of words. The co-occurrence, therefore, lies within a whole text, or a few paragraphs, but not within the short span used in studies of collocation. These groups of co-occurring words characterise 'topics', and researchers may choose to refer to them using topic-like titles, but these are only convenient abstractions from lists of words. The 'topics' may be of very different kinds. For example, here are the co-occurring sets of words in four topic-lists extracted from our corpus:

(*a*) *Forest*, *carbon*, *deforest*, *tropic*, *land*, *area*, *cover*, *conservation*, *forestry*, *timber*
(*b*) *Risk*, *health*, *disaster*, *effect*, *hazard*, *disease*, *people*, *affect*, *reduce*, *potential*
(*c*) *Should*, *right*, *principle*, *this*, *distribution*, *not*, *equitable*, *which*, *justice*, *or*
(*d*) *More*, *than*, *less*, *not*, *greater*, *also*, *much*, *other*, *however*, *rather*

List or Topic (*a*) appears to belong to an objective description of forestry conservation and deforestation activities, and is quite easily labelled as 'forest conservation'. List or Topic (*b*) takes a more value-driven assessment of physical risk and could be labelled as 'hazards'. List or Topic (*c*) includes more grammatical words (*should*, *this*, *not*, etc.) and is again value-driven but focusses on moral equity and the distribution of resources. It could be labelled 'equity'. List or Topic (*d*) is much less obviously a 'topic', though it is related to the genre of research papers that constitute our corpus. The words in it can be shown to relate to evaluations of research findings, but a specific label is more difficult to find.

As we shall explain below, the number of 'topics' identified in a corpus is specified by the researcher. Choosing a larger or smaller number will give a greater or lesser degree of granularity in the topics. For example, the following three topic-list beginnings from our corpus could be considered as a single topic, 'agriculture/farming', but there is value in considering them separately:

(*e*) *Crop*, *production*, *agriculture*, *soil*, *food*, *yield*, *increase*, *fertility*, *use*, *plant*
(*f*) *Land*, *area*, *agriculture*, *use*, *cultivation*, *cattle*, *population*, *livestock*, *pasture*
(*g*) *Farmer*, *household*, *income*, *farm*, *village*, *migration*, *livelihood*, *food*, *rural*

List or Topic (*e*) might be said to be 'agriculture as an economic activity'. List or Topic (*f*) suggests 'agriculture or farming as a human activity'. List or Topic (*g*) might be said to construe farming on a smaller scale – the household or village rather than the nation. The differences between the lists are at the same time intuitively meaningful and difficult to capture in words.

Lists (*e*) to (*g*) demonstrate another feature of topic modelling: lists of words are not exclusive but overlap. The word *agriculture* appears in (*e*) and also in (*f*); *food* occurs in (*e*) and in (*g*). If more of the lists were shown, the overlaps would be greater.

In the next section of this paper, we describe the background to topic modelling. In Section 3, we describe the corpus and method we used in our study. Section 4 gives some results that, we suggest, demonstrate the usefulness of this way of studying 'aboutness'. In Section 5, we compare topic modelling with other ways of identifying 'aboutness' and consider the advantages and disadvantages of each. In the final section, we consider the implications for corpus linguistics and its use in the characterisation of specialised discourse.

2. Probabilistic topic models: an overview

Probabilistic topic modelling is a machine learning technique that automatically identifies 'topics' in a given corpus (Blei, 2012). It has been applied in various areas, including sociology (e.g., DiMaggio *et al*., 2013), digital humanities (Meeks and Weingart, 2012), political science (Grimmer, 2010), literary studies (e.g., Jockers and Mimno, 2013), and most notably, academic discourse (e.g., Blei and Lafferty, 2006, 2007), among others. Latent Dirichlet allocation (LDA; Blei *et al*., 2003) is the approach to topic modelling that has been most frequently employed recently. The following explanation of topic models describes LDA and is largely based on Blei (2012). In topic modelling, each word type in each text is assigned to one topic-list. A text consists of multiple topics of different probability (e.g., 30 percent Topic A, 15 percent Topic B, 20 percent Topic C), approximately following the proportion of word tokens in the text that are assigned to each topic. All of the texts in a corpus share the same set of topics, but with different proportions.

As noted above, a topic in turn is construed by a probability distribution over a fixed vocabulary. Certain words (e.g., *pollution*) are more likely to occur under a certain topic (e.g., 'environment') than under another topic (e.g., 'Shakespeare'). Topic-lists of words can be ordered by the strength of the probability of co-occurrence. The characteristic words of a topic can be viewed as keywords of the topic, and, similarly, the texts with a high probability of a topic can be viewed as key texts of the topic. In this sense, a topic is a recurring pattern of word co-occurrence (Brett, 2012).

Topic modelling works on a 'bag of words' principle. That is, it is linguistically naïve and pays no attention to the grammatical or semantic connections between words. Multiple estimation procedures have been proposed for topic models. Below, we explain how the estimation procedure called collapsed Gibbs sampling (Griffiths and Steyvers, 2004) works. In assigning a topic to a token, the following two principles apply:

(*1*) Tokens in a text receive as few topics as possible.
(*2*) Tokens of the same word type receive as few topics as possible across the texts.

Point 1 means that if a word in a text is assigned to Topic X, the other words in the same text are more likely to be assigned Topic X. Point 2 means that if a word is assigned Topic X, the other occurrences of the same word in the corpus are more likely to be assigned Topic X. These two principles compete with each other. For example, let us consider the case in Table 1.

==Insert Table 1 about here==

This tiny corpus contains six word types with three tokens each across three texts. In topic modelling, analysts need to decide the number of topics identified in the corpus. Let us say that we want to identify two topics in the corpus, and suppose that Topic 1 was assigned to *romeo* in Text 1. Based on Principle 1, we should then also assign Topic 1 to the other two words in the same text (*juliet* and *hamlet*). Now, since *hamlet* was assigned to Topic 1, based on Principle 2, all the other occurrences of *hamlet* should also be assigned to Topic 1. There is only one other occurrence of the word, and so we assign Topic 1 to *hamlet* in Text 2. Again, following Principle 1, all the other words in the same text should be assigned to the same topic. Thus, *environment* and *ozon* in Text 2 are assigned to Topic 1. Finally, based on Principle 2 again, all the other occurrences of the two words should also be assigned to the same topic, and so should the other words in the same text as them. This leads to the assignment of Topic 1 to *climate* in Text 3. Notice that following the two principles led to a single topic with all the words in the corpus, although we wanted to identify two topics.

Topic modelling balances the two principles. In the above case, for example, *environment* and *ozon* in Text 2, as well as all the words in Text 3, may be assigned Topic 2. This violates Principle 1 because Text 2 includes multiple topics. With this sacrifice, however, we can identify two topics as we wished. Note that this is achieved through trial and error. In the process, for instance, all of the words in Texts 2 and 3 may receive Topic 2, which violates Principle 2 as the two occurrences of *hamlet* receive different topics, but otherwise satisfies the two principles. This possibility is likely to be rejected on the grounds that, in further texts, *hamlet* co-occurs much more frequently with *romet* and *juliet* than with *environment*, *ozon* and *climate*, which makes it more reasonable to assign the same topic to *romeo*, *juliet* and *hamlet*.

Despite the apparent relevance of a topic-identifying technique to corpus linguistics, topic modelling has gained little attention in the field. This is perhaps not surprising given that the technique is linguistically naïve. Linguists are typically, and justifiably, suspicious of methods based on a ‘bag of words’ hypothesis. Nonetheless, this paper demonstrates the usefulness of topic models in corpus linguistics in the context of the investigation of a particular academic discourse.

3. A topic model of academic discourse

3.1 Corpus

Our corpus consists of research papers published in the journal Global Environmental Change (GEC). The corpus includes all of the articles from the first volume (1990/1991) to Volume 20 (2010). In compiling the corpus we targeted only full-length articles and did not include non-research papers, such as book reviews. The corpus includes the main body text but excludes other sections of research papers such as the abstract or appendices. Tables and figures are not included, either. Mathematical symbols and equations have been replaced with the non-word *EQSYM*. The corpus includes 675 research papers and consists of 4.1 million words.

Based on external criteria, it is a specialised corpus; it is large enough to preclude hands-on reading of all the texts in it as a way of surveying the corpus.

In taking a topic modelling approach, an initial decision we have to make is what to conceive as a text. In our study, we wished to take a more fine-grained approach than would be captured by considering each research paper as a single text. A research paper may contain several topics. A paper, for instance, might have a high frequency of theme-related words such as *environment* or *pollution* at the beginning, while the same paper may have a high frequency of method-related words such as *analyze* or *experiment* in the middle. To capture the within-paper shift of topic of this kind, each paper was divided into multiple blocks, each constituting a 'text' as defined in the topic model. More specifically, each block included a minimum of 300 words, and a block was not permitted to cut across a paragraph boundary, on the assumption that paragraphs themselves are topic-based units. For example, suppose that a paper consists of six paragraphs, and the number of words in each paragraph is as follows:

Paragraph 1:	240 words
Paragraph 2:	150 words
Paragraph 3:	80 words
Paragraph 4:	200 words
Paragraph 5:	50 words
Paragraph 6:	100 words

In this case, the first block includes Paragraphs 1 and 2 because Paragraph 1 alone does not include 300 words, whereas Paragraphs 1 and 2 combined do. Similarly, the second block has to contain Paragraph 3 to Paragraph 5 because Paragraph 3 alone or Paragraphs 3 and 4 combined do not reach 300 words, while Paragraphs 3 to 5 combined do. However, the only remaining paragraph, Paragraph 6, does not include 300 words, and it is unwise to exclude this paragraph from the analysis because we may miss potentially interesting information about the ending paragraph of research papers. Therefore, in the example above, the preceding block was extended to the final paragraph; as a result, this hypothetical paper has two blocks in total which cover all six paragraphs. The division of papers into blocks allows us to investigate topic transition within papers.

Topic transition within text-blocks is assumed to be smaller than that between text-blocks because neighbouring paragraphs tend to belong to the same section and are likely to be topically related. This, however, does not mean that topic modelling assumes topically uniform text-blocks. Rather, a strength of the technique is that each text-block includes a mixture of topics (Blei, 2012). Therefore, even if a text-block includes paragraphs with different topics, it does not significantly affect the identification of topics in the corpus.

Most of the previous literature using topic models excludes the words in the standard stop words list (e.g., Marshall, 2013) because function words and pronouns provide little information about the topic of texts. Those words, however, are potentially informative in characterising research papers linguistically; pronouns, for example, enact engagement between writer and reader (Hyland, 2005). We therefore excluded far fewer words as stop words: only prepositions, articles, *and*, *it*, *as*, *that*, *be*, *have* and *do* (and the inflected forms of the last three verbs). This ensured that we retain the potentially important insights brought by closed-class words. We also excluded one-letter words because they included much noise such as various abbreviations (e.g., *p* for 'page') and statistical values (e.g., *t*, *F* and *r*).

