

Tags, Trash & Trigrams: a Tale from the Text Mines

(Linguistic Computing Methods for Analysing
Digital Records of Learning)

Richard Forsyth, Shaaron
Ainsworth, David Clarke,
Claire O'Malley & Pat
Brundell

School of Psychology 0115-951 5281
rsf@psychology.nottingham.ac.uk

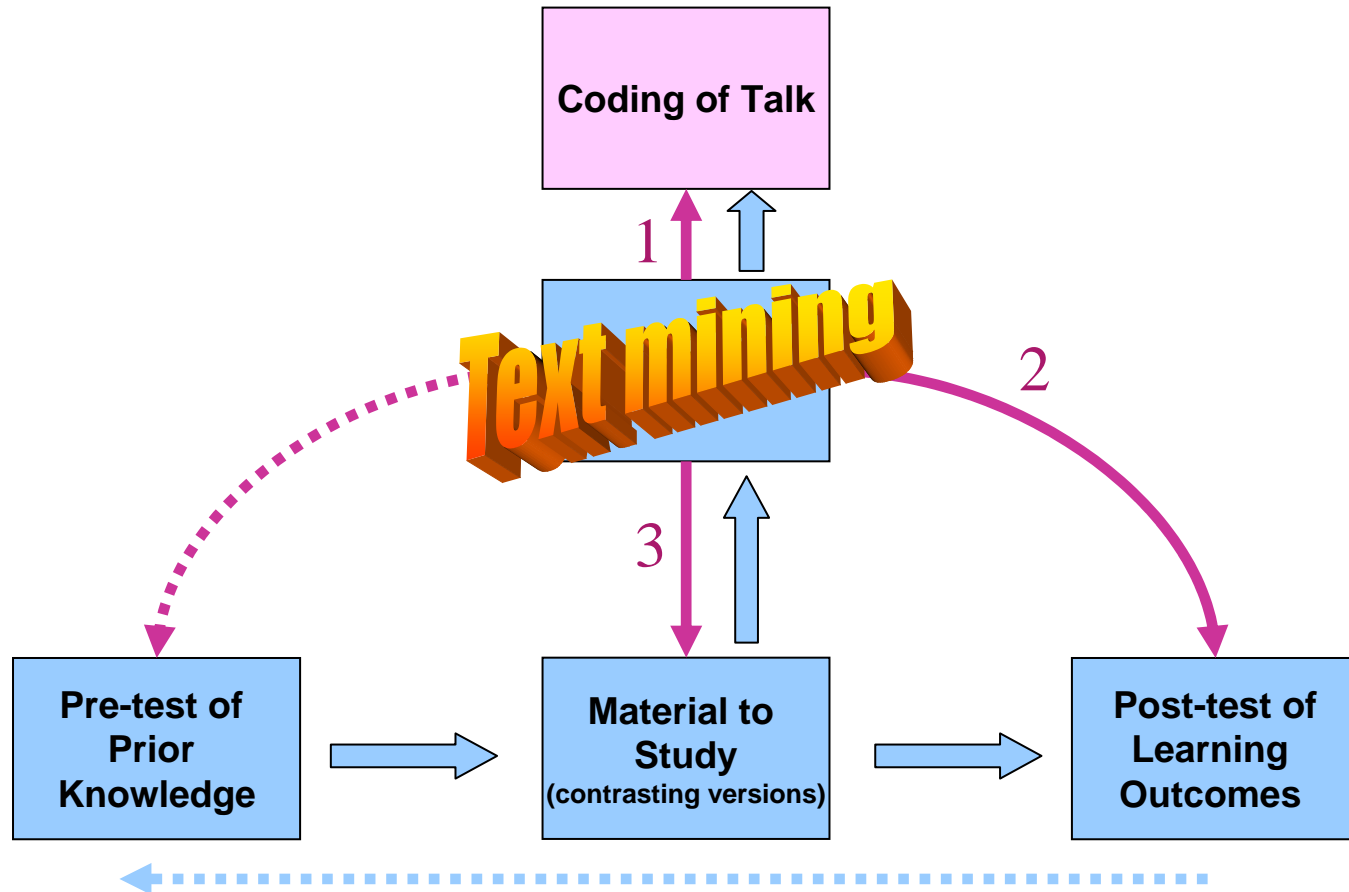
Themes

- Relating linguistic to extra-linguistic information
- Advances in Learning Sciences
- Exploiting digital records (legacy data) of verbalizations
- e-Social Science
 - text-mining as a means of quantifying qualitative information
- Computational coding
 - social scientists spend huge amounts of time hand-coding verbal data

Talking and Learning

- Talking to yourself can help:
e.g. Self explaining
- Talking with others can help:
e.g., Collaborative learning, Argumentation-based learning, Socratic dialogue
- But not all talk is good....So learning scientists need to analyse what people say when they talk..
 - This is both time consuming and can be unreliable.
 - Hence our interest in whether text mining can help a) speed things up, b) increase reliability and c) help us deal with the increasing numbers of digital records of learning

