Day 5: Automated dictionary-based approaches

Kenneth Benoit

Essex Summer School 2014

July 25, 2014

Rationale for dictionaries

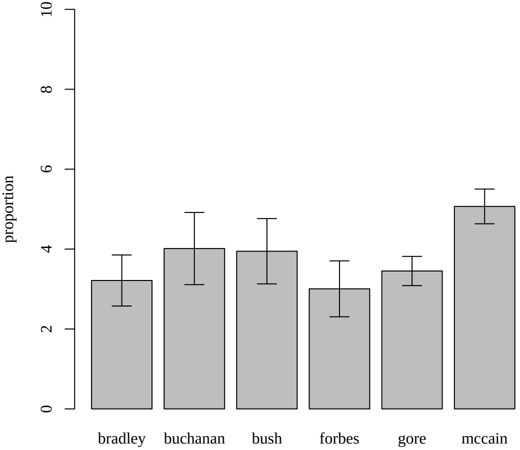
- Rather than count words that occur, pre-define words associated with specific meanings
- Another move toward the fully automated end of the text analysis spectrum, since involves no human decision making as part of the text analysis procedure
- Frequently involves lemmatization: transformation of all inflected word forms to their “dictionary look-up form” — more powerful than stemming
- Example: General Inquirer codes *I*, *me*, *my*, *mine*, *myself* as self, and *we*, *us*, *our*, *ours*, *ourselves* as selves

Well-known dictionaries: General Inquirer

- General Inquirer (Stone et al 1966)
- Maps texts to counts from an extensive dictionary
- Latest version contains 182 categories – the "Harvard IV-4" dictionary, the "Lasswell" dictionary, and five categories based on the social cognition work of Semin and Fiedler
- Examples: "self references", containing mostly pronouns; "negatives", the largest category with 2291 entries
- Uses stemming
- Also uses disambiguation, for example to distinguishes between *race* as a contest, *race* as moving rapidly, *race* as a group of people of common descent, and *race* in the idiom "rat race"
- Output example: `http://www.wjh.harvard.edu/~inquirer/Spreadsheet.html`

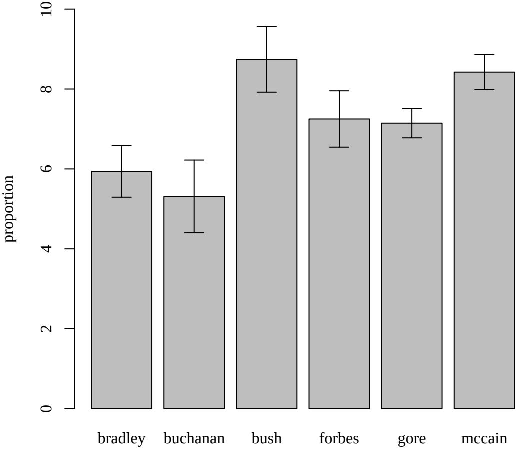
General Inquirer Applied to US Presidential Candidate Speeches (2000)

Negative language



General Inquirer Applied to US Presidential Candidate Speeches (2000)

Positive language



Well-known dictionaries: Regressive Imagery Dictionary

- Consists of about 3,200 words and roots, assigned to 29 categories of primary process cognition, 7 categories of secondary process cognition, and 7 categories of emotions
- designed to measure primordial vs. conceptual thinking
  - Conceptual thought is abstract, logical, reality oriented, and aimed at problem solving
  - Primordial thought is associative, concrete, and takes little account of reality – the type of thinking found in fantasy, reverie, and dreams
- Categories were derived from the theoretical and empirical literature on regressive thought by Martindale (1975, 1990)