All the words were stemmed (e.g., *require* → *requir*, *analysis* → *analysi*) with the Porter stemming algorithm (Porter, 1980). Stemming was employed rather than lemmatisation because lemmatisation requires a dictionary, and specialised corpora tend to include words that are too infrequent to be included in a typical lemmatisation dictionary. For instance, *acequias*, *biogenics* and *Carpathians* are not lemmatised as *acequia*, *biogenic* and *Carpathian*, respectively, in Someya's list.[3] Removing inflectional and derivational morphemes through stemming may collapse the words that should ideally be distinguished and lead to information loss (Sinclair, 1991). Without stemming, however, input data (i.e., document-term matrix) can be too sparse and we may not be able to target as many words as we can when stemming is applied (see the paragraph below). Here, we follow a common practice in topic modelling and opt for retaining as many root words as possible. The effect of stemming, however, has not been investigated in topic modelling literature, and there is no doubt that the issue should be addressed in future research.

To ensure the reliability of results, the topic model only targeted the 7,758 word types that occurred in at least 0.1 percent of all the text-blocks. Stemming and the removal of short or infrequent words are common pre-processing steps in topic models (e.g., DiMaggio *et al*., 2013; and Marshall, 2013). Each 300+ word text-block was assigned with information on where in the paper it appeared (e.g., 70 percent from the beginning of the paper).

Table 2 shows the numbers of papers and text-blocks across publication years. After excluding the stop words mentioned above, the corpus included 10,555 text-blocks with the average length of 242 words (SD=50).

==Insert Table 2 about here==

[3] Someya's list can be downloaded from, for example: http://www.laurenceanthony.net/software/antconc/.

3.2 Model building and selection

We used the topicmodels package (Grün and Hornik, 2011) in R (R Core Team, 2015) to build the topic models. There is no agreed way to automatically decide the number of topics (but see, for example, Ponweiser, 2012, for attempts). In other words, the decision on how many topics a corpus will be deemed to contain is a subjective one and the answer may be defended on the grounds of usefulness but not on the grounds of accuracy. As an exploratory step, therefore, we built models with 40, 50, 60, 70, 80, 90 and 100 topics, and we inspected them to decide the appropriate level of granularity with which to explore the data. We decided to use the model with sixty topics because the model with fifty topics lacked some of the potentially interesting topics that we observed in the sixty-topic model and the one with seventy topics included some apparently redundant topics. Appendix A shows the ten words (or word stems) with the highest probability of occurrence in each of the sixty topics.

This list of topics and other findings in this paper were deduced from the following pieces of information:

(*i*) The probability of each topic in each text-block (e.g., Topic 1 occupies 2.2 percent of Text-block 4);
(*ii*) The probability distribution of each topic over word types (e.g., the stemmed word *environ* occupies 3.2 percent of Topic 15); and,
(*iii*) The assignment of the topic to each word type in each text-block (e.g., the word *water* was assigned to Topic 10 in Text-block 7).

Figure 1 demonstrates topic distribution in some of the text-blocks and papers. The horizontal axis represents sixty topics and the vertical axis represents the corresponding probability in each text-block (2A) or paper (2B). In Figure 1A, the panel label is the identifier of the text-block. For instance, '1993_3_2_Glantz_0.91' indicates it is a block of paragraphs taken from the paper whose first author is Glantz and which was published in 1993 in Volume 3 Issue 2 of GEC, and the block is located at the point 91 percent away from the beginning of the paper (i.e., towards the end). This within-paper location indicates where the middle word in the block falls in the paper and was calculated by dividing the sum of the number of words before the block and half of the number of words in the block divided by the number of words in the paper. Figure 1A shows that different topics are prominent in different text-blocks, and that while some text-blocks have one very prominent topic and the other topics are weak (e.g., Topic 37 in the final panel), others have multiple prominent topics (e.g., Topics 1, 22 and 25 in the second panel).

==Figure 1 about here==

Figure 1B shows topic distribution at the level of papers. For this purpose, we averaged topic probability across all the text-blocks taken from the paper. Here, we chose four papers that have Topic 10 as the most prominent topic. The top ten keywords of Topic 10 are as follows: *water*, *river*, *basin*, *suppli*, *flow*, *irrig*, *resourc*, *avail*, *use* and *stress*. The topic can justifiably be summarised as 'water', and indeed, the titles of the four papers signal that water is their main topic:

> *Climate change*, *water resources and security in the Middle East*
> (1991_1_4_Lonergan)
>
> *Equilibrium and non-equilibrium theories of sustainable water resources management*: *Dynamic river basin and irrigation behaviour in Tanzania*
> (2007_17_2_Lankford)
>
> *A revised approach to water footprinting to make transparent the impacts of consumption and production on global freshwater scarcity*
> (2010_20_1_Ridoutt)
>
> *Virtual water 'flows' of the Nike Basin*, *1998–2004*: *A first approximation and implications for water security*
> (2010_20_2_Zeitoun)

Given the above, we labelled this topic 'water systems, supplies, and trade'.

Figure 1C visually represents the topic probability of individual words in a selection of the topics. Each row represents a topic with its interpretative label as given on the left (see Appendix A for the complete list of topics). The shading in each cell indicates probability, with a darker shade corresponding to higher probability. We can tell that topic probability for any given word is highly skewed: a word has a few prominent topics at most and has negligible probability for most of the topics. The word *polici* (*policy*), for instance, is highly probable in Topic 36 (Environmental policy actors, makers) but is practically absent in the other topics. Although still skewed, some words have a decent level of probability in multiple topics. For the word *area*, a fair amount of probability mass is allocated to Topics 19 (Forestry management), 23 (Wetlands, coastal, flooding) and 39 (About ecosystems and biodiversity). This means that the word is relatively frequent in the three topics compared to the other topics. The figure also shows that individual topics are characterised by just a few keywords. Topic 3 (Emission regulations), for example, includes a high frequency of *carbon*, *emiss* (*emission*), *greenhous* (*greenhouse*) and *level*, but not other words. Thus, these are the distinctive keywords of the topic. In this manner, topic models link topics and their keywords.

Appendix A shows the labels and keywords of our sixty topics. Many topics are straightforwardly thematic topics, such as Topic 3 labelled as ‘Emissions regulation’ and with keywords like *emiss*, *reduct*, *greenhous* and *co2*. Not every topic is thematic, however. Topic 30, for instance, has been labelled ‘Hypothetical discussion’ and captures the co-occurrence of the words that are often used in expressing speculation, such as *if*, *would*, *could*, *possibl* and *potenti*. This topic does not correspond to a topic in its usual sense of the word but represents the manner in which people write. In this way, topic models go beyond indicating textual ‘aboutness’, and give additional information about register and style (see Rhody, 2012).

The type of co-occurrence in topic models can be well understood in comparison to multidimensional analysis (MDA; Biber, 1988), another latent model that is often used in corpus linguistics. In MDA, analysts assume that there are latent (i.e., unobserved) dimensions that give rise to the co-occurrences of linguistic features. In topic models, we assume that latent topics invite word co-occurrences. In both cases, ‘co-occurrence’ takes the span of a few hundred to a few thousand words. In this sense, topic models differ from collocations, where the span is typically much shorter. Co-occurrences in topic models and in MDA often have situational reasons. In topic models, words co-occur typically because they are topically related and words under the same topic tend to co-occur, while in MDA, linguistic features co-occur because they are functionally related and those that serve the same function tend to co-occur (Biber, 1995).

4. Exploration of the model

As suggested above, many methods of manipulating corpus data essentially re-organise the word types in the corpus – for example, in order of frequency or significance or strength of co-occurrence – to give the researcher an alternative view to that which may be obtained from reading individual texts. In some cases, the research question that is posed will determine what organisation is most appropriate. For example, if the aim is to track diachronic change in the way an entity is represented, the starting point may be to identify the word or phrase types that are most significantly different in frequency between texts published in time (t) and those appearing at time t+1, t+2 and so on. These types then constitute the starting point for more detailed investigation. Questions such as these are predicated on there being relevant external criteria for identifying sub-corpora, such as the year in which a constituent text was published.

In some cases, however, the investigation may be more exploratory and a word type organisation may be sought that is not dependent on the prior identification of sub-corpora. Perhaps the most general question to ask of a corpus is: ‘what is this corpus about?’ The lists of words that are the outcome of the organising principle of topic modelling offer insights into the nature of the corpus under investigation without reliance on prior hypotheses. In this section, we firstly (Section 4.1) offer an interpretative overview of the sixty lists or topics shown Appendix A and then answer a series of more specific questions.

4.1 Surveying the topics in the corpus

As discussed above, the sixty 'topics' identified in the GEC corpus give different kinds of information about the corpus. Appendix A shows the ten words (or word stems) with the highest probability of occurrence in each topic. For convenience, each topic is also given a mnemonic label. Between them, the topics encapsulate and delineate what might be called the themes of the corpus. These include (in no particular order):

- Kinds of natural environment; for example, [*forest*, *carbon*, *deforest*, *tropic*, *land*, *area*, *cover*, *conserv*, *forestri*, *timber* = Topic 19]; [*flood*, *sea*, *rise*, *coastal*, *area*, *level*, *protect*, *impact*, *loss*, *sealevel* = Topic 23]; [*speci*, *biodivers*, *conserv*, *area*, *ecosystem*, *plant*, *divers*, *protect*, *veget*, *site* = Topic 39]
- Geographical locations; for example, [*local*, *scale*, *level*, *region*, *differ*, *spatial*, *nation*, *these*, *across*, *which* = Topic 32]; [*countri*, *develop*, *nation*, *world*, *intern*, *their*, *india*, *global*, *industri*, *most* = Topic 35]; [*region*, *africa*, *south*, *southern*, *europ*, *area*, *central*, *north*, *most*, *asia* = Topic 60]
- Kinds of human economic activity; for example, [*crop*, *product*, *agricultur*, *soil*, *food*, *yield*, *increas*, *fertil*, *use*, *plant* = Topic 4]; [*energi*, *use*, *fuel*, *effici*, *technolog*, *power*, *sector*, *transport*, *consumpt*, *industry* = Topic 5]; [*product*, *sector*, *trade*, *import*, *increas*, *export*, *consumpt*, *fish*, *market*, *economy* = Topic 34]
- Political institutions and actions; for example, [*govern*, *institut*, *actor*, *state*, *network*, *power*, *polit*, *author*, *their*, *role* = Topic 6]; [*polici*, *polit*, *this*, *issu*, *maker*, *question*, *decis*, *make*, *what*, *which* = Topic 36]; [*program*, *state*, *it*, *us*, *govern*, *agenc*, *nation*, *committe*, *offici*, *support* = Topic 52]
- Aspects of risk; for example, [*adapt*, *vulner*, *capac*, *or*, *sensit*, *social*, *cope*, *exposur*, *measur*, *abil* = Topic 9]; [*environment*, *global*, *problem*, *environ*, *econom*, *concern*, *issu*, *chang*, *secur*, *polit* = Topic 15];[*risk*, *health*, *disast*, *effect*, *hazard*, *diseas*, *peopl*, *affect*, *reduc*, *potenti* = Topic 20]
- Research actions; for example, [*group*, *respond*, *particip*, *interview*, *survey*, *their*, *question*, *they*, *respons*, *inform* = Topic 26]; [*studi*, *this*, *analysi*, *paper*, *approach*, *section*, *discuss*, *case*, *how*, *present* = Topic 38]; [*indic*, *variabl*, *measur*, *eqsym*, *valu*, *signific*, *index*, *effect*, *correl*, *relationship* = Topic 44]
- Groups of people; for example, [*individu*, *their*, *public*, *respons*, *action*, *peopl*, *they*, *behaviour*, *perceiv*, *percept* = Topic 16];

