

Day 3: Quantitative methods for comparing texts

Kenneth Benoit

Spring 2014

February 10, 2014

Session 2 Basic Outline

- ▶ Building blocks/foundations of quantitative text analysis
- ▶ Justifying a term/feature frequency approach
- ▶ Selecting texts
- ▶ Selecting features
- ▶ Practical issues working with texts
- ▶ Demonstrations
- ▶ Examples

RELIABILITY IN TEXT ANALYSIS

Tradeoff: Reliability versus validity

- ▶ **Reliability** refers to the dependability and replicability of the data generated by the text analysis method
- ▶ **Validity** is the quality of the data that leads us to accept it as “true,” insofar as it measures what it is claimed to measure
- ▶ In text analysis, these two objectives frequently **trade off** with one another, since only human judgment can (ultimately) ensure validity, but human judgment is inherently unreliable
- ▶ Each concept has many variations, and in the case of reliability, several measures that can be applied
- ▶ Validity is the hardest to establish, since questions can always be raised about human judgment

Examples of tradeoffs

- ▶ Examples in coding text units:
 - ▶ Perfectly reliable procedure: Code all text units as pertaining to “Economic growth: positive”
 - ▶ Perfectly valid: Get a Nobel Prize laureate in economics to classify each text unit
- ▶ Examples in unitizing a text:
 - ▶ Perfectly reliable: Have a computer parse all texts into n -grams, such as words, pairs of adjacent words, etc. based on pre-defined rules (space is a delimiter, etc.)
 - ▶ Perfectly (?) valid: Have expertly trained humans parse the text into “quasi-sentences”

Reliability: Definitions

Reliability in essence means getting the same answers each time an identical research procedure is conducted.

- ▶ The extent to which a research procedure yields the same results on repeated trials (Carmines and Zeller 1979)
- ▶ The assurance that data are obtained independently of the measuring event, instrument, or person, and that remain constant despite variations in the measuring process (Kaplan and Goldsen 1965)
- ▶ Interpretivist conception: Degree to which members of a designated community agree on the readings, interpretations, responses to, or uses of given texts or data (Krippendorff)

Importance of Reliability

- ▶ In text analysis (and most other forms of empirical analysis), unreliable procedures yield results which are meaningless.
- ▶ Typically measures in terms of **agreement** between two human coders, when referring to hand-coded content analysis
- ▶ Computerized methods have largely removed this concern, inasmuch as they are mechanical procedures that yield the same results each time the procedure is repeated.

Types of reliability

Distinguished by the way the reliability data is obtained.

Type	Test Design	Causes of Disagreements	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intraobserver inconsistencies + interobserver disagreements	medium
Accuracy	test-standard	intraobserver inconsistencies + interobserver disagreements + deviations from a standard	strongest

Reliability test designs

- Test-retest** The same text is reanalyzed/reread/reclassified, or the same measurement is repeatedly applied to the same set of texts. Goal is to establish inconsistencies. (Establishes *stability*)
- Test-test** Two or more individuals, working independently, apply the same analysis instructions to the same texts, to compare intraobserver differences. (Establishes *reproducibility*).
- Test-standard** The performance of one or more procedures is compared to a procedure that is taken to be correct. Deviations from a (“gold”) standard are then recorded. (Establishes *accuracy*.) Typically used in coder training, or training of automated (computer-based) procedures.

Designing reliability checks in practice

- ▶ Repeating the procedure on the sample data
- ▶ Using independent tests from separate coders
- ▶ Can a “gold standard” be identified?
- ▶ Split-design tests
- ▶ Example: CMP
 - ▶ Same coders repeat own codings
 - ▶ Different coders code same test
 - ▶ The “reliability” coefficient reported in the dataset is correlation of category percentages obtained by a coder on the training document used by CMP versus the master “gold standard” version of the coding done by Andrea Volkens

