Day 7: Words as Data Approaches

Kenneth Benoit

Essex Summer School 2011

July 19, 2011

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

Training, validation, and test sets

- We can steal some useful terminology from Machine Learning:

Training set	documents you use to build the dictionary
Validation set	documents you use to tell how well you're doing
Test set	documents you use to quantify external validity

- This scheme is intended to avoid 'over-fitting' — building a dictionary that is highly specific to a set of documents

- A problem if you only sampled the population of texts, or want to use the dictionary on new data

Prior probabilities and updating

A test is devised to automatically flag news stories about terrorism

- 1% of news stories in general pertain to terrorism
- 80% of news stories will be flagged by the test as about terrorism
- 10% of non-terrorism stories will also be flagged

We run the test on a new news story, and it is *flagged as* being about terrorism.

Question: What is probability that the story is *actually* about terrorism?

Prior probabilities and updating

- What about without the test?
  - Imagine we run 1,000 news stories through the test
  - We expect that 10 will be about terrorism
- With the test, we expect:
  - Of the 10 found to be about terrorism, 8 should be flagged as terrorism
  - Of the 990 non-terrorism stories, 99 will be wrongly flagged as terrorism
  - That's a total of 107 stories flagged as terrorism
- So: the updated probability of a story being terrorism, conditional on being flagged as terrorism, is $\frac{8}{107} = 0.075$
- The *prior* probability of 0.01 is updated to only 0.075 by the positive test result
- This is an example of Bayes' Rule,

$$\Pr(R = 1|T = 1) = \frac{\Pr(T{=}1|R{=}1)\Pr(R{=}1)}{\Pr(T{=}1)}$$

A Sketch of the Statistical Framework

Bayes Theorem:

$$P(\theta \mid W) = \frac{P(W \mid \theta)P(\theta)}{P(W)}$$

So if $P(\theta = \text{`agriculture}') = 0.5$ then

	θ		
	agriculture	security	
nuclear	0	1	1
tractor	1	0	1
revolution	0.78	0.22	1

Proportions

Compute category proportions (as before):

$$\hat{\theta}_i = \frac{C_i}{\sum_j C_j}$$

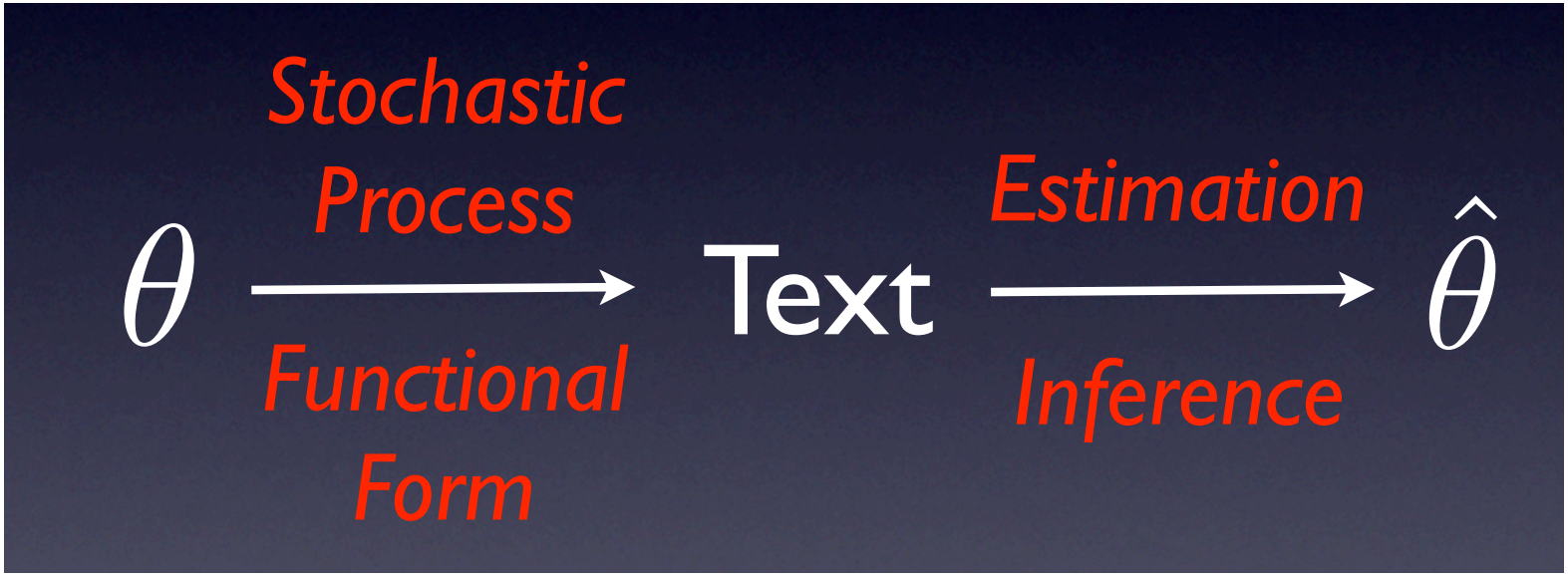
C_i is a sum of $P(\theta = i \mid W)$s which can now be fractional