The experimental situation: outline



Questions Investigated

- Q1 : How well can human-assigned categories be predicted (semi-)automatically from linguistic data?
 - potential for
 - time-saving; larger data sets; new applications; higher reliability
- Q2 : How well can learning outcomes be predicted from learners' verbalizations?
 - potential for
 - time-saving; semi-automated assessment; new applications; insight into cognitive/social processes
- Q3 : Can the material students are learning from be identified from their language?
 - relevant to
 - observational (non-laboratory) studies

Interim Answers

- Q1 : How well can human-assigned categories be predicted (semi-)automatically from linguistic data?
 - fairly well
 - 62-69% correct assignments, recently 74% (chance level 33%)
- Q2 : How well can learning outcomes be predicted from learners' verbalizations?
 - well
 - 69% variance accounted for
- Q3 : Can the material students are learning from be identified from their language?
 - yes
 - 96% correct decisions

Dataset Details (N.B. legacy data) :

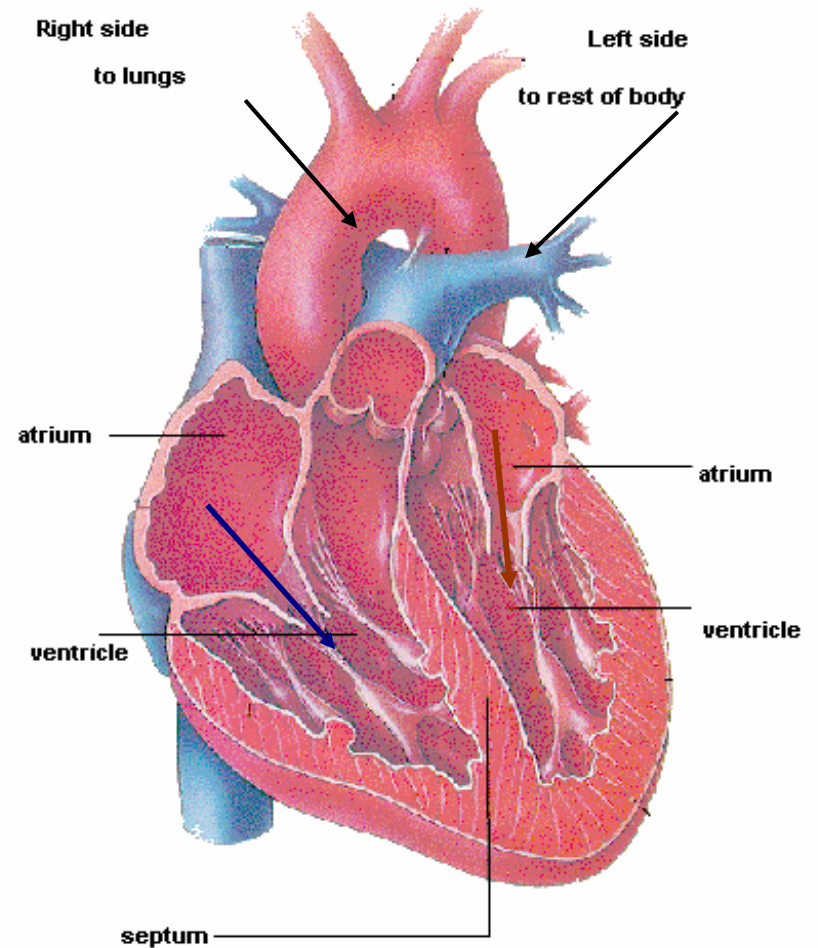
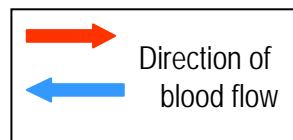
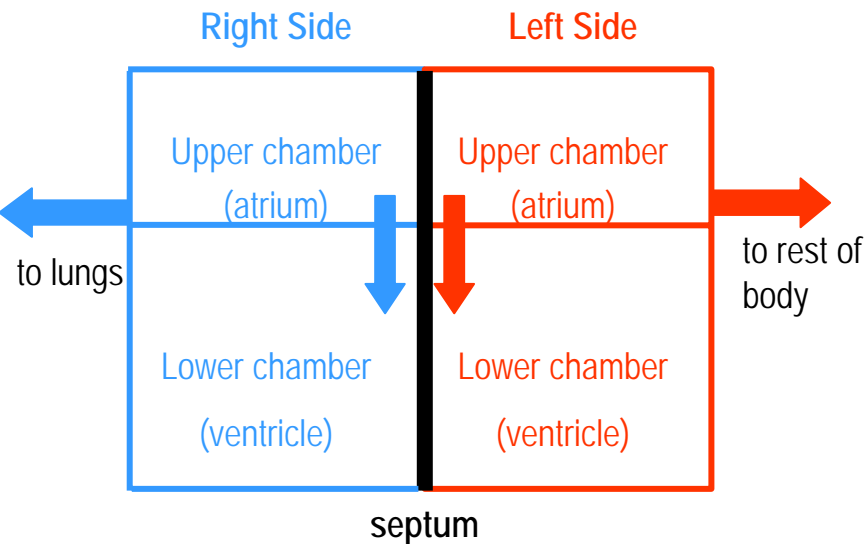
Dataset:	AB (Ainsworth & Burcham, 2004)	AR (Robertson, 2004)
Participants:	24 (13 female, 11 male)	24 (13 female, 11 male)
Words:	44,388	23,330
Segments:	2071	1784
Material:	text	diagrams
Conditions:	high versus low coherence	abstract versus realistic
Assessment:	3 pre-test & 5 post-test measures	3 pre-test & 5 post-test measures

- Both learning tasks on same topic = cardiovascular system
- (Plenty of data “cleansing” & reorganizing done.)

