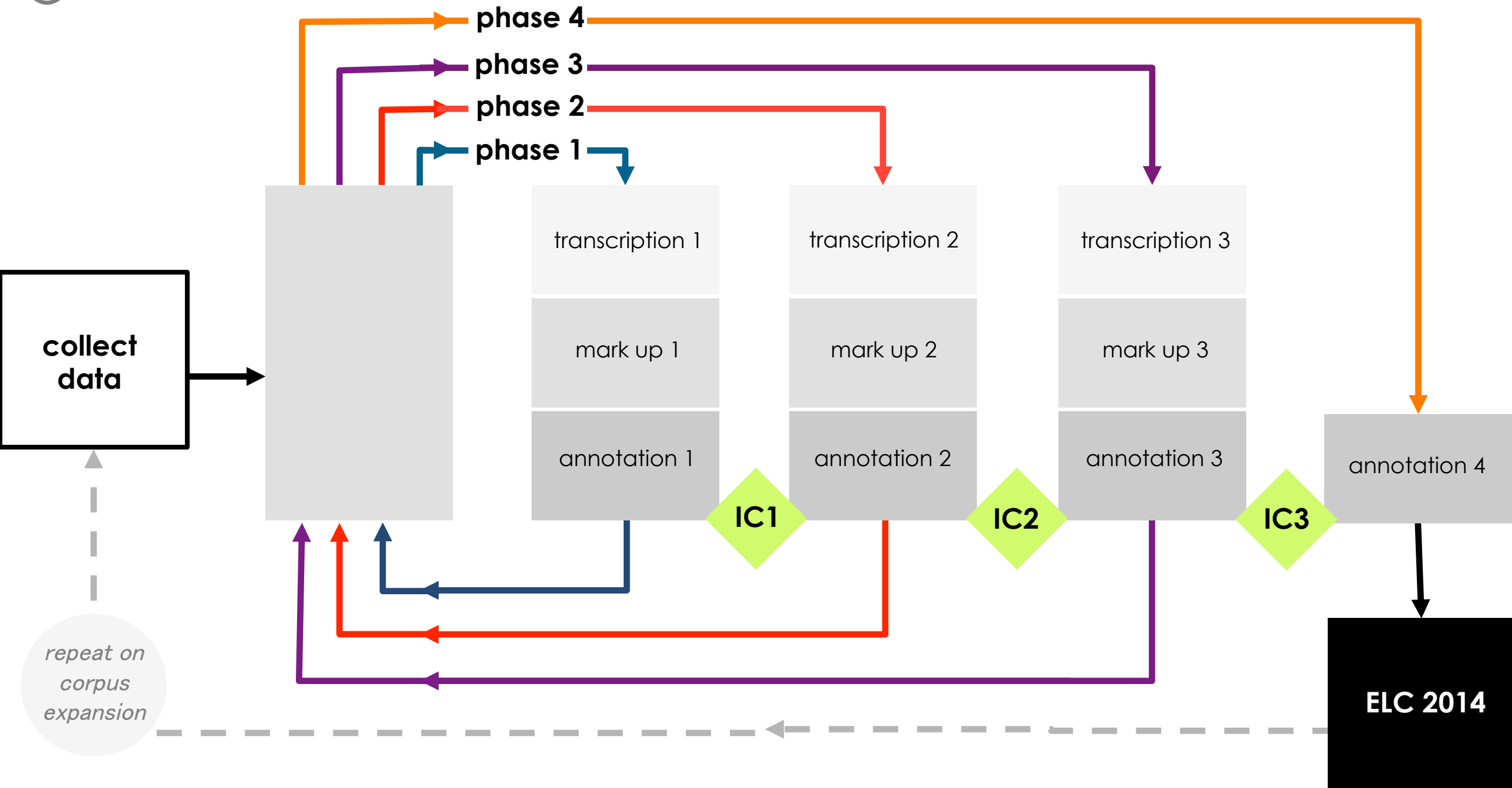
A cyclical approach to pragmatic annotation:

Our experience with the Engineering Lecture Corpus (ELC)