Regressive Imagery Dictionary categories

- Full listing of categories

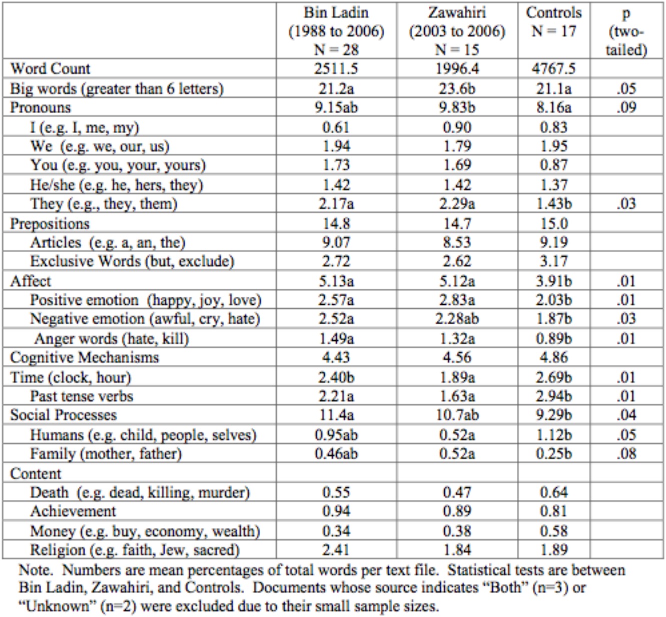
1 orality
2 anality
3 sex
4 touch
5 taste
6 odour
7 general sensation
8 sound
9 vision
10 cold
11 hard
12 soft
13 passivity
14 voyage
15 random movement
16 diffusion
17 chaos
18 unknown
19 timelessness
20 counscious
21 brink-passage
22 narcissism
23 concreteness
24 ascend
25 height
26 descent
27 depth
28 fire
29 water
30 abstract thought
31 social behaviour
32 instrumental behaviour
33 restraint
34 order
35 temporal references
36 moral imperative
37 positive affect
38 anxiety
39 sadness
40 affection
41 aggression
42 expressive behaviour
43 glory
44 female role
45 male fole
46 self
47 related others
48 diabolic
49 aspiration
50 angelic
51 flowers
52 synthesize
53 streight
54 weakness
55 good
56 bad
57 activity
58 being
59 analogy
61 integrative con
62 novelty
63 negation
64 triviality
65 transmute

- More on categories:
`http://www.kovcomp.co.uk/wordstat/RID.html`

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- Exact dictionary is proprietary (e.g. *secret*) but you can view a summary here:
  `http://www.liwc.net/descriptiontable1.php`

Example: Terrorist speech



	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two-tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

Example: Laver and Garry (2000)

- A *hierarchical* set of categories to distinguish policy domains and policy positions – similar in spirit to the CMP
- Five domains at the top level of hierarchy
  - economy
  - political system
  - social system
  - external relations
  - a “ ‘general’ domain that has to do with the cut and thurst of specific party competition as well as uncodable pap and waffle”
- Looked for word occurences within “word strings with an average length of ten words”
- Built the dictionary on a set of specific UK manifestos

