Automated prototypical text detection for corpus and critical discourse studies using *KeyAnt*

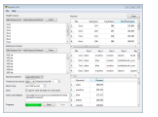
Laurence Anthony
Faculty of Science and Engineering, Waseda University, Japan
anthony@waseda.jp

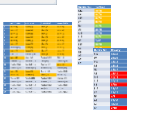
Paul Baker
Dept. of Linguistics and English Language, Lancaster University, UK.
j.p.baker@lancaster.ac.uk

UCREL Seminar Series, Lancaster University, January 15, 2015

Overview

- **Background**
  - The importance of prototypicality
  - Identification of prototypical texts in CDA and NLP studies
- ***KeyAnt* Approach**
  - Overview
  - Method
- **Validation Experiments**
  - Prototypical short/long texts
  - Prototypical texts in a small corpus
  - Outlier texts





Background

Background
The importance of prototypicality

- Prototypicality – a definition
  - "having the typical qualities of a particular group or kind of person or thing" (Merriam-Webster, 2014)
- Prototypicality types
  - **lexical**, grammatical, structural, semantic, contextual, functional, thematic, …
  - **lexical – single words** vs multi-word units

Background
The importance of prototypicality

- Prototypical text identification has many applications
  - classifying texts according to genre
  - down-sampling a large corpus before conducting a qualitative analysis of a few typical files
  - finding typical student essays at a particular level (e.g. CEFR C1)
  - flagging texts for further analysis (e.g. extremist writing)
  - identifying atypical/outlier or "resistant" discourses on a topic

Background
Identification of prototypical texts in CDA and NLP studies

- CDA (and other qualitative) studies
  - opportunistic selection
    - e.g. Caldas-Coulthard et al. (2003)
      - "…we purchased all the 15 bear books available in a local children's book store in London."
  - limitations
    - non-principled
    - possible bias of researcher ('cherry picking')
    - difficult to replicate the results

Background
Identification of prototypical texts in CDA and NLP studies

- CDA (and other qualitative) studies
  - selective downsizing
    - e.g. Khosravinik (2010)
      - in a corpus of 170,000 articles, select articles from five one-week periods where the number of articles about immigration peak (resulting in 439 articles)
  - limitations
    - can still result in a large number of sample texts
    - 'cherry picking' criticism is not completely addressed

7

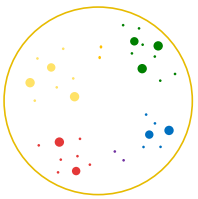
Background
Identification of prototypical texts in CDA and NLP studies

- NLP (and other statistical) studies
  - supervised learning approaches
    - 1) pick prototypical texts representing target classes
    - 2) analyze the selected texts in terms of words, sentence length, …
    - 3) create a statistical model of similarity (e.g. Nearest-Neighbor Classifier)
    - 4) classify target texts into particular classes using the model
  - unsupervised learning (clustering) approaches
    - 1) pre-select features of interest (words, sentence length, …)
    - 2a) group neighbor texts based on features of interest
    - 2b) split a set of texts into parts based on features of interest
    - 3) continue until all texts are assigned a category

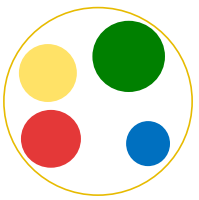
8

Background
Identification of prototypical texts in CDA and NLP studies

- NLP (and other statistical) studies



Supervised learning



Unsupervised learning

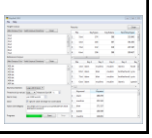
9

Background
Identification of prototypical texts in CDA and NLP studies

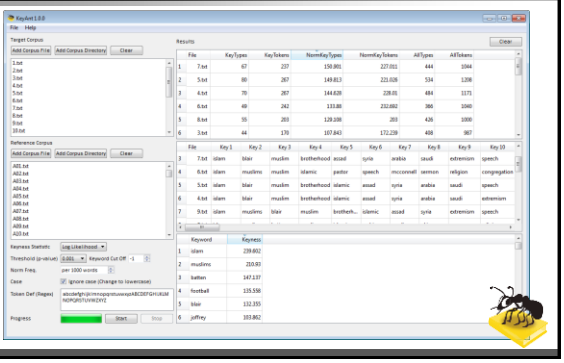
- Limitations of NLP approaches
  - supervised approaches require pre-selected 'typical' texts
  - unsupervised approaches are highly dependent on similarity measures and grouping criteria
  - both approaches may employ 'black box' mathematical methods for classification/clustering

10

***KeyAnt* Approach**



KeyAnt Approach
Version 1.0.0 (beta)

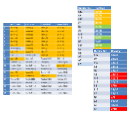


KeyAnt Approach
Version 1.0.0 (beta)