Siân Alsop and Hilary Nesi

Coventry University

15th Corpus Linguistics in the South
Faculty of Education, University of Cambridge, 28/10/17


phase 4
phase 3
phase 2
phase 1
collect data
transcription 1
transcription 2
transcription 3
mark up 1
mark up 2
mark up 3
annotation 1
annotation 2
annotation 3
annotation 4
IC1
IC2
IC3
repeat on corpus expansion
ELC 2014

elc.coventry.domains/about

data collection

- Hilary Nesi (Coventry University), Lynn Grant (Auckland University of Technology), and Ummul Khair Ahmad (Universiti Teknologi Malaysia), funded by the British Council (PMI 2 Connect Research Cooperation, British Council (RC 90)).
- Lectures are mostly in civil, mechanical and electrical engineering - similar topics are often covered in the different cultural/educational contexts.
- Lectures recorded using AV equipment and vary in duration between 41-104 minutes.
- Delivered across undergraduate degree programmes (from years 1-3), largely as part of courses that are mandatory to the programmes.
- A range of lecturers were filmed from each institution.

elc.coventry.domains/about

data collection

Summary of ELC holdings

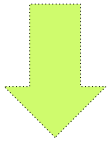
		Coventry University (United Kingdom)	Universiti Teknologi Malaysia (Malaysia)	Auckland University of Technology (New Zealand)
abbreviation		UK	MS	NZ
identifier series		1000-1030	2000-2018	3000-3028
token size		156838	120211	251108
engineering type	civil	27	6	0
	electrical	0	0	17
	graphics	0	0	3
	mechanical	3	12	2
	fluid mechanics	0	0	3
	solid mechanics	0	0	3
total lectures		30	18	28
total lecturers		4	9	4
average lecture length (tokens)		5228	6678	8968

data annotation

Categories: prayer | housekeeping | defining | summary | story | humour

- The ELC is manually annotated using a fairly experimental taxonomy – this presents specific challenges in terms of defining (as well as measuring) agreement.
- The focus is on the reproducibility of annotation decisions - to ensure consistency and to provide a foundation for future expansion and comparability.
- The ELC annotation does not pose a classical classification problem because it indexes strings of text rather than single lexical items.
- Exact matches are not the norm, so calculating the fuzziness of the boundaries that different annotators identify is a more valuable test of the reliability of ELC annotations.
- The three steps of IAR testing correspond to various needs at each point between the four Phases. They also represent an increasingly coarse approach to agreement at each pass as confidence in both the tagset and indexing grew.

phase 1



Transcription 1	- Raw text - Local language experts and subject specialists - 461136 tokens
Mark up 1	- Structural metadata (e.g. utterance tags, pauses) - Guided by principles of reuse, merging and comparison across systems - TEI-compliant - 26473 tags
Annotation 1	- 354 examples annotated - 3 significant changes to the working list categories

phase 1

Annotation adjustment 1

Working list
prayer
housekeeping
defining term
review lecture content
preview lecture content
personal narratives
teasing
self-recovery
self-denigration
black humour
disparagement of out-group member
mock threat
register and wordplay
greetings
reference to students' future profession

Adjustment 1	
element	*attribute*
prayer	
housekeeping	
defining	
summary	review lecture content preview lecture content
story	personal narrative
humour	bawdy black disparagement irony jokes mock threat playful sarcasm self-denigration teasing wordplay

inter-coder reliability 1

$$agreement(A,B) = \frac{(\text{indices\_A} \cap \text{indices\_B})}{(\text{indices\_A} \cup \text{indices\_B})}$$

Calculating intersection agreement

4 independent annotators: lead annotator (rf1002), 2 local experts (rm1001 and rf3001) who did partial markup and partial annotation of transcripts in two subcorpora (one each), and an independent language expert (rm1003)

<story type="exemplum"_ANN1>ok the other one if we look <story type="exemplum"_ANN2>remember last time when I show you the er Sampoong er which is the one which happen in Korea this is what happen when you don't properly consider design yeah ok so they put up a new equipment on the roof top of the building and this is very heavy they push along the slab ok which is not practical ok so I push forward a bit ok this is the the events yeah that lead to the failure ok ok see what happen alright so that is a more serious case related to failures yeah hopefully none of our students in the future will be what we call it er will be relat- will be what we call it er link with such event yeah if you look at the video just now er many people were found guilty and some of them were sentence to several years in jail er including the the C E O I believe some of the engineers as well negligence yeah greedy</story_ANN1> ok those are the thing that happen at the at the what we call it at the projects</story_ANN2>