[*group*, *respond*, *particip*, *interview*, *survey*, *their*, *question*, *they*, *respons*, *inform* = Topic 26]

- Modelling the future; for example, [*model*, *use*, *simul*, *base*, *paramet*, *each*, *which*, *result*, *repres*, *function* = Topic 1]; [*will*, *futur*, *may*, *this*, *can*, *if*, *more*, *like*, *current*, *need* = Topic 11]; [*would*, *could*, *not*, *if*, *might*, *or*, *this*, *but*, *ani*, *should* = Topic 30]

This is by no means a comprehensive listing. It confirms and expands on the information given on the journal website:[4] this is a research journal about the natural world, human beings, and the interactions between them. The sixty topics encompass the scope of the journal, and between them give the observer a good intuitive 'feel for' the journal content.

As with any list of words, some more specific observations might be made. For example, a relatively large number of words refer to individuals or groups of people, but these tend to be at a high level of generality (e.g., *people*) or abstraction (e.g., *actor*, *decision-maker*, *committee* and *stakeholder*). More importantly, the topic lists serve to organise the words so that each word type is nuanced by the words it co-occurs with. For example, the natural entities of rivers, forests and oceans (Topics 10, 19 and 23) are transformed into entities used by or impacting on humankind: *river* co-occurs with *irrigate*/*irrigation* (10); *forest* co-occurs with *conservation* and *timber* (19); *sea* co-occurs with *flood* and *impact* (23). Most strikingly, perhaps, words to do with risk and its mitigation (problems and solutions) occur in no fewer than fifteen out of the sixty topics. One topic (20) connects general negative words such as *risk*, *hazard* and *disaster* with the human-related words *health* and *people* and with *reduce*/*reduction* – words associated with the mitigation of negative effects. As examples of other topics, Topic 3 connects *carbon* with *mitigation*, Topic 10 connects *water* and *river* with *stress*, Topic 15 connects *environment* with *problem*. *Forest* is connected with *conservation* (Topic 19). *Sea* and *coastal* are linked with both *protect* and *loss* (Topic 23). *Vegetation* is linked with *conservation* (Topic 39) and *pollution* is linked with *control* (Topic 45). Topic 54 connects *ecology* with *resilience*, while Topic 55 links *climate change* with *response*, *adapt*(*ation*) and *mitigation*. These co-occurrences provide detail of how natural and human entities are connected in the journal, and how entities are connected both with problems and the ways they may be addressed.

4.2 Within-paper topic distribution

We now turn to the question of how the topics are distributed within papers. This gives information about the organisation of papers in the journal. Figure 2A shows the distribution of each of the sixty topics. The

[4] See: http://www.journals.elsevier.com/global-environmental-change/.

horizontal axis represents the within-paper position, where 0 is the beginning of the paper and 1 is the end of it. Each line indicates the predicted probability of each topic based on the generalised additive model (Wood, 2006) that models topic probability based on text-block position. The cubic regression spline was used as the smoothing basis. We can see that different topics behave differently. Some topics are prominent at the beginning of the paper, while others are prominent at the end. Yet others show a U-shaped pattern or randomly fluctuate.

==Insert Figure 2 about here==

Figure 2B illustrates the distribution of the six topics whose relative probability decreases most radically from the beginning to the end of the paper (i.e., the topics with the lowest standardised slopes). The panels were ordered such that the topic with the most dramatic probability decrease (Topic 50: *et*, *al*, *2005*, *2003*, etc.) comes first, followed by the topic with the second most dramatic decrease (Topic 53: *al*, *et*, *1996*, *1995*, etc.), and so forth.

The figure also demonstrates the 95 percent confidence interval of the probability. The first two topics (50 and 53) are related to in-text citations, as exemplified by such keywords as *et* and *al*, and numbers representing years, and, thus, it is natural that their probability is high at the beginning of papers, where a literature review is typically located. Topic 38 appears to cover the overview of the paper, with such keywords as *studi*, *this*, *paper*, *approach* and *discuss*. Topics 27, 15 and 49 are more directly related to the contextualisation of papers. The keywords of Topic 27 include temporal expressions such as *year*, *period*, *recent*, *centuri*, *decad* and *past*, which provide the historical context of the paper. Topics 15 and 49 are similar in that they are both on specific issues (global environmental security issues and global warming, respectively). All six topics help to situate the paper in a wider context, and, thus, are more prominent at the beginning of papers.

Figure 2C similarly shows the topics whose relative probability most radically increases towards the end of the paper. They are all prominent in the discussion and 'future research' sections of the paper. Topic 40 directly discusses findings with such keywords as *more*, *than*, *less*, *rather*, *signific*, *high* and *differ*. Topic 30 relates to hypothetical discussion, as mentioned earlier, and is used to offer implications and speculations of the paper. Topic 11 is similarly related to discussion of the future, while Topic 12 encompasses the overall implication of the paper with words like *manag*, *plan*, *strategi*, *institut*, *learn* and *implement* as keywords. Topic 42 is another non-thematic topic that includes words related to discussion and evaluation. Here, we succeeded in identifying the paper structure with the topic model.

4.3 Chronological change of GEC

Topic models can also inform us of the chronological change within a journal (Blei and Lafferty, 2006; and Priva and Austerweil, 2015). Figure 3 shows the chronological topic transition obtained in a similar manner to Figure 2. Instead of the smooth curve based on generalised additive models, however, Figure 3 draws the average probability of each topic in each year. Figure 3A illustrates the topic transition of all of the sixty topics. We can observe a variety of patterns: different topics tend to be prominent in different years.

==Insert Figure 3 about here==

Figure 3B shows the transition of the six topics that are prominent in early years but decline in later years. These were identified in the same manner as in Figure 2B. The prominent topics in early years tend to describe particular problems. Topic 45 deals with pollution issues, while Topic 15 addresses environmental security. Topic 35 discusses problems in developing and developed countries, and Topics 49 and 5 are related to global warming and energy use, respectively. Topic 11 describes the predicted and potential impacts of the issues.

Figure 3C shows the transition of the topics that are prominent only in recent years. Topic 50 is prominent in the latter half only because it characterises in-text citations after 2000. The other prominent topics tend to address the people vulnerable to environmental change and the ways in which humans tackle environmental issues. Topic 9 is about how people can adapt to climate change and who are vulnerable to it. Topic 56 discusses the impact of environmental change on farmers, while Topic 12 is related to environmental management. Topic 24 deals with how environmental issues are discussed in the media, and Topic 18 pertains to how local communities adapt to environmental change with their local knowledge and traditions. The shift of the prominent topics above suggests that GEC set its research agenda in the first years by identifying environmental problems and in later years started to address those research agendas.

4.4 Identifying different types of papers

Topic models can also help to identify different types of papers. GEC is an interdisciplinary journal and includes a wide range of topics. We hypothesised that some papers focus on a single, perhaps specialised, topic, while others may address a variety of topics. To examine this possibility, we identified two papers with the highest and the lowest relative entropy (Gries, 2013), which in this case selects a paper whose topic distribution is heavily skewed and one where the distribution is relatively even.

Figure 4 illustrates the topic distribution of the two papers. The upper panels show the topic profiles of the whole papers, while the lower panels show the topic profiles at the level of text-blocks. In one paper, 2008_18_3_Hof, there is one very prominent topic (Topic 22: Explaining cost–benefit analyses in figures, especially damage), and all the others are nearly negligible. This tendency applies to individual text-blocks as well. In the other paper, 1992_2_2_Dahlberg, although a few topics tend to be more prominent than others, there is no single topic that is the strongest throughout the paper. The topic profiles, thus, suggest that Hof *et al.*'s paper focusses on a single topic throughout the paper, while Dahlberg's paper includes a number of topics. This is indeed what we find.

==Insert Figure 4 about here==

Hof *et al.*'s paper is entitled 'Analysing the costs and benefits of climate policy: value judgements and scientific uncertainties'. The paper, as the title suggests, addresses the costs and benefits of climate policy, and more specifically, computationally models the impacts of climate policy under various parameter settings. The paper heavily draws on an earlier modelling work, called the Stern Review, that also computationally modelled the economic impacts of climate policy, and regards its results as the benchmark. The paper is closely focussed on the reporting and discussion of their modelling work.

Dahlberg's paper is titled 'Renewable resources systems and regimes: key missing links in global change studies'. The paper, as mentioned earlier, contains a variety of topics, which are well illustrated in the abstract:

> The author argues that:
> - as we move towards a post fossil fuel era, societies will become more dependent on renewable resource systems;
> - current food and fibre systems at national and subnational levels are only partially understood because of the great emphasis placed on their production aspects;
> - at regional and international scales, agriculture, grazing, forestry, and fisheries overlap in multiple-use renewable resource regimes which are not captured with current concepts and data sets;
> - just as with other aspects of industrial society, hierarchical approaches and contextual analysis are needed to capture the full environmental, social, and technological dimensions of these systems and regimes; and,
> - only through a reconceptualization and rethinking along these lines will we be able to restructure current industrial systems in ways designed to develop more sustainable and regenerative systems.
>
> (1992_2_2_Dahlberg; list-formatting added)

Notice that the individual points above are not necessarily on the same theme. Thus, the abstract already signals that the paper includes a number of topics. In the main body, too, we can observe a topic shift by looking at the first sentences of two successive text-blocks:

> Multiple-use problems in categorization are especially difficult in defining grazing lands.
>
> (1992_2_2_Dahlberg_0.47)

> Coastal wetlands link directly into fisheries. Of the 11 million acres of coastal wetlands in 1780, half were gone by the mid-1970s.
>
> (1992_2_2_Dahlberg_0.49)

It is not surprising that the most prominent topic of the former text-block is Topic 33 labelled ‘Land use description’ and that of the latter is Topic 34 labelled ‘Fishing trade’. The example here thus illustrates that topic models can identify papers with radically different thematic structures.

