Predicting Outcomes from Patient-Clinician Dialogue

Matthew Purver
(with Christine Howes, Rose McCabe)

Centre for Intelligent Sensing
Computer Science / Medicine and Dentistry
Queen Mary University of London

CIS centre for intelligent sensing

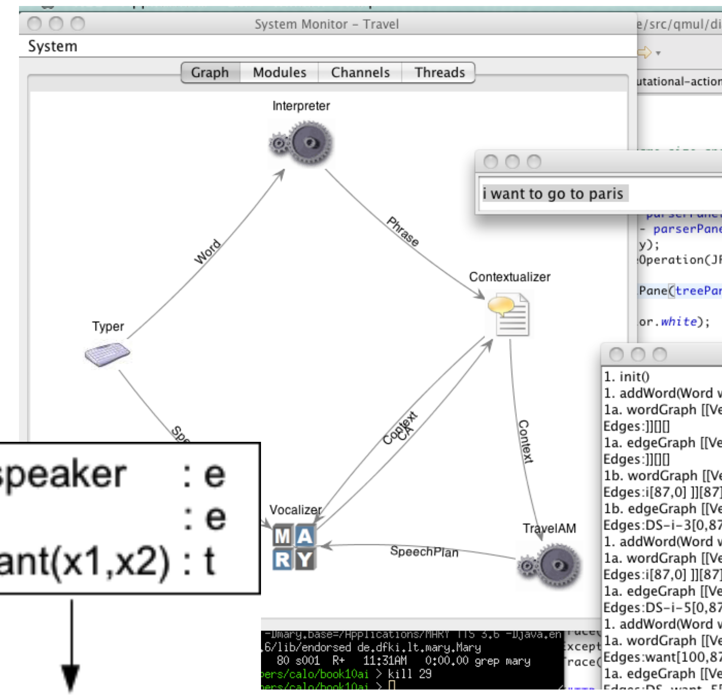
Queen Mary University of London

Human-Computer Dialogue

- Coordination and repair
  - Self-repair, other-repair, clarification

> A: I want to go to er
> B: yes
> A: to London
> B: London?
> A: sorry no Paris, in March

- Incrementality
  - Semantic parsing and generation
  - DYLAN dialogue system



x1 = speaker	: e
x2	: e
p = want(x1,x2)	: t

↓

x1 = speaker	: e
x2 = ticket	: e
p = want(x1,x2)	: t

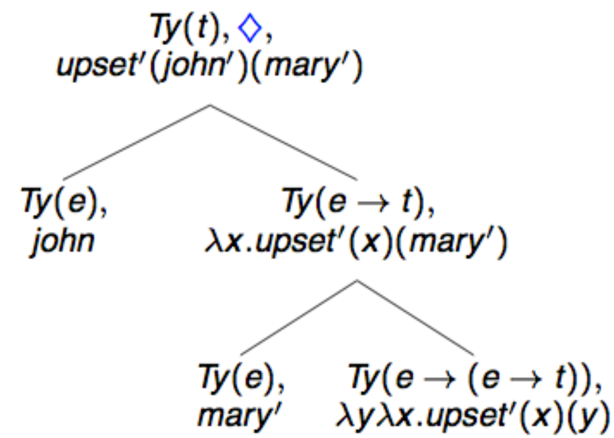
↓

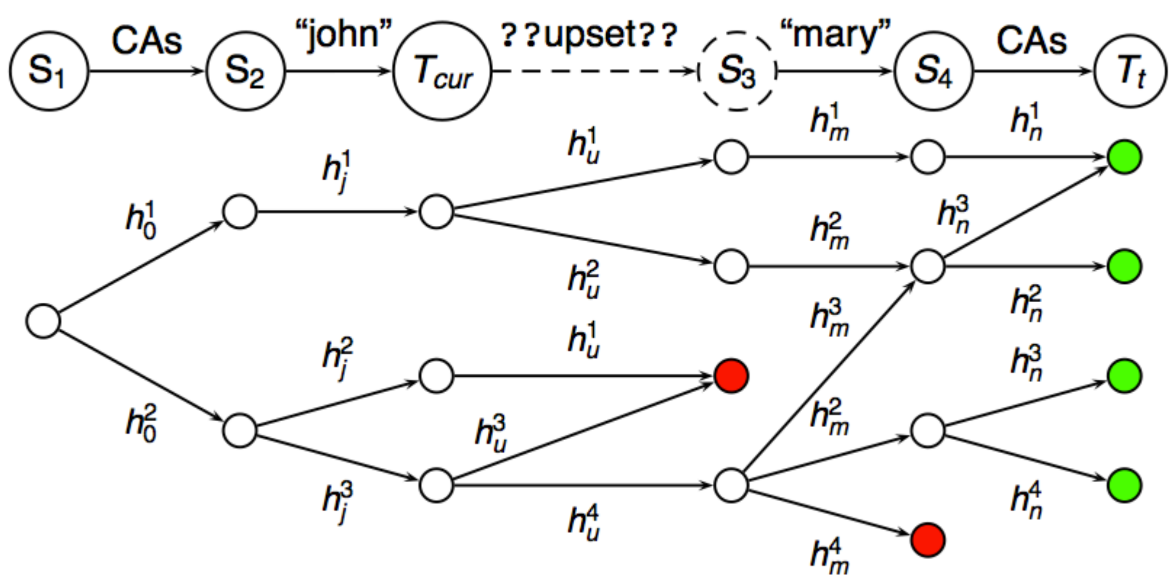
x1 = speaker	: e
x2 = ticket	: e
x3 = paris	: e
p' = to(x3,x2)	: t
p = want(x1,x2)	: t