- e.g. two tokens of 'revolution' adds 1.56 to agriculture and 0.44 to security

Text as Data: Basic Principles

- Data are observed characteristics of underlying tendencies to be estimated – and therefore not *intrinsically* interesting
- Analysis inherit properties of statistics:
  - Precise characterizations of uncertainty (efficiency of estimators)
  - Concerns with reliability (consistency of estimators)
  - Concerns with validity (unbiasedness of estimators)
- We must be concerned with the stochastic processes generating the data
- We must be concerned with functional relationships between characteristics of texts and authors and observed words

Text generation as a stochastic process



Problem 1: Interpret this data!

	party	spend_~l	votes1st	electo~e	gender	margin~d	m
1.	ind	6544.23	335	95060	m	Safe	5
2.	ind	14558.26	1614	95060	m	Safe	5
3.	fg	19153.71	5468	95060	m	Winnable	5
4.	lab	10658.21	4272	95060	m	Winnable	5
5.	ff	19648.3	9343	95060	m	Safe	5
6.	ff	16968.18	12489	95060	m	Winnable	5
7.	ff	24100.27	8711	95060	m	Unlikely	5
8.	gp	12110.11	4961	95060	f	Unlikely	5
9.	lab	8404.43	3732	95060	m	Unlikely	5
10.	fg	19743.1	7841	95060	m	Unlikely	5
11.	lab	.	.	95060	m	Unlikely	5
12.	sf	6633.45	2078	95060	m	Unlikely	5
13.	fg	11217.01	4819	87087	m	Unlikely	5
14.	ff	22383.53	10679	87087	m	Unlikely	5
15.	sf	28953.32	10832	87087	m	Safe	5

	party	spend_~l	votes1st	electo~e	gender	margin~d	m
16.	lab	8756.67	550	87087	m	Safe	5
17.	pd	30573.12	1131	87087	m	Safe	5
18.	ind	17196.73	1026	87087	m	Safe	5
19.	gp	10699.87	1100	87087	m	Unlikely	5
20.	fg	12839.2	4639	87087	m	Unlikely	5
21.	ind	17934.79	7722	87087	m	Unlikely	5
22.	ff	20122.19	3731	87087	m	Unlikely	5
23.	ff	21483.37	7204	87087	m	Unlikely	5
24.	fg	11124	6113	87087	m	Unlikely	5
25.	csp	3141.27	358	87087	m	Unlikely	5
26.	ind	34542.73	1943	87087	m	Unlikely	5
27.	ff	13120.48	6717	78643	m	Unlikely	4
28.	gp	7771.53	2903	78643	m	Safe	4
29.	csp	140	176	78643	m	Safe	4
30.	fg	14195.46	4015	78643	m	Safe	4


54900
54793
52161
52903
92656
53672
50850
85326
72639
58662
80924
53202
65199
68396
79329
77126
80219
73511
57568
8340
67631
57048
87909
58205
80109
54670
91666
87737
87087
95060
69172
90219
107894
61194
78643
47394
94969
94096
51793
71907
56386
71679
53132

Problem 2: Interpret this data!

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. The fight against increases in the cost of living is the most important single issue in economic management.

People without jobs represent waste of productive effort: National supports a policy of full employment and the dignity of labour. We do not accept unemployment as a balancing factor in economic management.

