Day 2: Selecting Texts and Features

Kenneth Benoit

Spring 2014

January 27, 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

BUILDING BLOCKS

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

word frequency refers to the number of times that words occur in a text or in a *corpus* of texts

concordance a(n alphabetical) list of the principal words used in a text, with their immediate contexts

lemmas the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached.

“key” words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

VALIDITY OF FEATURE FREQUENCY APPROACHES

Word frequency as an indicator of substantive content

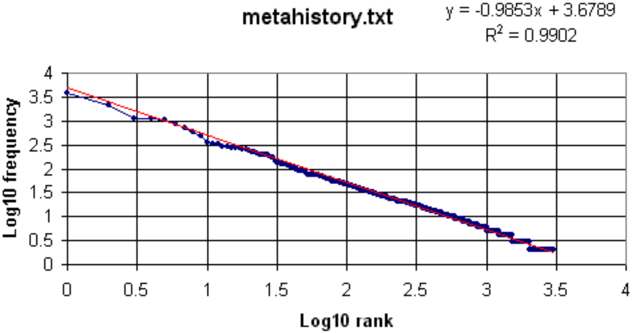
- Individual word usage tends to be associated with a particular degree of affect, position, etc. without regard to context of word usage
- Atomic words have been found to be far more informative than *n*-grams in this regard (Benoit and Laver 2003, Midwest paper)
- Some approaches focus on occurrence of a word as a binary variable, irrespective of frequency: a binary outcome (e.g. Hopkins and King 2008)
- Other approaches use frequencies: Poisson, multinomial, and related distributions (e.g. Laver, Benoit and Garry 2003)

Word frequency: Zipf's Law

- Zipf's law: Given some corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.
- The simplest case of Zipf's law is a "1/f function". Given a set of Zipfian distributed frequencies, sorted from most common to least common, the second most common frequency will occur 1/2 as often as the first. The third most common frequency will occur 1/3 as often as the first. The nth most common frequency will occur $1/n$ as often as the first.
- In the English language, the probability of encountering the the most common word is given roughly by $P(r) = 0.1/r$ for up to 1000 or so
- The assumption is that words and phrases mentioned most often are those reflecting important concerns in every communication

Word frequency: Zipf's Law

- Formulaically: if a word occurs f times and has a rank r in a list of frequencies, then for all words $f = \frac{a}{r^b}$ where a and b are constants and b is close to 1
- So if we log both sides, $\log(f) = \log(a) - b\log(r)$
- If we plot $\log(f)$ against $\log(r)$ then we should see a straight line with a slope of approximately -1.



Concordances

- Lists of most frequently appearing words in a text or corpus
- Often these filter out stop words (recall the word cloud algorithms from Session 1)
- Rationale behind filtering out words based on frequency
  - Substantive: Non-discriminating words (articles, conjunctions, pronouns, etc.) are non-informative
  - Practical: Non-discriminating words may strain computational abilities of particular statistical or computational techniques, esp. those requiring word frequency matrix analysis
  - Substantive: Low-frequency words may simply not be worth bothering about

Word concordances on popular web sites

- Amazon word statistics example `http://www.amazon.com/Innovative-Comparative-Methods-Policy-Analysis/dp/0387288287/ref=sr_1_1?ie=UTF8&s=books&qid=1249293340&sr=8-1`

- New York Times inaugural address example: `http://www.nytimes.com/interactive/2009/01/17/washington/20090117_ADDRESSES.html`

Word frequency examples

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

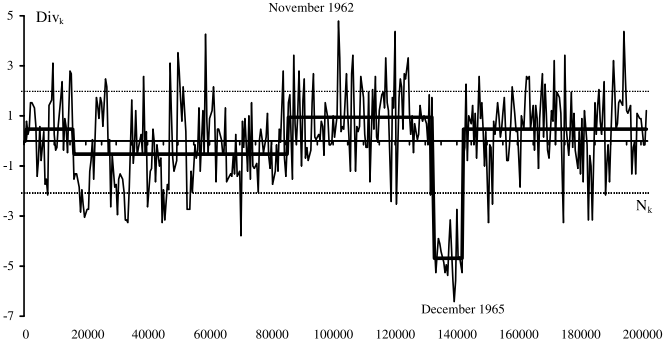
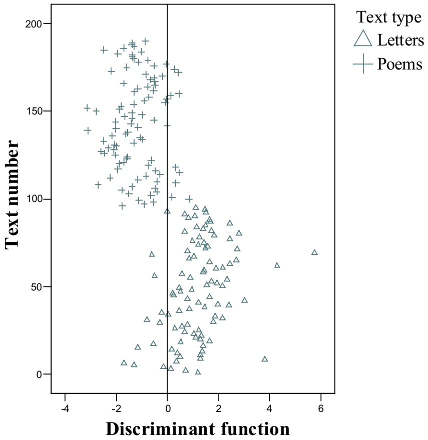


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

Examples continued

- Word *length* (defined as number of syllables) can be indicative of genre, if not necessarily authorship (Kelih et. al. 2004)



SELECTING TEXTS AND UNITS

Data types

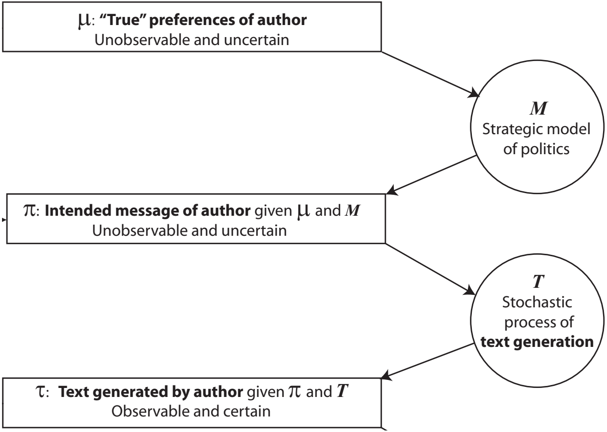
- Texts you've "created" yourself
  - Interview transcripts
  - Focus group transcripts
  - Open-ended survey questions
- "Natural" texts
  - speeches
  - documents
  - essays
  - literature
- Conversations

Strategies for selecting units of textual analysis

- Words
- *n*-word sequences
- pages
- paragraphs
- Themes
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

Sample v. “population”

- Basic Idea: Observed text is a stochastic realization
- Systematic features shape most of observed verbal content
- Non-systematic, random features also shape verbal content


μ: **“True” preferences of author**
Unobservable and uncertain
M
Strategic model of politics
π: **Intended message of author** given μ and *M*
Unobservable and uncertain
T
Stochastic process of **text generation**
τ: **Text generated by author** given π and *T*
Observable and certain

Sampling strategies for selecting texts

- Difference between a sample and a population
- May not be feasible to perform any sampling
- May not be necessary to perform any sampling
- Be wary of sampling that is a feature of the social system: “social bookkeeping”
- Different types of sampling vary from random to purposive
  - random sampling
  - non-random sampling
- Key is to make sure that what is being analyzed is a valid representation of the phenomenon as a whole – a question of research design

Random versus “Constructed” Sampling

- Based on a study by Riffe, Aust and Lacy (1993), who compared sampling from newspaper articles randomly versus “constructed”
- Either randomly sample 7 consecutive days, or between 2–4 consecutive weeks, and compare to “known” quantities
- Study showed that constructed sampling is much more efficient
- Why? Because cyclic variation in newspaper content occurs according to the day of the week – not every day contains equal proportions of different content

SELECTING FEATURES

Strategies for feature selection

- document frequency How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- purposive selection Use of a dictionary of words or phrases
- deliberate disregard Use of “stop words”: words excluded because they represent linguistic connectors of no substantive content

Common English stop words

`a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your`

- But no list should be considered universal

A more comprehensive list of stop words

as, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, aint, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, arent, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, cmon, cs, came, can, cant, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldnt, course, currently, definitely, described, despite, did, didnt, different, do, does, doesnt, doing, dont, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadnt, happens, hardly, has, hasnt, have, havent, having, he, hes, hello, help, hence, her, here, heres, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, id, ill, im, ive, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isnt, it, itd, itll, its, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, lets, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldnt, since, six, so, some, somebody,

Strategies for feature *weighting*: tf-idf

- $tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
  where $n_{i,j}$ is number of occurences of term t_i in document d_j, k is total number of terms in document d_j

- $idf_i = \ln \frac{|D|}{|\{d_j : t_i \in d_j\}|}$
  where
  - $|D|$ is the total number of documents in the set
  - $|\{d_j : t_i \in d_j\}|$ is the number of documents where the term t_i appears (i.e. $n_{i,j} \neq 0$)

- $tf\text{-}idf_i = tf_{i,j} \cdot idf_i$

Computation of tf-idf: Example

Example: We have 100 political party manifestos, each with 1000 words. The first document contains 16 instances of the word "environment"; 40 of the manifestos contain the word "environment".

- The *term frequency* is $16/1000 = 0.016$
- The *document frequency* is $100/40 = 2.5$, or $\ln(2.5) = 0.916$
- The *tf-idf* will then be $0.016 * 0.916 = 0.0147$
- If the word had only appeared in 15 of the 100 manifestos, then the *tf-idf* would be 0.0304 (three times higher).
- A high weight in tf-idf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; hence the weights hence tend to filter out common terms

Stemming words

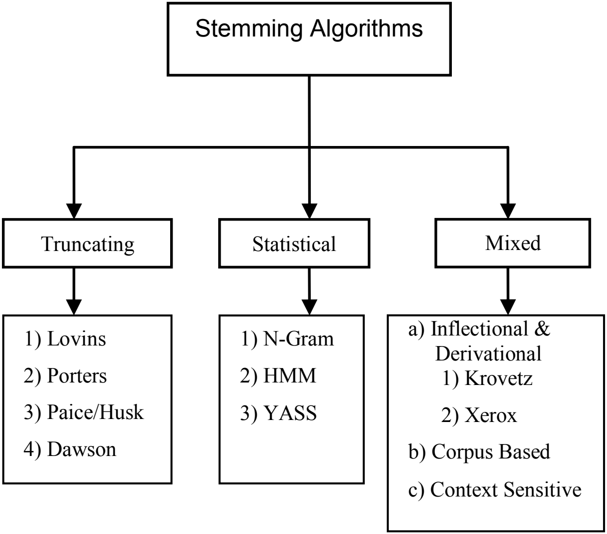
Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

example: `produc` from `production, producer, produce, produces, produced`

Varieties of stemming algorithms



Stemming Algorithms

Truncating

Statistical

Mixed

1) Lovins

2) Porters

3) Paice/Husk

4) Dawson

1) N-Gram

2) HMM

3) YASS

a) Inflectional & Derivational

1) Krovetz

2) Xerox

b) Corpus Based

c) Context Sensitive

Issues with stemming approaches

- The most common is proably the Porter stemmer
- But this set of rules gets many stems wrong, e.g.
  - `policy` and `police` considered (wrongly) equivalent
  - `general` becomes `gener`, `iteration` becomes `iter`
- Other corpus-based, statistical, and mixed appraoches designed to overcome these limitations (good review in Jirvani article)
- Key for you is to be careful through inspection of morphological variants and their stemmed versions

Selecting more than words: collocations

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

how to detect: pairs occuring 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.

RELIABILITY IN TEXT ANALYSIS

Tradeoff: Reliability *contra* 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 perfomance or one or more procedures is compared to a procedure that is taken to be correct. Deviations from a ("gold") standard are then recorded. (Establshes *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 occured by chance*
  - Cohen's κ is most common
  - Krippendorf'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 *dis*agreement: 4
- Expected *dis*agreement (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...