Incremental Grammar Induction

- Induction from semantics
  - for an incremental grammar
  - with incremental learning
  - (see IWCS 2013)

x = john
y = mary
p = upset(x,y)



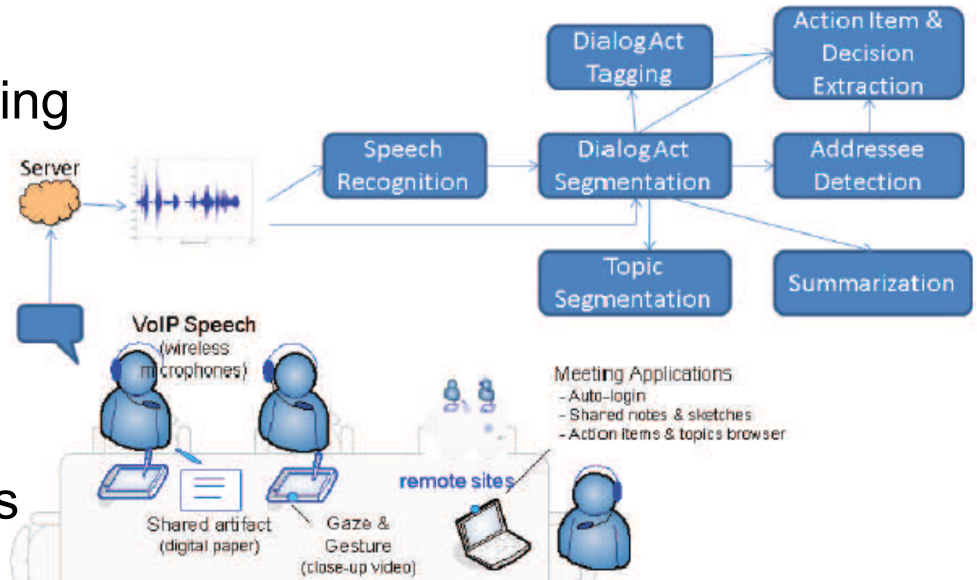


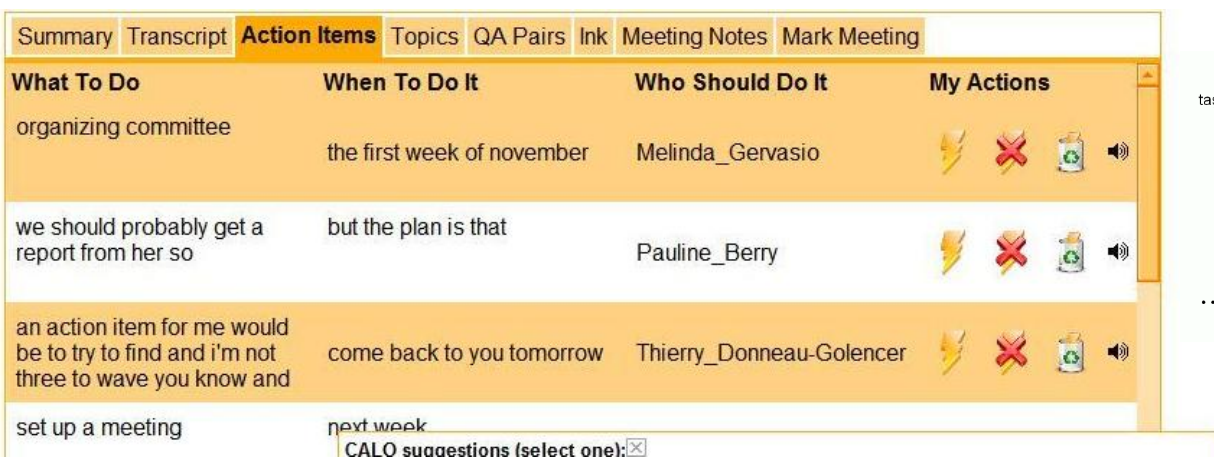
CIS centre for intelligent se

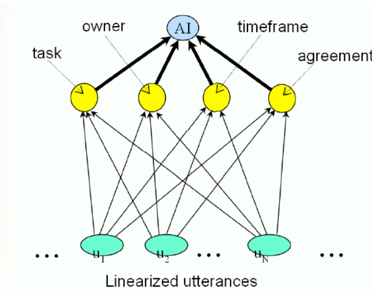
Queen Mary University of London

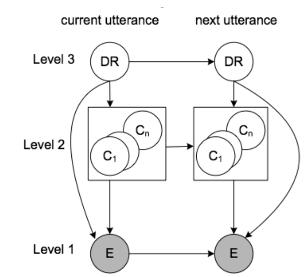
Human-Human Dialogue

- Dialogue modelling
  - Coordination and repair
  - Conversation & topic modelling
- Meeting assistance
  - e.g. decision detection
  - Robust structural modelling outperforms other approaches