- **Step 1**: Calculate the keywords (unusually frequent types or tokens) in the target corpus based on a suitable reference corpus
  - e.g. using log-likelihood + (log) relative frequency
- **Step 2**: Rank the target corpus texts by the number of keywords they contain (normalized by length of text)
  - the highest ranked texts contain the most characteristic words in the corpus, thus defined "the most typical" texts
- **Step 3**: Display each corpus text according to its rank and show the keywords it contains
  - clicking on any file name displays the full text

13

Validation Experiments



Experiment 1
Prototypical short texts

- Target Corpus
  - 20 newspaper texts (about 1100 tokens each)
    - 10 about 'Islam'
    - 5 about 'football'
    - 5 random
- Reference Corpus
  - BE06 (Baker, 2009)
    - 500 texts representative of British English in 2006

15

	0.05 (1069)	0.01 (610)	0.001 (234)	0.0001 (150)
1	Islam (5)	Islam (5)	Islam (7)	Islam (5)
2	Islam (2)	Islam (7)	Islam (5)	Islam (4)
3	Islam (7)	Islam (6)	Islam (4)	Islam (7)
4	Islam (4)	Islam (4)	Islam (6)	Islam (8)
5	Islam (6)	Islam (2)	Islam (8)	Islam (6)
6	Islam (3)	Islam(3)	Islam (3)	Islam (2)
7	Review (19)	Islam (1)	Islam (1)	Islam (1)
8	Islam (1)	Islam (8)	Islam (2)	Islam (3)
9	Obituary (16)	Review (19)	Islam (9)	Islam (9)
10	Islam (8)	Obituary (16)	Football (11)	Obituary (16)
11	Science (17)	Football (14)	Obituary (16)	Football (11)
12	Football (11)	Football (11)	Islam (10)	Islam (10)
13	Football (14)	Science (17)	Review (19)	Review (19)
14	Islam (9)	Islam (9)	Football (14)	Football (14)
15	Islam (10)	Islam (10)	Science (17)	Tennis (18)
16	Tennis (18)	Tennis (18)	Tennis (18)	Science (17)
17	Football (12)	Football (13)	Football (12)	Football (13)
18	Football (13)	Football (12)	Football (13)	Football (12)
19	Art (20)	Art (20)	Football (15)	Art (20)
20	Football (15)	Football (15)	Art (20)	Football (15)

Considering key types

	0.05 (1069)	0.01 (610)	0.001 (234)	0.0001 (150)
1	Islam (6)	Islam (6)	Islam (6)	Islam (6)
2	Islam (5)	Islam (5)	Islam (4)	Islam (5)
3	Islam (4)	Islam (7)	Islam (7)	Islam (4)
4	Islam (7)	Islam (4)	Islam (5)	Islam (7)
5	Islam (3)	Islam (3)	Islam (8)	Islam (8)
6	Islam (1)	Islam (8)	Islam (3)	Football (11)
7	Islam (2)	Islam (2)	Football (11)	Islam (3)
8	Obit (16)	Football (11)	Obituary (16)	Islam (9)
9	Islam (8)	Obituary (16)	Islam (9)	Islam (2)
10	Review (19)	Science (17)	Islam (2)	Football (14)
11	Football (11)	Islam (1)	Islam (1)	Islam (1)
12	Science (17)	Football (14)	Football (14)	Review (19)
13	Football (14)	Islam (9)	Science (17)	Obituary (16)
14	Islam (9)	Review (19)	Review (19)	Science (17)
15	Tennis (18)	Tennis (18)	Islam (10)	Football (12)
16	Islam (10)	Islam (10)	Tennis (18)	Tennis (18)
17	Football (12)	Football (12)	Football (12)	Islam (10)
18	Football (13)	Art (20)	Art (20)	Art (20)
19	Art (20)	Football (13)	Football (15)	Football (13)
20	Football (15)	Football (15)	Football (13)	Football (15)

Considering key tokens

Experiment 1 - Summary
Prototypical short texts

- All methods identified Islam texts as the top 5 most typical (also about the same story involving Tony Blair)
- Islam text 10 wasn't ranked as very typical (a different news item)
- The football texts are on relatively diverse topics
- Review text 19 came relatively high – 2 of the top 20 keywords were a result of this text:
  - Islam, Muslims, Batten, Blair, football, Joffrey, Muslim, Mara, Brotherhood, Islamic, Assad, Kundnani, Syria, Arabia, Saudi, manager, Sansa, UKIP, pastor

18

Experiment 2
Prototypical long texts

- Target Corpus
  - 20 fictional excerpts (about 2000 tokens each)
    - 10 from 'Dracula' novels
    - 5 from 'Frankenstein' novels
    - 5 random novels
- Reference Corpus
  - BE06 (Baker, 2009)

