

# Language Processing for Diagnosis and Treatment in Mental Health

Matthew Purver  
Queen Mary University of London

with Niall Gunter, Christine Howes, Rose McCabe

# Acknowledgements

The CMSI project was supported by CreativeWorks London, a Knowledge Exchange Hub for the Creative Economy funded by the Arts and Humanities Research Council; and completed in collaboration with Chatterbox Labs Ltd & the Barbican

The PPAT & AOTD projects were supported by Queen Mary University of London's EPSRC-funded Pump-Priming and Innovation Funds, and PsychologyOnline Ltd; and completed in collaboration with PsychologyOnline Ltd and iLexIR Ltd.

The project ConCreTe acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET grant number 611733§



Arts & Humanities  
Research Council



Queen Mary  
University of London



CONCRETE

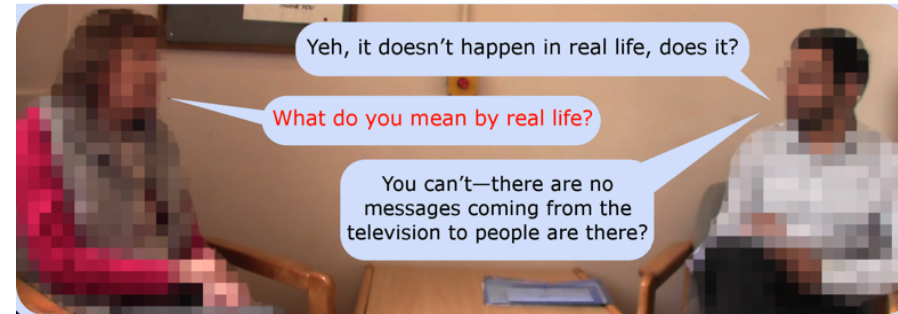


Pioneering research  
and skills



# Questions

- What features of language correlate with / predict symptoms & outcomes?
  - Topic?
  - Sentiment/emotional content?
  - Specific words/phrases?
- Can we use them to help diagnosis and/or treatment?
- Can we detect them automatically?
  - Accurately
  - Robustly
  - Using existing NLP techniques/tools
- How can we do better?



# Mental Health & Language

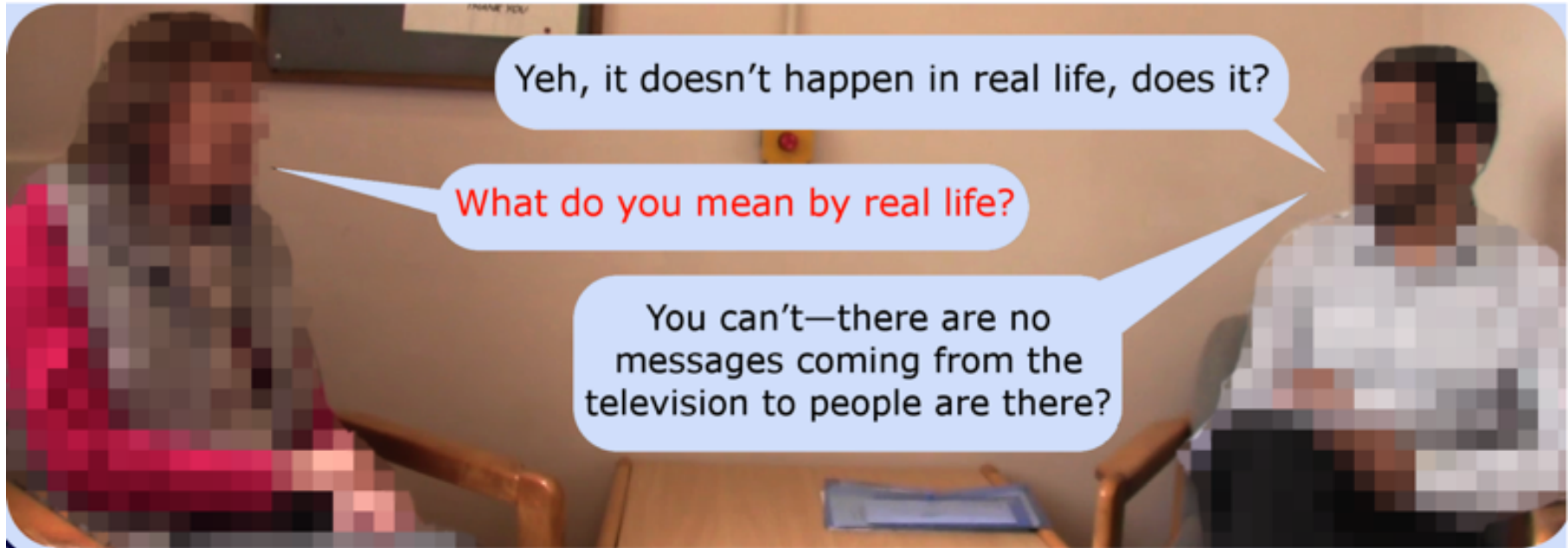


- Communication is important in mental health:
  - Communication quality associated with outcomes
    - (Ong et al, 1995; McCabe et al, 2013)
  - Communication *during treatment*:
    - Conversation structure (how)
    - Conversation content (what)
- Can NLP techniques help us analyse & understand therapy?
- PPAT project:
  - transcripts of face-to-face therapy for schizophrenia
- AOTD project:
  - online text-based therapy for depression & anxiety
- (Howes, McCabe, Purver, SIGDIAL 2012, IWCS 2013, ACL 2014)
- SLADE project:
  - transcripts of face-to-face diagnosis meetings for dementia

# PPAT: Face-to-Face Dialogue

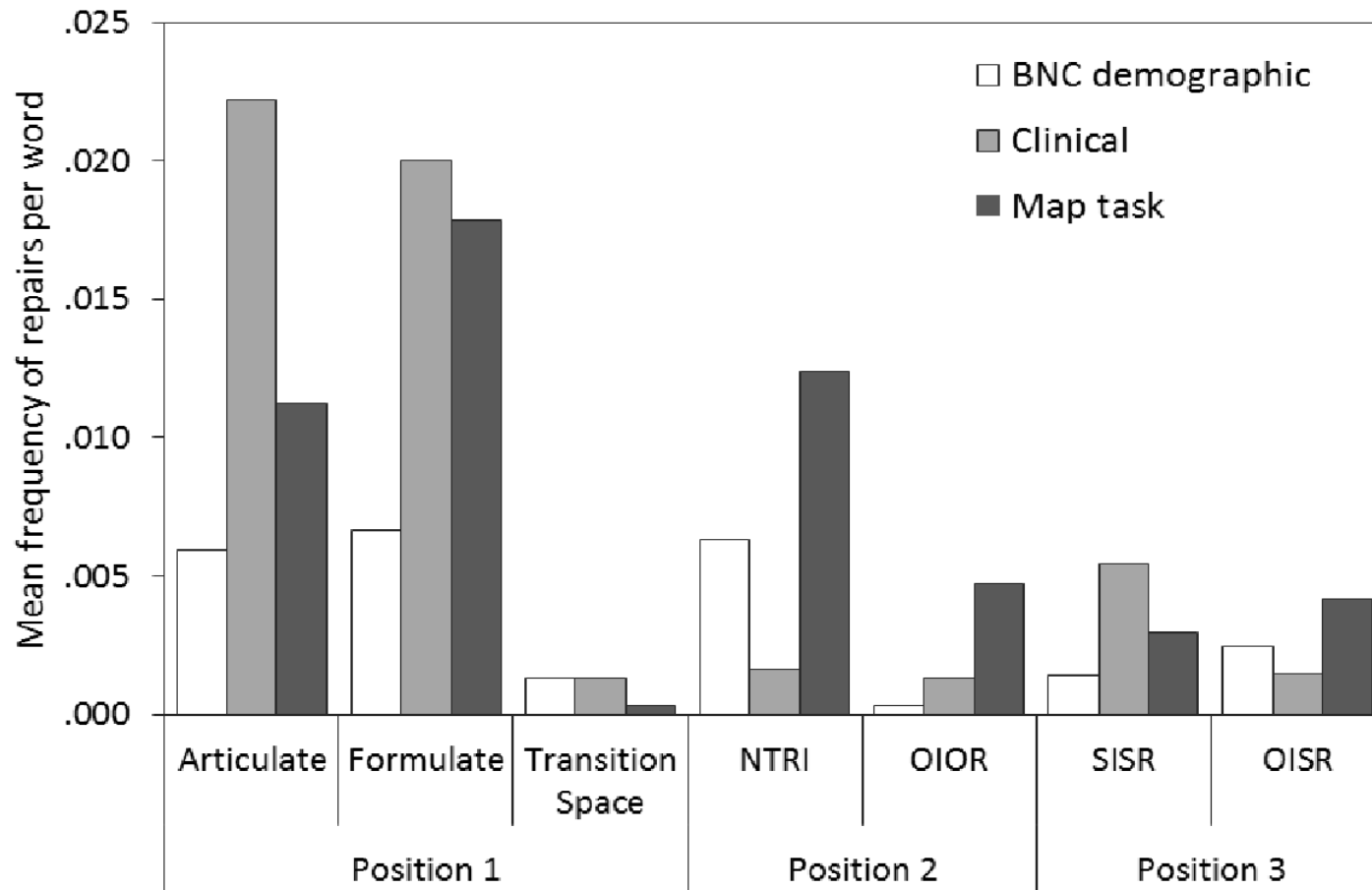
- Transcripts of therapy for schizophrenia
- Manual annotation & statistical analysis
  - McCabe et al (2013)
- Automatic NLP processing & machine learning
  - Howes et al (2012; 2013)
- Detecting symptoms
  - *positive* (delusions, hallucinations, beliefs)
  - *negative* (withdrawal, blunted affect, alogia)
- Predicting related outcomes
  - ratings of communication quality
  - future adherence to treatment:
    - non-adherence: risk of relapse 3.7 times higher
  - shared understanding shown to be a related factor