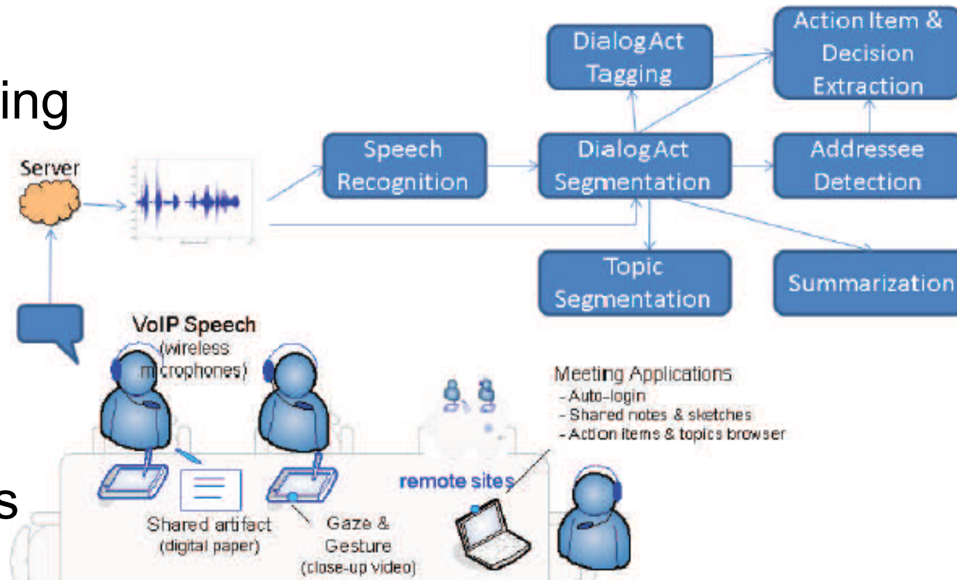
University of London

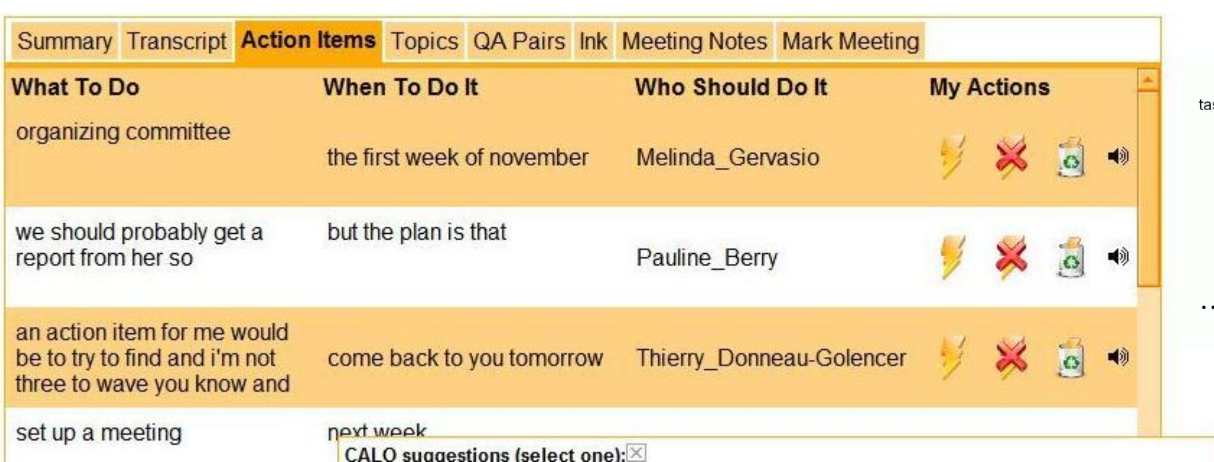
Human-Human Dialogue

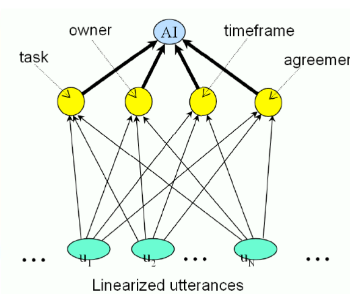
A:	not really. So there was the notion of the preliminary patent that uh
B:	yeah it is a cheap patent
…	…
A:	yeah and it is really broad you er don't have to
B:	yeah
C:	I actually think we should apply right away
D:	yeah I think that is a good idea
C:	I think you should I mean like this week start moving in that direction
…	…
A:	mhmm
D:	right