Data Details : Samples of High & Low Coherence Text

Original	Minimal	Maximal
the oxygenated blood is then pumped through the bicuspid valve into the left ventricle	it is pumped through the bicuspid valve into the left ventricle	this oxygenated blood is then pumped through the bicuspid valve (the a-v valve on the left side of the heart) from the atria to the left ventricle

Data Details : Abstract and Concrete



Support Software

- several data-gluing programs (mainly in Python)
 - also starting to develop machine-learning s/w in Python
- some statistical functions (in R)
- Main external system: WMatrix
 - syntactic tagging, e.g.
 - AT article; Il preposition; JJ adjective; VV lexical verb
 - semantic tagging, e.g.
 - A5 evaluation; B2 health/disease; O1 substances; Q2 speech acts

Specimen of WMatrix tagging output (AR27; 35, 36)

Syntactic Code	Word	Semantic Codes (in decreasing order of predicted likelihood)
■ CC	And	Z5
■ PPH1	it	Z8
■ VBZ	's	Z5 A3+
■ VVG	showing	A10+ S1.1.1
■ RR	obviously	Z4 A11.2+
■ AT	the	Z5
■ JJ	main	A11.1+ N5+++
■ NN1	pump	O2 B5
■ VBZ	is	A3+ Z5
■ AT	the	Z5
■ NN1	heart	B1 M6 A11.1+ E1 X5.2+
■	,	
■ PPH1	it	Z8
■ VBZ	's	A3+ Z5
■ RG	very	A13.3
■ JJ	big	N3.2+ N5+ A11.1+ S1.2.5+ X5.2+
■ CC	and	Z5
■ DD1	that	Z8 Z5
■ VM	will	T1.1.3
■ VVI	pump	M2 A1.1.1 A9- Q2.2 S8+ B3

Q1 : Utterance Classification

Textual Material

“The septum divides the heart lengthwise into two sides”

3 codes:

- Paraphrase,
 - The septum is what goes down the middle of the heart
 - Self-explanation,
 - Septum is what separates the two ... some sort of control
 - Monitoring-statement,
 - I'm not sure why
- 3 variable types:
- Absolute Counts / Proportional Rates / Transitions

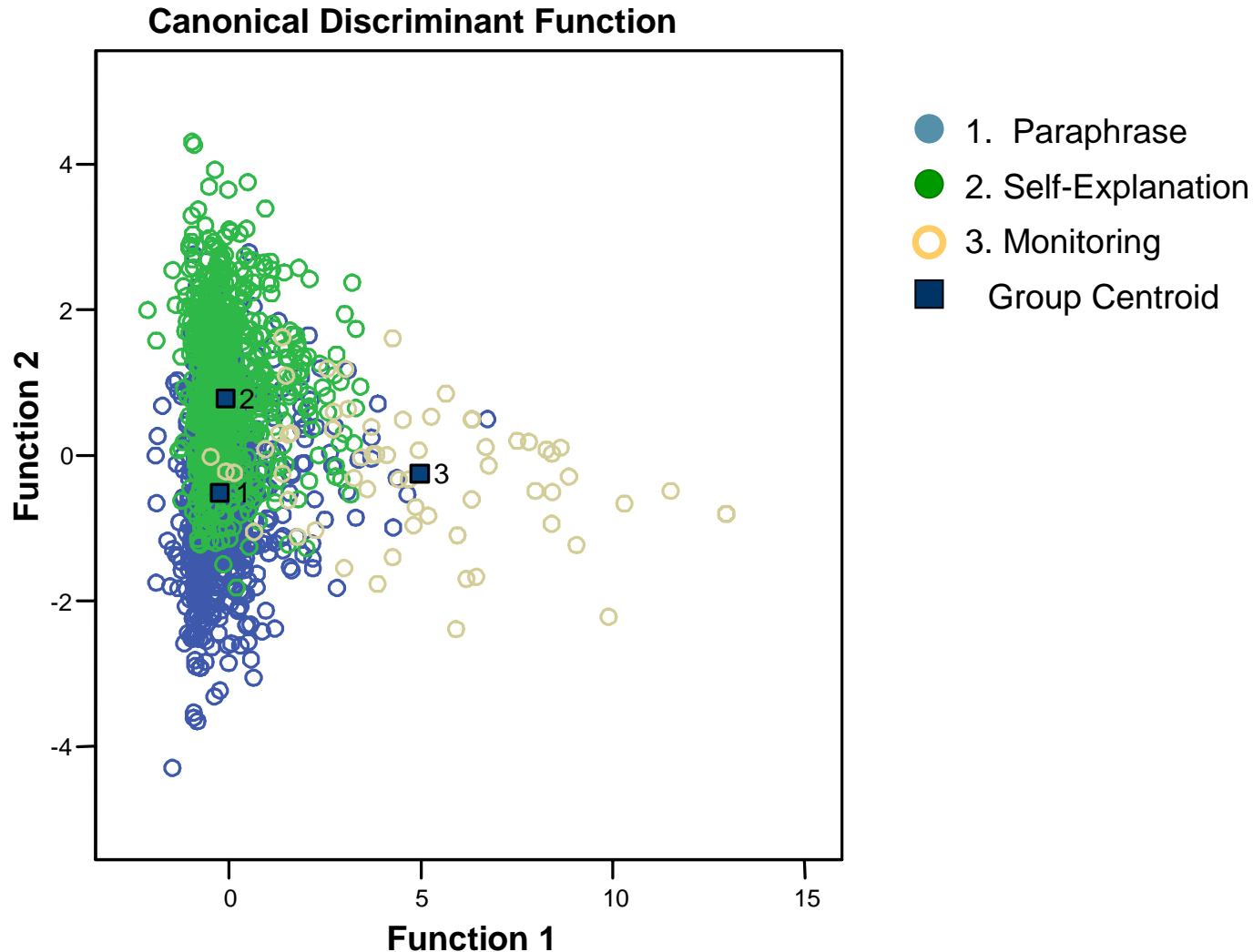
Utterance Classification : Initial Results