# Linguistic analysis: Repair



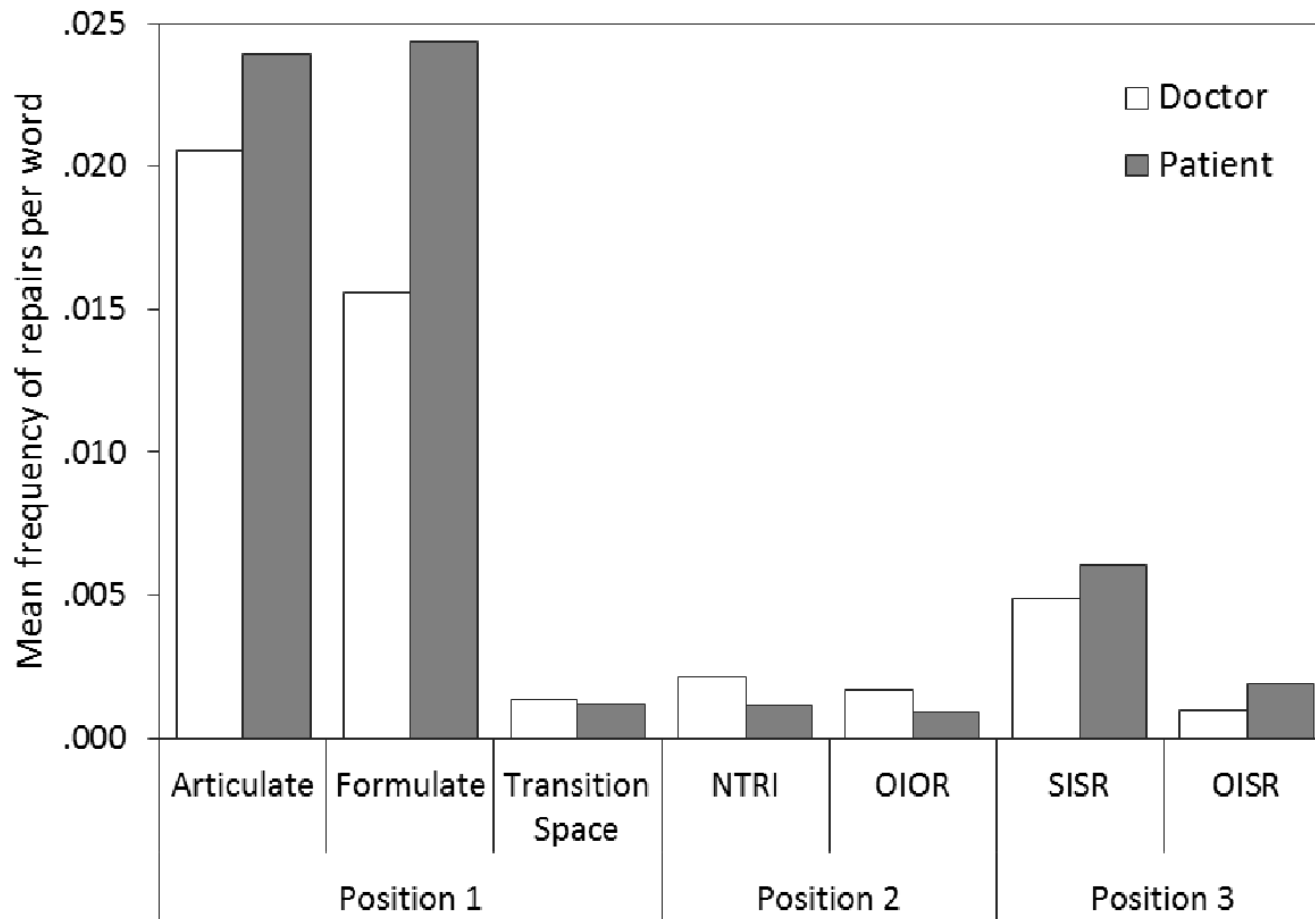
- Manual linguistic analysis
  - Significant role of *repair*
  - Patient-initiated other-repair & self-repair