Results (in tokens, not including annotation/markup)
String length = 194
Ann1 = 176
Ann2 = 187
Intersection = 169 tokens / 87% (0.87 probability)

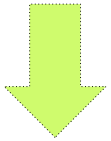
inter-coder reliability 1

IAR test 1: agreement probability based on annotator pairs per pragmatic element

annotator pairs		humour	story	summary
rf1002 vs rm1001	overlap probability	0.46	0.50	0.62
	rf1002 total	262	64	433
	rm1001 total	59	10	34
	overlaps	29	7	30
	misses	231	60	407
rf1002 vs rf3001	overlap probability	0.49	0.71	0.69
	rf1002 total	59	45	327
	rf3001 total	29	21	117
	overlaps	0.46	15	88
	misses	231	36	268
rf1002 vs rm1003	overlap probability	0.25	0.40	0.53
	rf1002 total	18	18	93
	rm1003 total	30	67	99
	overlaps	6	13	50
	misses	36	59	92
rf1002 vs all	overlap probability	**0.40**	**0.54**	**0.61**

The results were low, but encouraging. The average of 52% between the lead and other annotators falls within the “fair to good beyond chance” range (cf. Capozzoli, McSweeney and Sinha 1999). The complete misses were particularly useful.

phase 2



Transcription 2	- 17.69% change (81575 tokens)
Mark up 2	- 11119 tags changed / added / removed (42%)
Annotation 2	- 1 significant change to the categories - 671 instances changed / added / removed (190%)

phase 2

Annotation adjustment 2

Adjustment 1	
element	*attribute*
prayer	
housekeeping	
defining	
summary	review lecture content preview lecture content
story	personal narrative
humour	bawdy black disparagement irony jokes mock threat playful sarcasm self-denigration teasing wordplay

Adjustment 2	
element	*attribute*
prayer	
housekeeping	
defining	
summary	review lecture content preview lecture content
story	personal narrative professional narrative
humour	bawdy black disparagement irony jokes mock threat playful sarcasm self-denigration teasing wordplay

inter-coder reliability 2

IAR test 2: checking specific agreement

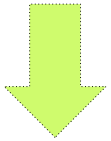
- 14 participants (two groups of 8 and 6) – all language experts with no prior connection to the project
- manually encoded a hardcopy sample of ELC transcripts
- 3 x 500 word strings (1 per subcorpus)
- the only selection criteria were that at least two sets of annotation indices (start and end) as identified by the lead annotator were present in the sample
- all participants given a description of the categories and annotation principles
- clips of video data aligned to the samples shown

inter-coder reliability 2

	sample 1 (1010)			**sample 2** (2017)			**sample 3** (3024)			*average*
	humour	*story*	*summary*	*humour*	*story*	*summary*	*humour*	*story*	*summary*	
p1	0.59	0.68	0.92	0.41	0.69	1.00	0.85	0.71	0.79	0.74
p2	0.67	0.78	0.85	0.67	0.77	0.95	0.52	1.00	1.00	0.80
p3	0.82	0.21	1.00	0.60	0.65	1.00	0.60	0.67	1.00	0.73
p4	0.76	0.69	1.00	0.76	0.70	0.81	0.64	0.90	0.85	0.79
p5	1.00	0.72	0.76	0.87	0.85	0.62	0.59	0.82	0.42	0.74
p6	0.43	0.80	0.72	0.56	0.78	0.90	0.89	0.95	0.80	0.76
p7	0.87	0.56	0.95	0.89	0.81	0.82	0.45	0.84	0.51	0.74
p8	1.00	0.70	0.81	1.00	0.90	0.57	0.53	1.00	0.93	0.83
p9	0.87	1.00	0.68	1.00	1.00	0.87	0.75	0.80	0.95	0.88
p10	0.65	0.65	0.92	0.74	0.81	0.98	0.67	0.80	0.90	0.79
p11	1.00	0.94	0.67	0.65	0.83	0.62	0.66	0.59	1.00	0.77
p12	0.82	0.79	1.00	0.89	0.33	0.85	0.76	1.00	0.74	0.80
p13	0.65	0.23	0.83	0.85	1.00	0.73	0.74	0.96	0.94	0.77
p14	0.90	0.80	0.82	0.59	0.96	0.79	0.55	0.70	0.81	0.77
average	0.79	0.68	0.85	0.75	0.79	0.82	0.66	0.84	0.83	0.78