- Percentage correct assignments (LDA with best 8 variables)

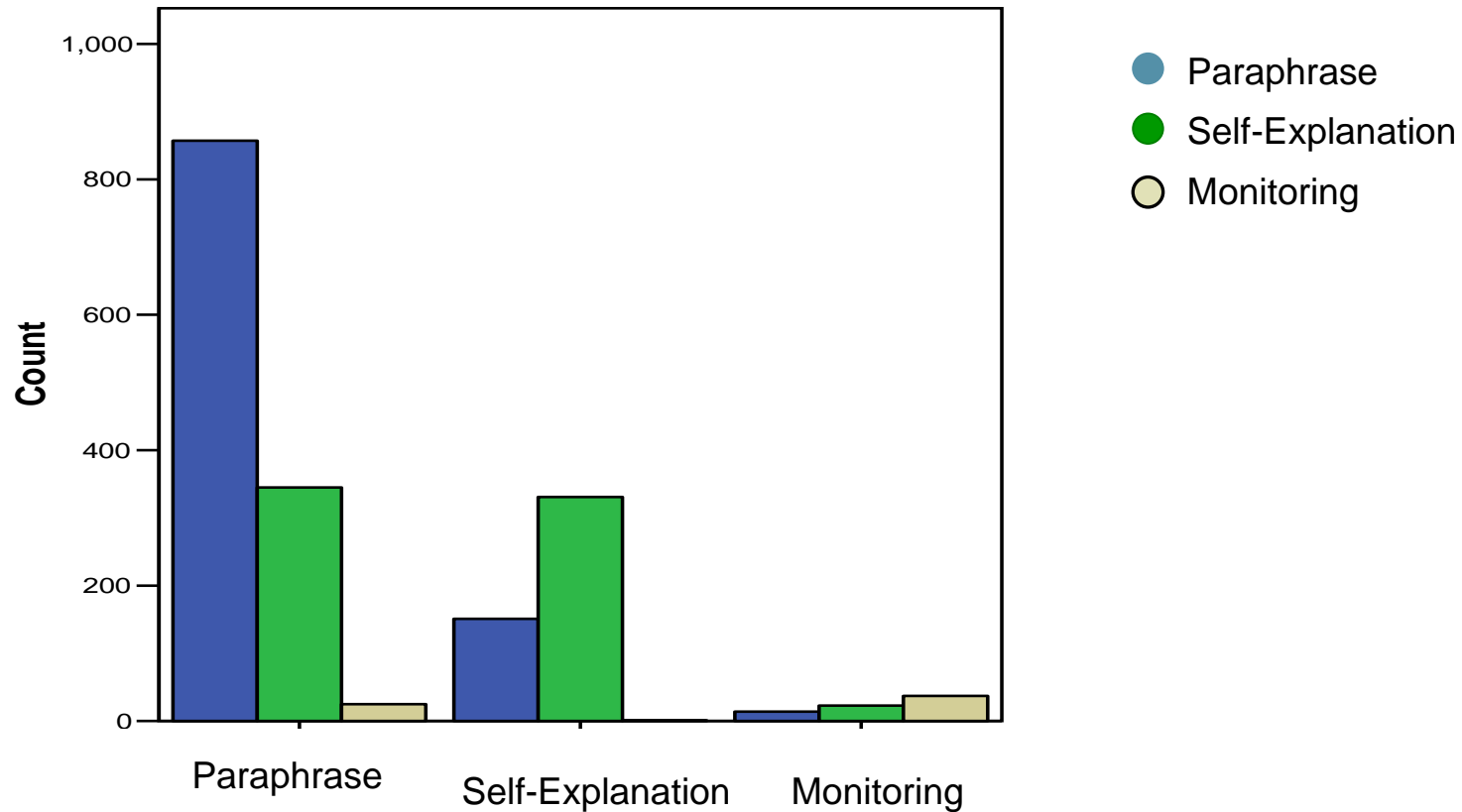
Dataset:	AB	AR
Counts:	61.9%	68.7%
Rates:	62.5%	66.9%
Transitions:	61.7%	59.4%