Measures of agreement

- ▶ **Percent agreement** Very simple: (number of agreeing ratings) / (total ratings) * 100%
- ▶ **Correlation**
 - ▶ (usually) Pearson's r , aka product-moment correlation
 - ▶ Formula: $r_{AB} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{A_i - \bar{A}}{s_A} \right) \left(\frac{B_i - \bar{B}}{s_B} \right)$
 - ▶ May also be ordinal, such as Spearman's rho or Kendall's tau-b
 - ▶ Range is [0,1]
- ▶ **Agreement measures**
 - ▶ Take into account not only observed agreement, but also *agreement that would have occurred by chance*
 - ▶ **Cohen's κ** is most common
 - ▶ **Krippendorff's α** is a generalization of Cohen's κ
 - ▶ Both range from [0,1]

Reliability data matrixes

Example here used binary data (from Krippendorff)

Article:	1	2	3	4	5	6	7	8	9	10
Coder A	1	1	0	0	0	0	0	0	0	0
Coder B	0	1	1	0	0	1	0	1	0	0

- ▶ A and B agree on 60% of the articles: 60% agreement
- ▶ Correlation is (approximately) 0.10
- ▶ Observed *disagreement*: 4
- ▶ Expected *disagreement* (by chance): 4.4211
- ▶ Krippendorff's $\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{4}{4.4211} = 0.095$
- ▶ Cohen's κ (nearly) identical

Reliability and validity differences

- ▶ Reliability can be established through tests as a part of a research procedure; validity cannot be established through the same sort of (repetition) tests.
- ▶ Validity concerns substantive *truths*, whereas reliability is mainly procedural.
- ▶ Unreliability limits the chance of obtaining valid results, in the sense that procedures whose results cannot be trusted are less likely to be true.
- ▶ Reliability is no guarantee of validity, since reliable procedures can be consistently wrong, even when these procedures involve human judgment.

The design of the experiment

- ▶ Data: 14 speeches from the debate on Irelands 2010 budget (FF+Greens vs FG+Lab+SF)
- ▶ Subjects: 18 human readers, mostly PhD students (LSE and TCD)
- ▶ Task: Identify speaker positions, directly and by pairwise comparison and indicate uncertainty
- ▶ Questions: Does the model recover human positioning? What is appropriate certainty?

Walk through the paper...

RELIABILITY IN TEXT ANALYSIS

Some useful linguistic terms

From a field known as *corpus linguistics*

type for our purposes, a unique word

token any word – so token count is total words

hapax legomena (or just *hapax*) are types that occur just once

Key Words in Context

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.

lime (14)

79[C.10] 4 /Which was builded of **lime** and sand;/Until they came to
247A.6 4 /That was well biggit with **lime** and stane.
303A.1 2 bower./Well built wi **lime** and stane./And Willie came
247A.9 2 /That was well biggit wi **lime** and stane./Nor has he stoln
305A.2 1 a castell biggit with **lime** and stane./O gin it stands not
305A.71 2 is my awin./I biggit it wi **lime** and stane./The Tinnies and
79[C.10] 6 /Which was builded with **lime** and stone.
305A.30 1 a prittie castell of **lime** and stone./O gif it stands not
108.15 2 /Which was made both of **lime** and stone./Shee tooke him by
175A.33 2 castle then./Was made of **lime** and stone./The vttermost
178[H.2] 2 near by./Well built with **lime** and stone./There is a lady
178F.18 2 built with stone and **lime**!/But far mair pittie on Lady
178G.35 2 was biggit wi stane and **lime**!/But far mair pity o Lady
2D.16 1 big a cart o stane and **lime**./Gar Robin Redbreast trail it

Another KWIC Example (Seale et al (2006))

Table 3
Example of Keyword in Context (KWIC) and associated word clusters display

Extracts from Keyword in Context (KWIC) list for the word 'scan'

An MRI **scan** then indicated it had spread slightly

Fortunately, the MRI **scan** didn't show any involvement of the lymph nodes

3 very worrying weeks later, a bone **scan** also showed up clear.