More consistent intersections at this stage (0.78 average overlap), some variation, and no misses. This round pushed the reliability just into the “excellent beyond chance” category (Capozzoli, McSweeney and Sinha 1999), or at least towards acceptable reliability (Lombard, Snyder-Duch and Bracken 2002).

phase 3



Transcription 3	- 0.2% change (1106 tokens)
Mark up 3	- 876 tags changed / added / removed (2.33%)
Annotation 3	- 3 significant changes to the categories (the big one!) - 588 instances changed / added / removed (55%)

phase 3

Annotation adjustment 3

Adjustment 2	
element	*attribute*
prayer	
housekeeping	
defining	
summary	review lecture content preview lecture content
story	personal narrative professional narrative
humour	bawdy black disparagement irony jokes mock threat playful sarcasm self-denigration teasing wordplay

Adjustment 3	
element	*attribute*
prayer	
housekeeping	
explaining	defining reasoning translating
summary	review previous lecture content review current lecture content preview current lecture content preview future lecture content
story	anecdote exemplum narrative recount
humour	bawdy black disparagement irony / sarcasm jokes mock threat / playful self-denigration teasing wordplay

inter-coder reliability 3

IAR test 3: hit or miss

- intended to eliminate remaining examples of misses to ensure the highest possible overall reliability
- all instances of pragmatically annotated text at the finest level of element and attribute type were extracted for review by the project lead
- instead of identifying boundaries, the primary purpose of this final test was to accept or reject annotations
- some examples were queried for further discussion with the lead annotator

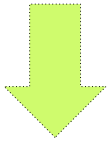
inter-coder reliability 3

IAR test 3: hit or miss

	humour	story	summary	total	average
a. initial number of annotated strings	695	161	1306	2162	
b. miscategorisation identified by rf1004 resulting in the removal of annotation	65	9	60	134	
c. partial miscategorisation identified by rf1004 resulting in either adjustment to attribute type, or contraction/expansion of annotation boundaries	92	12	86	190	
d. query raised by rf1004 resulting in no change to annotation	12	11	87	110	
e. final number of annotated strings (a-b)	630	153	1246	2029	
f. % boundary adjustment = (c/a)*100	14.60	7.89	6.90		9.80
g. % rejection = (100-(e/a*100))	9.35	5.59	4.59		6.51

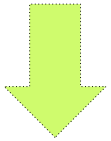
High agreement in this final failsafe test (0.84-0.93): the difference in the number of categories originally identified and those rejected by rf1004 (no overlap) is less than 10% in each category and less than 7% on average (row g), and the percentage of instances where boundaries were adjusted is between 7-15% and less than 10% overall (row f).

phase 4



Transcription 4	- No change
Mark up 4	- No change
Annotation 4	- 1 change to the categories - 305 instances changed / added / removed (20%)