4.5 Disambiguating the senses of polysemous words

A further strength of the topic model is that it often reveals how different senses of polysemous words behave (see DiMaggio *et al.*, 2013). We will illustrate this below with the word *level* as an example. Figure 5A demonstrates the within-paper change in the frequency of the word *level*. The line represents the fitted value of the generalised additive model that predicts the relative frequency of *level* in each text-block based on the within-paper position. Each observation (or text-block) was weighted by the number of words in the text-block. While the frequency of *level* fluctuates somewhat, we cannot observe a systematic pattern of change. This, however, is merely the aggregated pattern. Figure 5B illustrates the change in the probability of the seven topics where *level* is one of the top twenty keywords. Their interpretive labels are given below:

Topic 3:	Emissions regulation
Topic 22:	Explaining cost-benefit analyses in figures, esp damage
Topic 23:	Wetlands, coastal, flooding
Topic 32:	Spatial scope of human activities and decisions at different levels
Topic 44:	Variables and correlations
Topic 49:	Greenhouse gases, climate changes
Topic 59:	Population and other growth trends

Although *level* is a keyword in these seven topics, its sense varies across the topics. In Topics 3 and 49, the word is used to refer to the degree of concentration, as in the case of '[t]he present base level of atmospheric CO_2 concentration' (1993_3_4_Schulze_0.29847182425979). These topics behave similarly in Figure 5B in that they are clearly more prominent at the beginning of the paper than at the end. This is probably because the topics are on greenhouse gas emissions and the CO_2 level is often discussed to contextualise the paper.

In Topics 22, 44 and 59, the word refers to a position on a scale, as in 'societies tend to dematerialize above a certain level of wealth' (2010_20_4_Schandl_0.478278251599147). These topics show similar patterns in Figure 5B as well: their probability tends to be highest at approximately 60–70 percent from the beginning of the paper. This is because it is associated with the results section of the paper. For instance, a variable can be correlated with the levels of income or education. As in the case above, we can observe that similar senses of the word behave similarly within papers.

In Topic 23, *level* refers to a height or distance, as in the case of 'Wetlands are sensitive to sea-level rise as their location is intimately linked to sea level' (2004_14_1_Nicholls_0.420021895146576). The probability of this topic is high at around 60 percent as well. Here, again, the word, particularly in connection to the rise of sea levels, occurs in the results section of papers.

Finally, in Topic 32, the word refers to a relative rank on a scale, as in 'local governments may feel they are left little option but to use their powers at the local level to respond to regional level concerns' (1995_5_4_Millette_0.489480090419058). The probability of the topic is highest towards the end of the paper. This is presumably because the roles that the multiple levels of actors (e.g., international, national and regional) play are discussed in the conclusion of the papers. The discussion here illustrates that the topic model reveals the systematic pattern of the individual senses of a word that cannot be observed when the senses are aggregated and that, more generally, topic models can discriminate different senses of a word without any semantic information (see DiMaggio *et al*., 2013).

5. Contrasting with existing techniques

In this section we will contrast the topic model with existing methods in corpus linguistics. While we know of no technique that is directly comparable to topic models, we will attempt to highlight the differences with the following four techniques: (*i*) semantic tagging, (*ii*) keywords analysis, (*iii*) collocation networks, and (*iv*) concgrams.

The first two are often used to achieve the same goal as topic models, which is to gain insights into textual aboutness. The two techniques will thus be compared to topic models from this perspective. The demonstrative task we will tackle is the identification of chronological change in GEC. The latter two techniques are similar to topic models from a more methodological perspective: They identify word co-occurrence patterns. In the comparison below, therefore, we will discuss the differences between the techniques from a methodological perspective.

5.1 Topic models and semantic tagging

This subsection compares the topic model to the UCREL Semantic Analysis System (USAS) accessed through Wmatrix (Rayson, 2008). A critical difference between supervised semantic tagging such as USAS and the unsupervised topic model such as the one introduced in this paper is that the former assigns pre-specified categories to words whereas the latter finds (typically) semantically related groups of words in a bottom-up way. The granularity of semantic categories needs to be pre-determined in semantic tagging. Semantic tagging, thus, requires a sophisticated tagset and a dictionary. In topic models, however, the specified number of topics is identified inductively, and thus the granularity depends on the topical heterogeneity of the corpus and the number of topics identified in it. They do not require tagsets or dictionaries. Indeed, topic models can be run on any language, provided that the text can be tokenised.

To compare empirically the results of semantic tagging and topic models, we annotated our GEC corpus with USAS through Wmatrix. USAS assigns multiple tags to a token, but only the first candidate was retained. When the first candidate included two tags (i.e., double membership), both were retained as separate tags. Markers of the position on semantic scales (+ and –), those of semantic templates indicating multi-word units, and other symbols following the main tag (e.g., *f* standing for 'female') were removed.

To investigate chronological change in GEC, we identified key semantic fields in the first decade (1991–2000) and those in the second decade (2001–2010). This was performed through the Keyword List function in AntConc (Anthony, 2014) with one sub-corpus (e.g., papers published between 1991 and 2000) as the target corpus and the other (e.g., papers published between 2001 and 2010) as the reference corpus. Log-likelihood was used as the keyword statistic. To capture USAS tags with AntConc, a token was defined as a sequence of English alphabet characters, digits and full-stops.

Tables 3a and 3b list the resulting top ten key semantic tags in each decade and the five most frequent words for each tag in the target decade. For instance, the tag O1.3, which corresponds to the semantic category labelled as 'Substances and materials generally: Gas', was nearly three times as frequent in 1991–2000 papers as in 2001–2010 papers (33.7 *versus*

12.1 per 10,000 words), and the most frequent words with the tag in 1991–2000 papers were *CO_2*, *gas*, *gases*, *methane* and *ozone*.

==Insert Table 3a about here==
==Insert Table 3b about here==

Some findings match with the findings based on the topic model (e.g., 'substance and materials' in semantic tagging as the key semantic fields in the first decade *versus* Topic 45 labelled as 'Toxic substance and pollution management' as the key topic). The semantic category in USAS, however, is sometimes too coarse for our corpus. W5, labelled as 'Green issues' and including environmental terms, was the third key category in 1991–2000 papers and the five most frequent words were *environmental*, *environment*, *conservation*, *nature* and *pollution*. When we look at which topics those five words are the keywords of, we notice that, as expected, *environment*(*al*) and *pollution* are included in Topics 15 and 45 – the two topics that showed the most dramatic decline. The word *conservation*, however, is included in Topic 19 ('Forestry management') and Topic 39 ('About ecosystems and biodiversity') as one of the top ten keywords, and the probability of these topics remains relatively unchanged. This suggests that while some green issues such as air pollution are on a declining trend, the trend does not apply to other issues such as forest conservation.

A similar observation can also be made regarding key semantic fields in the latter decade. A key semantic domain in 2001–2010 papers is 'weather' and includes such words as *climate*, *rainfall*, *flood* and *climatic*. When we look at the topics whose keywords include these words, we notice that while the occurrence of the word *climate* in Topic 55 ('Mitigation, adaptation') increases over time, the reverse is true for Topic 49 ('Greenhouse gases, climate changes').

Therefore, there are sub-patterns within the single semantic category that the topic model can distinguish but the USAS model cannot. The topic model can thus provide a more fine-grained view of the thematic structure of the corpus.

5.2 Topic models and keywords analysis

Another potentially comparable technique is keywords analysis. Keywords in keywords analysis are a list of words that are more frequent in a corpus than in the reference corpus, and have been often associated with textual 'aboutness' (Bondi, 2010; and Scott, 2010). Aboutness here, however, is defined with reference to the reference corpus, and represents how the target corpus is different from the reference corpus. The need to pre-specify the reference corpus is potentially a drawback of the technique.

To compare empirically the topic model to keywords analysis, we identified keywords of the GEC papers published in the first decade and those in the second decade. As in the identification of key semantic fields earlier, we had one sub-corpus (e.g., 1991–2000) as our target corpus and identified the keywords using the other sub-corpus (e.g., 2001–2010) as the reference corpus. Log-likelihood was employed as the keywords statistic, and the analysis was undertaken using AntConc. All the words were stemmed. Digits were included in the token definition because numbers were occasionally present in the keywords of our topic model and among the words that are frequent in the key semantic tags identified earlier.

Table 4a and Table 4b show the top twenty keywords of each decade. The word *figur* is the top keyword in 1991–2000 papers only because figures were referred to as, for instance, *Figure 1*, until the 1998 volume but as *Fig. 1* afterwards. Numbers representing years after 1999 occupy eleven out of the twenty keywords in the second decade because they represent in-text citations after 2000. The keywords suggest that over time GEC came to deal less with the emission of greenhouse gases, such as methane, CFC and CO_2, and its impact on the global environment and more with vulnerability, adaptation and resilience related to environmental change. These are in line with our observation based on the topic model.

==Insert Table 4a about here==
==Insert Table 4b about here==

The topic model, however, often brings us more easily interpretable findings. From keyword analysis, we can observe that one of the keywords in the latter decade is *household*. The word, however, is difficult to interpret because there is no other keyword in the list that seems to be related to it in the first instance. When we look at Topic 56 (‘Households, village level’), which includes *household* as one of the keywords and is a topics that is prominent in later years, we can see that *household* co-occurs with such words as *farmer*, *farm*, *village* and *livelihood*. These words suggest that households in this context refer to those of farmers in villages. Combined with the increasing probability of Topic 9 (‘Vulnerability, adaptive capacity’) in later years, we can hypothesise that the households of farmers in villages are vulnerable to environmental change and need to adapt, and that this topic is on an increasing trend.

Part of the difficulty in interpreting the word *household* is due to the small number of keywords considered. The twenty-first keyword is *livelihood* and the thirty-second is *farmer*, both of which facilitate the interpretation of *household* in the same manner as above. However, it is only the topic model that automatically groups related words. In keywords analysis, researchers still need to reason that *household* is perhaps not related to some keywords like *fig* or *water*, but *vulner* and *adapt* are closely relevant. The topic model automates this process.

5.3 Topic models compared against collocation networks and concgrams

In collocation networks (Brezina *et al.*, 2015; and Williams, 1998, 2002), analysts specify a node word as the starting point and investigate the network of words where the edges represent collocation. Collocation networks are similar to topic models in that they both identify word co-occurrences.

There are, however, notable differences between the two techniques. First of all, collocation typically looks at co-occurrence patterns within an immediate environment around the node word (e.g., five words to the left and the right of the node word), whereas topic models target co-occurrences within more extensive texts. As a result, each method captures different aspects of meaning: collocation studies reflect the distributed or prosodic meaning associated with phraseology whereas topic models identify thematic meaning. Related to this, in collocation networks it is necessary to specify a node word, and the potentially subjective choice of the node word influences the aspect of the corpus that the technique can reveal. On the other hand, topic models target the entire corpus, reducing the arbitrariness of the analysis.