# Compare other dialogue contexts



- Therapy: more self-repair, less other-repair & initiation

# Patient-doctor comparison



- Patients: more self-repair, less other-repair & initiation



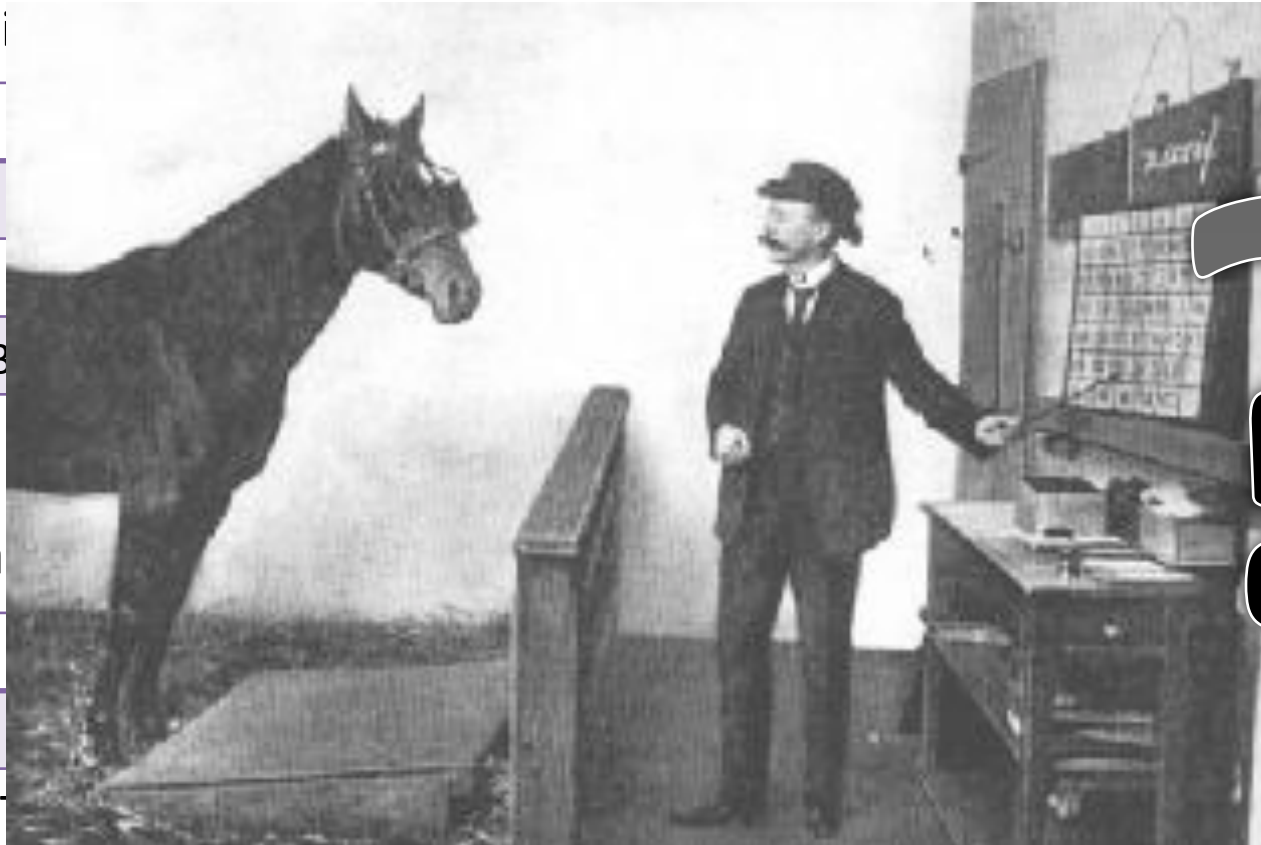
# NLP: brute force

- Classify entire dialogues (patient turns only) with SVMs, ngrams

– Predi



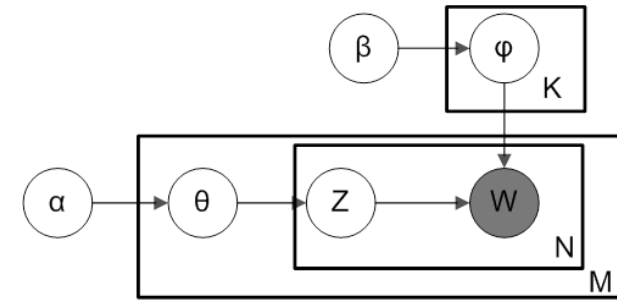
- Similar
- Human



- But how well will this generalise? And what does it **mean**?

# LDA topic modelling

- Infer 20 lexical “topics”:



Topic 0	feel low alright mood long drug feeling tired time confider
Topic 4	voices pills mood cannabis telly voice shaking chris contro
Topic 5	letter health advice letters council copy send dla cpn prob
Topic 7	church voice voices hear medication sister bad hearing tak
Topic 9	school children kids back september oclock gonna phone
Topic 10	weight months medication stone risk lose eat write gp has
Topic 11	place support work centre gotta job stress feel psychologis
Topic 12	door house police thought ring knew worse wall hadnt sat
Topic 13	doctor alright years nice ill anxious write long sit eye hear
Topic 14	drug taking milligrams hundred doctor night time medicat
Topic 15	sort medication work drugs kind team issues drink alcohol
Topic 16	mum place brother tablets died dad depot house meet mo
Topic 17	people life drug make care lot friends dry camera live cop
Topic 18	alright house drink drinking money alcohol god drugs livin

# LDA topic modelling

- LDA topics given manual “interpretations”:
  - (including sentiment aspect)

Interpretation	Example words from top 20
0 Sectioning/crisis	hospital, police, locked
1 Physical health - side-effects of medication and other	gp, injection, operation
2 Non-medical services - liaising with other services	letter, dla, housing
3 Ranting - negative descriptions of lifestyle etc	bloody, cope, mental
4 Meaningful activities - social functioning	progress, work, friends
5 Making sense of psychosis	god, talking, reason
6 Sleep patterns	sleep, bed, night
7 Social stressors - other people stressors/helpful	home, thought, told
8 Physical symptoms - e.g. pain, hyperventilating	breathing, breathe, burning
9 Physical tests - Anxiety/stress arising from tests	blood, tests, stress
10 Psychotic symptoms - e.g. voices, etc.	voices, hearing, evil
11 Reassurance/positive feedback/progress	sort, work, sense
12 Substance use - alcohol/drugs	drinking, alcohol, cannabis
13 Family/lifestyle	mum, brother, shopping
14 Non-psychotic symptoms - incl. mood, paranoia	feel, mood, depression

# Outcome prediction using topics

- Include topic weight per dialogue, with general Dr/P factors, as features:

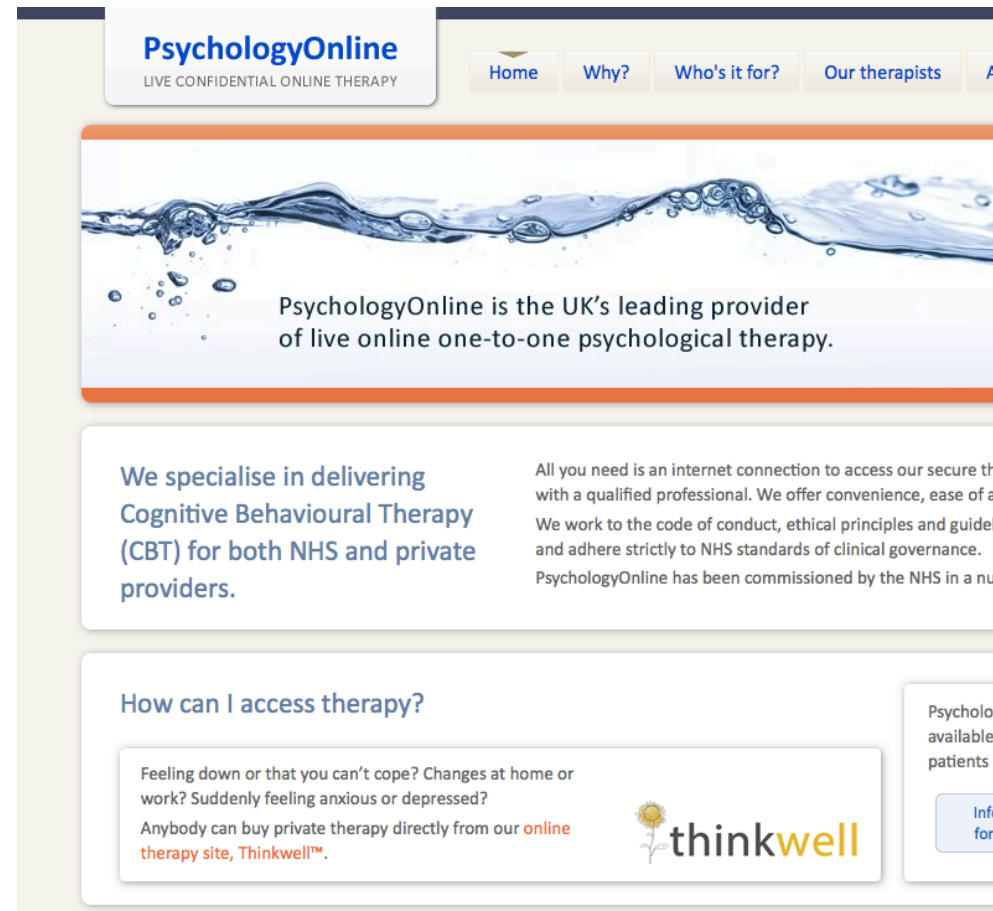
Measure	Manual Acc (%)	LDA Acc (%)
HAS Dr	<b>75.8</b>	<b>75.0</b>
HAS P	59.0	53.7
PANSS positive	<b>61.1</b>	58.8
PANSS negative	<b>62.1</b>	56.1
PANSS general	59.5	53.4
PEQ communication	59.7	56.7
PEQ comm barriers	<b>61.9</b>	<b>60.4</b>
PEQ emotion	57.5	57.5
Adherence (balanced)	<b>66.2</b>	54.1

# Schizophrenia: Summary

- Repair correlates with adherence
  - automatic detection difficult (if not impossible?)
  - ... but a very sparse phenomenon
- Topic modelling provides useful features:
  - topics correlate well with human-annotated topics
  - topics predict symptom severity
  - topics predict therapeutic relationship ratings
  - topics & emotion/stance interrelate
- Predicting future adherence to treatment:
  - topics: 66% (manual), 54% (auto)
  - words & ngrams (phrases): 70%
  - humans: 70% (transcripts), 80% (video)

# Online Text-based Therapy

- Text-based therapy for depression & anxiety
  - PsychologyOnline Ltd
- Cognitive behavioural therapy
  - 2,000 sessions, 500 patients, mean 5.65 sessions/patient
- Anonymisation using RASP
  - (Briscoe et al, 2006)
  - Non-trivial
- Outcome measure
  - Patient Health Questionnaire (PHQ-9)
  - Current severity, progress since start