phase 4



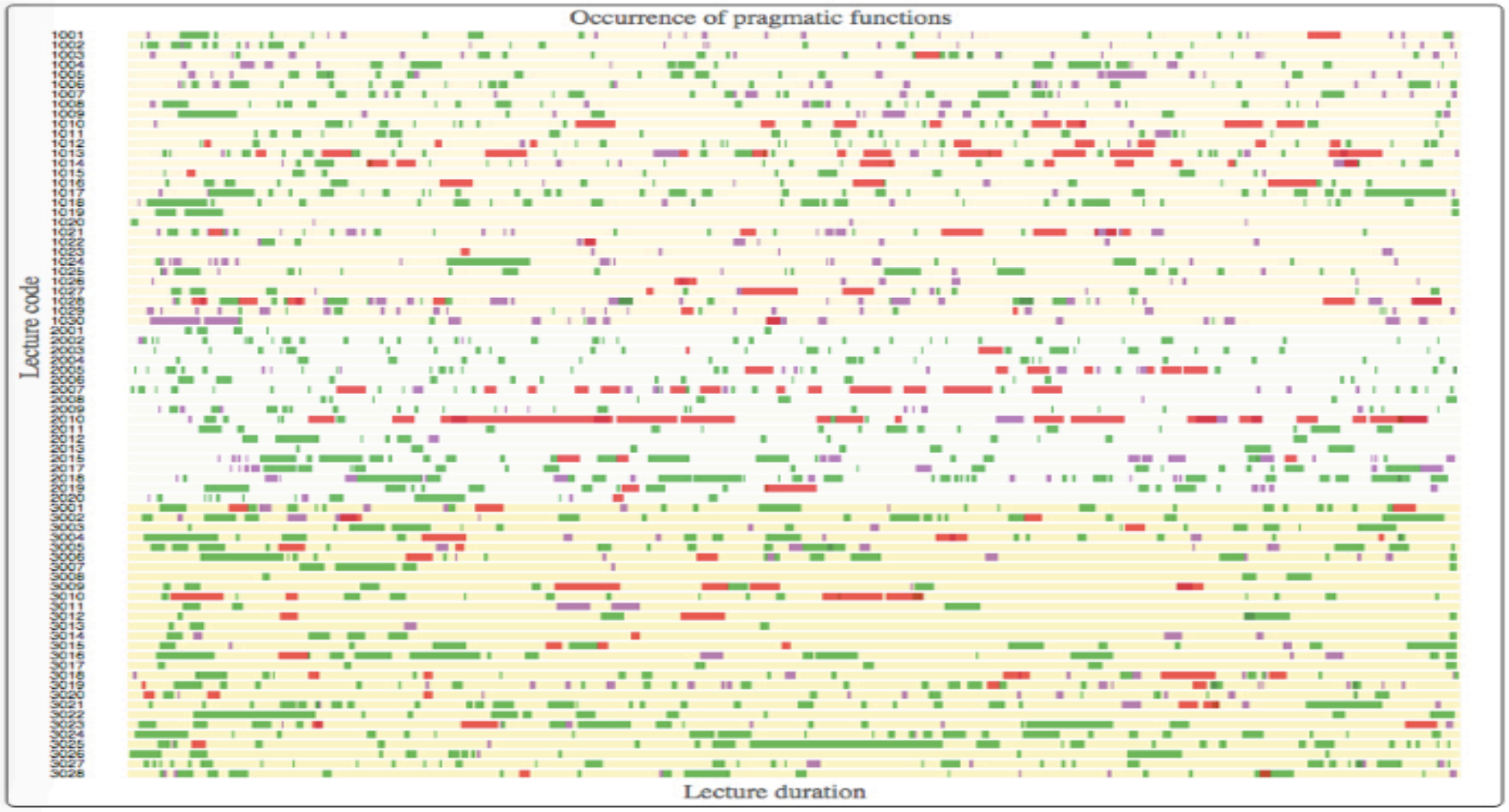
Annotation adjustment 4

Adjustment 3	
element	*attribute*
prayer	
housekeeping	
explaining	defining reasoning translating
summary	review previous lecture content review current lecture content preview current lecture content preview future lecture content
story	anecdote exemplum narrative recount
humour	bawdy black disparagement irony / sarcasm jokes mock threat / playful self-denigration teasing wordplay

Adjustment 4	
element	*attribute*
prayer	
housekeeping	
explaining	defining reasoning translating equating
summary	review previous lecture content review current lecture content preview current lecture content preview future lecture content
story	anecdote exemplum narrative recount
humour	bawdy black disparagement irony / sarcasm jokes mock threat / playful self-denigration teasing wordplay

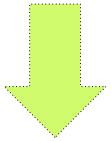
elc.coventry.domains/about

occurrence and duration of three elements



KEY: humour | story | summary

lessons learnt



1. Don’t annotate unstable texts
2. Use some form of version control
3. Manual annotation of pragmatic features is time-consuming
4. Humour is particularly tricky to classify
5. Establishing a taxonomy requires multiple cycles
6. The categories identified describe this set of lectures
7. We think the results are worth the effort …

references

Capozzoli, M., McSweeney, L., and Sinha, D. (1999) 'Beyond Kappa: A Review of Interrater Agreement Measures'. *The Canadian Journal of Statistics* 27 (1), 3-23

Lombard, M., Snyder-Duch, J., and Campanella Bracken, C. (2010) Intercoder Reliability: *Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects* [online] available from <http://matthewlombard.com/reliability/> [27/06/14 2014]