- Are these figures any good? For comparison
 - Crude chance rate = 33%;
 - Chance expectation using category frequencies = 48%;
 - **(Cross-validated) Success rate using LDF = 59 to 69%.**
- For those who like p-values:
 - from 13.03 to 17.21 standard deviations above the mean;
 - from $p < 10^{-40}$ to $p < 10^{-67}$ (Null Hypothesis).

Utterance Classification : LDF 3-group Plot, AR data:



Utterance Classification : LDF 3-by-3, AR data (cross-validated)



Classification : Variables used (AR data)

- 8 variables maximum (counts): picked: 6 syntactic, 2 semantic

Step entered	Function 1	Function 2	WMatrix category (* = semantic)	Four commonest tokens coded thus in AR data texts
1	2.043	.291	*X2: mental acts & processes	know, suppose, think
2	2.582	-.129	XX: negation	n't, not
3	-.139	.552	VV: lexical verb	goes, got, go, get
4	-.065	-1.038	*H2: parts of buildings	atrium, portal, chambers, atriums
5	-.164	.587	CS: subordinating conjunction	that, so, because, when
6	-.209	.112	NN: noun	blood, heart, body, lungs
7	.103	-.367	JJ: adjective	left, right, pulmonary, other
8	-.699	.777	VM: modal verb	can, will, could, ca [n't]

Q2: Can we Predict Learning Outcomes?

PCA : 5 post-test measures (e.g. multiple choice, blood path diagram, implicit knowledge) → 2 “dimensions”

- Dim 1: Overall Learning
- Dim 2: Types of Learning

■ AB data :

- PC1 ~ 50% variance
- PC2 ~ 24% variance

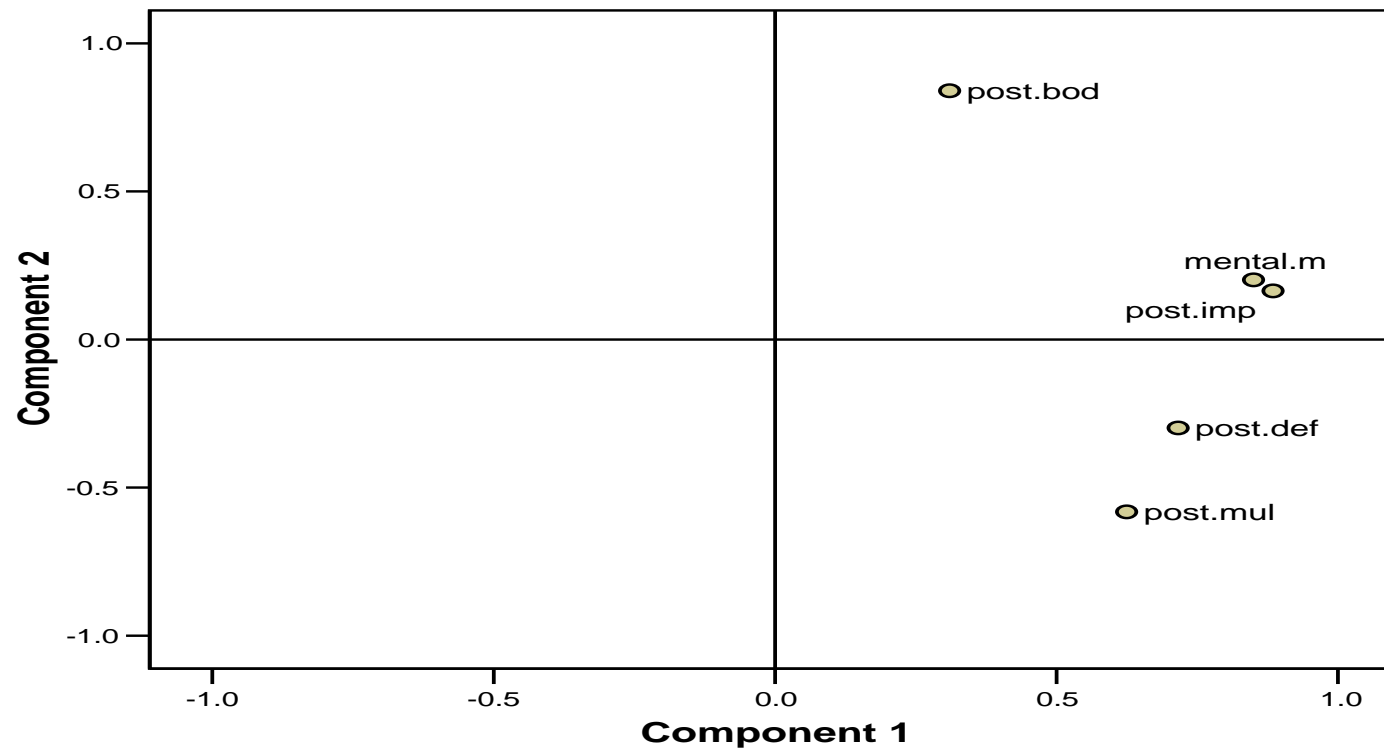
AR data

- PC1 ~ 55% variance
- PC2 ~ 21% variance

Learning Outcomes : First 2 Principal Components (AB dataset)