The screenshot shows the PsychologyOnline website. At the top, the logo 'PsychologyOnline' is displayed with the tagline 'LIVE CONFIDENTIAL ONLINE THERAPY'. A navigation menu includes 'Home', 'Why?', 'Who's it for?', and 'Our therapists'. Below the navigation is a large banner featuring a water splash graphic and the text: 'PsychologyOnline is the UK's leading provider of live online one-to-one psychological therapy.' Underneath the banner, there are two columns of text. The left column states: 'We specialise in delivering Cognitive Behavioural Therapy (CBT) for both NHS and private providers.' The right column contains two paragraphs: 'All you need is an internet connection to access our secure therapy with a qualified professional. We offer convenience, ease of access...' and 'We work to the code of conduct, ethical principles and guidelines and adhere strictly to NHS standards of clinical governance. PsychologyOnline has been commissioned by the NHS in a number of...' Below this is a section titled 'How can I access therapy?' which includes a text box with the question 'Feeling down or that you can't cope? Changes at home or work? Suddenly feeling anxious or depressed?' and the answer 'Anybody can buy private therapy directly from our online therapy site, Thinkwell™.' To the right of this text is the 'thinkwell' logo, which features a sunflower icon. On the far right edge of the screenshot, a partial sidebar is visible with the text 'PsychologyOnline available for patients' and a button labeled 'Info for...'.

# Patient Health Questionnaire (PHQ-9)

- Collected before each session
  - 0-27 scale: higher score = more severe depression
  - moderate/severe  $\geq 10$  (*in/out-of-caseness*)
  - $\Delta$  since start

PATIENT HEALTH QUESTIONNAIRE (PHQ-9)				
NAME: _____		DATE: _____		
Over the last 2 weeks, how often have you been bothered by any of the following problems? (use "✓" to indicate your answer)				
	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0	1	2	3

# Topics

- Themes include family, sleep, symptoms, progress, process:

0	time session sorry today great send next now one work thanks see thank please help make able perhaps look
1	feel life think know way things now like want make self feelings people change maybe someone much need others
2	right well great sure appointment feel thank just lol tonight please know get sorry say bye meeting last though
3	eating eat food weight sick drink meal now lunch control great chocolate absolutely day healthy dinner put use really
4	time husband mum family feel children now dad want see said friends also kids home life got school daughter
5	people say angry situation anger situations said way social others like one friends talk someone person behaviour saying know
6	get go know like need things going just think try want one something time good now make day start



# Topic vs Schizophrenia

Sleep patterns	day sleep week time bed work mood night get things days	sleep day time feel bed bit things hours morning sleeping night
Family	time husband mum family feel children now dad want see said friends also kids home life	mum money dad brother shopping died enjoy tablets blood bad daughter sister
Food / weight	eating eat food weight sick drink meal now lunch control great chocolate absolutely day healthy	weight stone eat medication gain hospital twelve weigh exercise cut gym
Negative feelings	feel life think know way things now like want make self feelings	feel medication feeling thoughts time mood low head past illness
Crises	get help gp depression pain know medication health therapy sorry appointment last face moment	remember doctor hospital reason police people memory ring shaking headaches door
Social stress	work job time good stress working get school life money wife issues	things back place years thought bit ago home put day coming

# Topic vs severity & progress

0	Materials, self-help, procedures	-		10	Unhelpful thinking/habits		
1	Feelings/effects of relationships on sense of self	+	+	11	Work/training/education issues/goals		
2	Positive reactions/encouragement			12	Agenda/goal setting & review		
3	Issues around food			13	Panic attack description/explanation	-	-
4	Family/relationships & issues with (mostly negative)	+		14	Other healthcare professionals, crises, risk, interventions	++	
5	Responses to social situations			15	Sleep/daily routine	+	
6	Breaking things down into steps	+		16	Positive progress, improvements	--	-
7	Worries/fears/anxieties	-		17	Feelings, specific occasions/thoughts		
8	Managing negative thoughts/mindfulness			18	Explaining/framing in terms of CBT model		+
9	Fears, checking, rituals, phobias	-	-	19	Techniques for taking control	-	-

# Sentiment/Emotion Detection

- Detect positive & negative sentiment
  - see e.g. (DeVault et al, 2013)
- Detect anger
  - challenge & emotion elicitation in CBT process
- Compared existing tools
  - Manually annotated 85 utterances in 1 session
    - *positive / negative / neutral*
    - Inter-annotator agreement  $\kappa = 0.66$
- Dictionary-based LIWC
  - sentiment 34-45%; anger recall = 0
- Data-based (RNNs) Stanford
  - sentiment 51-54% (no anger)

# Distant Supervision



- A common technique for sentiment detection

Best day in ages! #Happy :)

just because people are celebs they dont  
reply to your tweets! NOT FAIR :(

# Distant Supervision



- A common technique for sentiment detection

Best day in ages!

just because people are celebs they dont  
reply to your tweets! NOT FAIR

# Distant Supervision



- A common technique for sentiment detection

Best day in ages!

just because people are celebs they dont  
reply to your tweets! NOT FAIR

再做个梦如果明天我中奖了该怎么支配呢每次想这个问题都觉得很美\*^\_^\*

离队倒计时,期待奇迹的发生 (T\_T)

# Distant Supervision



- A common technique for sentiment detection

Best day in ages!

just because people are celebs they dont  
reply to your tweets! NOT FAIR

再做个梦如果明天我中奖了该怎么支配呢每次想这个问题都觉得很美

离队倒计时,期待奇迹的发生

- e.g. Go et al (2009): works well *if* you have a reliable but (semi-)independent label to hand

# Distant Supervision

- Can be applied to finer-grained emotions (Purver & Battersby, EACL 2012)
  - But quite bad for some ... how reliable are these?:

: - O

: - @

: - \$

: - P

- Can also get supervision from responses:

`_AggieGirl16:`     `@captain_lizard lol yeaaaah. I'm pretty lucky! Haha!`

`captain_lizard:`     `@_AggieGirl16 I'm glad you're happy, Monica! :)`



# Sentiment/Emotion Detection

- Detect positive & negative sentiment
  - see e.g. (DeVault et al, 2013)
- Detect anger
  - challenge & emotion elicitation in CBT process
- Compared 3 existing tools
  - 1 dictionary-based: LIWC
  - 2 data-based: Stanford (news), Sentimental (social media)
- Manually annotated 85 utterances in 1 session
  - *positive / negative / neutral*
  - Inter-annotator agreement  $\kappa = 0.66$
  - LIWC 34-45%; Stanford 51-54%; Sentimental 63-80%

# Sentiment/Emotion vs PHQ

	Severity (PHQ)	Progress ( $\Delta$ PHQ)
Sentiment mean	--	-
Sentiment std dev		+
Anger mean/max	+	
Anger std dev	+	

- More positive sentiment → better PHQ, progress
- More variable sentiment → worse progress
- More/more variable anger → worse PHQ

# Predicting final outcomes

- Changes in levels help predicting final in/out-of-caseness:
  - using features from initial and/or final sessions:

	Final In-caseness
<i>Baseline proportion</i>	<i>26.8%</i>
First + last session features, incl deltas	<b>0.71 (0.48)</b>
Including early PHQ scores	<b>0.76 (0.51)</b>

- Features chosen are informative:
  - Levels of sentiment & anger, progress & crisis/risk topics
    - Deltas between sessions
  - PHQ scores at assessment and initial treatment sessions

# Predicting dropout

- Can we predict dropout & non-engagement?
  - 148 of 500 did not enter or stay in treatment

	Dropout
<i>Baseline proportion</i>	<i>29.6%</i>
Assessment session features	0.65 (0.26)
Treatment session features	<b>0.70 (0.59)</b>
Both sessions	<b>0.73 (0.64)</b>

- >70% accuracy using initial session features
  - But only by including fine-grained word features

# Predicting therapy quality

- Can we distinguish “good” from “bad” therapists?
  - Top 25% vs bottom 25% based on number of patients recovered

	Dropout
<i>Baseline proportion</i>	<i>50%</i>
Only high-level features	0.67 (0.63)
Including lexical features	<b>0.78 (0.74)</b>

- Good accuracy using initial & final session features
  - But mostly by including fine-grained word features

# Depression: Summary

- Topic modelling provides useful features:
  - topics correlate well with human-annotated topics and previous study
  - topics correlate with symptom severity and progress
- Emotion detection provides useful features:
  - levels and variability predict symptoms and progress
  - needs care choosing & training tools
- Predicting useful outcome measures:
  - recovery: 71%, 76% with PHQ information
  - dropout: 73%
  - therapist quality: 78%
  - but perhaps we don't understand the last two ...