Yet another way to capture word co-occurrence patterns is through concgrams (Cheng *et al.*, 2006, 2009; and Warren, 2010). Concgrams identify the co-occurrence of words (e.g., *environmental problems*) that may be intervened by other words (e.g., *environmental health problems*) or may vary in position (e.g., *problems in environmental policy*). Comparison between concgrams and topic models is more or less similar to the comparison between collocation networks and topic models. Since the typical span used in concgrams is much smaller than the span used in topic models (i.e., text), the topic model can identify what is similar to themes in a corpus while the congram discloses more local meaning. Also, concgrams require analysts to choose a target word to analyse, which potentially introduces arbitrariness.

6. Conclusion

In this paper, we have demonstrated the use of topic models to explore a corpus of specialised English discourse. We gave some consideration to the role of topic models as a general exploratory technique. More specifically, however, we employed topic models to (*i*) investigate within-paper topical change, (*ii*) examine the chronological change of a journal, (*iii*) identify different types of papers, and (*iv*) differentiate multiple senses of words.

In our view, topic models are particularly useful in the initial exploration of corpora that are of large enough scale to preclude a manual approach such as reading each text. Corpus linguists often start exploring a large corpus by reading a sample of the texts in it or by making a word list of the corpus. The quantity of data may be managed by applying an annotation system, as in semantic tagging. This serves to classify the individual word forms and to provide an overview of the semantic content of the corpus. Alternatively, the corpus may be compared with a reference corpus to identify the words that are significantly more frequent in the target corpus. All these explorations have disadvantages. It is time-consuming to read sufficient texts adequately to understand the corpus. Word lists efficiently summarise the whole corpus, but the most frequent words tend to be grammatical words that give little information about what the corpus includes. Semantic annotation relies on pre-prepared semantic sets, and it is difficult to adjust it for level of granularity. Keywords presuppose that maximum distinctiveness is the most significant aspect of the content of a corpus, and can thus lead to a form of textual stereotyping; moreover this method presents the researcher with a simple rather than an organised list.

We have argued that topic models comprise a useful, bottom-up approach to a novel corpus that avoids the disadvantages of the other methods. They are a computational lens into the thematic structure of the corpus (DiMaggio *et al*., 2013), and each topic gives a sense of what the corpus is about. Consequently, topic models can help analysts to narrow down what specifically to look at in the corpus.

Furthermore, topic models are a relatively objective data-driven technique. Topic models receive very simple data as their input; a document-term matrix, which can be computed based on the bag of words of each text. Although the bag-of-words approach may seem too simple as means of representing a corpus, it works well in topic models to identify the thematic structure of the corpus. Furthermore, the bag of words does not require pre-specified categories. The meaningful outcome is achieved by the suite of sophisticated algorithms, and topic models can, thus, be described as linguistically naïve, relatively objective and computationally sophisticated.

Topic models are not free of limitations. In topic models, analysts need to consider carefully how to define a text because it is within texts that word co-occurrence patterns are identified. On the one hand, the fact that the concept of 'text' matters in topic models means that they take richer information into account than techniques that ignore text, such as keywords analysis, and as a result help us to identify multiple co-occurrence patterns of the same word. On the other hand, however, the topic probability distribution over texts and over words, as well as the keywords of each topic, changes when the definition of texts changes (for example, if we changed the minimum length required of our text word count from 300 to 200). This is potentially undesirable because the same corpus will then yield different summaries when texts are defined differently. A similar concern may be noted in the case for the choice of the number of topics. The number of topics determines the granularity of the model, and it is up to analysts to decide the number. Furthermore, topics in topic models require interpretive labels, which need to be assigned manually. Therefore, whereas topic models are objective in the sense that they do not require pre-specified categories and dictionaries, they still require analysts to make a number of decisions. This limitation, however, can also be seen as a strength. Identifying ‘topics’ or ‘aboutness’ is inevitably an act of interpretation. It is essentially qualitative and should not be disguised by quantitative methods. The fact that topic modelling demands two relatively arbitrary decisions at its outset means that the analytical subjectivity cannot be masked.

In this paper, we have only shown the use of the most basic type of topic models. Topic models have been extensively researched in machine learning and computational linguistics in recent years, and a number of improvements proposed. Here, we introduce a few of them. Firstly, a potential limitation of the topic model explored in this study is that it only targeted single words. While the bag-of-words approach combined with the sophisticated inductive technique can be illuminating, individual words alone may not capture all the themes of a corpus. To overcome the issue, several algorithms have been proposed to achieve *n*-gram topic models (e.g., El-Kishky, 2014). Further modifications to the original model include correlated topic models (Blei and Lafferty, 2007), which allow topics to be correlated, and dynamic topic models (Blei and Lafferty, 2006), which account for the chronological change of keywords within topics.

References

Anthony, L. 2014. AntConc. (Version 3.4.0.) Tokyo: Waseda University. Available online at: http://www.laurenceanthony.net/.

Biber, D. 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. 1995. Dimensions of Register Variation: A Cross-linguistic Comparison. Cambridge: Cambridge University Press.

Blei, D. 2012. ‘Probabilistic topic models: surveying a suite of algorithms that offer a solution to managing large document archives’, Communications of the ACM 55 (4), pp. 77–84.

Blei, D. and J. Lafferty. 2006. ‘Dynamic topic models’ in proceedings of the 23rd International Conference on Machine Learning, pp. 113–120.

Blei, D. and J. Lafferty. 2007. ‘A correlated topic model of Science’, Annals of Applied Statistics 1 (1), pp. 17–35.

Blei, D., A. Ng and M. Jordan. 2003. ‘Latent Dirichlet Allocation’, Journal of Machine Learning Research 3 (Jan), pp. 993–1022.

Bondi, M. 2010. ‘Perspectives on keywords and keyness: an introduction’ in M. Bondi and M. Scott (eds) Keyness in Texts, pp. 1–18.

Amsterdam: John Benjamins.

Brett, M.R. 2012. 'Topic modeling: a basic introduction', Journal of Digital Humanities 2 (1), pp. 12–16.

Brezina, V., T. McEnery and S. Wattam. 2015. 'Collocations in context: a new perspective on collocation networks', International Journal of Corpus Linguistics 20 (2), pp. 139–73.

Cheng, W., C. Greaves, J.M.H. Sinclair and M. Warren. 2009. 'Uncovering the extent of the phraseological tendency: towards a systematic analysis of concgrams', Applied Linguistics 30 (2), pp. 236–52.

Cheng, W., C. Greaves and M. Warren. 2006. 'From n-gram to skipgram to concgram', International Journal of Corpus Linguistics 11 (4), pp. 411–33.

DiMaggio, P., M. Nag and D. Blei. 2013. 'Exploiting affinities between topic modeling and the sociological perspective on culture: application to newspaper coverage of US government arts funding', Poetics 41 (6), pp. 570–606.

El-Kishky, A., Y. Song, C. Wang, C.R. Voss and J. Han. 2014. 'Scalable topical phrase mining from text corpora' in proceedings of the VLDB Endowment 8 (3), pp. 305–16.

Fletcher, W.H. 2007. kfNgram. Available online at: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

Gries, StTh. 2013. Statistics for Linguistics with R: A Practical Introduction. (Second edition.) Berlin: De Gruyter Mouton.

Griffiths, T.L. and M. Steyvers. 2004. 'Finding scientific topics' in proceedings of the National Academy of Sciences of the United States of America, 101 (supplementary 1), pp. 5228–35.

Grimmer, J. 2010. 'A Bayesian hierarchical topic model for political texts: measuring expressed agendas in Senate press releases', Political Analysis 18 (1), pp. 1–35.

Grün, B. and K. Hornik. 2011. 'Topicmodels: an R package for fitting topic models', Journal of Statistical Software 40 (13). Available online at: http://www.jstatsoft.org/v40/i13.

Hyland, K. 2005. 'Stance and engagement: a model of interaction in academic discourse', Discourse Studies 7 (2), pp. 173–92.

Jockers, M.L. and D. Mimno. 2013. 'Significant themes in 19th-century literature', Poetics 41 (6), pp. 750–69.

Marshall, E.A. 2013. 'Defining population problems: using topic models for cross-national comparison of disciplinary development', Poetics 41 (6), pp. 701–24.

Meeks, E. and S.B. Weingart. 2012. 'The digital humanities contribution to topic modeling', Journal of Digital Humanities 2 (1), pp. 2–6.

Ponweiser, M. 2012. Latent Dirichlet allocation in R. Vienna University of Business and Economics.

Porter, M.F. 1980. 'An algorithm for suffix stripping', Program 14 (3), pp. 130–7.

Priva, U.C. and J.L. Austerweil. 2015. 'Analyzing the history of Cognition using Topic Models', Cognition 135, pp. 4–9.

Rayson, P. 2008. 'From key words to key semantic domains', International Journal of Corpus Linguistics 4 (2008), pp. 519–49.

Rhody, L.M. 2012. 'Topic modeling and figurative language', Journal of Digital Humanities 2 (1), pp. 19–35.

Scott, M. 1996. WordSmith Tools. Oxford: Oxford University Press.

Scott, M. 2010. 'Problems in investigating keyness, or clearing the undergrowth and marking out trails...' in M. Bondi and M. Scott (eds) Keyness in Texts, pp. 43–57. Amsterdam: John Benjamins.

Sinclair, J. 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Team, R.C. 2013. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. Available online at: http://www.r-project.org/.

Warren, M. 2010. 'Identifying aboutgrams in engineering texts' in M. Bondi and M. Scott (eds) Keyness in Texts, pp. 113–26. Amsterdam: John Benjamins.

Williams, G. 1998. 'Collocational networks: interlocking patterns of lexis in a corpus of plant biology research articles', International Journal of Corpus Linguistics 3 (1), pp. 151–71.

Williams, G. 2002. 'In search of representativity in specialised corpora: Categorisation through collocation', International Journal of Corpus Linguistics 7 (1), pp. 43–64.

Wood, S. 2006. Generalized Additive Models: An Introduction with R. Boca Raton, Florida: Chapman and Hall/CRC.

Table 1: A very small corpus.

Text 1	*romeo*
	juliet
	hamlet
Text 2	*hamlet*
	environment
	ozon
Text 3	*environment*
	ozon
	climate

Table 2: Numbers of papers and text-blocks across years in the GEC Corpus.

Year	Papers	Text-blocks
1990/1991	24	361
1992	28	351
1993	21	414
1994	20	294
1995	33	414
1996	21	319
1997	20	334
1998	21	308
1999	30	472
2000	24	366
2001	25	390
2002	25	354
2003	25	357
2004	38	559
2005	36	508
2006	38	561
2007	42	689
2008	73	1,297
2009	53	870
2010	78	1,337
Total	675	10,555

Table 3a: Key semantic fields across time.