19

	0,.05 (1794)	0.01 (1175)	0.001 (442)	0.0001 (274)
1	Dracula (10)	Dracula (9)	Dracula (8)	Dracula (8)
2	Dracula (3)	Dracula (7)	Dracula (6)	Dracula (7)
3	Dracula (7)	Dracula (8)	Dracula (7)	Dracula (6)
4	Dracula (8)	Dracula (10)	Dracula (9)	Dracula (9)
5	Dracula (9)	Dracula (6)	Dracula (10)	Dracula (10)
6	Frankenstein (12)	Dracula (3)	Dracula (2)	Dracula (5)
7	Dracula (6)	Dracula (2)	Dracula (5)	Dracula (2)
8	Frankenstein (13)	Dracula (5)	Frankenstein (15)	Dracula (4)
9	Dracula (2)	Frankenstein (13)	Dracula (4)	Dracula (3)
10	Frankenstein (11)	Frankenstein (12)	Dracula (3)	Frankenstein (15)
11	Dracula (5)	Dracula (4)	Frankenstein (14)	Frankenstein (13)
12	Dracula(4)	Frankenstein (14)	Frankenstein (13)	Dracula (1)
13	Frankenstein (14)	Frankenstein (15)	Dracula (1)	Frankenstein (14)
14	Dracula (1)	Dracula (1)	Frankenstein (12)	The Moonstone (17)
15	Frankenstein (15)	Frankenstein (11)	The Moonstone (17)	Frankenstein (12)
16	Jane Eyre (16)	Jane Eyre (16)	Frankenstein (11)	Frankenstein (11)
17	The Moonstone (17)	The Moonstone (17)	Jane Eyre (16)	Jane Eyre (16)
18	Harry Potter (20)	Harry Potter (20)	Harry Potter (20)	Harry Potter (20)
19	Belle De Jour (19)	It (18)	Belle De Jour (19)	Belle De Jour (19)
20	It (18)	Belle De Jour (19)	It (18)	It (18)

Considering key types

Experiment 2 - Summary
Prototypical long texts

- Dracula excepts were ranked highest, with Frankenstein excepts appearing next
- Dracula except 1 was the least typical Dracula text
  - this is from the earliest version of the novel involving Harker's stay at the castle (the other excerpts are set in London)
- The 3 most recent files were placed as least typical
  - Belle de Jour, Harry Potter and the Deathly Hallows, Stephen King's It

21

Experiment 3
Prototypical "American" texts in a standard corpus

- Target Corpus
  - AmE06 (Potts and Baker, 2012)
    - 500 texts representative of American English in 2006
- Reference Corpus
  - BE06 (Baker, 2009)
- Research question
  - Which files are the "most American" when compared against the British reference texts

22

Experiment 3
Prototypical "American" texts in a standard corpus

- The top 3 files
  - All "H" files: Miscellaneous: Government documents, industrial reports etc.
    - H24: references to the Department of Treasury, lots of references to American states and cities
    - H17: descriptions of appoints to positions in US Court offices
    - H13: a Congressional Record

23

Experiment 3
Prototypical "American" texts in a standard corpus

- The bottom 3 files
  - All files from different registers
    - N06: an adventure novel set in Vietnam in 1975 which describes the main character jumping out of an aeroplane
    - P27: a historical romance novel set in France 1885
    - P19: a description of an African safari

24

Experiment 4
Prototypical "2006-like" texts in a standard corpus

- Target Corpus
  - AmE06 (Potts and Baker, 2012)
- Reference Corpus
  - Brown Corpus (Francis & Kucera, 1961)
- Research question
  - Which files are the "most 2006-like" when compared against the older American reference texts

25

Experiment 4
Prototypical "2006-like" texts in a standard corpus

- The top 3 files
  - H24: a file from the Department of Treasury (same as before) makes direct reference to "reader", "you", "your"
  - H21: a file with references to "Hurricane Katrina" (mentioned in 14 other files) and other words relating to political mood: "terrorism", "preparedness", "Palestinian"
  - G40: a first person autobiography from a woman who grew up in Mississippi describing racial segregation and gender issues

26

Experiment 4
Prototypical "2006-like" texts in a standard corpus

- The bottom 3 files
  - N02: an adventure story set in a jungle (same as before)
  - G21: a text about the American civil war
  - G71: a text about a 20th century artist called Nozkowski with no references to the time period

27

Experiment 5
Identification of 'outlier' texts

- Target Corpus
  - AmE06 (Potts and Baker, 2012)
- Reference Corpus
  - BE06 (Baker, 2009)
- Research design
  - Add one random text to each set of texts per corpus register
  - Perform the *KeyAnt* analysis for each register (15 in total)
  - Record the ***lowest*** ranked text as the outlier
    - e.g. With all 39 files from register A (press reportage) + one other file selected at random (in this case K12), does KeyAnt rank K12 at the bottom?