The bone **scan** is to check whether or not the cancer has spread to the bones.

The bone **scan** is done using a type of X-ray machine.

The results were terrific, CT **scan** and pelvic X-ray looked good

Your next step appears to be to await the result of the **scan** and I wish you well there.

I should go and have an MRI **scan** and a bone **scan**

Three-word clusters most frequently associated with keyword 'scan'

<i>N</i>	<i>Cluster</i>	<i>Freq</i>
1	A bone scan	28
2	Bone scan and	25
3	An MRI scan	18
4	My bone scan	15
5	The MRI scan	15
6	The bone scan	14
7	MRI scan and	12
8	And Mri scan	9
9	Scan and MRI	9

Another KWIC Example: Irish Budget Speeches

WordStat 6.1.7 - IRISH BUDGETS.DBF

Dictionaries Options Frequencies Phrase finder Crosstab **Keyword-In-Context**

List: User defined Sort by: Case number

Word: CHRISTMAS Context delimiter: None

CASENO		KEYWORD	
2	nally disappointed by what we have seen today. Instead of the Minister taking the radica	Christmas	in the hope of something better in the new year? The Minister has failed those employees.
3	nts, people on disability and even blind people. The Minister has some nerve quoting Ted	Christmas	hit single. Fianna Fáil's hit single for Christmas will be, "I saw NAMA killing Santa Claus". Pa
3	Minister has some nerve quoting Ted Kennedy, the champion of the poor and fairness in A	Christmas	will be, "I saw NAMA killing Santa Claus". Parents should know that child benefit is being cu
3	ications, how much worse is it for the early school leaver and young unemployed person?	Christmas	because they must take the decision to leave, as people all over rural Ireland and every tov
3	r reminding everyone that Fianna Fáil was the party that looked after child benefit. It woul	Christmas	.With a possible election next year, one never knows when a club might come in handy to
3	is. The Minister should ask Tiger Woods about it. I have read scores of articles by people	Christmas	? Is the Society of St. Vincent de Paul out of touch? Are they saying social welfare in Irela
3	elusive but most vital ingredient of economic policy. One cannot bottle it or buy it and there	Christmas	line people were laden down with shopping bags. If one walks over to Grafton Street one
4	al effect on the economy and society. Social welfare payments are always returned to the	Christmas	bonus, a double payment which affected 1.3 million people, is money that would have beer
4	hey are spent on rent, mortgages, food, utilities and other essentials. Cutting welfare expe	Christmas	food. The Government's Scrooge measures will come back to haunt it when it counts its V.
4	nsiderable difference to the paltry few millions of euro offered to job creation and retento	Christmas	in debt, in poverty and with the prospect of the very small payments made to them by the S
4	embers of the Government spoken to people in rural Ireland about how even as we speak	Christmas	bonus. Of course, that is not too complicated and it can easily be accomplished. The Gover
4	nents will have a detrimental effect on the economy and society. Social welfare payments	Christmas	. The loss of the Christmas bonus, a double payment which affected 1.3 million people, is ri
6	is is not happening. Day after day, Deputies, including those opposite, are receiving eviden	Christmas	. I do not know whether Deputy Perry heard a woman from Sligo speaking on radio this mo
7	but the Government did not see fit to remove it. Such countries as Holland realised the ero	Christmas	period. We suggested that the lower rate of VAT should be reduced. That would not be as
8	o poverty. Every family is today paying the price for 12 years of incompetent, reckless, dis	Christmas	payment. A couple on invalidity pension suffers a cut of €1,100. Carer's benefit is cut by €
8	cal parties for an adjustment of €6 billion. However, choices had to be made. What were th	Christmas	payment is gone. Earnest lectures on price statistics will not feed a hungry child or clothe ?
8	have been put onto the dole queue. Fianna Fáil has created one of the longest and deepes	Christmas	, we will witness the scenes of heartbreak and loss at airports and ferry ports as the cre
13	fiscal crisis, as Deputy Gilmore pointed out. The policies within this budget will get us throu	Christmas	cess work will be done in Leinster House to replace gas boilers with biomass boilers. Th
14	it is over and that this is "the last big push". I was expecting him to say it will all be over by	Christmas	. If it is the last big push, we know who he's sending over the top — the low paid workers

I hear sports shops are doing a roaring trade in single golf clubs this **Christmas**. With a possible election next year, one never knows when a club might come in handy to deal with men who break their promises. The Minister should ask Tiger Woods about it.