Key semantic fields in 1991–2000					
Freq. per 10,000 words		Keyness	Semantic tag	Semantic category	Most frequent words (freq. per 10,000 words)
1991–2000	2001–2010				
33.7	12.1	1,953.7	O1.3	Substances and materials generally: Gas	*CO2* (8.4), *gas* (6.2), *gases* (3.9), *methane* (3.2), *ozone* (2.3)
26.4	11.5	1,112.0	O1	Substances and materials generally	*fuel* (4.2), *biomass* (2.2), *fuels* (2.1), *CFCs* (1.9), *chemical* (1.2)
78.1	54.9	734.8	W5	Green issues	*environmental* (30.4), *environment* (7.0), *conservation* (4.3), *nature* (4.2), *pollution* (3.8)
48.2	31.4	645.7	Y1	Science and technology in general	*scientific* (9.0), *science* (5.5), *scientists* (3.9), *technology* (3.4), *technologies* (2.8)
24.1	12.8	642.6	W1	The universe	*world* (8.9), *atmospheric* (3.7), *World* (3.0), *worlds* (1.5), *layer* (1.0)
27.6	15.9	584.8	O4.6	Temperature	*warming* (7.0), *temperature* (5.6), *temperatures* (1.8), *burning* (1.2), *fire* (1.0)
33.3	21.3	475.8	X5.2	Interest/boredom/excited/ energetic	*energy* (15.1), *interest* (2.8), *interests* (2.4), *incentives* (1.4), *active* (1.1)
118.0	97.9	334.3	W3	Geographical terms	*global* (28.2), *land* (12.3), *forest* (7.8), *soil* (4.4), *forests* (4.2)
102.4	84.9	291.6	Z2	Geographical names	*Europe* (3.5), *USA* (3.5), *UK* (2.5), *China* (2.5), *Africa* (2.3)
17.7	11.0	286.7	I4	Industry	*industrial* (3.6), *industry* (2.9), *industrialized* (2.1), *industries* (1.1), *GNP* (0.8)

Table 3b: Key semantic fields across time.

Key semantic fields in 2001–2010					
Freq. per 10,000 words		Keyness	Semantic tag	Semantic category	Most frequent words (freq. per 10,000 words)
1991–2000	2001–2010				
40.1	85.4	2,862.7	Z1	Personal names	*al*. (28.7), *Turner* (0.8), *van* (0.7), *Smith* (0.7), *de* (0.6)
160.4	237.2	2,595.0	N1	Numbers	*2001* (9.3), *2000* (9.1), *2005* (9.1), *2002* (8.9), *2003* (8.7)
224.5	300.6	1,931.2	Z99	Unmatched	*IPCC* (4.8), *EQSYM* (2.2), *SRES* (2.0), *capita* (1.6), *Adger* (1.5)
14.0	33.5	1,411.7	S1.2.5	Toughness; strong/weak	*vulnerability* (14.0), *resilience* (4.5), *vulnerable* (2.9), *strong* (2.5), *abatement* (0.8)
47.6	76.8	1,202.3	P1	Education in general	*et* (34.1), *study* (7.6), *studies* (6.4), *al* (4.8), *education* (1.0)
21.5	33.1	427.0	A15	Safety/Danger	*risk* (9.1), *risks* (4.2), *protection* (2.3), *hazards* (2.2), *exposure* (2.1)
58.3	74.9	361.5	W4	Weather	*climate* (44.5), *rainfall* (3.4), *Climate* (3.3), *flood* (3.2), *climatic* (2.8)
33.8	45.3	292.8	Q1.2	Paper documents and writing	*et* (4.9), *al* (4.9), *address* (2.4), *application* (1.6), *addressed* (1.3)
7.3	13.0	273.8	S4	Kin	*households* (3.3), *household* (3.1), *family* (0.8), *fertility* (0.8), *families* (0.5)
67.6	81.0	211.6	X2.4	Investigate, examine, test, search	*data* (11.1), *research* (10.5), *analysis* (9.2), *assessment* (6.7), *assessments* (3.9)

Table 4a: Keywords across time.

Keywords in 1991–2000			
Freq. per 10,000 words		Keyness	Keyword
1991–2000	2001–2010		
7.2	1.2	894.0	*figur*
35.3	19.6	813.7	*environment*
8.1	1.8	792.5	*atmospher*
18.5	8.2	726.4	*energi*
10.0	3.0	712.4	*greenhous*
32.7	19.0	655.0	*global*
28.9	16.1	649.1	*emiss*
3.6	0.3	620.9	*methan*
2.9	0.2	576.3	*cfc*
2.1	0.0	543.2	*ec*
86.9	66.3	479.4	*be*
6.9	2.2	465.3	*fuel*
3.7	0.7	424.3	*usa*
6.8	2.5	405.3	*sea*
11.4	5.3	397.0	*industri*
3.7	0.7	396.8	*ozon*
6.6	2.4	380.7	*ga*
9.5	4.2	373.0	*co2*
2.9	0.4	372.9	*coal*
2.1	0.2	356.0	*aral*

Table 4b: Keywords across time.

Keywords in 2001–2010			
Freq. per 10,000 words		Keyness	Keyword
1991–2000	2001–2010		
9.4	37.6	2895.7	*al*
9.5	37.8	2874.7	*et*
< 0.1	10.1	2028.9	*2003*
< 0.1	9.7	1998.2	*2006*
< 0.1	9.9	1988.1	*2002*
0.1	10.0	1987.1	*2001*
< 0.1	8.8	1803.2	*2004*
0.4	10.6	1799.6	*2005*
< 0.1	8.3	1705.2	*2007*
4.9	19.8	1514.7	*vulner*
9.1	25.2	1273.7	*adapt*
1.8	10.5	1046.6	*2000*
0.2	5.7	1000.5	*2008*
3.0	11.4	815.3	*capac*
2.7	9.3	618.4	*fig*
< 0.1	2.9	602.0	*2009*
0.6	5.0	597.6	*resili*
2.0	7.7	578.6	*household*
1.7	6.9	543.5	*1999*
12.9	23.4	510.7	*water*

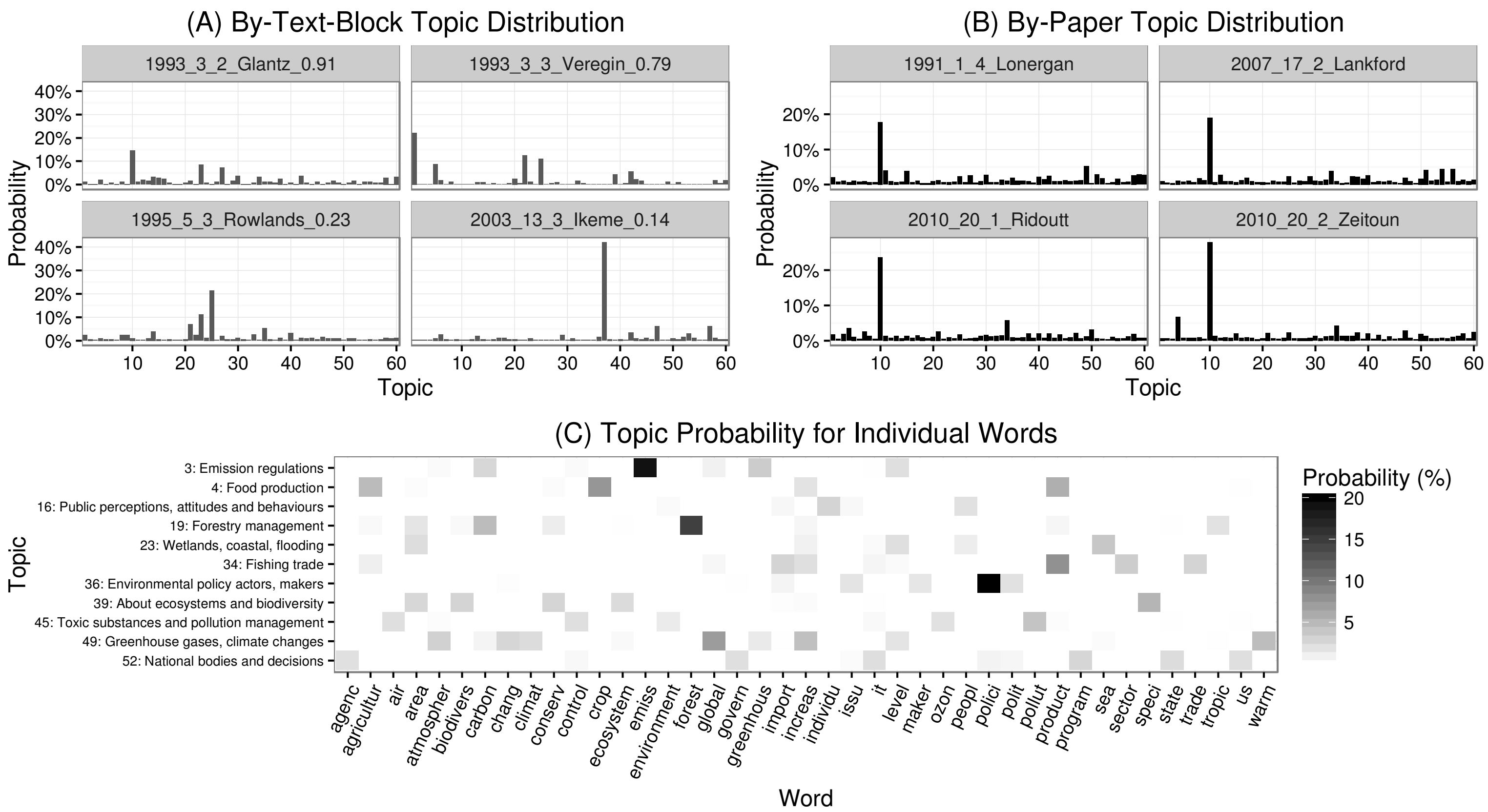

Figure 1: Topic distribution in text-blocks, papers and words.

