DATA FOR ESRC PROJECT (ES/K007300/1), 2013–2015

"Investigating interdisciplinary research discourse: the case of Global Environmental Change".

PI: Dr Paul Thompson

Co-I: Professor Susan Hunston

Research Fellows: Dr Akira Murakami; Dr Dominik Vajn

Institution: University of Birmingham

This project investigated the discourse of interdisciplinary research (IDR) through comprehensive linguistic analyses of the full holdings of a successful IDR journal, *Global Environmental Change* (GEC) in the period 1990-2010, and of ten other comparison journals published by Elsevier. The ten were chosen to represent other interdisciplinary (ID) journals and monodisciplinary (MD) journals. The corpus data cannot be included in the repository as it belongs to Elsevier – individual files can all be consulted through the Elsevier website.

The main lines of analysis were multidimensional analysis (MDA) for which Doug Biber (Northern Arizona University) acted as a consultant. From the MDA, we derived six constellations in which papers with similar MDA profiles clustered. We then examined the N-grams and P-frames in each constellation – the raw numerical data are available in this repository.

A second computational approach taken was to use topic modelling to establish, in an inductive manner, what the papers in the GEC corpus are 'about'. The TopicModel folder contains data for this investigation, some of which are discussed in our paper that appears in the *Corpora* journal (publication mid 2016).

We also conducted survey and interview data analysis and the data are presented here.

For details on the lines of investigations, please read the papers that we have published or are in the process of writing. Details can be found in the ResearchFish database for the project.

Repository Contents

The **MDA_Constellations folder** includes two files.

1. **MDA.xlsx** with the following five sheets:
    - FeatureSet includes the list of 148 linguistic features that were initially targeted in our MD analysis, their mean value, standard deviation, and the proportion of the files in which the value is zero.
    - FactorLoadings includes the factor loadings of the 53 features we included in our final MD model.
    - FactorialStructure includes the factorial structure of each factor.
    - DimensionScore includes the dimension score of each dimension in each paper
    - RandomForestContingencyTable includes the cross-tabulated table between the observed and predicted journals of each paper based on random forests. The features were the dimension scores of each paper.
2. **NGramsAndP-FramesByConstellation.xlsx** includes the most frequent n-grams and phrase-frames for each constellation. "Range" in the n-gram sheets indicates the number of constellations in which the frequency of the n-gram was three or above. "Variants" in the phrase-frame sheets indicate the number of variants of the phrase frame.

The **TopicModel folder** includes three files.

1. **TopicModel.xlsx** with the following four sheets:
    - ProbabilityOfEachText shows the probability of each topic in each text.
    - ProbabilityOfEachWord shows the probability of each topic for each word.
    - Keywords&KeyPapers includes the top 20 keywords and the top 20 key papers of each topic.
    - TopicLabels lists the labels of each topic.
2. **KeySemanticFields.xlsx** includes the key semantic tags (based on USAS) in each decade of GEC computed with the other decade as the reference corpus. It further includes the most frequent words in each key semantic tag.
3. **KeywordsAnalysisOfEachDecadeInGEC.xlsx** includes the keywords of each decade of GEC computed with the other decade as the reference corpus.

The **Interview folder** includes the transcripts of each interview (the Transcripts folder) and the summary of interview questions (Interview topic and questions.docx).

SurveyResults.xlsx includes four sheets:

- RawData includes the raw data of individual respondents after excluding the information that may help to identify them.
- QuestionDetail includes the details of each question.
- Summary1 shows the basic descriptive data of the responses.
- Summary2 also show the basic descriptive data of the responses, but is tabulated so that the response to the same question can be compared between monodisciplinary journals and interdisciplinary journals.

GECBasicData.xls includes two sheets:

- CorpusSize shows the data size of each volume and each issue in the GEC corpus.
- EachArticle shows various information for each paper in GEC, including the list of authors, the number of authors, the title, keywords, the number of words in the abstract, that in the body, among other things.

Labelling.xlsx includes three sheets:

- RawData shows the paper label assigned by each of the two researchers for each paper, and the agreed final label.
- ContingencyTable cross-tabulates the initially assigned label by the two researchers and demonstrates the agreement in each label.
- ChronologicalChange shows the chronological change of the frequency of each label.

Contact:

Dr Paul Thompson, Centre for Corpus Research, University of Birmingham, B15 2TT

p.thompson@bham.ac.uk