

Quantitative Linguistics

- Chi square as a statistical test
- Chi square as a similarity metric
- Mutual information
- Association ratio

Information theory

- The *entropy* of a distribution indicates how even it is:

$$H(p) = - \sum_x p(x) \log p(x)$$

- For example, lottery numbers have a very high entropy, lottery outcomes (for a single player) have a very low entropy
- Mutual information measures the reduction in entropy
- Entropy itself can be used to investigate the properties of a corpus

Information theory

- Kondratov (1969) approached Russian using entropy
- Unique code for each metrical word type: 3^2 is a three syllable word with stress on the second syllable
- Compare the entropy per word and per syllable for five different genres of Russian texts
- Entropy per word is highest for scientific texts; entropy per word and per syllable are lowest for poetry
- Conclusion: Constraints on metrical structure are stronger for poetry than other genres, and are the weakest for scientific texts

Text types

- We have seen comparisons of presumed text types using individual features (*while/whilst, the/her, meter*)
- This assumes that 'text types' or 'genres' discrete and identifiable
- D. Biber and colleagues have taken a very different approach: start with properties, and identify text types from them

Text types

- Text type analysis uses 67 linguistic features:
 - ★ first-person pronouns
 - ★ time adverbials (*early, instantly, soon...*)
 - ★ existential *there*
 - ★ subject *wh*-relatives
 - ★ emphatics (*a lot, for sure, really...*)
 - ★ private verbs (*assume, believe, doubt, know*)
 - ★ clausal coordination
- Factor analysis identified five dimensions of variation
 - ★ information vs. involved production
 - ★ narrative vs. non-narrative
 - ★ explicit vs. situation-dependent reference
 - ★ overt expression of persuasion
 - ★ abstract vs. non-abstract style

Text types

abstract	conjuncts	0.48
	agentless passives	0.43
	past participial clauses	0.42
	<i>by</i> passives	0.41
	past participial WHIZ deletions	0.40
	other adverbial subordinators	0.39
	predicative adjectives	0.31
non-abstract	type/token ratio	-0.31

Text types

- Chi square, mutual information identify properties of particular given text types
- Factor analysis “discovers” text types from a set of given properties
- Dimensions of variation can be given a linguistic interpretation by examining texts and properties
- Genres are located in a continuously variable, multi-dimensional space
- Provides a formal, quantitative foundation for studies of text types

Web as corpus

- Texts available via the internet can be seen as an enormous though chaotic corpus
- Search via search engines ([google](#)) or web concordancers ([WebCorp](#))
- How big is the web?
 - ★ 162,128,493 hosts (Jun. 2002, [ISC](#))
 - ★ Google indexes 3,083,324,652 web pages (Dec. 2002)
- If we guess there are about 1,000 words per page, then that means the WWW is a corpus of 10^{12} words!
- Compare to the Brown corpus (10^6 words), BNC (10^8 words), North American News Text corpus (10^9 words)

Web as corpus

- Virtually anything can be found somewhere on the web
- Nonetheless, clear trends often can be seen:

<i>separate</i>	15,800,000	95.27%
<i>seperate</i>	775,000	4.67%
<i>seprate</i>	7,910	0.01%
<i>seporate</i>	660	0.00%
<i>sepirate</i>	410	0.00%
<i>sepurate</i>	18	0.00%
<i>sepwrate</i>	1	0.00%
<i>sepyrate</i>	0	0.00%

Web as corpus

	Longman		Brown		LOB		London-Lund	
<i>different from</i>	1,193	91.6%	40	76.9%	38	86.4%	31	83.8%
<i>different than</i>	34	2.6%	12	23.1%	2	4.5%	3	8.1%
<i>different to</i>	75	5.8%	0	0%	4	9.1%	3	8.1%

	Google		AltaVista	
<i>different from</i>	3,780,000	64.9%	3,263,226	72.1%
<i>different than</i>	1,460,000	25.1%	954,319	21.1%
<i>different to</i>	585,000	10.0%	309,215	6.8%

Multi-lingual web

- Most material on the web is in American English, but substantial amounts of text can be found in nearly any written language
- Estimating the number of words of a particular language available on the web
 - ★ using a monolingual corpus, calculate the relative frequency of a number of language 'clue words' (*that, przez, ikke, daß, não*)
 - ★ count occurrences of clue words returned by a search engine
 - ★ take trimmed mean of relative frequency \times count
- This method yields less than 10% error on a large corpus of known composition

Multi-lingual web

clue word	rel freq	prediction
oder	0.00561180	2,417,488,684
sind	0.00477555	2,501,132,644
auch	0.00581108	2,668,062,907
wird	0.00400690	2,816,750,605
nicht	0.00646585	2,829,353,294
eine	0.00691066	2,856,389,983
sich	0.00604594	2,902,363,900
ist	0.00886430	2,981,546,991
auf	0.00744444	3,338,438,082
und	0.02892370	3,500,617,348
average		3,068,760,356

Multi-lingual web

	Oct 1996	Aug 1999	Feb 2000
English	6,082,090,000	28,222,100,000	48,064,100,000
German	228,938,428	1,994,229,409	3,333,127,671
French	223,316,023	1,529,795,169	2,732,221,327
Spanish	104,319,158	1,125,646,460	1,894,966,981
Italian	123,555,682	817,270,444	1,338,351,674
Portuguese	106,167,245	589,391,943	1,161,898,076
Norwegian	106,497,066	669,331,120	947,486,593
Finnish	20,647,404	107,260,274	166,599,467

Language identification

- Identifying what language a text is in has obvious applications for both information retrieval and linguistic research.
- A simple but very effective method uses n -gram counts and takes advantage of Zipf's Law.
- To construct a language profile, break a small training text ($\sim 15,000$ words) into tokens, break tokens into n -grams, and take the top 300 or so most frequent n -grams.
- The most frequent will be unigrams, reflecting the relative frequency of the letters of the alphabet.
- The next most frequent n -grams will reflect common letter combinations (*th*, *sj*) and morphemes (*the*, *op*).

Language identification

- To identify the language of a text, construct a profile of the text and compare it to each of the language profiles.

rank	language	document	distance
1	th	th	0
2	er	ing	3
3	on	on	0
4	le	er	2
5	ing	and	1
6	and	ed	no match
	

- Pick the language whose profile matches the profile of the document most closely.
- Works well for texts longer than a couple of sentences, but can be lead astray for very short texts (e.g., loanwords)

Text type identification

- We may want to collect a *sub-corpus* from the web of texts in a particular language or genre
- Biber's method requires heavily annotated texts, and is difficult to apply automatically
- Statistical methods use mostly lexical, character, and derivative cues
 - ★ Structural: passives, nominalizations
 - ★ Lexical: titles, suffixes, adverbs
 - ★ Characters: punctuation
 - ★ Derivative: type/token ratio, avg word length, avg sentence length

Text type identification

- Kessler, Nunburg, and Schütze: Brown corpus
 - ★ Brow: popular, middle, upper-middle, high
 - ★ Narrative: yes, no
 - ★ Genre: reportage, editorial, scitech, legal, nonfiction, fiction
- Using 55 non-structural properties, they could identify the Brow, Narrative, and Genre facets of a text with better than 85% accuracy