

Information Organisation: Indexing, Retrieval & Text Analysis

Jagdish S. Gangolly,
Department of Informatics,
SUNY Albany

October 11, 2009

I have often pondered over the roles of knowledge or experience, on the one hand, and imagination or intuition, on the other, in the process of discovery. I believe that there is a certain fundamental conflict between the two, and knowledge, by advocating caution, tends to inhibit the flight of imagination. Therefore, a certain naivete, unburdened by conventional wisdom, can sometimes be a positive asset.

R. Langlands, in **Harish-Chandra**
Biographical Memoirs of Fellows of the Royal Society
31 (1985) 197 - 225.

Objectives

- ▶ Development of lexical resources such as dictionaries and thesauri
- ▶ Exploration of textual data to support analysis of its properties
- ▶ Indexing of text to retrieve information to answer queries
- ▶ Architecture of Search Engines

Dictionaries

Dictionaries can be built based on either of the following:

- ▶ historical principles (as in case of older versions of the Oxford English dictionaries: [http:// www.oed.com/](http://www.oed.com/))
 - ▶ usually prescriptive
 - ▶ all forms of usage may not be covered
 - ▶ may have biases based on the socio-economic conditions of the compilers
- ▶ based on the analysis of the text in a corpus (as in the case of the Collins Cobuild Dictionaries:
<http://www.collins.co.uk/books.aspx?group=140>)
 - ▶ usually descriptive
 - ▶ biases can be reduced by choosing a corpus that is representative of the usage

Dictionaries are usually based on the written word based, but they do not have to be.

Thesauri

Thesauri contain relationships between words

- ▶ For nouns and adjectives:
Antonyms & synonyms
- ▶ For verbs:
Troponyms: walk – amble, shuffle, file, race, roam, meander, stroll, strut, stump, tread, trudge, foot (informal), leg (informal)

Some good sites: www.dictionary.com, www.visualthesaurus.com

Thesauri

Other Lexical relations include:

▶ Nouns

- ▶ **hypernyms:** Y is a hypernym of X if every X is a (kind of) Y (canine is a hypernym of dog)
- ▶ **hyponyms:** Y is a hyponym of X if every Y is a (kind of) X (dog is a hyponym of canine)
- ▶ **coordinate terms:** Y is a coordinate term of X if X and Y share a hypernym (wolf is a coordinate term of dog, and dog is a coordinate term of wolf)
- ▶ **holonym:** Y is a holonym of X if X is a part of Y (building is a holonym of window)
- ▶ **meronym:** Y is a meronym of X if Y is a part of X (window is a meronym of building)

Thesauri

Other Lexical relations include:

▶ Verbs

- ▶ **hypernym:** the verb Y is a hypernym of the verb X if the activity X is a (kind of) Y (to perceive is an hypernym of to listen)
- ▶ **troponym:** the verb Y is a troponym of the verb X if the activity Y is doing X in some manner (to lisp is a troponym of to talk)
- ▶ **entailment:** the verb Y is entailed by X if by doing X you must be doing Y (to sleep is entailed by to snore)
- ▶ **coordinate terms:** those verbs sharing a common hypernym (to lisp and to yell)

Thesauri

Other Lexical relations include:

▶ Adjectives

- ▶ **related nouns:**
- ▶ **similar to:** Hard – firm, solid; adamantine; woody; stony, granitic, granitelike, rocklike; unyielding; stonelike, petrous; hardened, case-hardened, hardboiled; steely; calculative, calculating, conniving, scheming, shrewd;
- ▶ **participle of verb:** *rolling* stone; *smiling* crocodile; *written* apology;

▶ Adverbs

- ▶ **root adjectives:** *extremely* interesting; *faster* walk;

Based on Wordnet in the Wikipedia
(<http://en.wikipedia.org/wiki/WordNet>)

You shall know a word by the company it keeps

*John Rupert Firth. 1957:11
Papers in Linguistics 1934-1951 (1957)
London: Oxford University Press.*

Exploration of Textual data

Why?:

- ▶ Voluminous data. Medline alone has well over 15 million records
- ▶ Traditional literature mining impossible
- ▶ Manual digital curation (preservation, maintenance) and provenance management very costly
- ▶ Information retrieval methods are often inadequate for aiding knowledge discovery

Exploration of textual data

- ▶ Construction of corpus
- ▶ Markup of documents in the corpus
- ▶ Text pre-processing
- ▶ Concordances
- ▶ Interpretation

Source: Developing Linguistic Corpora: a Guide to Good Practice
John Sinclair, Tuscan Word Centre
<http://ahds.ac.uk/guides/linguistic-corpora/index.htm>

Definition: A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Principles:

- ▶ The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.
- ▶ Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen.

