

Computer-Assisted Content Analysis

Andrew Wilson

Linguistics and English Language
Lancaster University

eiaaw@exchange.lancs.ac.uk

Outline

- What content analysis is
 - Five main types of computer-aided content analysis
 - how they work
 - pros and cons
 - Two sets of examples showing the techniques in action and highlighting some of the issues raised
-
-

Content Analysis

‘The technique known as content analysis ... attempts to characterize the meanings in a given body of discourse in a systematic and quantitative fashion’.

(Kaplan 1943: 230)

Computer-Assisted Content Analysis

- Content analysis itself is old – 17th century (*Songs of Sion*), WW2 (propaganda)
 - Computer-aided variants date from the 1960s
 - Two main 1960s programs exemplifying two main traditions:
 - Dictionary-based:
Stone et al: The General Inquirer
 - Contiguity-based:
Iker and Harway: The WORDS system
-
-

Four (or five) main types of CACA

- Conceptual analysis (exhaustive dictionary of conceptual fields)
 - Theoretically-grounded dictionary-based analysis (selected concepts)
 - Contiguity-based multivariate analysis
 - 'Enhanced' methods – semantic grammars, parsing, etc.
 - ? "Keyword" analysis
-
-

Conceptual analysis

- Uses a dictionary to group words into higher-level categories ("concepts")
 - Guided only by theories of linguistic relatedness
 - Tries to cover most of the vocabulary (95%+)
 - Useful for IR tasks, if you want to retrieve everything that refers to a given concept
 - ? May add relatively little of value in a straightforward text comparison framework, as categories often weighted by one or two lexical items
 - Ambiguity and overgeneralization can be problematic
-
-

Theory-based dictionaries (1)

- Two main kinds of dictionaries:
 - category-based
 - Regressive Imagery Dictionary (RID)
 - primary and secondary process thought; emotions
 - Motive Dictionary
 - need for affiliation, achievement, and power
 - Dresdner Angstwörterbuch (DAW)
 - anxiety types
 - norm-based
 - Heise's word norms (evaluation, potency, activity)
 - Osgood's semantic differential
 - Whissell's dictionary of affect (pleasantness)

Theory-based dictionaries (2)

- Dictionary construction phases:
 - start with a theory
 - e.g. McClelland's theory of motivation
 - create relevant category system
 - populate categories
 - thesauri; WordNet; other CACA dictionaries
 - test for exceptions and highly ambiguous items
 - remove causes of error, unless these can be corrected in other ways
 - validate
 - other measures (experiments); exemplary texts
-
-

Contiguity analysis

- Thesis: The main themes of a text can be captured through regularities of co-occurrence of frequently used words
 - This is what many commercial programs do (very simple to implement)
 - A plethora of different statistical methods, but the outcomes are very comparable
 - Issues of category interpretation – just a bunch of words
-
-

'Enhanced' methods

- Resists a more explicit categorization
 - The most technologically advanced methods, making use of automated syntactic analysis, etc.
 - Kansas Event Data System (KEDS)
 - Subject-Object-Action triplets
 - PCAD2000 for psychiatric text analysis
 - But note how good Berth's DAW dictionary is, minus parsing!
-
-

Keywords

- Statistical tests of significant differences in word frequencies between 2 or more texts:
 - text of interest + reference corpus
 - two (or more) texts of interest
- Familiar measures: chi-square, log-likelihood, Mann-Whitney, Fisher ...
- Shows up the characteristic words of a text



Examples (1)

Easter and Christmas sermons by George Carey (Archbishop of Canterbury) and Pope John Paul II

9 texts from each preacher = 18 in total

Conceptual analysis

Tag	JP	JP %	GC	GC %	LL	Category name
S9	358	4.09	261	1.90 +	90.71	Religion and the supernatural
S4	88	1.01	33	0.24 +	57.02	Kin
Q2.1	23	0.26	124	0.90 -	37.98	Speech etc:- Communicative
A3+	175	2.00	464	3.37 -	37.15	Being
Z2	154	1.76	128	0.93 +	28.53	Geographical names
N1	96	1.10	66	0.48 +	27.42	Numbers
X2.2-	37	0.42	14	0.10 +	23.77	Knowledge
Z6	44	0.50	148	1.07 -	22.12	Negative
A13.6	0	0.00	21	0.15 -	20.67	Degree: Diminishers
T1.3	95	1.09	76	0.55 +	19.43	Time: Period
S2.2	51	0.58	30	0.22 +	19.14	People:- Male
A6.3+	10	0.11	0	0.00 +	18.90	Comparing:- Variety
E5-	10	0.11	57	0.41 -	18.53	Fear/bravery/shock
Z8	906	10.35	1695	12.31 -	18.03	Pronouns etc.
G2.1	40	0.46	21	0.15 +	17.73	Crime, law and order
X2.1	29	0.33	103	0.75 -	17.18	Thought, belief
T2+++	14	0.16	2	0.01 +	16.38	Time: Beginning and ending
A7+	51	0.58	149	1.08 -	15.93	Definite (+ modals)
A5.1-	0	0.00	16	0.12 -	15.75	Evaluation:- Good/bad

Keywords

	JP	JP %	GC	GC %	LL
the	908	10.37	867	6.30 +	109.86
night	43	0.49	1	0.01 +	72.72
mystery	37	0.42	0	0.00 +	69.94
Lord	28	0.32	0	0.00 +	52.93
holy	49	0.56	9	0.07 +	51.42
or	0	0.00	50	0.36 -	49.20
son	31	0.35	2	0.01 +	45.48
that	70	0.80	248	1.80 -	41.16
cf.	20	0.23	0	0.00 +	37.81
born	32	0.37	7	0.05 +	30.67
Christ	62	0.71	31	0.23 +	29.31
salvation	15	0.17	0	0.00 +	28.35
you	84	0.96	53	0.38 +	28.09
this	114	1.30	85	0.62 +	27.51
Bethlehem	18	0.21	1	0.01 +	27.17
vigil	14	0.16	0	0.00 +	26.46
it	57	0.65	185	1.34 -	25.60
apostles	17	0.19	1	0.01 +	25.39
but	19	0.22	92	0.67 -	24.83
veni	13	0.15	0	0.00 +	24.57
witness	22	0.25	4	0.03 +	23.20
Brothers	12	0.14	0	0.00 +	22.68

- Carey uses but and or more frequently than John Paul – suggests a more “argumentative” style, presenting options

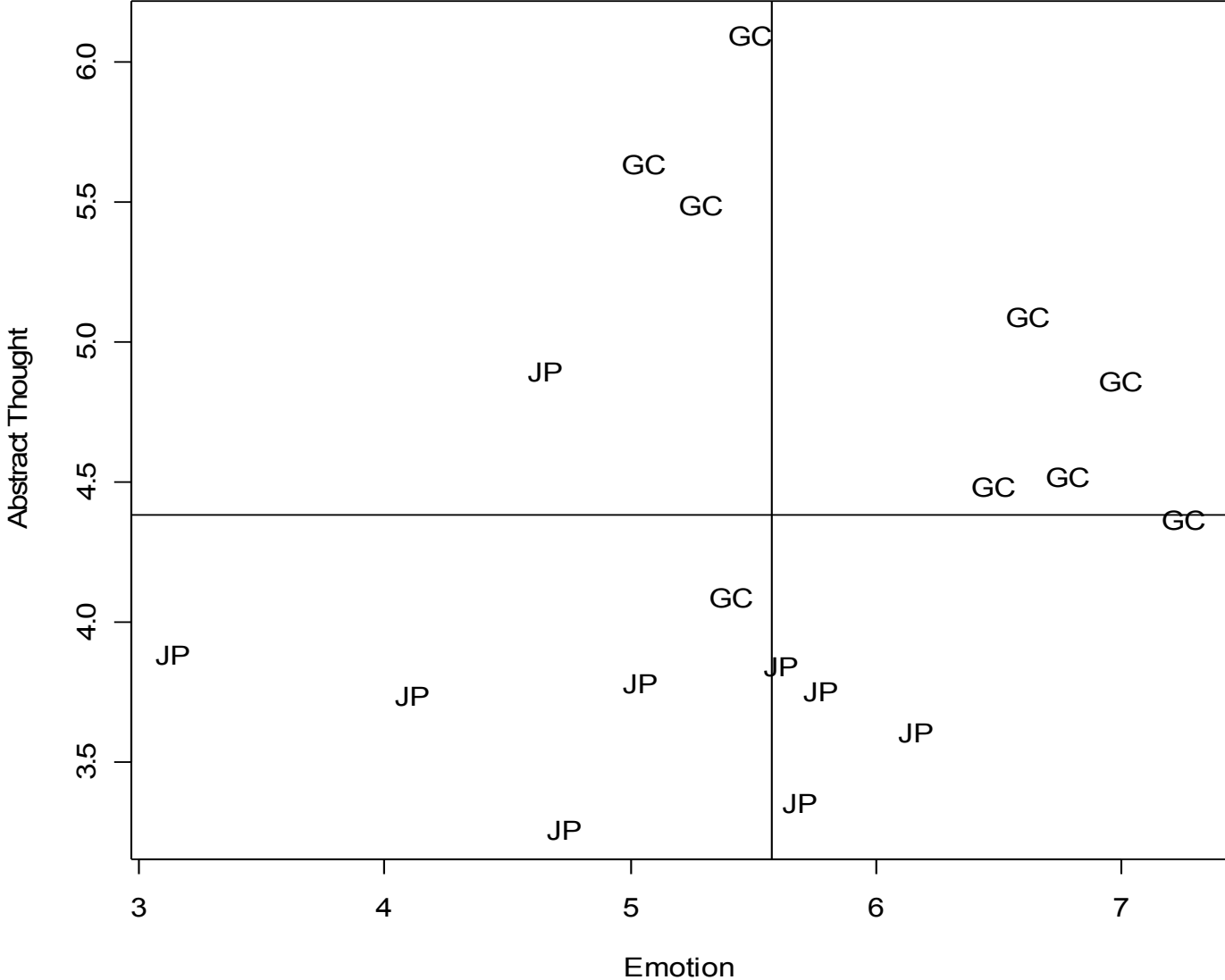


Theory-based (Thinking/Feeling)

- Jung – thinking/feeling personality dimension
- An influential dimension in various CACA studies
- Mergenthaler (1996) on psychotherapy sessions:
 - High emotion, high abstractness = connecting
 - High emotion, low abstractness = experiencing
 - Low emotion, high abstractness = reflecting
 - Low emotion, low abstractness = narrating
- A useful clue to individual styles of thought



GC = George Carey; JP = John Paul II



Examples (2)

Some data from the "Language of Shoes" project (cultural studies / non-verbal communication)

49 student riders from USA

22 student riders from UK

Short samples of free writing about riding boots (mean length = 221 words)

Conceptual analysis

Tag		US	US %	UK	UK %	LL	Category Name
T1.3+	12	0.11	30	0.74 -	35.50		Time: Period
O1.1	49	0.45	55	1.36 -	30.52		Substances and materials: Solid
N1	64	0.59	58	1.44 -	22.85		Numbers
A5.4+	12	0.11	17	0.42 -	12.56		Evaluation: Authenticity
A1.6	1	0.01	6	0.15 -	10.52		Physical/mental
N3.7+	79	0.73	12	0.30 +	10.43		Measurement: Length & height
N5.2+	10	0.09	14	0.35 -	10.21		Exceeding; waste
F4	39	0.36	4	0.10 +	8.56		Farming & Horticulture
N6+++	18	0.17	18	0.45 -	8.40		Frequency etc.
M6	127	1.17	27	0.67 +	7.94		Location and direction
A8	82	0.76	15	0.37 +	7.56		Seem
Z2	20	0.19	1	0.02 +	7.26		Geographical names
X3.4	31	0.29	3	0.07 +	7.20		Sensory: Sight
X4.1	6	0.06	9	0.22 -	7.06		Mental object: Conceptual object
S2.2	17	0.16	16	0.40 -	6.74		People: Male

Problem cases

paddock	F4	22	0.20
field	F4	13	0.12
paddocks	F4	2	0.02
cowboy	F4	1	0.01
crop	F4	1	0.01
leather	O1.1	30	0.28
rubber	O1.1	7	0.06
mud	O1.1	2	0.02
wax	O1.1	2	0.02
plastic	O1.1	2	0.02
nylon	O1.1	1	0.01
sand	O1.1	1	0.01
pvc	O1.1	1	0.01
cloth	O1.1	1	0.01
leather_thing	O1.1	1	0.01
chaps	S2.2	7	0.06
men	S2.2	2	0.02
man	S2.2	2	0.02
guy	S2.2	2	0.02
boys	S2.2	2	0.02
male	S2.2	1	0.01
guys	S2.2	1	0.01

- The categories for Farming and Male Persons are mainly referring to kinds of footwear – *paddock boots, field boots, and chaps*
 - The category for Substances and Materials (Solid) combines several kinds – the materials from which boots are made (*leather, rubber, PVC*); those in which they are worn (*sand, mud*); and a few others
-
-

Keywords

	US	US %	UK	UK %		LL
long	6	0.06	26	0.64	-	40.64
chaps	7	0.06	15	0.37	-	15.99
smart	0	0.00	6	0.15	-	15.63
leather	30	0.28	31	0.77	-	15.23
rubber	7	0.06	14	0.35	-	14.17
comfy	0	0.00	5	0.12	-	13.02
have	84	0.78	58	1.44	-	12.31
jodhpur	3	0.03	9	0.22	-	11.85
tall	62	0.57	8	0.20	+	10.43
trainers	0	0.00	4	0.10	-	10.42
warm	0	0.00	4	0.10	-	10.42
younger	0	0.00	4	0.10	-	10.42
wear	54	0.50	40	0.99	-	10.23
how	25	0.23	1	0.02	+	9.99
her	6	0.06	10	0.25	-	8.68
get	53	0.49	7	0.17	+	8.64
always	18	0.17	18	0.45	-	8.40
field	13	0.12	0	0.00	+	8.25
polish	13	0.12	0	0.00	+	8.25
practical	1	0.01	5	0.12	-	8.25

Contiguity analysis (factor analysis)

WORD	FAC1	FAC2	FAC3	FAC4	FAC5
6	-0.069	0.089	-0.193	0.118	0.189
ABLE	-0.233	-0.172	0.417	0.034	0.058
ACTUAL	-0.057	-0.209	0.028	-0.167	0.270
ALONG	-0.133	0.054	0.022	-0.123	-0.198
ANKLE	0.455	-0.040	0.003	0.004	-0.122
AREA	0.493	0.000	-0.003	0.122	0.066
AREN	0.061	-0.107	-0.190	-0.287	0.286
ARIAT	-0.046	-0.080	-0.414	-0.100	0.409
ASK	-0.109	-0.034	0.064	-0.010	0.093
BACK	0.102	-0.022	-0.080	-0.093	-0.052
BAD	-0.014	-0.122	-0.210	0.441	-0.051
BARN	0.028	0.009	0.004	-0.300	-0.202
BEFORE	-0.004	-0.055	-0.261	-0.310	-0.120
BEGIN	-0.084	-0.144	-0.164	0.032	-0.205
BIG	0.536	0.022	0.027	-0.283	0.110
BLACK	-0.022	0.010	-0.229	0.567	0.231
BOOT	-0.080	-0.296	0.082	0.131	-0.077
BRAND	-0.129	-0.069	-0.266	0.143	0.178