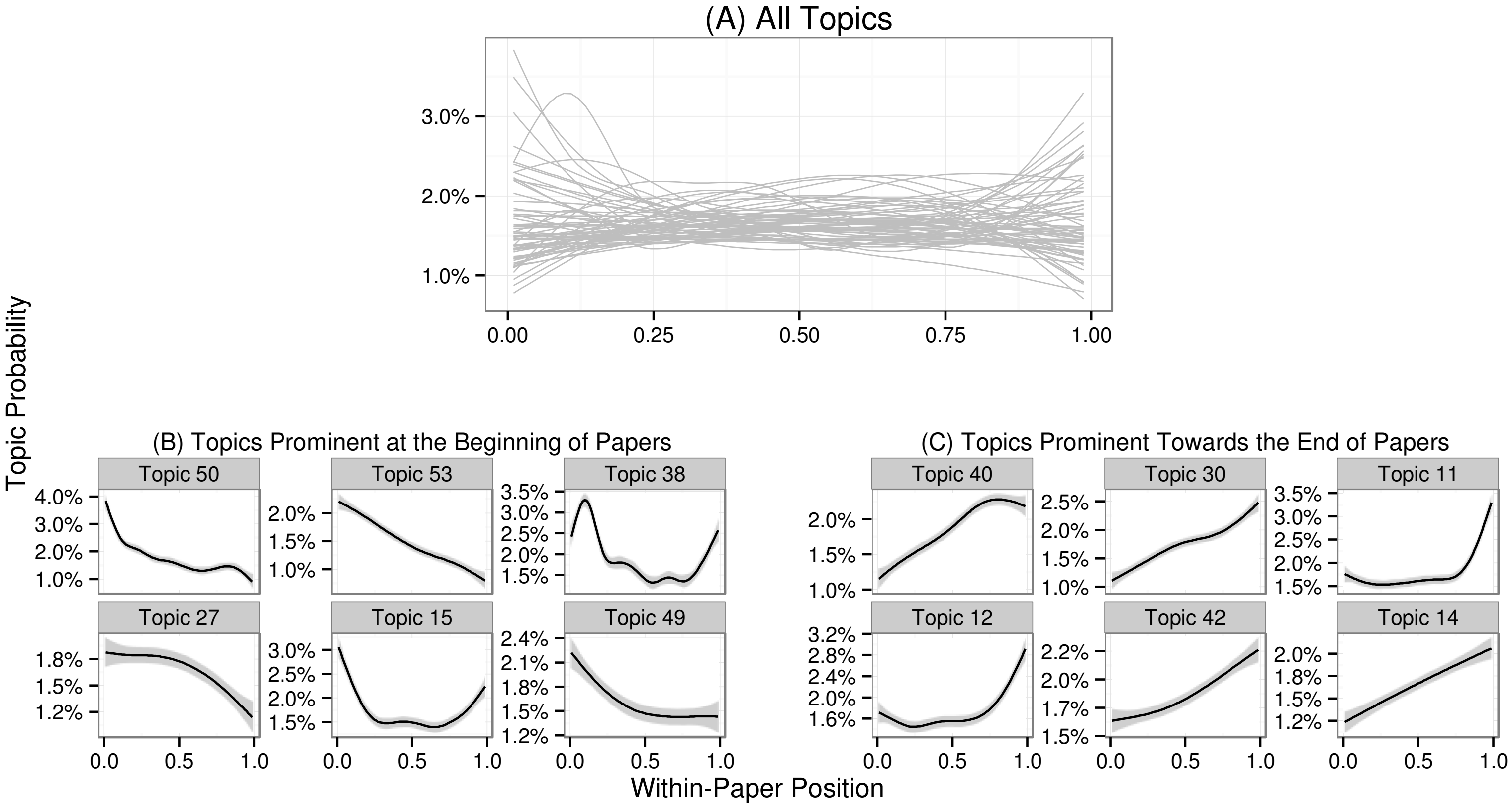

Figure 2: Within-paper distribution of topic probability.

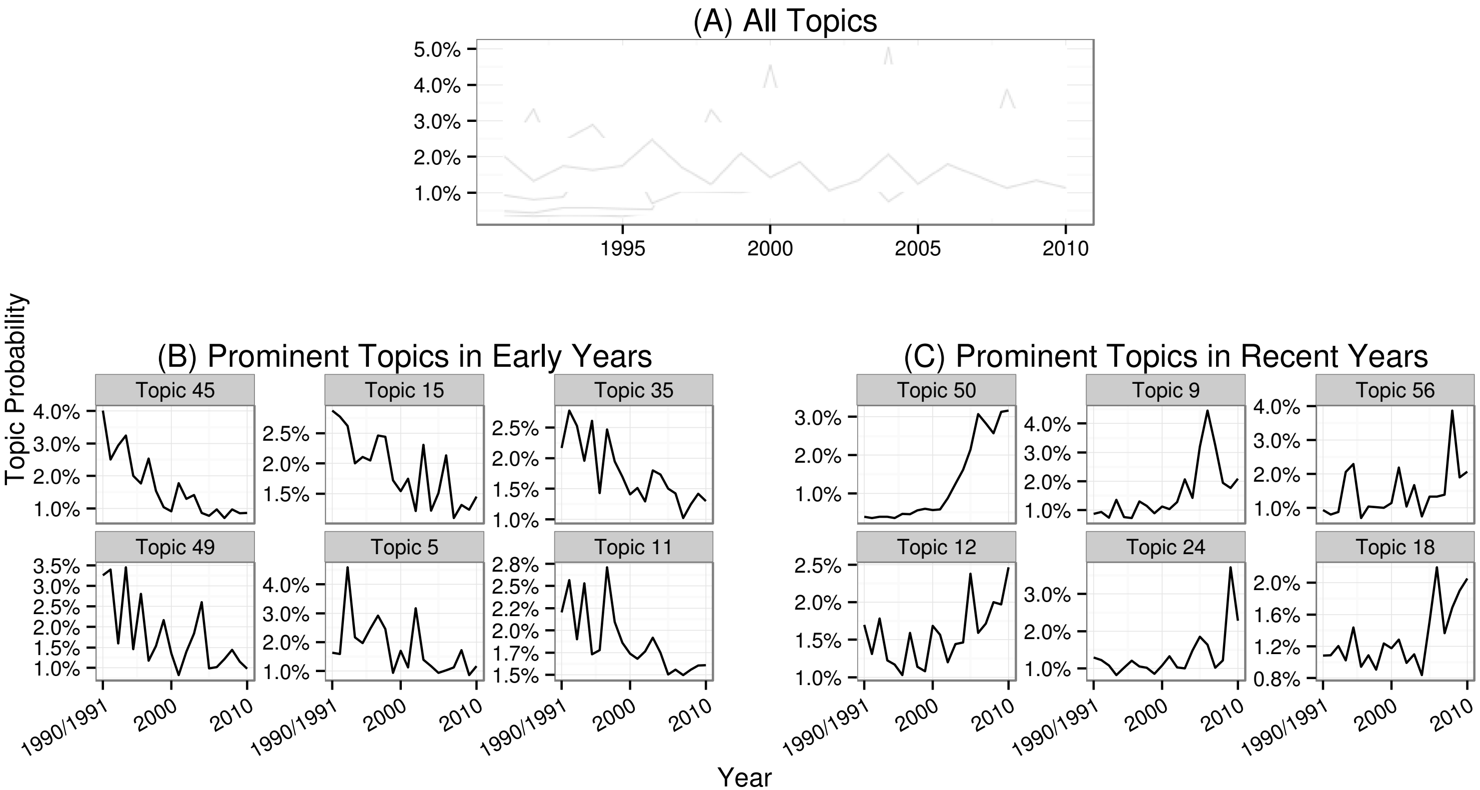

Figure 3: Chronological transition of topic probability.

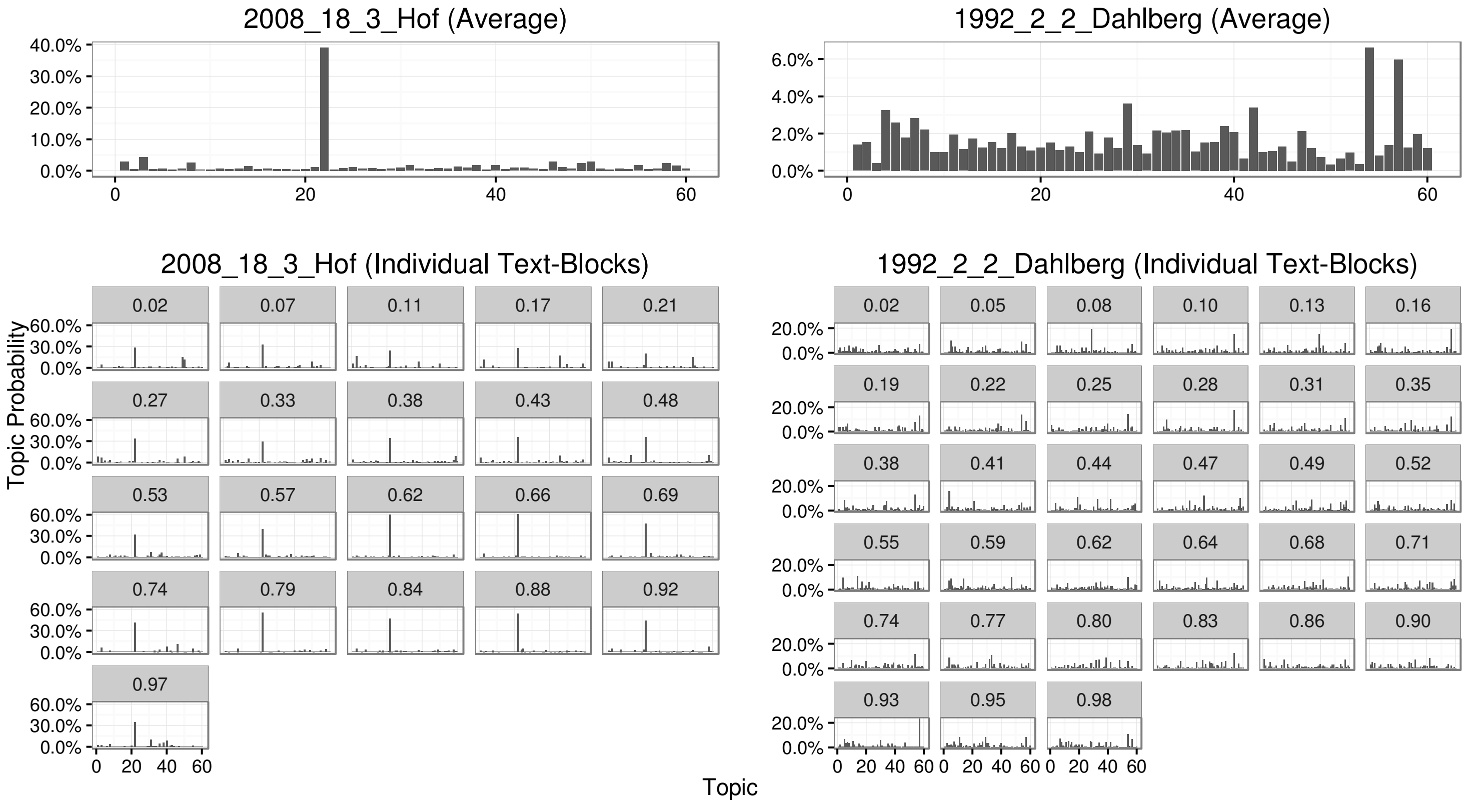

Figure 4: Topic profile of two distinct types of papers.