# SLADE: Dementia Diagnosis

- U. Exeter dataset
  - 148 diagnosis conversations with doctor (& carer)
    - 70 positive diagnosis of dementia
    - 78 negative diagnosis (Mild Cognitive Impairment in some cases)
  - After referral from GP, memory tests/scans
  - Given diagnosis, advice
- Relatively early stage
  - Can we aid diagnosis?



# Dementia & Language

- Vocabulary reduction (e.g. Hirst & Feng, 2012)
  - Authors over long timescales
- Content reduction (e.g. Orimaye et al 2014)
  - Fewer predicates
  - Fewer utterances, shorter sentences
  - DementiaBank: 74%
- Speech features (Jarrold et al, 2014)
  - Including lexical class features
  - Pronoun vs noun vs verb frequencies
  - Small set, healthy controls: 80-90%
- But we have short timescales, variable content ...



# Content vs Structure

- Content (topics, words) highly diagnosis-dependent
  - Advice on driving, legal requirements, future planning
  - And many other features e.g. length
  - Need content-*independent* features
- Qualitative investigation: dialogue structure
  - Question-answering (non-answering) behaviour
    - Strategies to avoid answering
  - Interaction structure (involvement of carer)
  - Feedback/coordination behaviour
    - Backchannels, hesitations, turn-taking patterns
    - Indications of understanding
    - Other- and self-repair

# Conversation-based studies

- Many CA-like studies
  - Watson et al 1999 ... Jones et al 2015
- Indicative dialogue-structural features
  - “Lack of fluency”
    - Self-repair
    - Lack of topic coherence
  - Other-repair
    - Types, appropriateness, answering behaviour, lack of corrections
  - Question-answering
    - Avoidance strategies, contentlessness
  - Pausing behaviour
    - Intra- and inter-utterance
  - Backchannel behaviour
    - More contentless utterances vs lower use of continuants?
  - Laughter

# Question-answering

- Watson et al (1999)

Normal 75: Can you remember the name of where you worked?

SDAT 76: Yeah

Normal 77: Mm? (as in 'Tell me').

SDAT 78: Oh yes.

Normal 79: Well what was the ...

SDAT 80: I remember those things love I-I can't remember his name now he was a big bully (laughs).

Normal 81: But what was the name of the place where you worked?

SDAT 82: And he used to hurry up the steps like this. You know hurry up the steps.

Normal 83: Yeah. What was the name of the place where you worked?

SDAT 84: No, that's one thing I'm sorry I can't remember. It was up near a hotel at Paddington anyway I know that much (laughs).

# Repair

- Watson et al (1999)

Normal 17: Don't you have any kids.  
SDAT 18: Huh?  
Normal 19: Any children?  
SDAT 20: Children, well my my fam-family were a heavy family three three or four.  
Normal 21: Oh, you mean a large family.

.....

Normal 25: Did you have your own children?  
SDAT 26: Ah no I used to go to school right (unintelligible) in reality about that size.  
Normal 27: What would be that size?  
SDAT 28: Can't go to school in the morning 'cause all depends on where you are.

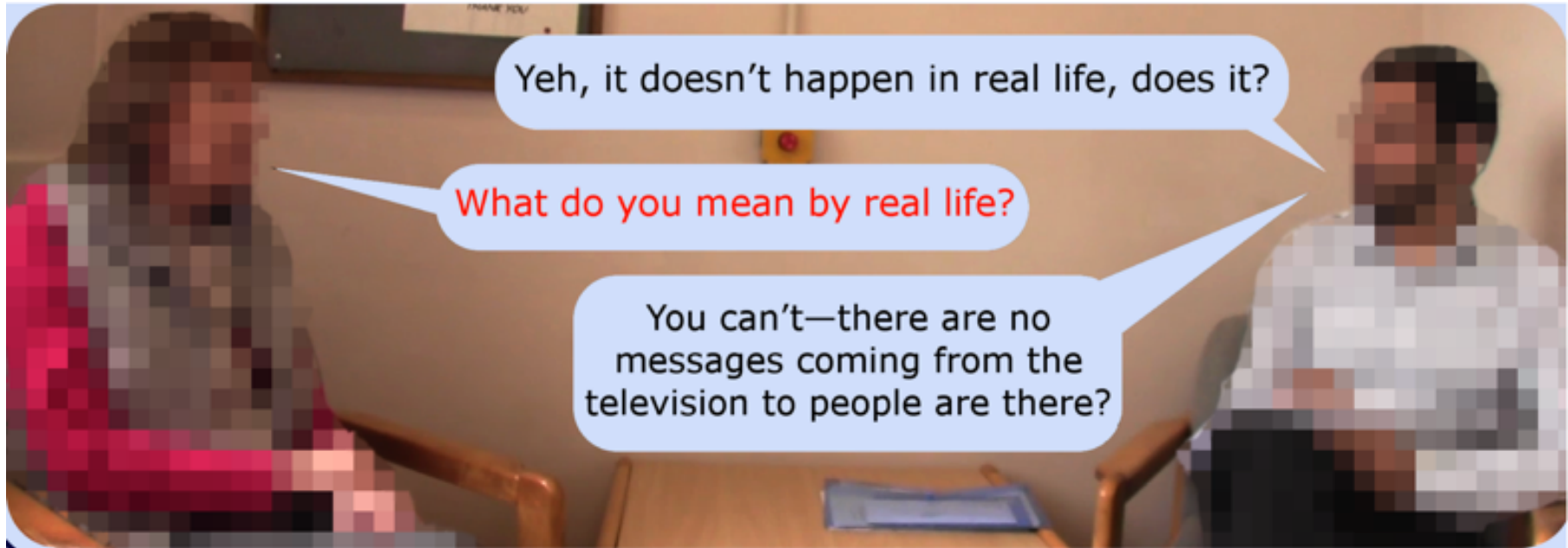
# First try

- Simple “content”-independent features:
  - Social indicators:
    - Greetings
    - Contribution of carer
  - Self-repair indices:
    - Pauses, filled pauses, incomplete words
  - Other-repair indices:
    - Repetition, conventional forms
  - Dialogue structure features:
    - Non-answering (short answers, “don’t know” keywords, pauses)
    - Carer answering (turn sequences)
- Accuracy 72% ... but:
  - Simple repetition indices not useful (inter/intra-utterance)
  - Simple answering (keywords/speaker changes) only marginal
  - Pauses very helpful

# Dialogue act detection

- Specific dialogue acts:
  - Questions, answers, backchannels, repair
- Big sparseness problem
  - Tagging for repair-related DAs (Surendran & Levow, 2006)
    - `check` 8% turns, 45% f-score, `clarify` 4% turns, 19% f-score
  - Fragment detection in dialogue (Schlangen, 2005)
    - Fragments 5% of turns, 30-40% f-score
- Individual classifiers from Switchboard
  - Questions, backchannels c.80% accuracy

# Repair



- Schizophrenia data:
  - Significant role of *repair*
  - Patient-initiated other-repair (above)
  - And self-repair:
    - Did you feel that – *did you despair so much that* – you wondered if you could carry on