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / The fight against increases in the cost of living is the most important single issue in economic management. /

People without jobs represent waste of productive effort: / National supports a policy of full employment / and the dignity of labour. / We do not accept unemployment as a balancing factor in economic management. /

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / The fight against increases in the cost of living is the most important single issue in economic management. /

People without jobs represent waste of productive effort: / National supports a policy of full employment / **and the dignity of labour.** / We do not accept unemployment as a balancing factor in economic management. /

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

??????

We believe that continued double-figure inflation will destroy the basis of the New Zealand economy and cause untold misery. / The fight against increases in the cost of living is the most important single issue in economic management. /

People without jobs represent waste of productive effort: / National supports a policy of full employment / **and the dignity of labour.** / We do not accept unemployment as a balancing factor in economic management. /

Finally, the National Development Council will be restored and consultation resumed between Government departments, academic specialists and private industry, including farming and organised labour.

408 Economic Goals:

Statements of intent to pursue any economic goals not covered by other categories.

Problem 3: Now try this one!

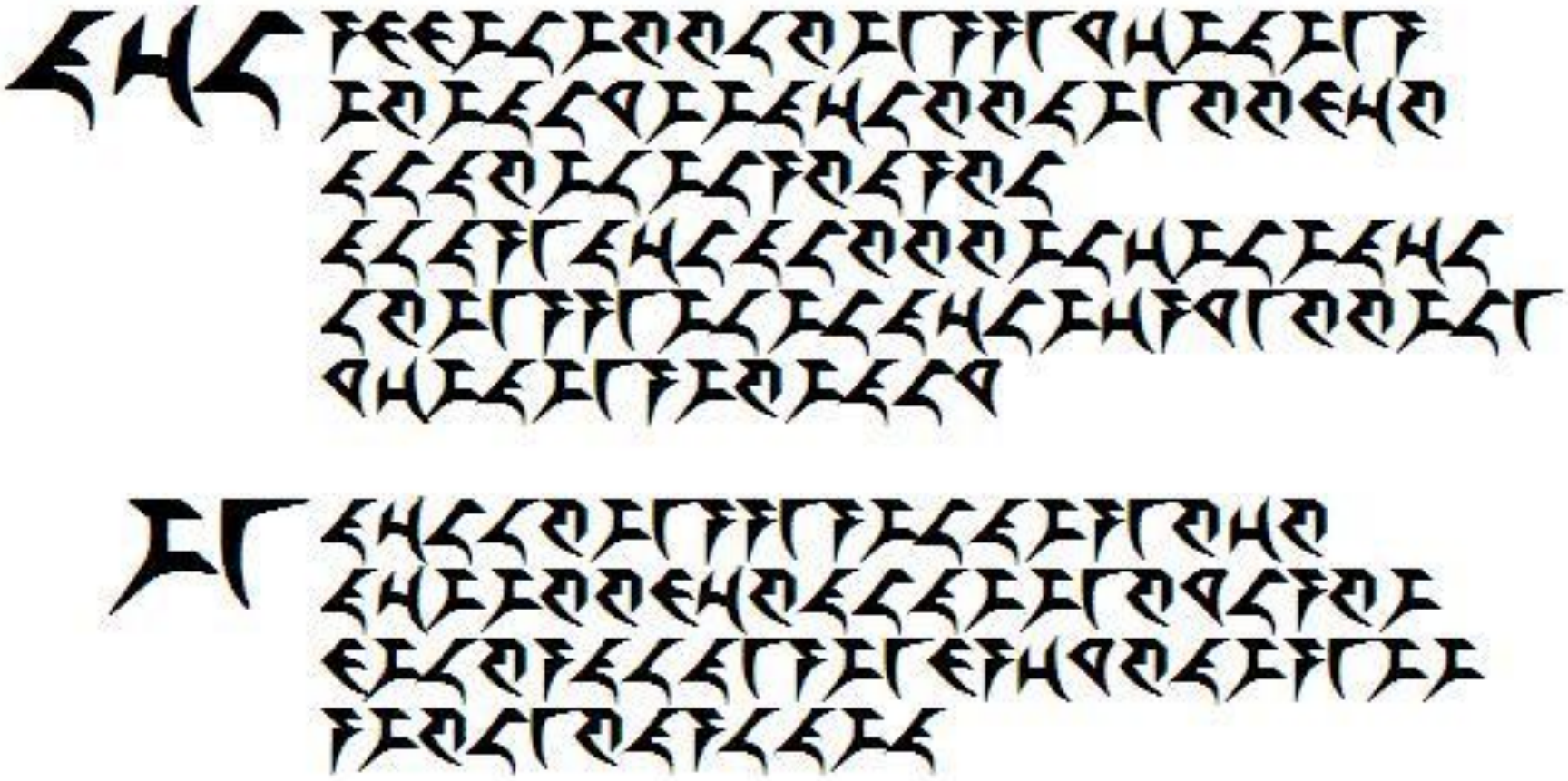
Kansainväliset uraaniyhtiöt ovat olleet kiinnostuneita Kainuussa sijaitsevista esiintymistä. Kainuun maakunta-kuntayhtymä on Perussuomalaisten valtuustoryhmän aloitteen pohjalta selvittänyt kainuulaisten suhtautumista mahdollisiin uraanikaivoksiin.