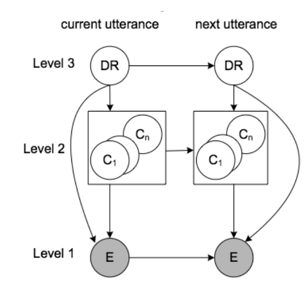
Human-Human Dialogue

- Dialogue modelling
  - Coordination and repair
  - Conversation & topic modelling
- Meeting assistance
  - e.g. decision detection
  - Robust structural modelling outperforms other approaches





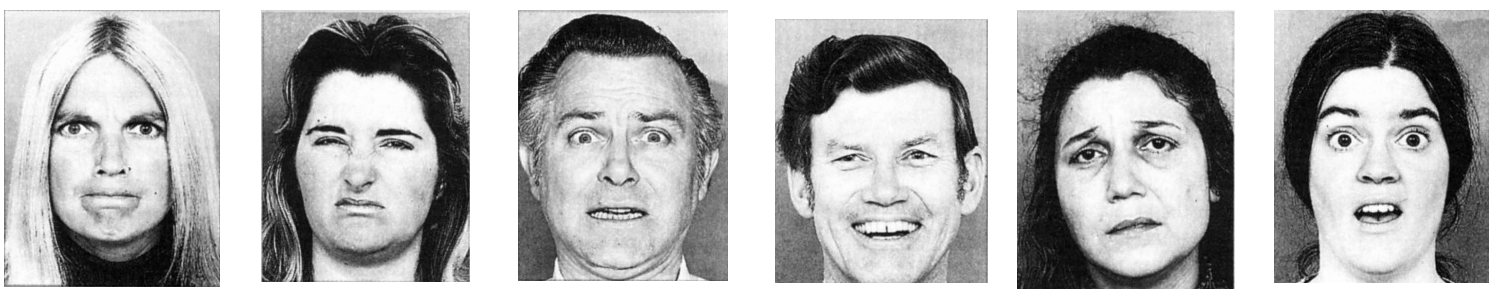




Queen Mary
University of London

Human-Human Dialogue - Online

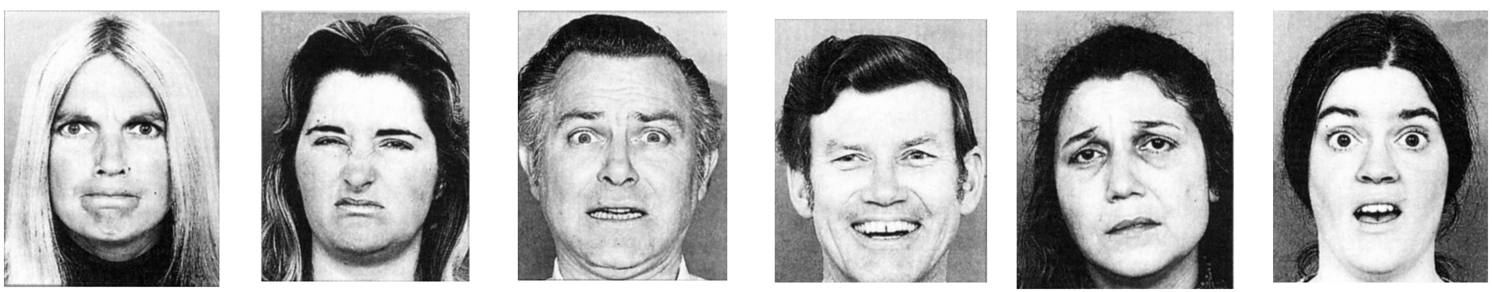
- Social media: Twitter, Facebook, Sina Weibo
  - Light/distant supervision, scalability
- Sentiment, emotions, …
  - `Nyt alexx tweetdreamsh RT @JDBAustralia: Goodnight everyone, i will tweet you all tomorrow <3`
  - `Gets so #angry when tutors don't email back... Do you job idiots! :@`
  - 考完它我就能回家啦~[鼓掌][鼓掌][鼓掌][鼓掌]开心O(∩_∩)O~~



- Conversation: (dis)agreement, support, opinions …
  - Conversation-based engagement & influence metrics

Human-Human Dialogue - Online

- Social media: Twitter, Facebook, Sina Weibo
  - Light/distant supervision, scalability
- Sentiment, emotions, …
  - Nyt alexx tweetdreamsh RT @JDBAustralia: Goodnight everyone, i will tweet you all tomorrow
  - Gets so angry when tutors don’t email back... Do you job idiots!
  - 考完它我就能回家啦~ 开心 ~~