# Self-repair

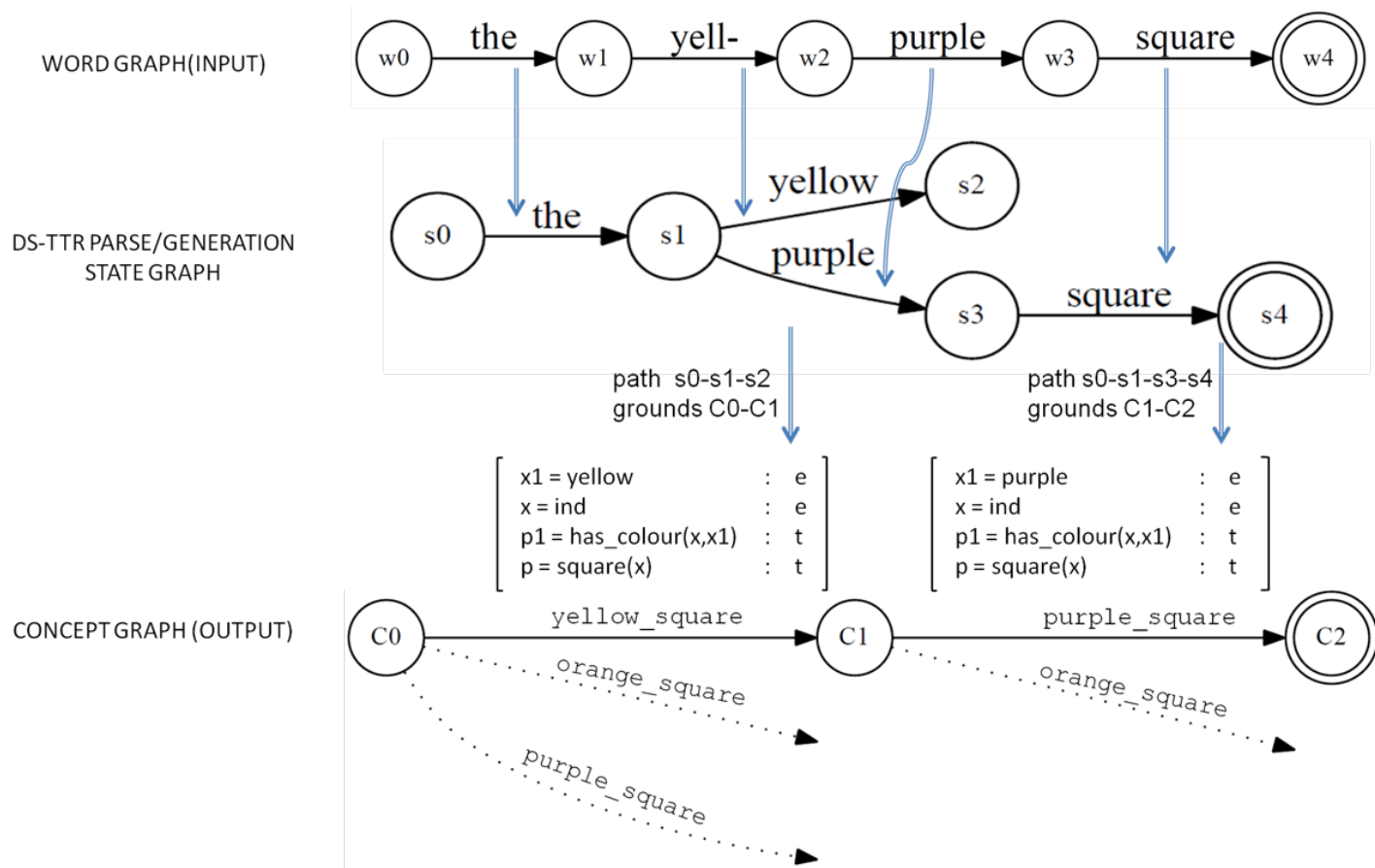


- (Hough & Purver, EMNLP 2014)
- “Disfluency detection” for speech recognition
  - A flight to Boston – uh, I mean, to Denver
    - A flight to Denver
  - John likes, uh, loves Mary
    - John loves Mary
- But what about:
  - The interview was – it was alright
  - I went swimming with Susan – or rather, surfing
- Incrementality & monotonicity:
  - Maintain semantic context, but with ...
  - incremental parsing & choice mechanisms
  - Using domain-general methods



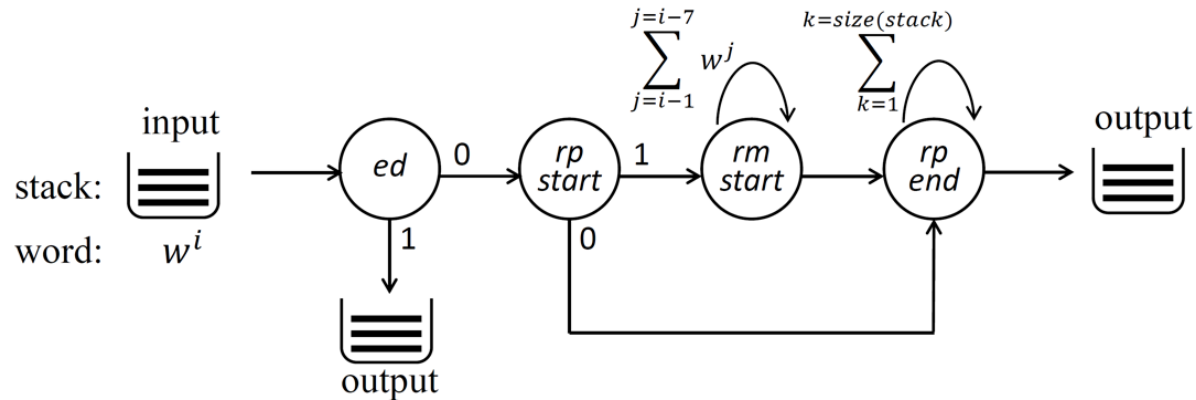
# Self-repair

- Incremental, monotonic context model



# Self-repair

- Incremental, information-theoretic repair point classifier



- Domain-general features:
  - Similarities between probability distributions
  - Changes in probability & entropy given repair hypotheses
  - Combined in random forest classifier
  - Near state-of-the-art F-score 0.81, with faster incremental performance
  - Transfer to mental health domain: 0.68, per-dialogue correlation 0.95

# Other-Repair

- (Howes, McCabe, Purver SIGDIAL 2012)
- Define features manually, extract automatically
  - Linguistically/observationally informed:
    - Wh-question words, closed class repair words
    - Repetition, parallelism
    - Backchannel behaviour, fillers, pauses, overlaps
  - Brute force: all unigrams
- Train SVMs to detect repairs (NTRIs & P2Rs)
  - 44,000 turns, only 567 NTRIs (159 patient), 830 P2Rs (262)
  - 80-86% on balanced data
  - but only 35-44% F-scores on real data (above 20-36% baselines)
- How can we do better?
  - Repair involves parallelism: not always lexical, but semantic
  - Self-repair model: language model distributions
  - Other-repair: lexical repetition

# Results – balanced data

- Balanced data (i.e. **small** dataset), patient only:

Target	Features	Accuracy (%)
NTRI	Repeated proportion	61.2
NTRI	All high-level	83.2
NTRI	All unigrams	82.4
NTRI	All features	86.3
P2R	Repeated proportion	61.5
P2R	All high-level	78.5
P2R	All unigrams	77.1
P2R	All features	79.8

- But of course the real data's not balanced ...

# Results – repair detection

- On balanced data: accuracy 80-86%
- Full dataset, patient only:

Target	Features	P (%)	R (%)	F (%)
NTRI	OCRProportion	85.7	22.6	<b>35.8</b>
NTRI	All high-level	42.8	40.6	<b>41.4</b>
NTRI	All features	44.9	43.6	<b>44.0</b>
P2R	OCRProportion	56.4	11.8	<b>19.6</b>
P2R	All high-level	36.2	28.4	<b>31.6</b>
P2R	All features	43.8	30.3	<b>35.4</b>

- We can probably do better:
  - Audio/video: intonation, non-verbal behaviour
  - Context: follow-up dialogue turns incl. other-person reaction
  - But: does it actually help anyway?

# Patient NTRI: Example 1



- PCA analysis shows adherence correlates with patient-led clarification

P I mean it's a worry a stress do you know what I mean but you know you know if I confronted it

Dr Right e:::rm who is your key worker again?

P My key worker is (name) o:::r?

Dr No no no from us is it Kathrin?

P Who?

Dr oh I'm sorry I thought you said Kathrin

P what my key worker from here?

Dr Yeah

P My CPN? () Me:::l –

Dr Melvo

P Melvin

# Patient NTRI: Example 2



- PCA analysis shows adherence correlates with patient-led clarification

P it's eh- my social life is better and I'm I'm communicating with people and there's a lot of people in the choir who I've talked to about sss- who had similar problems you know so it's- I'm not alone anymore so which is good.