Sotkamossa sijaitsevan Talvivaaran kaivoksen sivutuotteena tulee myös uraania, joka aiotaan ottaa jätelietteestä talteen. Tässä uraanin talteenotossa syntyy niin paljon ydinvoimalaitosten polttoainetta, että se riittäisi noin 80 prosenttisesti Suomessa toimivien ydinvoimaloiden tarpeisiin.

Talvivaaran tapauksessa ei kaivoksen johdon mukaan ole kysymys varsinaisen uraanikaivoksen avaamisesta, vaan vain sivutuotteen talteenotosta. Valtioneuvosto tullee päättämään Talvivaara-asiasta uraanin osalta tämän vuoden aikana.

Perussuomalaiset

and this one???



Wordscores in a nutshell

- *Wordscores* is a statistical method for extracting policy positions from political texts, implemented by computer
- Michael Laver, Kenneth Benoit, and John Garry. "Extracting Policy Positions From Political Texts Using Words as Data", *APSR* 2003
- Enables extraction of policy positions from texts without having to ascribe "meaning" to texts, or to read them, or even to be able to read them (works in English, German, French, and Italian so far)
- Because it is based on the statistics of relative word frequencies, Wordscores can generate estimates of uncertainty, something no existing methods of textual content analysis offer

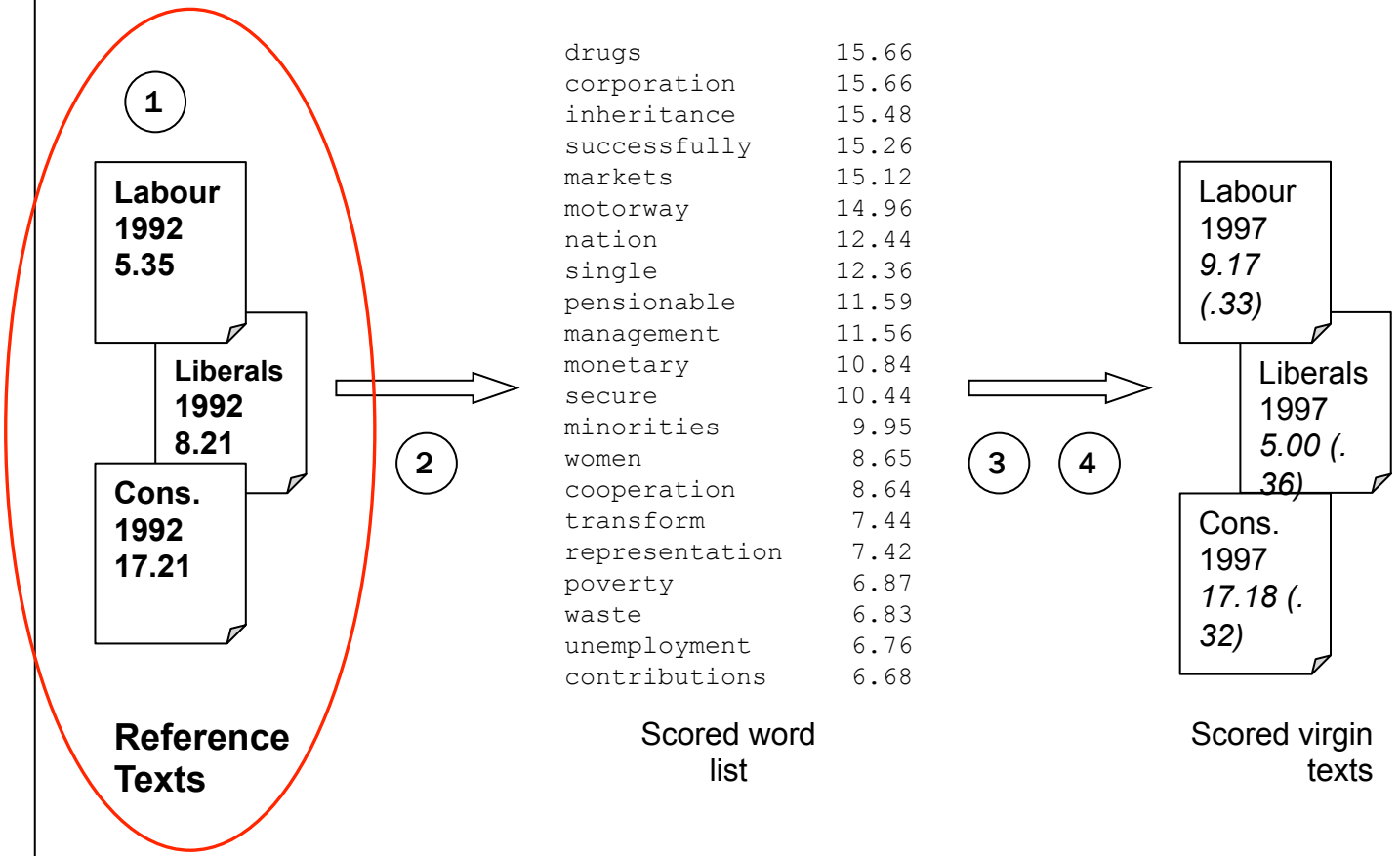
WORDSCORES conceptually

- "Reference" texts: texts about which we know something (a scalar dimensional score)
- "Virgin" texts: texts about which we know nothing (but whose dimensional score we'd like to know)
- Basic procedure:

1. Analyze reference texts to obtain word scores
2. Use word scores to score virgin texts

The Wordscore Procedure

(Using the UK 1997-2001 Example)

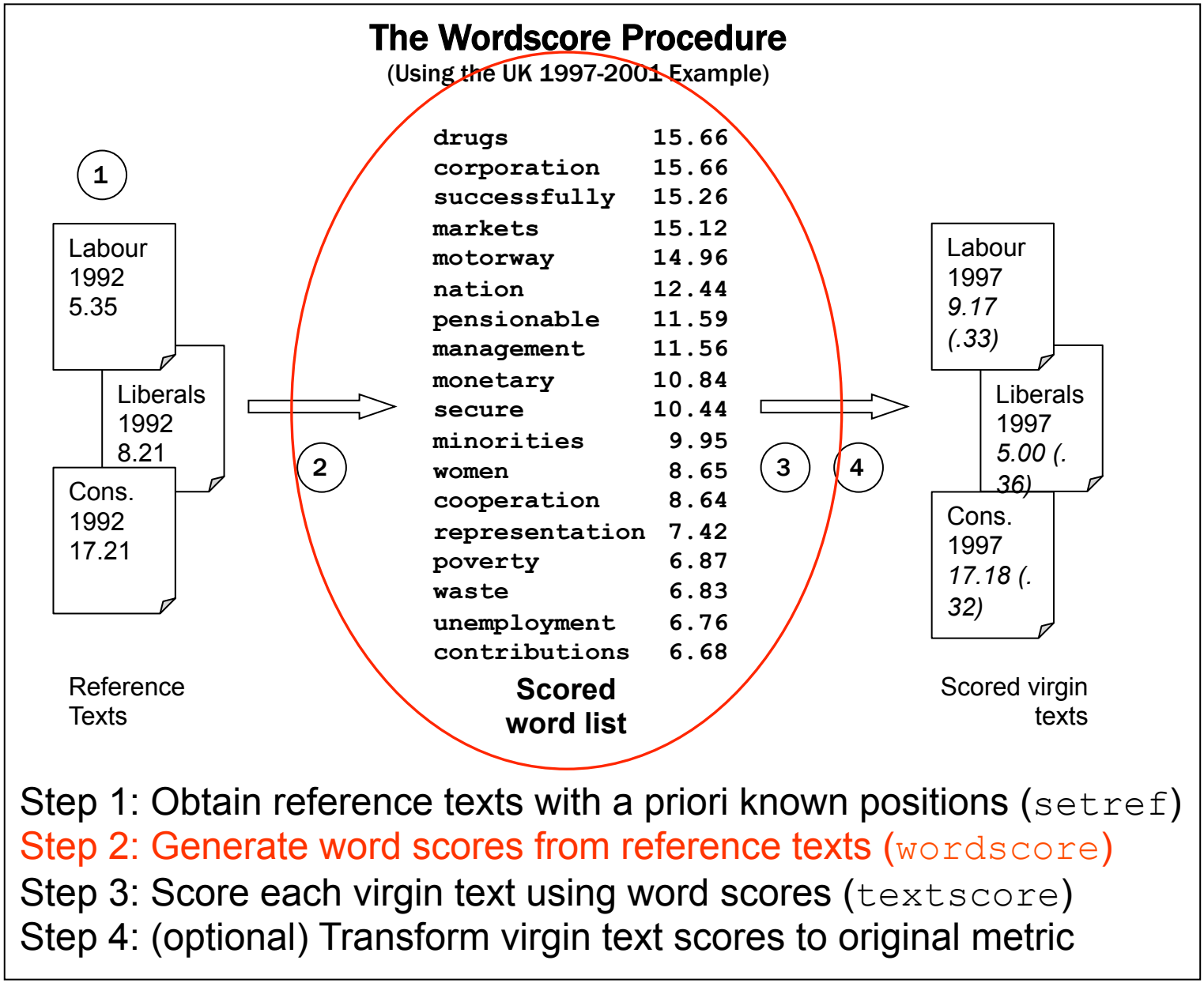


Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric



Step 1: Obtain reference texts with a priori known positions (`setref`)

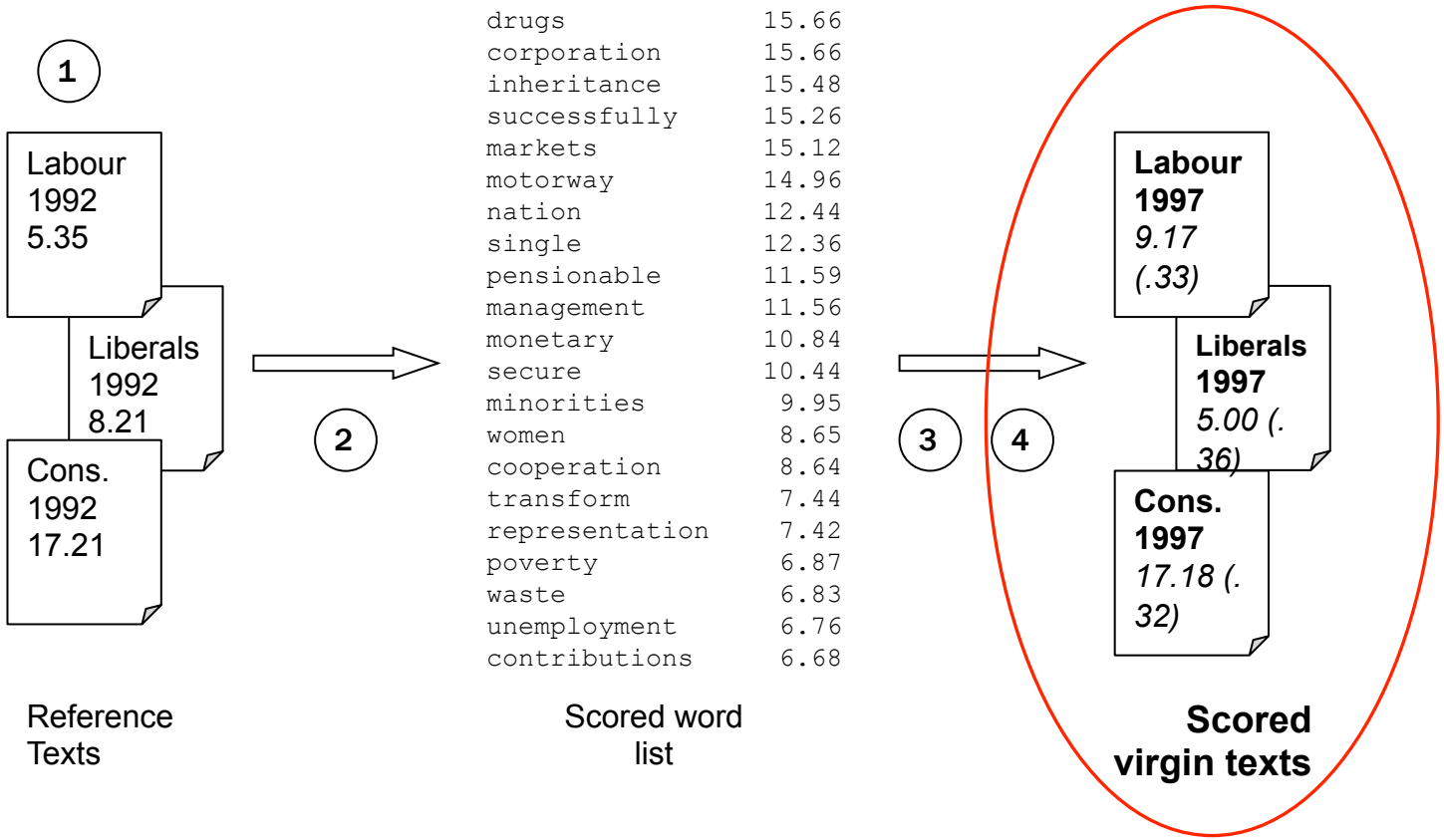
Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscore Procedure

(Using the UK 1997-2001 Example)



Step 1: Obtain reference texts with a priori known positions (`setref`)

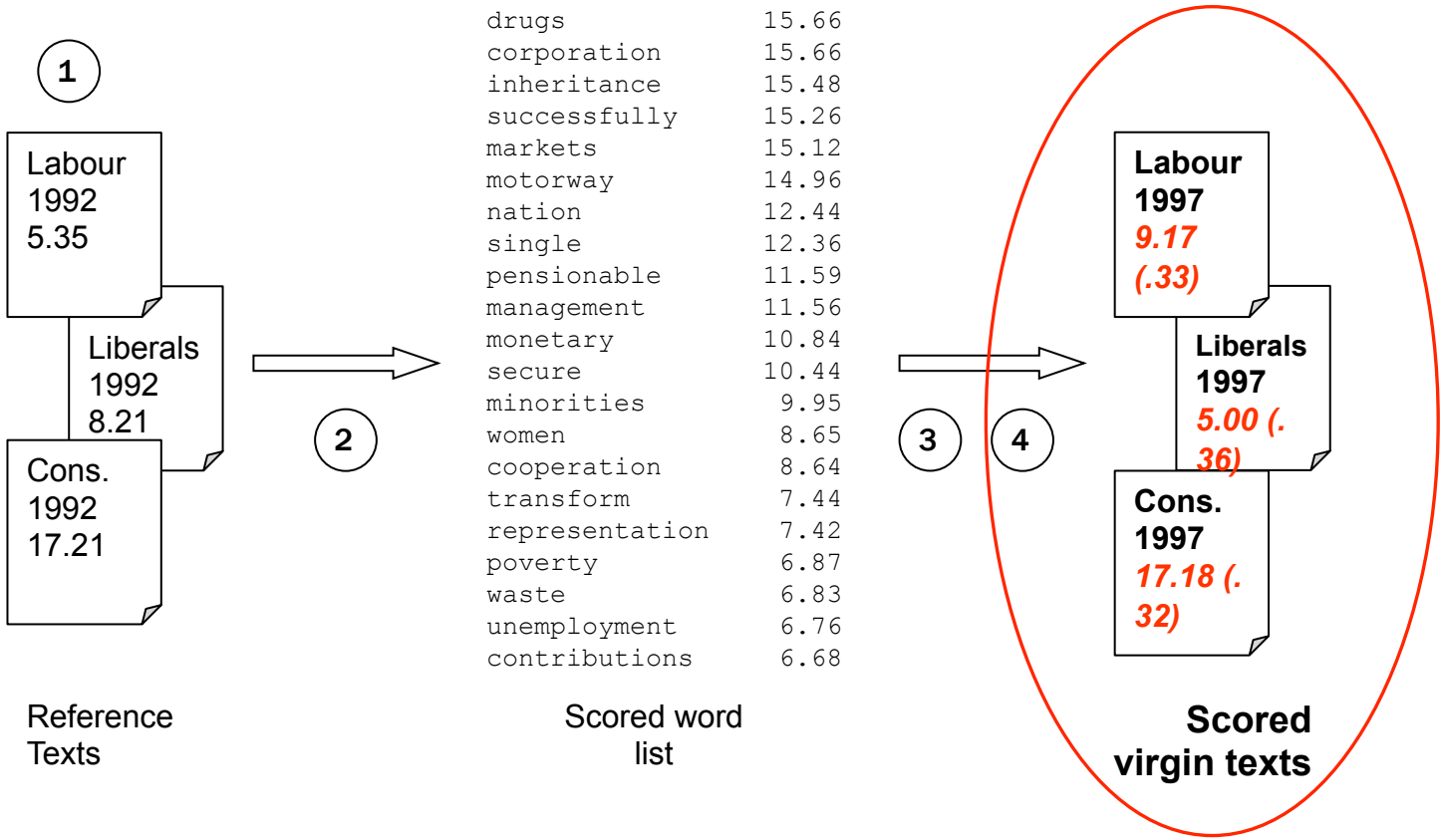
Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

The Wordscore Procedure

(Using the UK 1997-2001 Example)



Step 1: Obtain reference texts with a priori known positions (`setref`)

Step 2: Generate word scores from reference texts (`wordscore`)

Step 3: Score each virgin text using word scores (`textscore`)

Step 4: (optional) Transform virgin text scores to original metric

WORDSCORES mathematically

- Start with *R* reference texts and *V* virgin texts with with *W* words in common

  (using `wordcount` to generate a matrix of words and their relative frequencies in all reference texts)

- A_{rd} is assumed position of reference text *r* on policy dimension *d*

- F_{wr} is relative frequency of word *w* in text *r*

WORDSCORES mathematically

- Compute P_{wr} for each **reference text**: the probability that we are reading reference text *r* given that we are reading word *w*

$$P_{wr} = \frac{F_{wr}}{\sum_{r} F_{wr}}$$

- Example:

  Two reference texts, A and B. The word "choice" is used 10 times per 1,000 words in Text A and 30 times per 1,000 words in Text B.

  If we know only that we are reading the word "choice" in one of the two reference texts, then probability is 0.25 that we are reading Text A, and 0.75 that we are reading Text B.

WORDSCORES mathematically

- Compute S_{wd} for each **word**: the score of each word *w* on dimension *d*

$$S_{wd} = \Sigma_r (P_{wr} \cdot A_{rd})$$

- Example continued:

  We know (from independent sources) that Reference Text A has a position of –1.0 on dimension *d*, and Reference Text B has a position of +1.0.

  The score of the word "choice" is then:

  0.25 (– 1.0) + 0.75 (1.0) = – 0.25 + 0.75 = + 0.5

WORDSCORES mathematically

- Compute S_{vd} for each **virgin text**: the score of each virgin text *v* on dimension *d*

$$S_{vd} = \sum_{w} \left(F_{wv} \cdot S_{wd} \right)$$

- This score is the **mean** dimension score of all of the scored words that a virgin text contains, weighted by the frequency of the scored words
- *Uncertainty*: A weighted **variance** V_{vd} can also be computed for each virgin text, representing the uncertainty of the estimate S_{vd}. Because every words adds information to S_{vd}, more words reduce our uncertainty about S_{vd}. Also, the more consensus among the virgin words around S_{vd}, the more certain we are about S_{vd} .
- *Rescaling:* S_{vd} can be rescaled as S^*_{vd} for interpretation on the original metric of the reference text scores.