28

Experiment 5
Identification of 'outlier' texts

Cat	Register	Outlier File	Ranking
A	Press: Reportage	K12	40/40
B	Press: Editorial	L9	28/28
C	Press: Reviews	P13	18/18
D	Religion	C8	18/18
E	Skills, Trades and Hobbies	N7	34/37
F	Popular Lore	A3	28/49
G	Belles Lettres, Biographies, Essays	M6	40/76
H	Miscellaneous: Government documents, industrial reports etc	L13	30/31
J	Academic prose in various disciplines	R8	8/80
K	General Fiction	E15	28/30
L	Mystery and Detective Fiction	C6	25/25
M	Science Fiction	N8	4/7
N	Adventure and Western	A7	29/30
P	Romance and Love story	A5	30/30
R	Humour	L2	2/10

29

Experiment 5
Identification of 'outlier' texts

Cat	Register	Outlier File	Ranking
A	Press: Reportage	K12	40/40
B	Press: Editorial	L9	28/28
C	Press: Reviews	P13	18/18
D	Religion	C8	18/18
E	Skills, Trades and Hobbies	N7	34/37
F	Popular Lore	A3	28/49
G	Belles Lettres, Biographies, Essays	M6	40/76
H	Miscellaneous: Government documents, industrial reports etc	L13	30/31
J	Academic prose in various disciplines	R8	8/80
K	General Fiction	E15	28/30
L	Mystery and Detective Fiction	C6	25/25
M	Science Fiction	N8	4/7
N	Adventure and Western	A7	29/30
P	Romance and Love story	A5	30/30
R	Humour	L2	2/10

6 out of 15 cases ranked perfectly

30

Experiment 5
Identification of 'outlier' texts

Cat	Register	Outlier File	Ranking
A	Press: Reportage	K12	40/40
B	Press: Editorial	L9	28/28
C	Press: Reviews	P13	18/18
D	Religion	C8	18/18
E	Skills, Trades and Hobbies	N7	34/37
F	Popular Lore	A3	28/49
G	Belles Lettres, Biographies, Essays	M6	40/76
H	Miscellaneous: Government documents, industrial reports etc	L13	30/31
J	Academic prose in various disciplines	R8	8/80
K	General Fiction	E15	28/30
L	Mystery and Detective Fiction	C6	25/25
M	Science Fiction	N8	4/7
N	Adventure and Western	A7	29/30
P	Romance and Love story	A5	30/30
R	Humour	L2	2/10

(4 cases very close: 10 out of 15 cases ranked almost perfectly)

31

Experiment 5
Identification of 'outlier' texts

Cat	Register	Outlier File	Ranking
A	Press: Reportage	K12	40/40
B	Press: Editorial	L9	28/28
C	Press: Reviews	P13	18/18
D	Religion	C8	18/18
E	Skills, Trades and Hobbies	N7	34/37
F	Popular Lore	A3	28/49
G	Belles Lettres, Biographies, Essays	M6	40/76
H	Miscellaneous: Government documents, industrial reports etc	L13	30/31
J	Academic prose in various disciplines	R8	8/80
K	General Fiction	E15	28/30
L	Mystery and Detective Fiction	C6	25/25
M	Science Fiction	N8	4/7
N	Adventure and Western	A7	29/30
P	Romance and Love story	A5	30/30
R	Humour	L2	2/10

5 cases poorly ranked

32

Experiment 5
Identification of 'outlier' texts

- Why were some texts poorly ranked?
  - 2 of the cases involve only a small number of files (M=7, R=10). Not enough information?
  - Some registers are a bit vague and undefined in the Brown family (especially F Popular Lore and G Belles Lettres), so maybe typicality is more difficult to identify
  - J is academic writing, although the outlier text R is a "weird" text by a nine year old genius who writes about entomology, microphones, jujitsu, childbirth, music, the magazine National Geographic, skyscrapers, and limousines

33

Conclusions and future work

- Counting keywords seems to be a very good way of identifying typicality in a corpus
  - But…the choice of reference corpus matters
- *KeyAnt* is a freeware, multi-platform tool that can identify prototypical texts quickly and easily
- Future work
  - Implement a technique that requires no reference corpus
    - e.g. treat each file as a 'corpus' using the remainder of the target corpus as a reference corpus
  - Consider **all** lexis (not just keywords)
  - Test the *KeyAnt* approach on files of different sizes?
    - Are longer files seen as typical?

34

Automated prototypical text detection for corpus and critical discourse studies using *KeyAnt*

Laurence Anthony
Faculty of Science and Engineering, Waseda University, Japan
anthony@waseda.jp

Paul Baker
Dept. of Linguistics and English Language, Lancaster University, UK.
j.p.baker@lancaster.ac.uk

Lancaster University

UCREL Seminar Series, Lancaster University, January 15, 2015