Example: Laver and Garry (2000): Economy

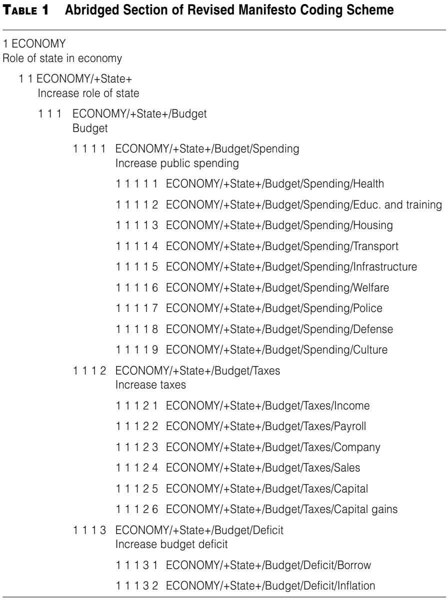


TABLE 1 Abridged Section of Revised Manifesto Coding Scheme

1	ECONOMY Role of state in economy
1 1	ECONOMY/+State+ Increase role of state
1 1 1	ECONOMY/+State+/Budget Budget
1 1 1 1	ECONOMY/+State+/Budget/Spending Increase public spending
1 1 1 1 1	ECONOMY/+State+/Budget/Spending/Health
1 1 1 1 2	ECONOMY/+State+/Budget/Spending/Educ. and training
1 1 1 1 3	ECONOMY/+State+/Budget/Spending/Housing
1 1 1 1 4	ECONOMY/+State+/Budget/Spending/Transport
1 1 1 1 5	ECONOMY/+State+/Budget/Spending/Infrastructure
1 1 1 1 6	ECONOMY/+State+/Budget/Spending/Welfare
1 1 1 1 7	ECONOMY/+State+/Budget/Spending/Police
1 1 1 1 8	ECONOMY/+State+/Budget/Spending/Defense
1 1 1 1 9	ECONOMY/+State+/Budget/Spending/Culture
1 1 1 2	ECONOMY/+State+/Budget/Taxes Increase taxes
1 1 1 2 1	ECONOMY/+State+/Budget/Taxes/Income
1 1 1 2 2	ECONOMY/+State+/Budget/Taxes/Payroll
1 1 1 2 3	ECONOMY/+State+/Budget/Taxes/Company
1 1 1 2 4	ECONOMY/+State+/Budget/Taxes/Sales
1 1 1 2 5	ECONOMY/+State+/Budget/Taxes/Capital
1 1 1 2 6	ECONOMY/+State+/Budget/Taxes/Capital gains
1 1 1 3	ECONOMY/+State+/Budget/Deficit Increase budget deficit
1 1 1 3 1	ECONOMY/+State+/Budget/Deficit/Borrow
1 1 1 3 2	ECONOMY/+State+/Budget/Deficit/Inflation

Example: Laver and Garry (2000)

```
ECONOMY / +STATE
    accommodation
    age
    ambulance
    assist
    ...

ECONOMY / -STATE
    choice*
    compet*
    constrain*
    ...
```

How to build a dictionary

- The ideal content analysis dictionary associates all and only the relevant words to each category in a perfectly valid scheme
- Three key issues:
  - Validity Is the dictionary's category scheme valid?
  - Sensitivity Does this dictionary identify *all* my content?
  - Specificity Does it identify *only* my content?

How to build a dictionary

Assume you want to construct an entry for the category ‘Terrorism’
Imagine two different dictionary entries:

- One contains all the words in the language (D1)
- The other contains the word ‘terrorist’ (D2)

D1 is *highly sensitive*: no language about terrorism is ever missed,
but *highly unspecific*: terrorism language is swamped
D2 is *highly specific*: the word occurs in discussions of terrorism,
but *highly insensitive*: much terrorism language is ignored
Of course, useful dictionaries lie in the middle

Coding scheme fundamentals

1. First key principle: Hierarchy
   1.1 First level: Domain
   1.2 Second level: subdomain
   1.3 (Third+ levels: may be additional sub-domains)
2. Second key principle: Confrontation
   Lowest-level categories should be for/against pairs, or "for/neutral/against"
3. On testing: Not necessary at design stage in the same way as for human coding – this is replaced by sensitivity/specificity testing in dictionary construction

How to build a dictionary

1. Identify “extreme texts” with “known” positions. Examples:
   - Opposition leader and Prime Minister in a no-confidence debate
   - Opposition leader and Finance Minister in a budget debate
   - Five-star review of a product (excellent) and a one-star review (terrible)
2. Search for differentially occuring words using word frequencies
3. Examine these words in context to check their sensitivity and specificity
4. Examine inflected forms to see whether stemming or wildcarding is required
5. Use these words (or their lemmas) for categories

Detecting “keywords”