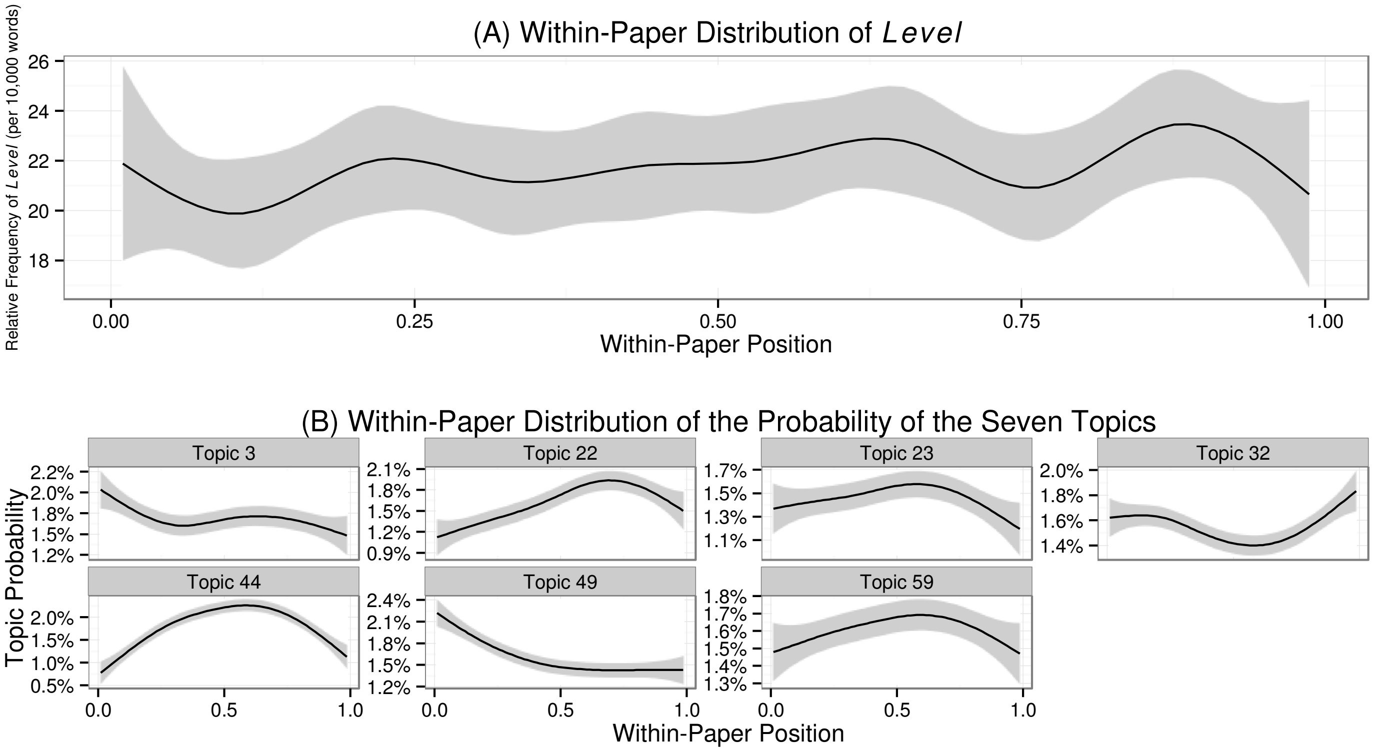

Figure 5: Within-paper distribution of the word *level* and the distribution of the probability in the seven topics where *level* is a keyword.

Appendix A (*continued on following page*): Topic labels and keywords.

Topic	Label	Keywords
1	Modelling	*model, use, simul, base, paramet, each, which, result, repres, function*
2	Global research community (planning, agenda, organisations)	*research, scienc, scientif, scientist, global, ipcc, work, assess, knowledg, intern*
3	Emissions regulation	*emiss, reduct, greenhous, co2, carbon, gas, level, reduc, target, mitig*
4	Food production	*crop, product, agricultur, soil, food, yield, increas, fertil, use, plant*
5	Energy use, efficiency	*energi, use, fuel, effici, technolog, power, sector, transport, consumpt, industri*
6	Network actor analysis	*govern, institut, actor, state, network, power, polit, author, their, role*
7	Commercial partnerships, competition	*technolog, industri, market, new, compani, busi, develop, regul, such, their*
8	'We' as researchers and our intention, evaluation and procedures	*we, our, this, these, can, which, not, import, both, first*
9	Vulnerability, adaptive capacity	*adapt, vulner, capac, or, sensit, social, cope, exposur, measur, abil*
10	Water systems, supplies, and trade	*water, river, basin, suppli, flow, irrig, resourc, avail, use, stress*
11	How we look at the future	*will, futur, may, this, can, if, more, like, current, need*
12	Learning and management	*manag, plan, strategi, institut, learn, implement, practic, improv, new, challeng*
13	Interviews, personal, quotes	*they, what, not, one, but, so, when, go, peopl, you*
14	Costs and market regulations	*cost, benefit, invest, econom, incent, price, market, reduc, measur, tax*
15	Global environmental security and other problems	*environment, global, problem, environ, econom, concern, issu, chang, secur, polit*
16	Public perceptions, attitudes and behaviours	*individu, their, public, respons, action, peopl, they, behaviour, perceiv, percept*
17	Property rights, access, genetic resources	*resourc, servic, natur, properti, or, access, ecosystem, these, manag, right*
18	Local knowledge, traditions, culture	*communiti, peopl, local, their, tradit, mani, live, indigen, which, knowledg*
19	Forestry management	*forest, carbon, deforest, tropic, land, area, cover, conserv, forestri, timber*
20	Health and disaster risks	*risk, health, disast, effect, hazard, diseas, peopl, affect, reduc, potenti*
21	Mapping, satellite imagery, GIS etc methods	*data, use, avail, estim, includ, base, inform, studi, this, sourc*
22	Explaining cost-benefit analyses in figures, esp damage	*cost, damag, valu, estim, rate, case, loss, time, level, increas*
23	Wetlands, coastal, flooding	*flood, sea, rise, coastal, area, level, protect, impact, loss, sealevel*
24	Media and public discourse, and reviews of scientific literature	*discours, point, articl, this, media, public, report, issu, frame, us*
25	References to figures and tables	*each, two, differ, three, these, all, fig, tabl, which, type*
26	Reports on interviews, focus groups, surveys	*group, respond, particip, interview, survey, their, question, they, respons, inform*
27	Historical contextualisation	*year, period, earli, sinc, time, recent, centuri, decad, this, past*
28	Assessment processes, participatory	*process, assess, inform, stakehold, particip, decisionmak, integr, involv, knowledg, issu*
29	Sustainable development	*develop, sustain, need, goal, econom, integr, object, this, achiev, it*
30	Hypothetical discussion	*would, could, not, if, might, or, this, but, ani, should*

Appendix A (*continued*): Topic labels and keywords.

Topic	Label	Keywords
31	Discussing models and scenarios	*uncertainti*, *decis*, *choic*, *can*, *approach*, *probabl*, *such*, *make*, *differ*, *altern*
32	Spatial scope of human activities and decisions at different levels	*local*, *scale*, *level*, *region*, *differ*, *spatial*, *nation*, *these*, *across*, *which*
33	Land use description	*land*, *area*, *agricultur*, *use*, *cultiv*, *cattl*, *popul*, *livestock*, *which*, *pastur*
34	Fishing trade	*product*, *sector*, *trade*, *import*, *increas*, *export*, *consumpt*, *fish*, *market*, *econom*
35	Developing and developed countries	*countri*, *develop*, *nation*, *world*, *intern*, *their*, *india*, *global*, *industri*, *most*
36	Environmental policy actors, makers	*polici*, *polit*, *this*, *issu*, *maker*, *question*, *decis*, *make*, *what*, *which*
37	Justice and ethics	*should*, *right*, *principl*, *this*, *distribut*, *not*, *equiti*, *which*, *justic*, *or*
38	Metatext, meta-analyses and case-studies	*studi*, *this*, *analysi*, *paper*, *approach*, *section*, *discuss*, *case*, *how*, *present*
39	About ecosystems and biodiversity	*speci*, *biodivers*, *conserv*, *area*, *ecosystem*, *plant*, *divers*, *protect*, *veget*, *site*
40	Discussing findings	*more*, *than*, *less*, *not*, *greater*, *also*, *much*, *other*, *howev*, *rather*
41	Project development and approval	*project*, *fund*, *activ*, *develop*, *organ*, *monitor*, *oper*, *this*, *implement*, *technic*
42	Discussion and evaluation personal	*not*, *but*, *there*, *onli*, *this*, *veri*, *even*, *they*, *mani*, *no*
43	Climate events and impacts on tourism	*event*, *may*, *extrem*, *or*, *island*, *these*, *exampl*, *tourism*, *infrastructur*, *expect*
44	Variables and correlations	*indic*, *variabl*, *measur*, *eqsym*, *valu*, *signific*, *index*, *effect*, *correl*, *relationship*
45	Toxic substances and pollution management	*pollut*, *control*, *air*, *ozon*, *environment*, *wast*, *effect*, *deplet*, *which*, *problem*
46	Comparing scenarios ref to figures	*scenario*, *futur*, *project*, *differ*, *use*, *result*, *region*, *rang*, *this*, *assum*
47	Discussing different cases and outcomes	*or*, *may*, *can*, *such*, *other*, *some*, *case*, *eg*, *exampl*, *not*
48	International protocols, agreements mainly historical	*intern*, *negoti*, *agreement*, *convent*, *nation*, *protocol*, *state*, *eu*, *issu*, *parti*
49	Greenhouse gases, climate changes	*global*, *warm*, *increas*, *atmospher*, *chang*, *climat*, *temperatur*, *effect*, *concentr*, *level*
50	2000 references	*et*, *al*, *2005*, *2003*, *2006*, *2002*, *2004*, *2007*, *2001*, *2008*
51	Precipitation models seasonal	*temperatur*, *precipit*, *season*, *increas*, *rainfal*, *period*, *annual*, *dri*, *year*, *averag*
52	National bodies and decisions	*program*, *state*, *it*, *us*, *govern*, *agenc*, *nation*, *committe*, *offici*, *support*
53	1990 references	*al*, *et*, *1996*, *1995*, *1998*, *1997*, *1999*, *1994*, *2000*, *1992*
54	Ecological systems and resilience	*system*, *chang*, *resili*, *complex*, *dynam*, *ecolog*, *human*, *or*, *interact*, *natur*
55	Mitigation, adaptation	*chang*, *climat*, *impact*, *effect*, *respons*, *mitig*, *futur*, *assess*, *potenti*, *adapt*
56	Households, village level	*farmer*, *household*, *their*, *incom*, *farm*, *villag*, *migrat*, *livelihood*, *food*, *rural*
57	Social and cultural theories	*social*, *cultur*, *which*, *natur*, *human*, *societi*, *polit*, *theori*, *perspect*, *view*
58	Large scale stats trends, rates	*year*, *total*, *per*, *million*, *estim*, *averag*, *10*, *than*, *annual*, *tabl*
59	Population and other growth trends	*popul*, *growth*, *increas*, *urban*, *econom*, *per*, *rate*, *citi*, *incom*, *capita*
60	Broad regional focus	*region*, *africa*, *south*, *southern*, *europ*, *area*, *central*, *north*, *most*, *asia*