Dr and eerr how is it with with your alcohol currently

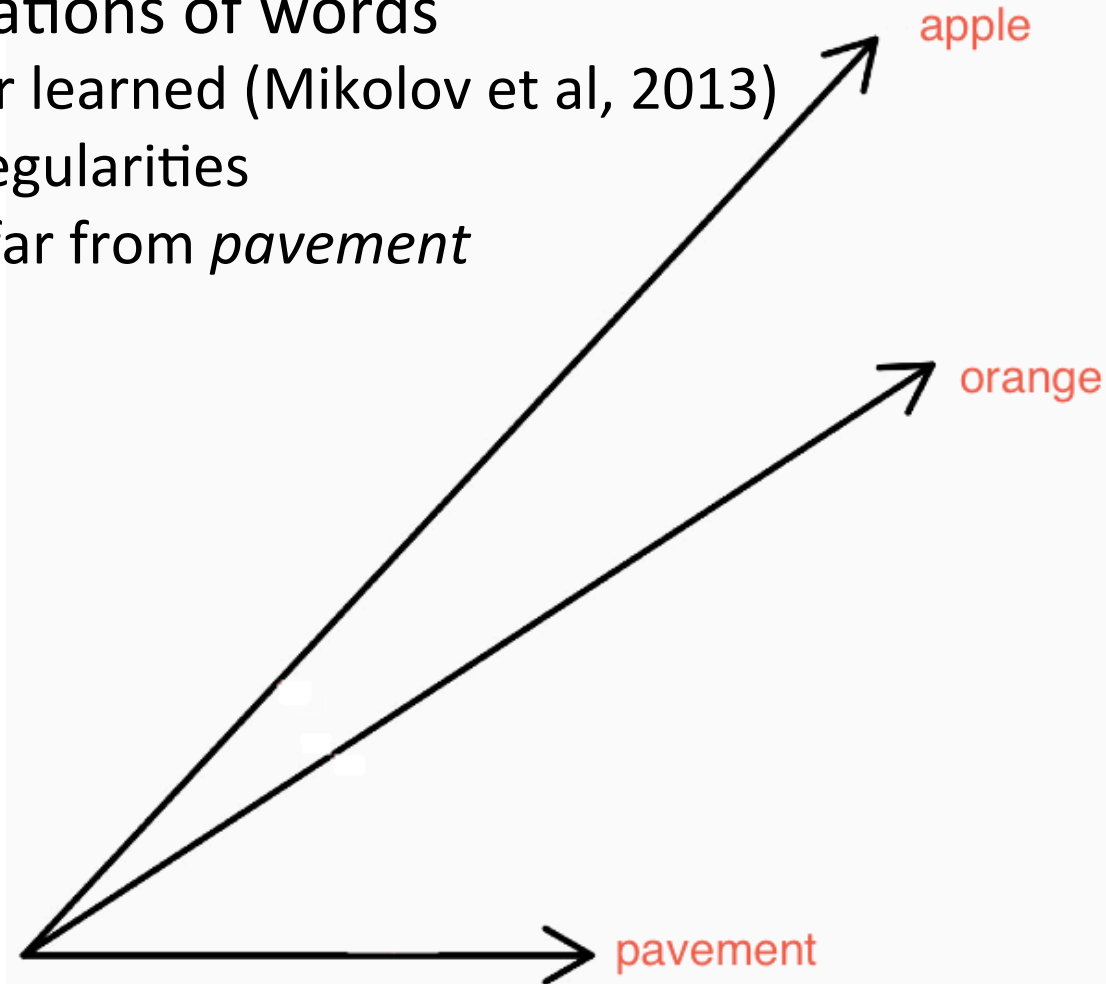
P with my

Dr alcohol drinking (words)

P uuuummm I've cut down eerrm a lot maybe eerrrm I don't get pissed anymore.

# Distributional Semantics

- Vector space representations of words
  - Co-occurrence-based or learned (Mikolov et al, 2013)
  - Semantic similarity & regularities
  - *apple* close to *orange*, far from *pavement*
  - *(king – queen)*
    - $\approx$  *(man – woman)*
    - $\approx$  *(uncle – aunt)*



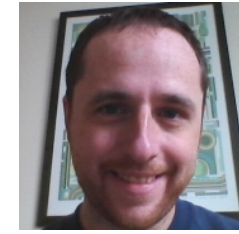


# Distributional Semantics

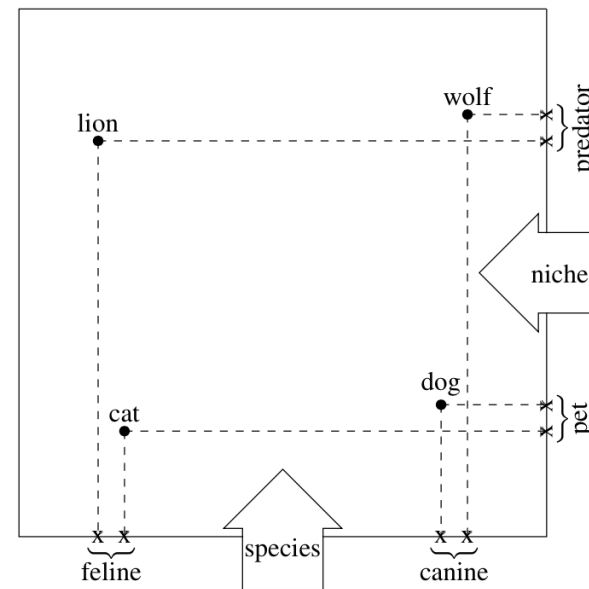
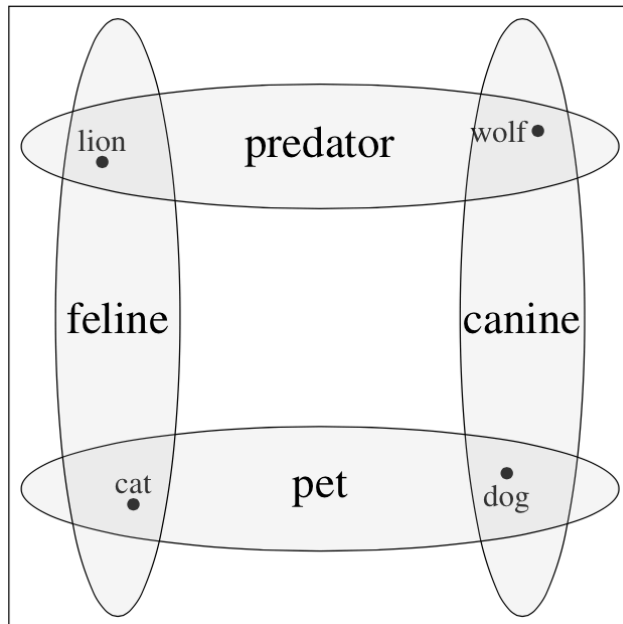


- Standard distributional models help DA tagging ...
  - (Milajevs et al, EMNLP 2014)
  - ... but not much! (0.60 -> 0.63 accuracy)
- Standard models reflect within-sentence distributions:
  - word2vec (Mikolov et al, 2013) on Google News 100bn wd
  - Closest neighbours of “hello”:
    - hi 0.654899
    - goodbye 0.639906
    - howdy 0.631096
    - goodnight 0.592058
- Training on dialogue data can help:
  - (Kalchbrenner & Blunsom, 2013) RCNNs: 0.74 accuracy
  - But gives a domain/task-specific model