I have read scores of articles by people who argue that child benefit payments are of little importance, including journalists and academics who argue it would make no difference if the payment were restricted. Most of these articles were written by men, none of whom could state absolutely that he spoke for his wife or partner. I have yet to meet a mother of young or teenage children who says casually that child benefit has no importance to her. Perhaps I do not mix in circles where this benefit is a trifle. Certainly, I do not represent a constituency that places no value on the advantages of universal child benefit.

Almost every day I hear the voice of Marian Finucane on radio advertisements for the Simon Community, as I am sure everyone here does. She tells us that the current crisis has brought community services to breaking point. I hear the same message from Professor John Monaghan of the Society of St. Vincent de Paul. Are these societies lying? Is the Simon Community faking its message this **Christmas**? Is the Society of St. Vincent de Paul out of touch? Are they saying social welfare in Ireland is so generous that it can be cut? I have

14 cases Number of items: 19

Basic descriptive summaries of text

Readability statistics Use a combination of syllables and sentence length to indicate “readability” in terms of complexity

Vocabulary diversity (At its simplest) involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens

Word (relative) frequency

Theme (relative) frequency

Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Flesch-Kincaid readability index

- ▶ F-K is a modification of the original **Flesch Reading Ease Index**:

$$206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right)$$

Interpretation: 0-30: university level; 60-70: understandable by 13-15 year olds; and 90-100 easily understood by an 11-year old student.

- ▶ **Flesch-Kincaid** rescales to the US educational grade levels (1-12):

$$0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59$$

Gunning fog index

- ▶ Measures the readability in terms of the years of formal education required for a person to easily understand the text on first reading
- ▶ Usually taken on a sample of around 100 words, not omitting any sentences or words
- ▶ Formula:

$$0.4 \left[\left(\frac{\text{total words}}{\text{total sentences}} \right) + 100 \left(\frac{\text{complex words}}{\text{total words}} \right) \right]$$

where complex words are defined as those having three or more syllables, not including proper nouns (for example, Ljubljana), familiar jargon or compound words, or counting common suffixes such as -es, -ed, or -ing as a syllable

Simple descriptive table about texts: Example

Speaker	Party	Tokens	Types
Brian Cowen	FF	5,842	1,466
Brian Lenihan	FF	7,737	1,644
Ciaran Cuffe	Green	1,141	421
John Gormley (Edited)	Green	919	361
John Gormley (Full)	Green	2,998	868
Eamon Ryan	Green	1,513	481
Richard Bruton	FG	4,043	947
Enda Kenny	FG	3,863	1,055
Kieran O'Donnell	FG	2,054	609
Joan Burton	LAB	5,728	1,471
Eamon Gilmore	LAB	3,780	1,082
Michael Higgins	LAB	1,139	437
Ruairi Quinn	LAB	1,182	413
Arthur Morgan	SF	6,448	1,452
Caoimhghin O'Caolain	SF	3,629	1,035
All Texts		49,019	4,840
<i>Min</i>		919	361
<i>Max</i>		7,737	1,644
<i>Median</i>		3,704	991
<i>Hapaxes with Gormley Edited</i>		67	
<i>Hapaxes with Gormley Full Speech</i>		69	

Quantifying similarity

Compare vectors of features for (binary) absence or presence – called (by Choi et al) “operational taxonomic units”