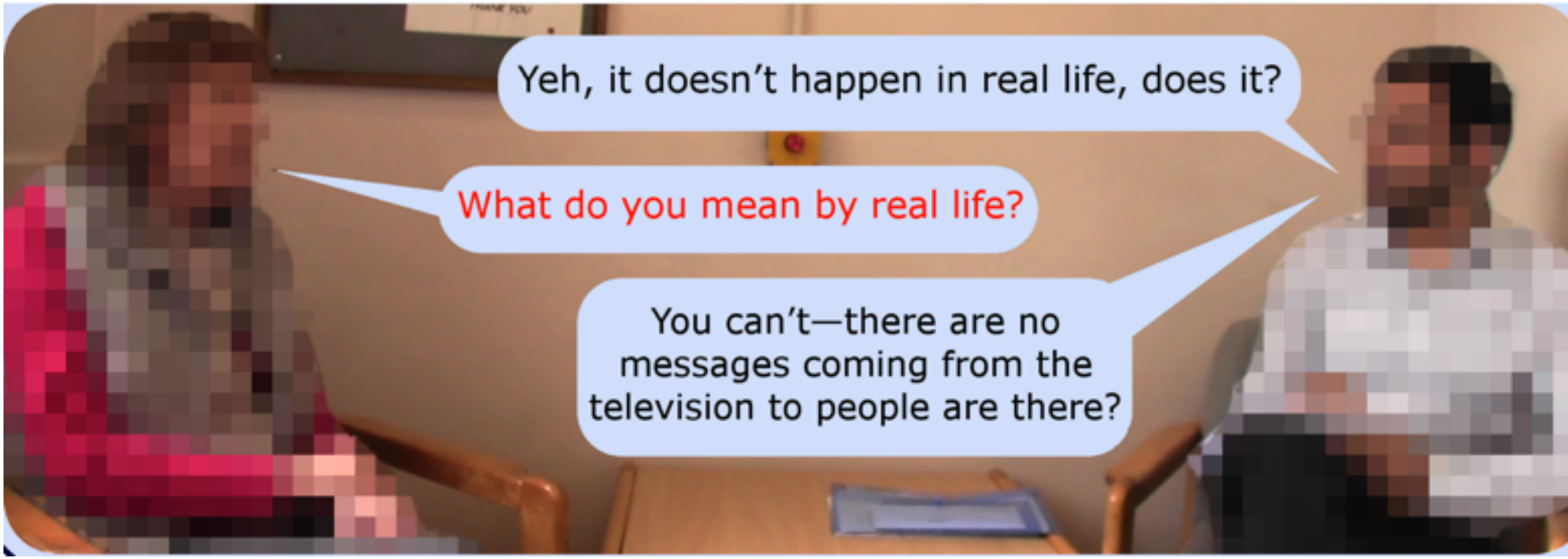


- Conversation: (dis)agreement, support, opinions …
  - Conversation-based engagement & influence metrics

Human-Human Dialogue - Online

chatterbox.co

Doctor-Patient Communication



- Language processing for psychiatric therapy:
  - Diagnosing symptoms
  - Predicting outcomes

Schizophrenia

- Mental ill health nearly half of all ill health in UK
  - (Layard et al, 2012)
- Schizophrenia: a serious but treatable condition
  - estimated to affect 400,000 people in England
  - Positive symptoms: delusions, hallucinations, beliefs
  - Negative symptoms: withdrawal, blunted affect, alogia
- *Non-adherence to treatment* a significant problem
  - Risk of relapse 3.7 times higher (Fenton et al, 1997)
  - About half of patients are non-adherent in the year after discharge from hospital (Weiden & Olfson, 1995)

Shared Understanding

- 60% of variance in treatment outcomes due to *non-specific effects*
  - Shared understanding is one (McCabe et al, 2002)
- Can dialogue features help us predict outcomes?
  - Including (non-)adherence
  - Help improve treatment
- Study of 138 patients: consultation dialogue & adherence 6 months later
- Dialogue structure: *repair*
  - Coordination and building shared understanding
- Dialogue content: *topics*
  - Patients focus on symptoms, doctors on treatment

Repair in Therapy Dialogue

- Self-repair (e.g. P1SISR)

Dr:	You probably have seen so many psychiatrists **o o over the years**

Dr:	**Did you feel that did you despair so much that** you wondered if you could carry on

Repair in Therapy Dialogue

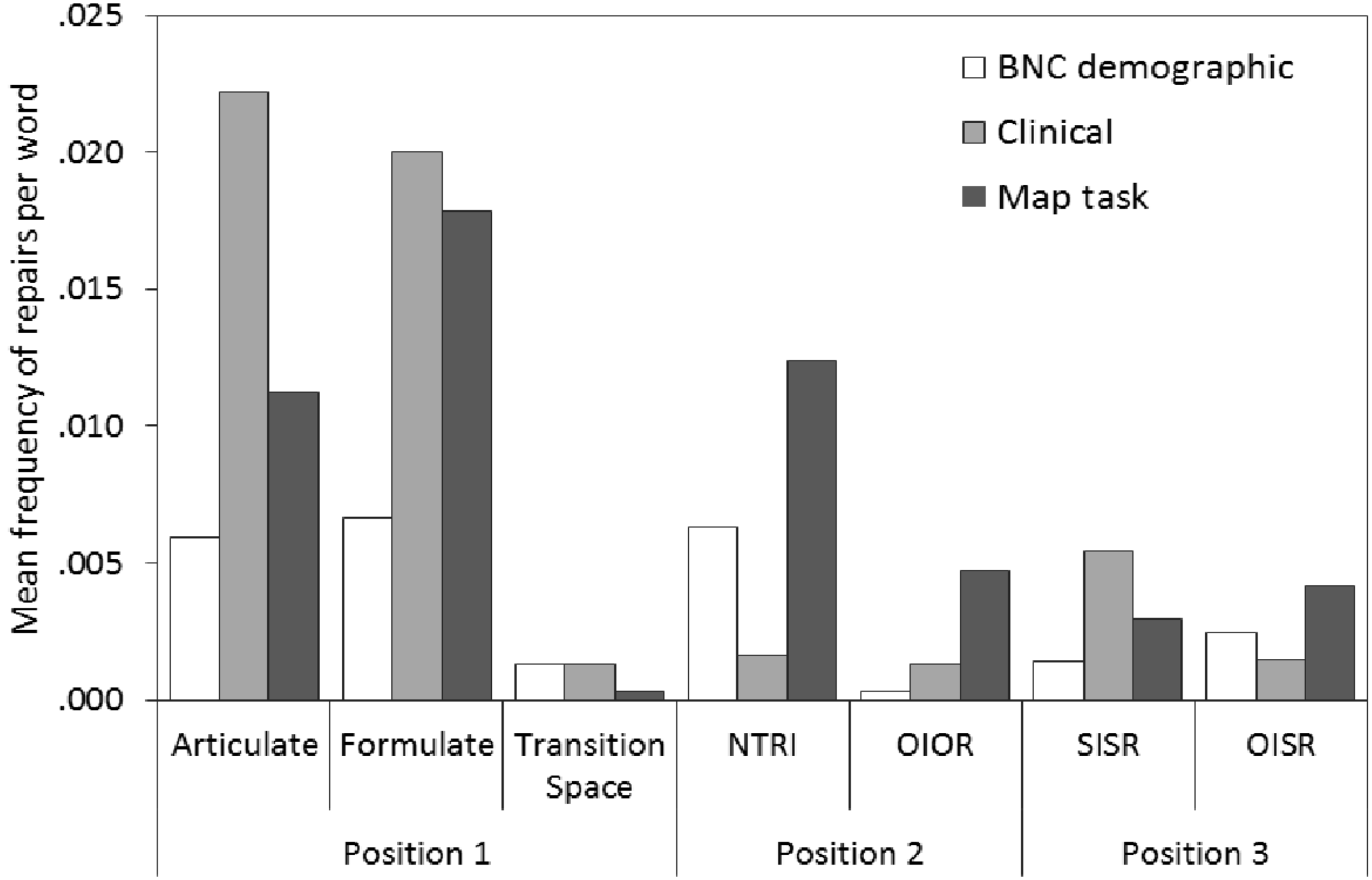
- Other-repair (e.g. P2OIOR)

Dr:	Rather than **the diazepam** which I don't think is going to do you any good
P:	**the valium**

- Repair initiation (e.g. P2NTRI then P3OISR)

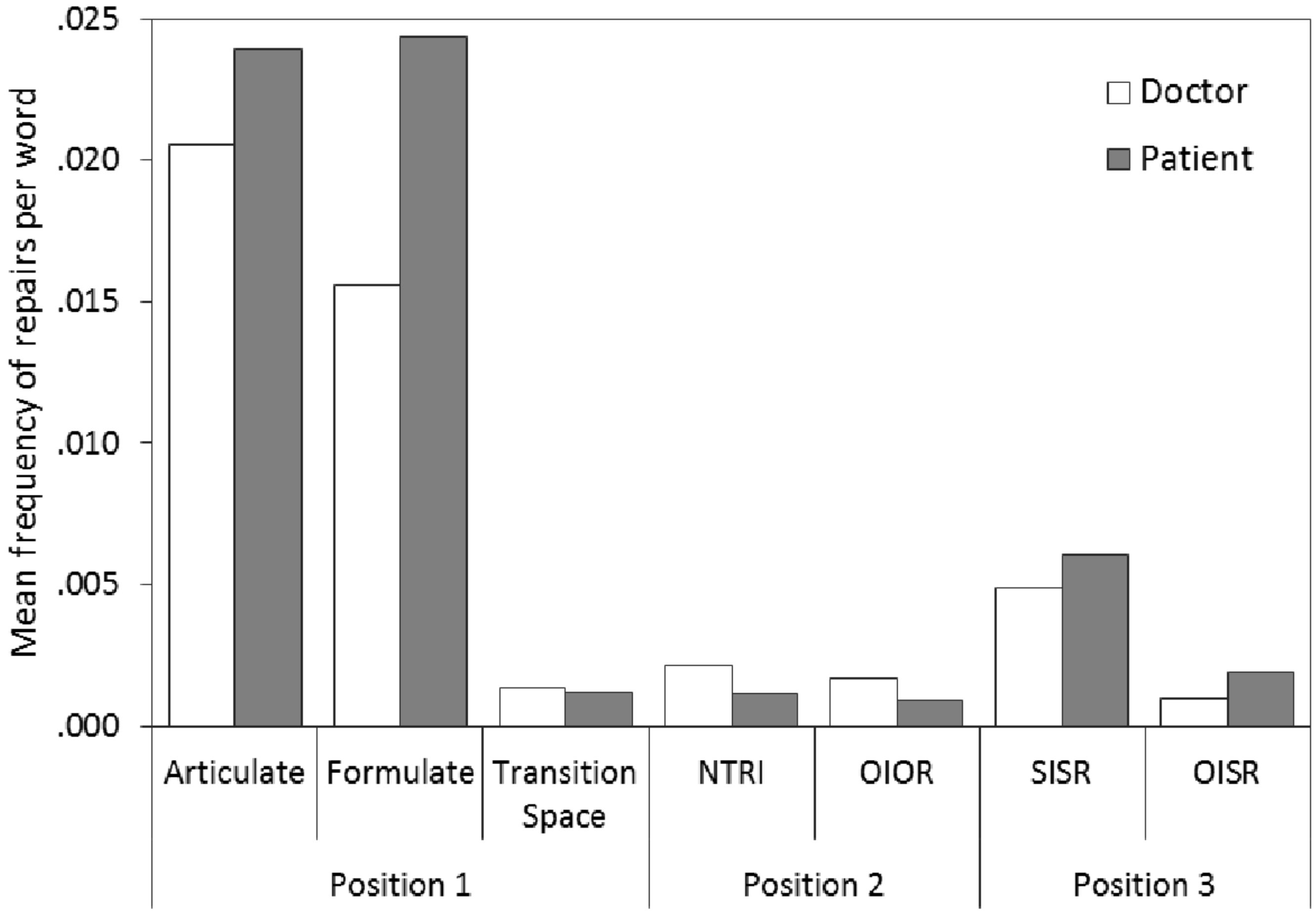
Dr:	Yeh, it doesn't happen in real life does it?
P:	**What do you mean by real life?**
Dr:	**You can't - there are no messages coming from the television to people are there?**

Comparison with other dialogue contexts



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

Patient-doctor comparison



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

Automatic repair detection

- Can we detect repair? And thus predict outcomes?
- Can we approach this like e.g. dialogue act tagging?
  - Supervised discriminative classification
- It's a very sparse phenomenon: **0.8% of turns**

- 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
- Emotion tagging in suicide notes (Liakata et al, 2012)
  - 45% f-score overall, as low as 5% for sparse categories

Features

- Sometimes we see specific lexical / phrasal items:

Dr:	Ok you have done it before
P:	**Pardon?**
Dr:	If you have done it before

Dr:	Who is your GP now
P:	**What?**
Dr:	Who is your GP

P:	They're not negative erm but they're positive as i eh erm um it's like imagining how your life will be
Dr:	Ok, ok, ok so thinking about how
P:	**Do you know what I'm talking about?**

Features

- Sometimes it's more complex than that

Dr:	Yep well that is a possible side effect
P:	**Side effect?**
Dr:	Of the err Haliperidol

Dr:	One thing that I ask you is when you were low in mood did you have suicidal thoughts
P:	**Did I have ...?**
Dr:	Suicidal thoughts

Dr:	Paroxitine
P:	**Fluoxitine**
Dr:	Ah Fluoxitine

Features

- Sometimes it's more complex than that

Dr:	Who's your key worker there do you know
P:	**Err the person who comes to see me?**
Dr:	Yeah the person you see most often

Dr:	Do you do you really feel it or is it a sensation
P:	**Is it what I'm thinking is that what you mean?**
Dr:	No is it just err the mind playing tricks on you or is it something

Dr:	Aaa so have you had any more thoughts about studying
P:	**What music?**

Method

- 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 the unigrams used (patient-only to avoid doctor specificity)

- Train SVMs to detect NTRIs & P2Rs
  - 44,000 turns of which 567 NTRIs (159 patient), 830 P2Rs (262)
  - 5-fold cross-validation

Results – repair detection

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

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

- 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?

Results – non-adherence prediction

- Apply to entire dialogues (patient turns only):

Features	P (%)	R (%)	F (%)
Class of interest	28.9	100.0	**44.8**
High-level	27.0	51.9	**35.5**
+ repair features	27.0	51.9	**35.5**
Best features	70.3	70.3	**70.3**

- Similar for symptoms, some outcomes e.g. HAS, PEQ
- Human psychiatrist given same task:

Data	P (%)	R (%)	F (%)
Text transcripts	60.3	79.6	**68.6**
Transcripts + video	69.6	88.6	**78.0**

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

Lexical features: topical content?

- Predicting non-adherence:

air	fill	mates	simply
anyone	finished	monthly	sodium
balanced	fish	mouse	stable
bleach	flashbacks	nowhere	stock
build	grass	pains	symptoms
building	grave	possibly	talks
busy	guitar	pr	teach
challenge	h	recent	terminology
chemical	hahaha	removed	throat
complaining	lager	ri	virtually
cup	laying	schizophrenic	was
dates	lifting	sensation	wave
en	lucky	sickness	worse

Lexical features: topical content?

- Predicting patient evaluation scores:

20th	electric	onto	sometime
ages	energy	overweight	son
angry	environment	oxygen	standing
anxiety	experiencing	packed	stomach
background	facilities	percent	suddenly
bladder	friendly	personally	sundays
booked	helps	picture	suppose
boy	ignore	played	table
broken	immediately	programs	team
bus	increased	progress	television
certificate	irritated	provide	thursdays
dead	kick	public	troubles
deep	later	quid	uhhm
drunk	lee	radio	upsetting
earn	loose	realised	walks
eeerrrr	low	reply	watchers

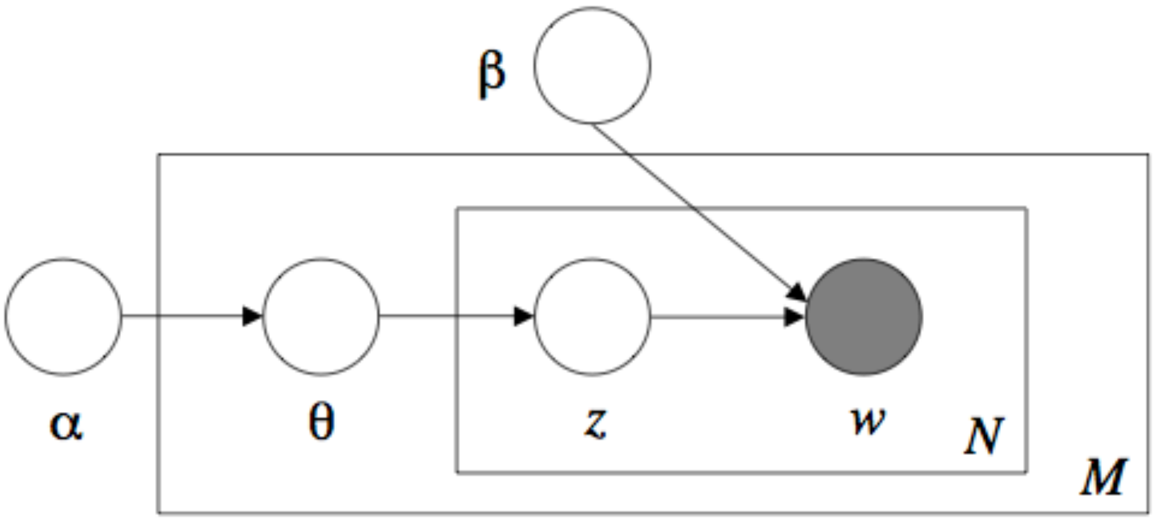
Topic modelling

- Higher-level information: topic
  - Could this be more generalisable, while providing more insight?
- Existing manual definition of 20 “topics”
  - Medication, side-effects, treatment, management
  - Symptoms, health, self-harm
  - Daily activities, living situation, relationships, …

Topic	Name	Description
01	Medication	Any discussion of medication, excluding side effects
02	Medication side effects	Side effects of medication
03	Daily activities	Includes activities such as education, employment, h
04	Living situation	The life situation of the patient, including housing, fi
05	Psychotic symptoms	Discussion on symptoms of psychosis such as halluc
06	Physical health	Any discussion on general physical health, physical i
07	Non-psychotic symptoms	Discussion of mood symptoms, anxiety, obsessions,
08	Suicide and self harm	Intent, attempts or thoughts of self harm or suicide (
09	Alcohol, drugs & smoking	Current or past use of alcohol, drugs or cigarettes an

Automatic topic modelling

- Can we learn topics from the data?
  - Latent Dirichlet Allocation (Blei et al, 2003)
  - Unsupervised generative approach



- Apply to dialogue data:
  - “document” = therapy dialogue = patient
  - “topic” = probability distribution over words

Automatic 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 livir

Automatic topic modelling

- LDA topics given manual “interpretations”:

	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	Reasurrance/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

Automatic & manual topics

- Cross-correlations across dialogues:

Manual	Automatic	R
Medication	Medication regimen	0.64
Psychotic symptoms	Making sense of psychosis	0.36
Psychotic symptoms	Psychotic symptoms	0.50
Physical health	Physical health	0.60
Non-psychotic symptoms	Sleep patterns	0.38
Alcohol, drugs and smoking	Substance use	0.65
General chat	Sectioning/crisis	0.36

- Correlations with symptoms:
  - Manual “psychotic symptoms” v PANSS pos: 0.49
  - Auto “making sense of psychosis” v PANSS pos: 0.38

Outcome prediction using topics

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

Measure	**Manual Acc (%)**	**LDA Acc (%)**
HAS Dr	**75.8**	**75.0**
HAS P	59.0	53.7
PANSS positive	**61.1**	58.8
PANSS negative	**62.1**	56.1
PANSS general	59.5	53.4
PEQ comm	59.7	56.7
PEQ comm barr	**61.9**	**60.4**
PEQ emo	57.5	57.5
Adherence (balanced)	**66.2**	54.1

Conclusions

- We can detect repair quite well
  - . . . but it's too sparse to predict outcomes
- We can predict some outcomes (including adherence)
  - . . . without knowing very much about the language
  - . . . quite well with specific lexical features
  - . . . OK (for some) with manually defined topics
- We can detect topics similar to manual topics
  - . . . but they're good for some things, bad for others

- Should we separate form from content? How?
- Can we do better with multi-modal processing?