- Detects words that *discriminate* between partitions of a corpus
- For instance, we could partition the Irish budget speech corpus into “government” and “opposition” speeches, and look for words that occur in one partition with higher relative frequency in opposition than in government speeches
- This is done by constructing a 2×2 table for each word, and testing association between that word and the partition categories

Detecting “keywords”: Constructing the association table

	Class A	**Class B**	Total
Word	a	b	$a+b$
~ Word	c	d	$c+d$
Total	$a+c$	$b+d$	$N = a+b+c+d$

Pearson's chi-squared statistic

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected} = \sum_{i=1}^{k} \frac{(Y_i - np_i)^2}{np_i}$$

$$d.f. = k - 1$$

Chi-squared test of independence

Basic intuition: if the two variables were independent of each other, the relative proportions should be similar to the marginal distributions.

E.g. a word would occur at equal relative frequencies in each subset of a corpus

Since we have two margins, we need to calculate the proportion as:

$$\hat{p}_{word,subset} = \hat{p}_{word} \times \hat{p}_{subset}$$

Generally:

$$\text{Expected Frequency} = \frac{r}{N} \cdot \frac{c}{N} \cdot n = \frac{rc}{N}$$

where r and c refer to row and column marginals

Chi-squared test of independence: example

Look for the association of "Christmas" with government or opposition in the Irish budget speeches (2010) corpus.

	Government	Opposition	
"Christmas"	1	18	19
Other word	17,126	31,752	48,878
	17,127	31,770	48,897

Next step: calculate expected proportions by multiplying marginal proportions.

Chi-squared test of independence: example

	Government	
“Christmas”	(19 * 17,127)/48,897	19
Other word	(48,878 * 17,127)/48,897	48,878
	17,127	48,897

	Opposition	
“Christmas”	(19 * 31,770)/48,897	19
Other word	(48,878 * 31,770)/48,897	48,878
	31,770	48,897

Next step: calculate this through.

Chi-squared test of independence: example

	Government	Opposition	
“Christmas”	6.66	12.34	19
Other word	17120.34	31,757.66	48,878
	17,127	31,770	48,897

Next step: compare expected to observed values.

Chi-squared test of independence: example

	Government	Opposition	
"Christmas"	$1 - 6.66 = -5.66$	$18 - 12.34 = 5.66$	19
Other word	$17127 - 17120.34 = 6.66$	$31752 - 31757.66 = 5.66$	48,878
	17,127	31,770	48,897

Next step: calculate χ^2.

Chi-squared test of independence

$$\chi^2 = \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} \frac{(Y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

$$d.f. = (n_r - 1)(n_c - 1)$$

Chi-squared test of independence

$$\begin{aligned}\chi^2 &= (-5.66)^2/6.66 + (5.66)^2/12.34 + \\ &\quad (6.66)^2/17120.34 + (5.66)^2/31757.66 \\ &= 7.41\end{aligned}$$

$$d.f. = (n_r - 1)(n_c - 1)$$

```
> 1 - pchisq(7.41, 1)
[1] 0.006486232
```

Likelihood ratio test of independence

Alternative test to χ^2 to test association (independence).

Basic intuition: We compute the likelihood of the observed data and divide this by the likelihood of the independence model. Higher values are less likely to have occurred by chance.

So with r and c referring to row and column marginals, with expected frequency

$$\hat{y}_{ij} = \frac{r}{N} \cdot \frac{c}{N} \cdot n = \frac{rc}{N}$$

then

$$G^2 = 2 \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} y_{ij} \log \frac{y_{ij}}{\hat{y}_{ij}}$$

aymptotically, $G^2 \sim \chi^2$ with $(n_r - 1)(n_c - 1)$ d.f.

What to do with dictionary results

- Describe the results
- Scale quantities: pro- v. anti-, left v. right, etc. Example: Laver and Garry
- Could use these as features to measure similarity using (e.g.) cosine similarity
- Treat as other features and use machine learning or data mining methods