Corpus Construction

- ▶ Only those components of corpora which have been designed to be independently contrastive should be contrasted.
(Normative, historical, monitor, varietal corpora, specialised corpora)
- ▶ Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
- ▶ Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
- ▶ Samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.

- ▶ The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
- ▶ The corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
- ▶ Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.

Markup of Documents in the corpus

- ▶ Part of speech tagging
- ▶ Sense tagging
- ▶ Semantic tagging
- ▶

Supervised and unsupervised algorithms are used for word-sense disambiguations for sense tagging. For details, see: *Sense Tagging: Semantic Tagging with a Lexicon* by Yorick Wilks and Mark Stevenson,
<http://www.aclweb.org/anthology-new/W/W97/W97-0208.pdf>

Text Preprocessing

- ▶ Tokenization
- ▶ Expansion of abbreviations and acronyms to their canonical orthographic representations
- ▶ Sentence boundary detection
- ▶ Morphological analysis/ Lemmatization

Concordances

An alphabetical arrangement of the principal words contained in a book, with citations of the passages in which they occur
Oxford English Dictionary

Headword	No.	Context...	Word	...Context	Reference
HEAR	15	That my own	heart	drifts and cries, having no...	Deep Analysis
HEARD	9	By the shout of the	heart	continually at work	And the wave
HEARING	7	Nothing to adapt the skill of the	heart	to, skill	And the wave
HEARS	3	The tread, the beat of it, it is my own	heart	,	Träumerei
HEARSE	1	Because I follow it to my own	heart		Many famous
HEART	25	My	heart	is ticking like the sun:	I am washed t
HEART'S	2	The vague	heart	sharpened to a candid co...	The March Pa:
HEART-SHAPED	1	Contract my	heart	by looking out of date.	Lines on a Yo
HEARTH	1	Having no	heart	to put aside the theft	Home is so Se
HEARTS	7	And the boy puking his	heart	out in the Gents	Essential Bea
HEARTY	1	A harbour for the	heart	against distress.	Bridge for the
HEAT	6	These I would choose my	heart	to lead	After-Dinner F
HEAT-HAZE	1	Time in his little cinema of the	heart		Time and Spax
HEATH	1	This petrified	heart	has taken,	A Stone Churr
HEATS	1	How should they sweep the girl clean...	heart	,	I see a girl dra
HEAVE	1	Hands that the	heart	can govern	Heaviest of fic
HEAVEN	4	For the	heart	to be loveless, and as col...	Dawn
HEAVEN-HOLDING	1	With the unguessed-at	heart	riding	One man walk
HEAVIER-THAN...	1	If hands could free you,	heart	,	If hands could
HEAVIEST	2	That overflows the	heart		Pour away the

Words: 7318 Tokens: 37070 At word: 2990 Deleted lines: 1 [24] Word sort: Asc alpha (string) Context sort: Asc occurrence order

Source: <http://www.concordancesoftware.co.uk/>

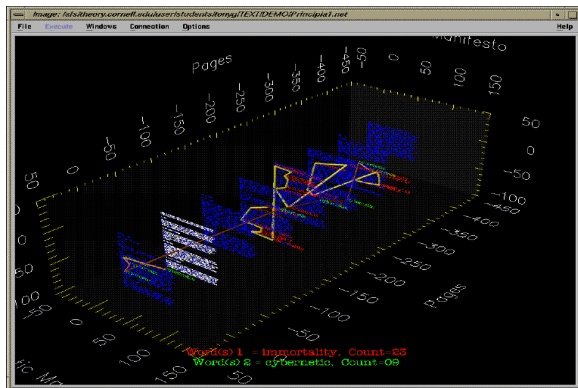
Text Visualisation



Source:

<http://www.nbb.cornell.edu/neurobio/land/OldStudentProjects/cs490-95to96/tonyg/Language.Viz1.html>

Text Visualisation



Source:

<http://www.nbb.cornell.edu/neurobio/land/OldStudentProjects/cs490-95to96/tonyg/Language.Viz1.html>

Analysis of Baden-Powell Farewell Addresses *Text and Corpus Analysis*, Michael Stubbs