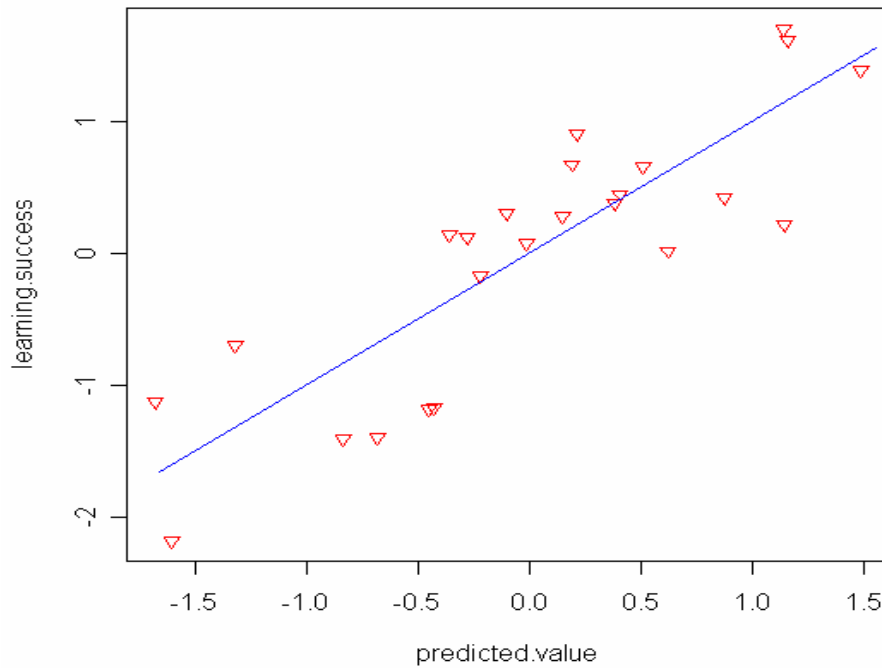
Component 1 = aggregate: how much they learn

Component Plot



Learning Outcomes : Simple Linear Regression

- Best 3 variables (over-fitting control) for AB data:
 - $\text{Dim1} = 0.722 + 0.105 \cdot \text{aVOCbody} - 64.761 \cdot \text{rVOCit} - 0.040 \cdot \text{aVOCof}$
 - [Adjusted R-square = 0.69.]



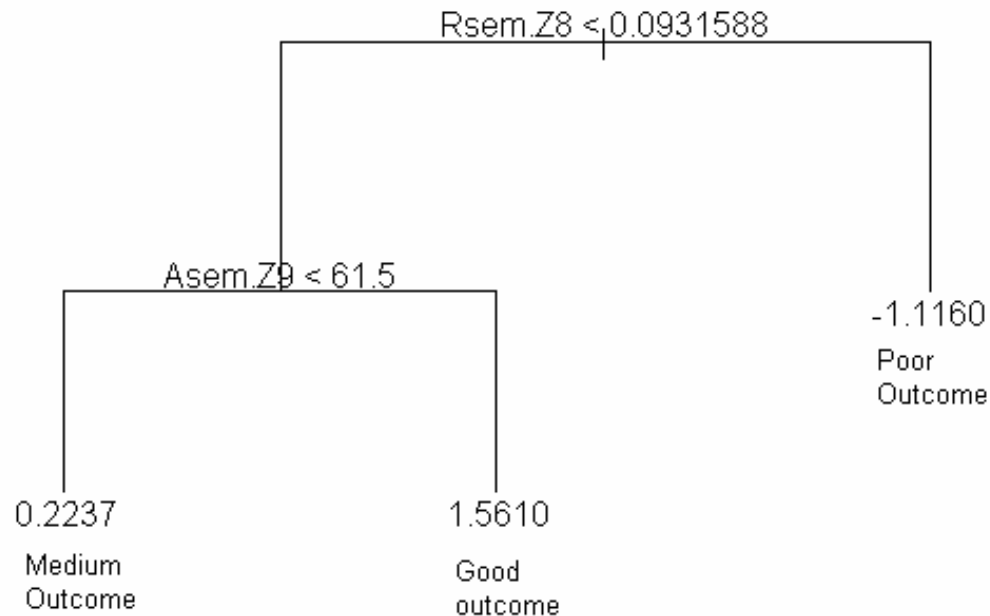
Good learners use:

lots of “body”
not much “of”
low rate of “it”!