Table 1 OTUs Expression of Binary Instances i and j

$j \setminus i$	1 (Presence)	0 (Absence)	Sum
1 (Presence)	$a = i \bullet j$	$b = \bar{i} \bullet j$	$a+b$
0 (Absence)	$c = i \bullet \bar{j}$	$d = \bar{i} \bullet \bar{j}$	$c+d$
Sum	$a+c$	$b+d$	$n=a+b+c+d$

- ▶ Cosine similarity:

$$S_{\text{cosine}} = \frac{a}{\sqrt{(a+b)(a+c)}} \quad (1)$$

- ▶ Jaccard similarity:

$$S_{\text{Jaccard}} = \frac{a}{\sqrt{(a+b+c)}} \quad (2)$$

Quantifying similarity: Edit distances

- ▶ Edit distance refers to the number of operations required to transform one string into another
- ▶ Common edit distance: the **Levenshtein distance**
- ▶ Example: the Levenshtein distance between "kitten" and "sitting" is 3
 - ▶ kitten → sitten (substitution of "s" for "k")
 - ▶ sitten → sittin (substitution of "i" for "e")
 - ▶ sittin → sitting (insertion of "g" at the end).
- ▶ Not common, as at a textual level this is hard to implement and possibly meaningless

Summarizing

- ▶ Involves characterizing the coded text units using additional quantification

- ▶ Examples

Category frequencies Coded category frequency measures, such as the proportion of times “economy” is mentioned in a speech, or the proportion of mentions of the environment

Type/token measures Frequency tabulations of token types and their frequencies

Range/variance Here we might be interested in the total number or the spread or variance of categories used in particular documents or by particular speakers

- ▶ May also involve scales or indexes constructed from summary information

Summarizing: Example

Democratic	Republican
iraq	consent
administration	ask
year	unanimous
health	bill
families	committee
program	senate
care	30
debt	2006
women	border
veterans	senator
help	vote
americans	law
country	hearing
children	authorized
new	further
education	states
funding	proceed
workers	order
programs	session
disaster	time

Top 20 Democratic and Republican words from the 2006 US Senate (source: Nicholas Beauchamp 2008)

Summarizing: Scale Example

- ▶ A very simple example comes from the CMP, using PER110 “European Union: Positive Mentions” and PER108 “European Union: Negative Mentions”
- ▶ The overall pro- versus anti- EU-ness can be assessed as $PER110 - PER108$. Theoretical range is $[-100, 100]$.
- ▶ A more complicated example is the CMP’s famous “rile” index, which adds 26 categories of the “right” and subtracts from this the sum of 13 categories of the “left”.

Vocabulary Diversity Example

- ▶ Variations use vocabulary diversity analysis (e.g. Labbé et. al. 2004)

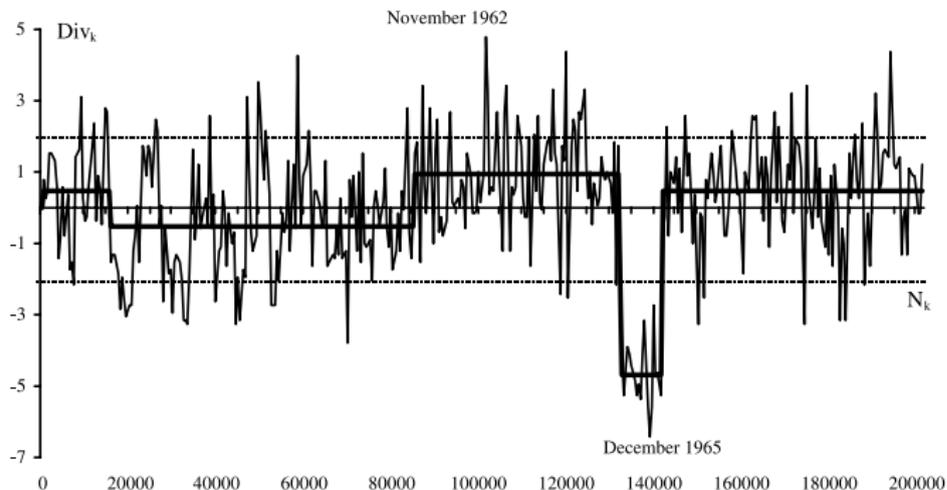


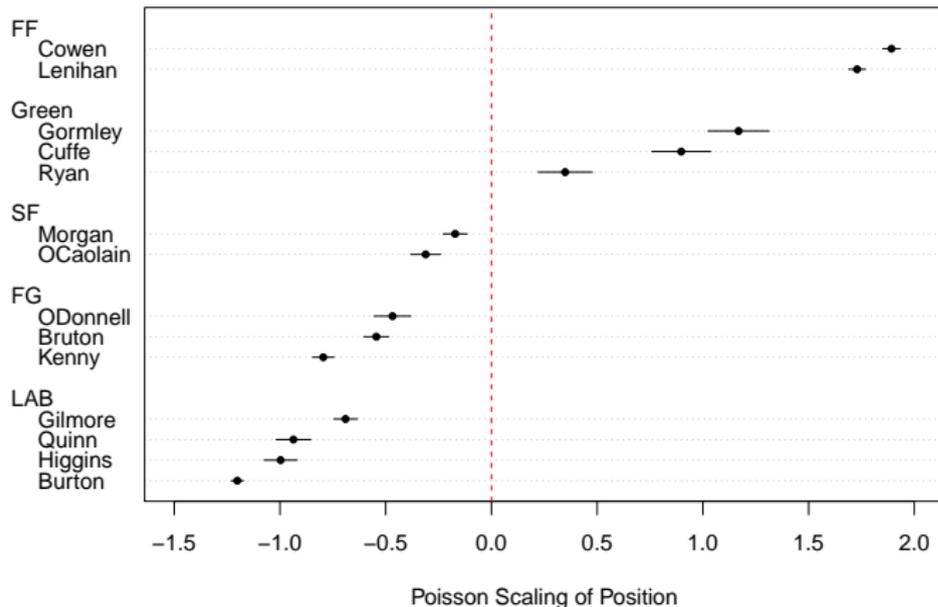
Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

Inference and Reporting

- ▶ This involves drawing conclusions from the research, and these conclusions will depend on the *validity* established by the research design
- ▶ Reporting means communicating the results in a clear and relevant fashion. (This can be challenging – see for instance the Schonhardt-Bailey article.)
- ▶ No iron-clad rules here – use your discretion as applied to a particular case

Graphical Methods: Example

- ▶ From a uni-dimensional scaling model from a term-document matrix (Poisson scaling)



LIWC Example

- From an application of the Linguistic Inquiry and Word Count dictionary to texts by Al Zawahiri and Bin Laden, benchmarked against a general corpus

	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.

Selecting more than words: collocations

collocations **bigrams**, or **trigrams** e.g. *capital gains tax*

how to detect: pairs occurring more than by chance, by measures of χ^2 or *mutual information* measures

example:

Summary Judgment	Silver Rudolph	Sheila Foster
prima facie	COLLECTED WORKS	Strict Scrutiny
Jim Crow	waiting lists	Trail Transp
stare decisis	Academic Freedom	Van Alstyne
Church Missouri	General Bldg	Writings Fehrenbacher
Gerhard Casper	Goodwin Liu	boot camp
Juan Williams	Kurland Gerhard	dated April
LANDMARK BRIEFS	Lee Appearance	extracurricular activities
Lutheran Church	Missouri Synod	financial aid
Narrowly Tailored	Planned Parenthood	scored sections

Table 5: Bigrams detected using the mutual information measure.

Scaling Issues

- ▶ Scaling becomes a major issue when we wish to construct quantities of interest from quantitative content analyses
- ▶ Simple example: Proportion of content of a given type (e.g. anti-Lisbon treaty)
- ▶ Complex example: Left-right policy positions (e.g. CMP “Rile”)
- ▶ Are the metrics “natural”?
- ▶ Does the output metric resemble the input metric (if any)?
- ▶ What properties should the scale have, such as boundaries, type of increase, etc?
- ▶ How can uncertainty be characterized for the given scale?

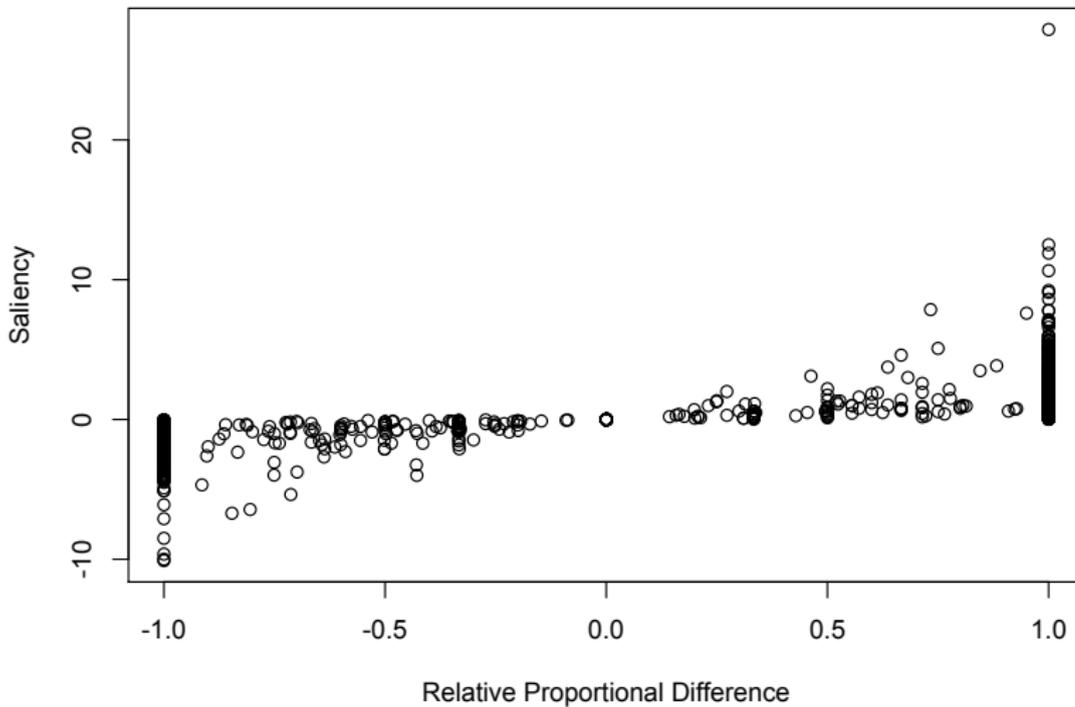
Logit scale for left-right

- ▶ The Comparative Manifesto Project scales policy positions as absolute proportional difference, measured by proportion of “Right” mentions less proportion of “Left” mentions: $\frac{(R-L)}{N}$
- ▶ Problems:
 - ▶ Addition of irrelevant content shifts the scale toward zero
 - ▶ Assumes the additional mentions increase emphasis in a linear scale
- ▶ The alternative is to scale $\frac{(R-L)}{(R+L)}$ (Kim and Fording 2002; Laver and Garry 2000), but this too has problems:
 - ▶ Still linear shift in position for increase in repetition
 - ▶ Quickly maxes out at the extremes
- ▶ Lowe, Benoit, Mikhaylov and Laver (2010) propose using a logistic odds-ratio scale $\log \frac{R}{L}$

Comparing scales:

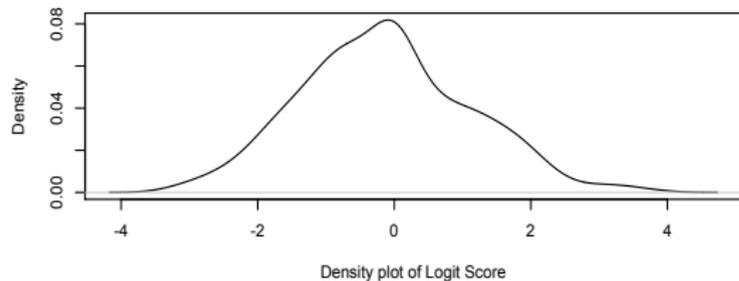
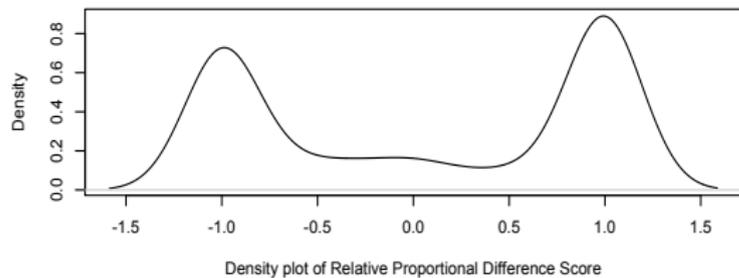
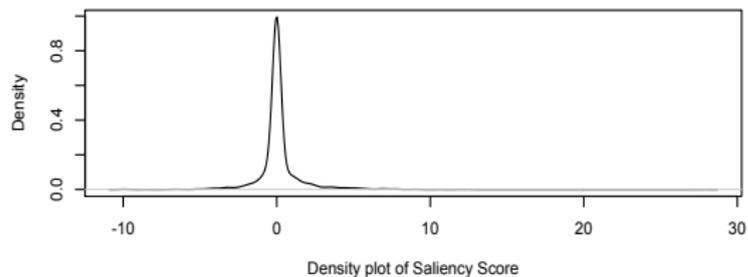
$$\hat{\theta}^{(S)} \text{ v. } \hat{\theta}^{(R)}$$

Protectionism



Comparing scales

Protectionism
distributions



And some results from natural v. quasi-sentences

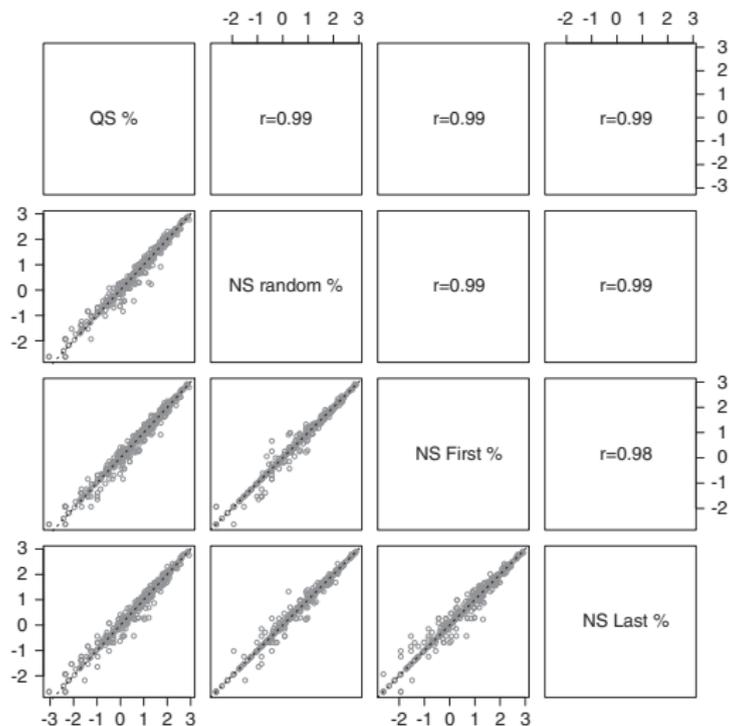


Fig. 2. Comparing quasi-sentence aggregate category percentages to natural sentence recodings

Notes: Three rules are compared: randomly assign the code based on constituent QSs; take the first QS code for the natural sentence; and take the last QS code for the natural sentences. Total manifestos analysed: 15.