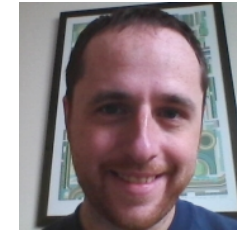
# Meaning is Contextual



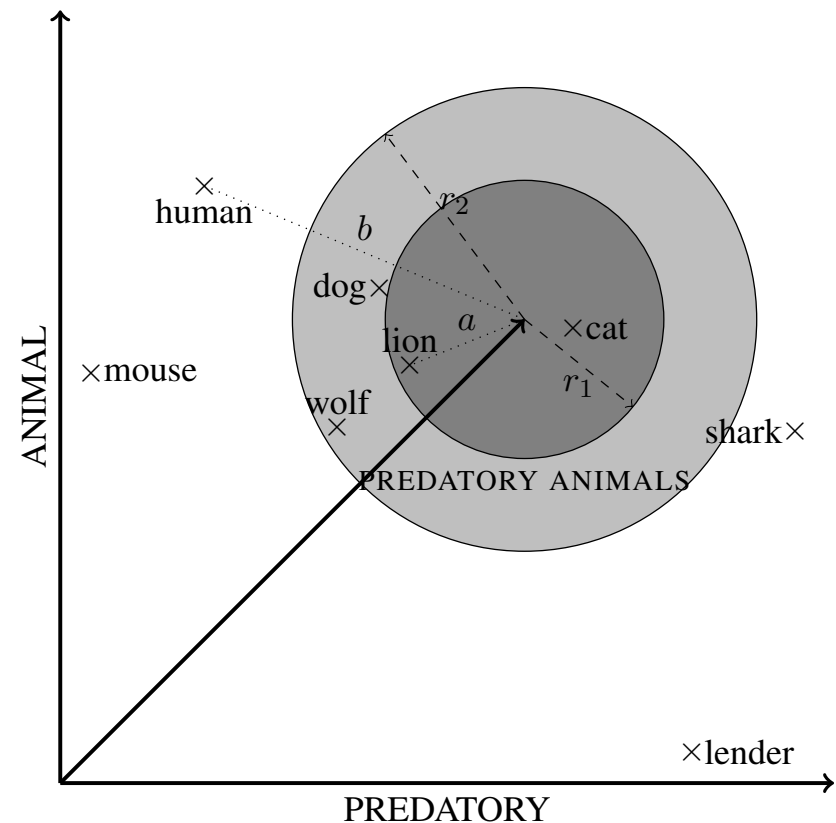
- Perhaps we need to account for **context**
- Distributional semantics & concept formation
  - (Agres, McGregor, Purver, Wiggins in prep)



# Meaning is Contextual



- Dynamic dimension selection from sparse space
- “Conceptual space” formation
- WordNet eval (precision @ 50):
  - 0.18
  - 0.17 (word2vec)
  - 0.12 (GloVe)
- Human eval (recall):
  - 0.24 (anchor)
  - 0.16 (norm)
  - 0.14 (word2vec)



# Sparse vs Reduced Spaces

- Lexical similarity



# Sentence similarity



# Addition survives

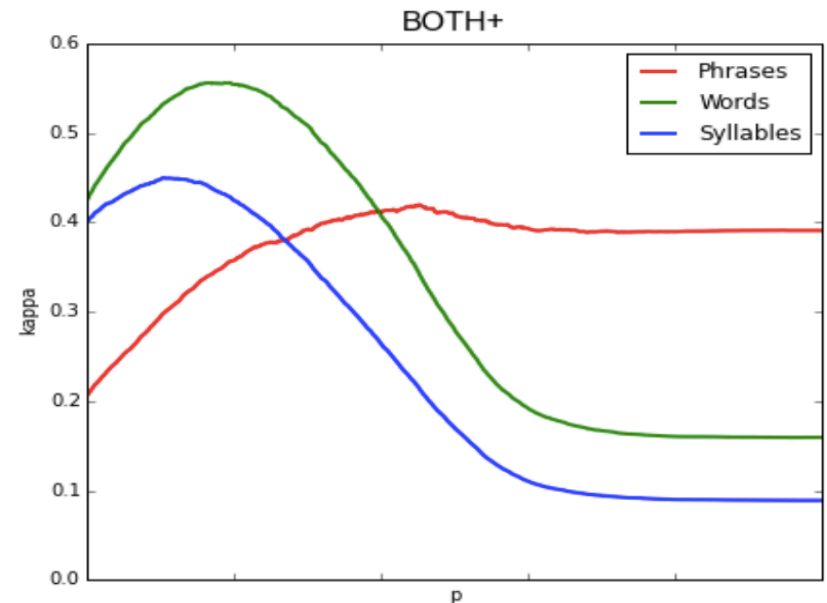
- Kron and mult degrade



# What are the right units?

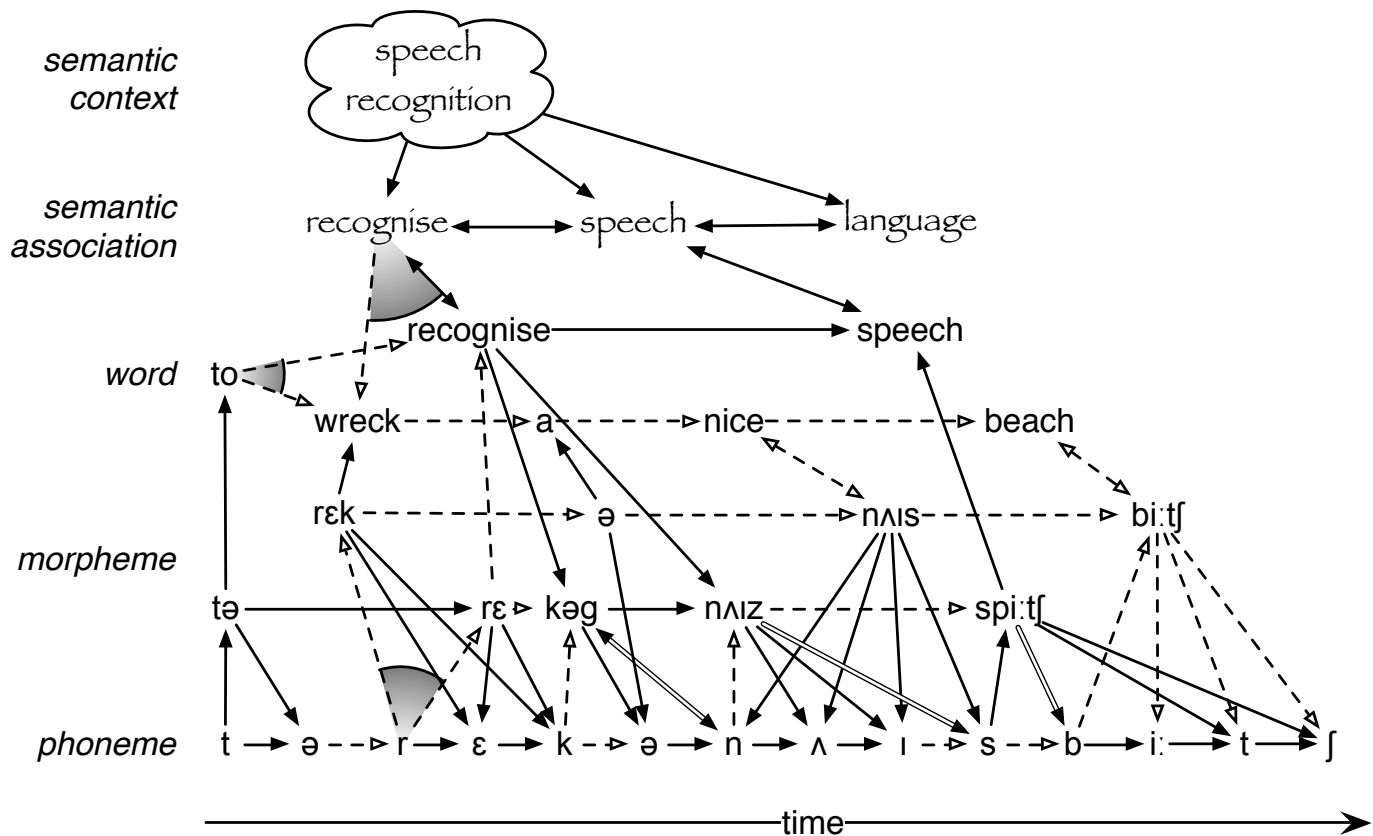


- Perhaps we need to learn from **contextual distributions**
- Which means we need to know the units of interest
  - (cf. Nishida “conversation quanta”?)
- Unsupervised, information-theoretic induction
  - (Griffiths et al, in prep)
  - Segment on changes:
    - information content
    - entropy
  - At different levels:
    - syllables 0.67 F1
    - words 0.71 F1



# What are the right units?

- Scaling up to a hierarchical model





# Thanks!

- To you and:

- Shauna Concannon
- Rose McCabe
- Julian Hough
- Arash Eshghi
- Niall Gunter
- Christine Howes
- Dmitrijs Milajevs
- Mehrnoosh Sadrzadeh
- Dimitri Kartsaklis
- Zheng Yuan
- Pat Healey
- Ruth Kempson
- Kat Agres
- Jamie Forth
- Stephen McGregor
- Geraint Wiggins
- Sascha Griffiths