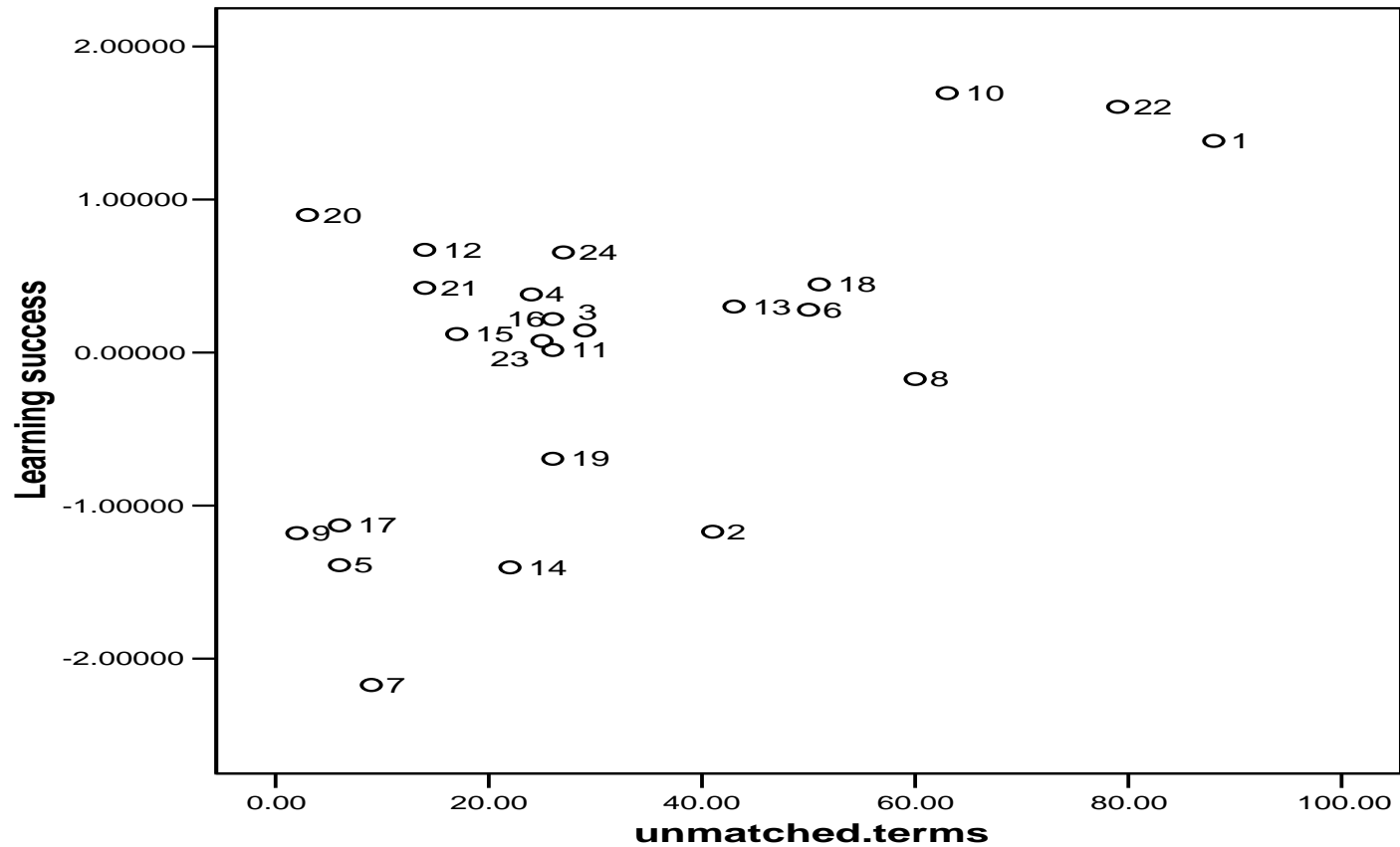
Learning Outcomes : Regression Tree (AB data)

- One of the best 3-leaf-node Regression trees
- (Dim1 ~ semantic variables):
- N.B.
 - High values go right
 - Low values go left

WMatrix semantic category	Most common words in AB corpus
Z8 [pronouns]	it, that, they, you, which, i, this, we
Z9 [unmatched ~ technical terms]	deoxygenated, oxygenated, arterioles, tricuspid, carbon-dioxide, venules, systole, bicuspid



Learning Outcomes : unmatched terms (AB data)