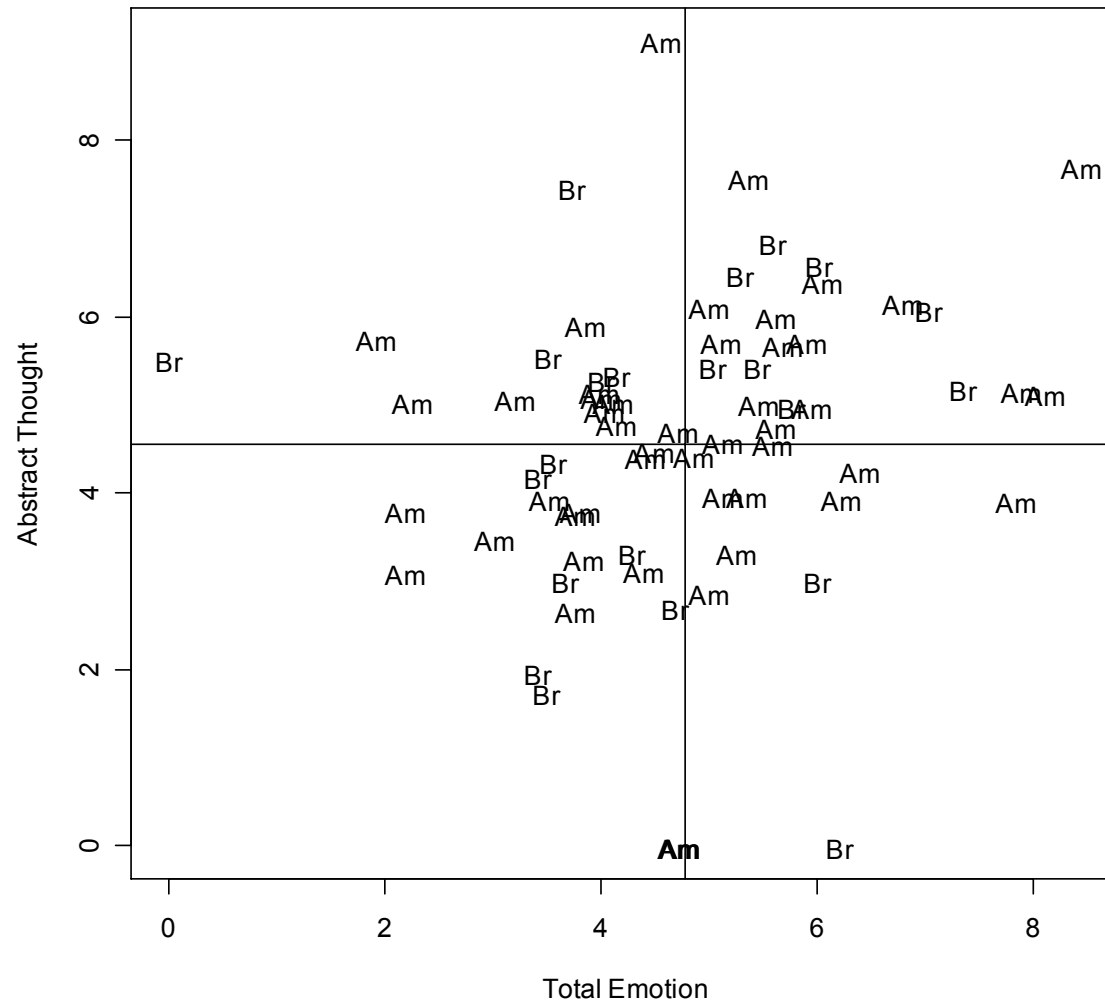
Interpretation

- Just looks like a list of words with numbers attached
 - As a rule of thumb, for each factor we take those with the highest loadings (e.g. ≥ 0.5)
 - Still a mixed bag of subjects, objects, verbs, modifiers – we need to know the texts in order to make sense of the word groups...
 - Burghard Rieger talks in terms of “connotative clouds” rather than “semantic fields” in such cases
-
-

Factor interpretations

- Factor 1 – Problems of fit
 - wide, low, thin, large, top, leg, ease, problem, calf, fit, big, woman, knee
 - Factor 2 – Price
 - note, money, shoe, spend, certain, stylish, say, good, saddle, English, expense, now
 - Factor 3 – Getting stuck
 - help, difficult, take
 - Factor 4 – Styles and colours
 - dress, field, brown, standard, black
-
-

Theory-based (Thinking/Feeling)



- The interpretation here is not as clear-cut as with the sermons example, but we can still use the quadrants to retrieve samples of language with the given characteristics
- Could be useful for market segmentation and advertising – similar approaches have been used in the past