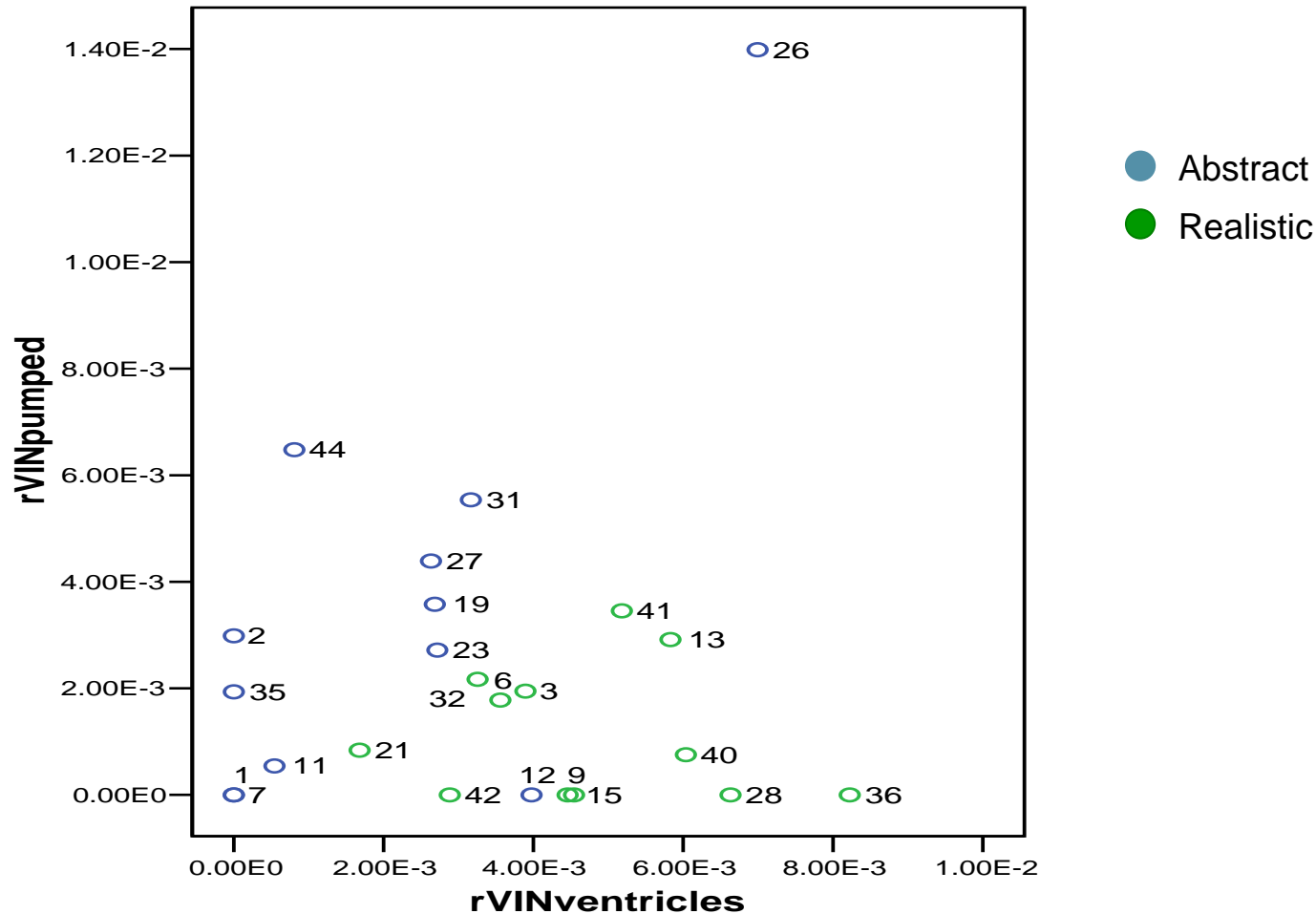
Q3 Can you identify material: 1) Text Coherence (AB data)

	Function
	1
semtF1.Z5	-3.483
comtNN.IZ	3.900
VINTsl.valve	3.530
VINTcirculatory	-1.608
.system	
(Constant)	.151

Coherence:	High	Low
Predicted High	12	1
Predicted Low	0	11

Negative discriminant function value implies High-Coherence condition.

Can you identify material: 2) Diagram Type (AR data)



Summary

- Q1 : How well can human-assigned categories be predicted (semi-)automatically from linguistic data?
 - fairly well
 - 69% correct (chance baseline = 33%)
 - late news: up to 74% using Naive Markov Model
- Q2 : How well can learning outcomes be predicted from learners' verbalizations?
 - well
 - 69% improvement on chance
- Q3 : Can the material students are learning from be identified from their language?
 - yes
 - 96% correct decisions

Implications & Complications

- Plenty left to do
- Extensions:
 - other types of textual data
 - further linguistic features
 - more sophisticated algorithms
 - e.g. evolutionary computing
 - Link to mined interaction data
 - Link to more generic tools (with nice GUI)
- Pitfalls:
 - overfitting (variables >> cases)
 - back-translation (interpreting rules discovered)
 - rule visualization??

What we Plan to do Next:

- Fresh data sets (video records)
- Additional linguistic features, e.g.
 - CohMetrix variables
 - variables derived from Biber's system
- More exciting algorithms, e.g.
 - CNG-wv (Keselj & Cercone, 2004)
 - k-nearest-neighbour regression
 - evolutionary rule-finding
 - Bayesian/Markovian language modelling
- Exploring Human/Machine Trade-Off....

For example (when to move from SS to e-SS?)....

Naive Markov Classifier, AR data:

