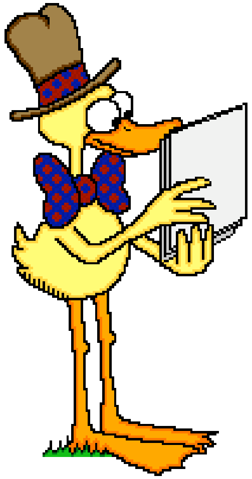
Document Understanding Conference

DUC 2007



Hoa Trang Dang

National Institute of Standards and Technology

April 26, 2007

NIST
National Institute of Standards and Technology

Thank You!

- 32 Participating teams from:
  - 11 countries
  - 5 continents (N. America, Europe, Asia, Africa, Australia)
- Assessors A, B, C, D, E, F, G, H, I, and J
- DUC 2007 Program Committee:
  - John Conroy, Donna Harman, Ed Hovy, Kathy McKeown, Drago Radev, Lucy Vanderwende
  - Karen Sparck-Jones

Document Understanding Conferences

- 2000 Summarization roadmap, progress:
  - simple genre ⇨ complex genre
  - simple tasks ⇨ demanding tasks
    - extract ⇨ abstract
    - single document ⇨ multiple documents
    - English ⇨ other language
    - generic summaries ⇨ focused or evolving summaries
  - intrinsic evaluation ⇨ extrinsic evaluation

DUC 2001-2006 Summarization

- for single, multiple newswire documents
- at various lengths (10 words, 100+ words)
- of various sorts (generic, viewpoint-oriented, query-oriented)
- comparing automatic summaries with manual ones
  - intrinsic: linguistic quality, content coverage, Rouge
  - extrinsic (simulated): usefulness, responsiveness

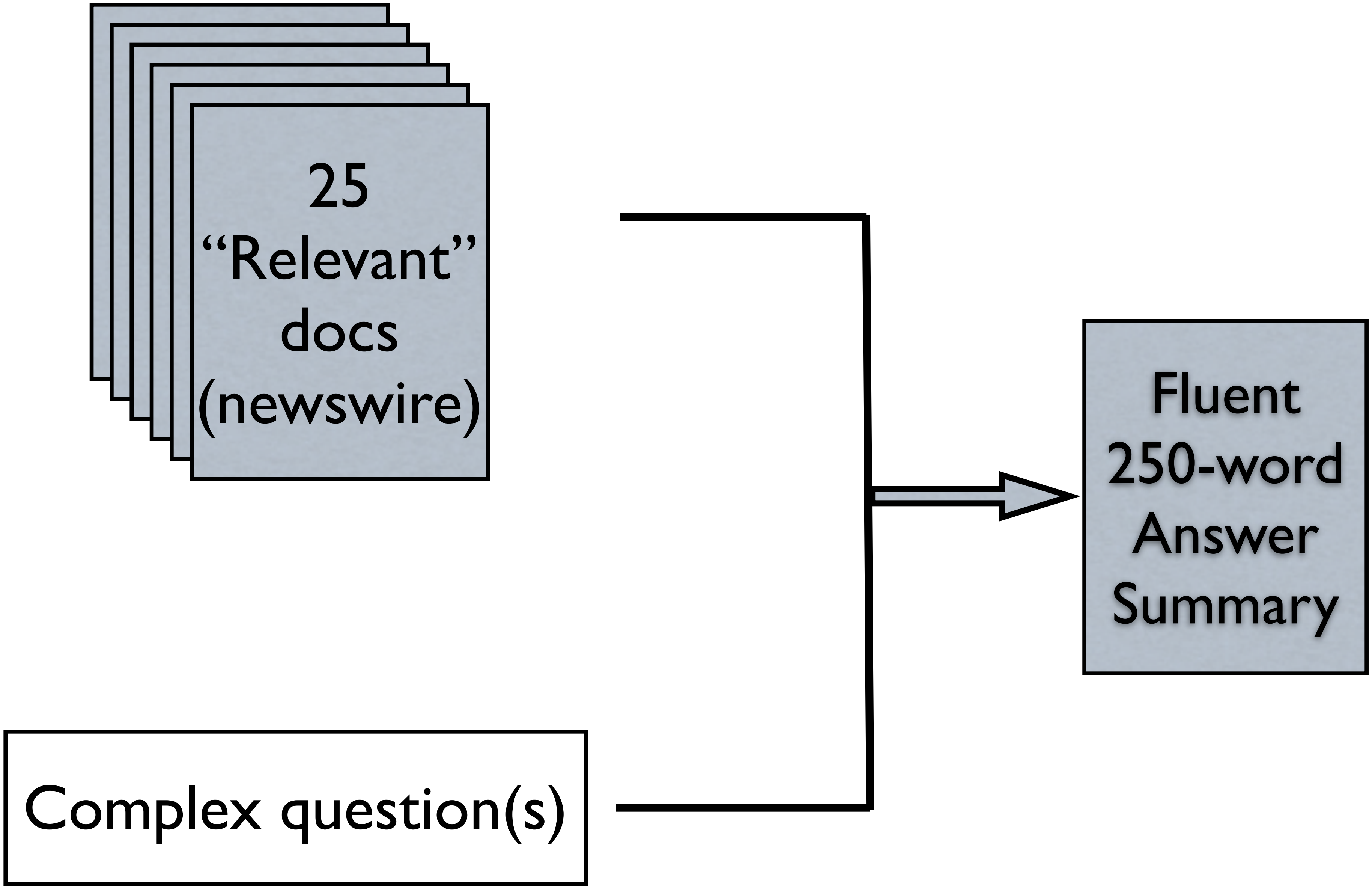
DUC 2007 Tasks and Evaluations

- Summaries focused by questions representing user need/interests

1. Main Task: 250 word-summary
    - length requires structuring of summary
    - evaluated for content, readability
2. Update Task: 100 word-summary
    - assumption of some user knowledge
    - evaluated for content

DUC 2007 Main Task

2005-2007 Question-focused task



25 “Relevant” docs (newswire)

Complex question(s)

Fluent 250-word Answer Summary

Example DUC 2007 Topic

- num: D0715D
- title: International Land Mine Ban Treaty
- narr: Which countries have signed the Ottawa Treaty for the elimination of anti-personnel land mines, and how many have ratified it? What countries have refused to sign, and why? How effective has the treaty been?

Main task: topics, documents, peers

- 45 topics developed by 10 NIST assessors
- Documents from AP, NYT, XIN newswire
- Model summaries written by 10 assessors (ID = A-J)
  - 4 model summaries per topic
- 30 participants (ID = 3-32)
- 2 Baselines (ID = 1-2):
  - Simple: first 250 words of most recent document
  - Generic: high-performance generic summarizer

Generic Baseline: CLASSY04

- Top in DUC 2004 (generic 100-word summary)
- Topic description is not used
- Sentence splitting/shortening taken from CLASSY07
- 5-state Hidden Markov Model
  - states represent hidden summary and non-summary sentences
- Observations: log(# signature terms + 1)
  - signature terms computed based on given clusters
- Pivoted QR to remove redundancy

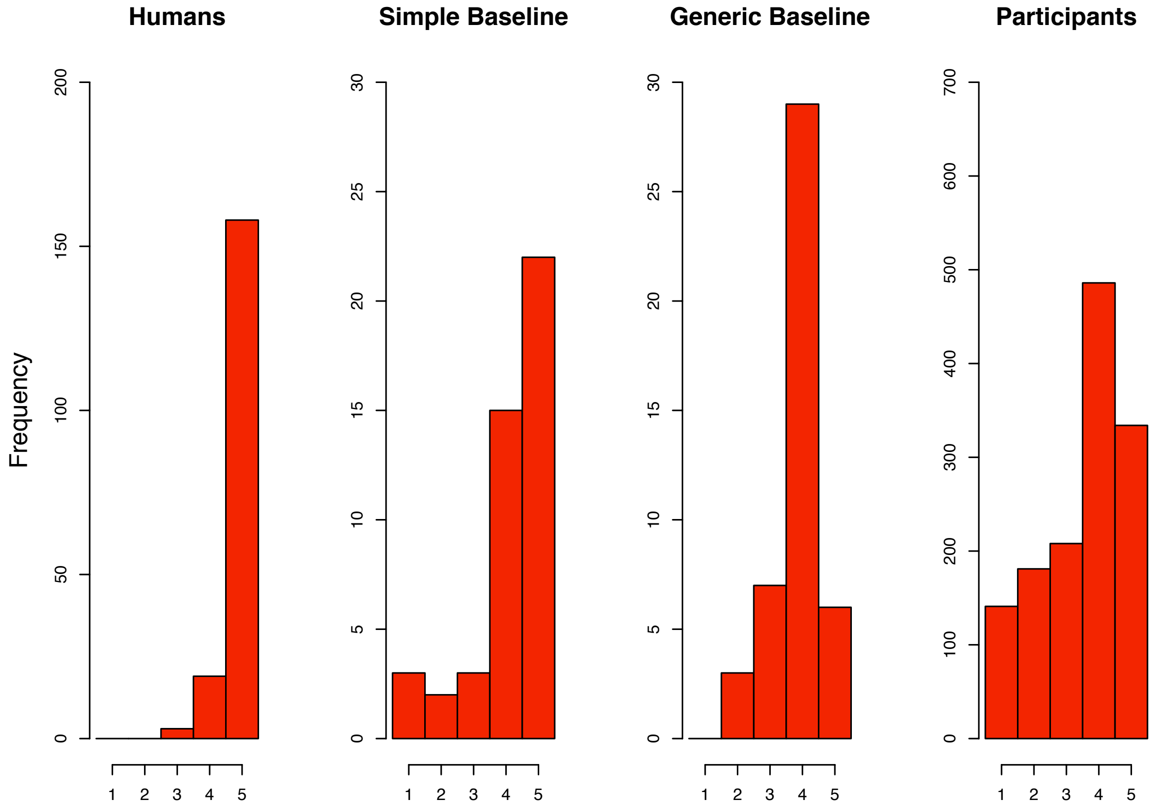
Evaluation methods

- Manual evaluation
  - Readability: 5 linguistic qualities
  - Content responsiveness
  - *Pyramids (optional, volunteer community effort)*
- Automatic evaluation of content
  - ROUGE-2, ROUGE-SU4 (stemmed, keep stopwords)
  - BE (HM)

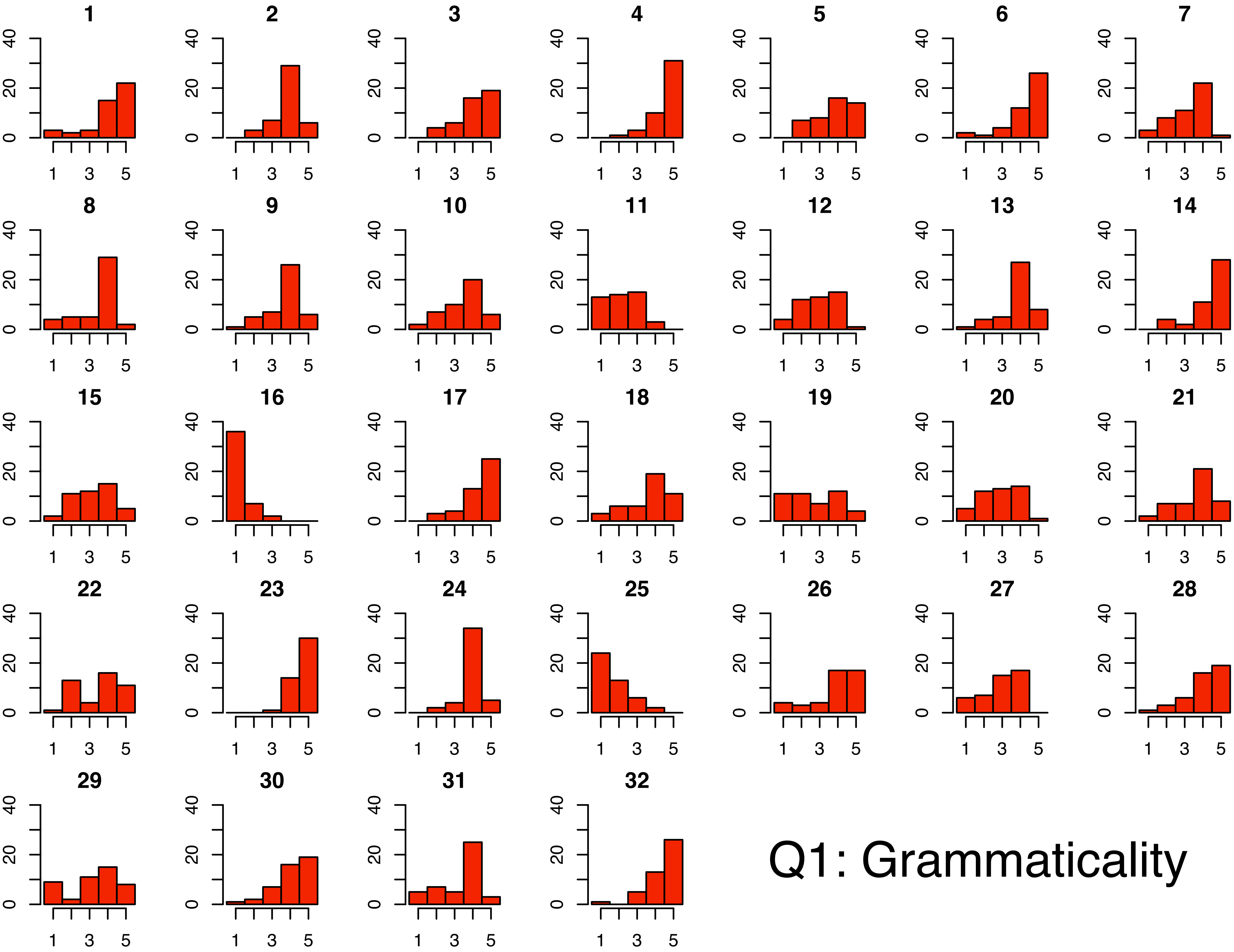
Manual evaluation

- 10 assessors
- One assessor/topic: linguistic quality, responsiveness
- Assessor is topic developer, a summarizer for topic
- Each score based on a 5-point scale
  - (1=very poor ... 5=very good)
- No manual assessment of overall responsiveness (content + linguistic quality)

Q1: Grammaticality

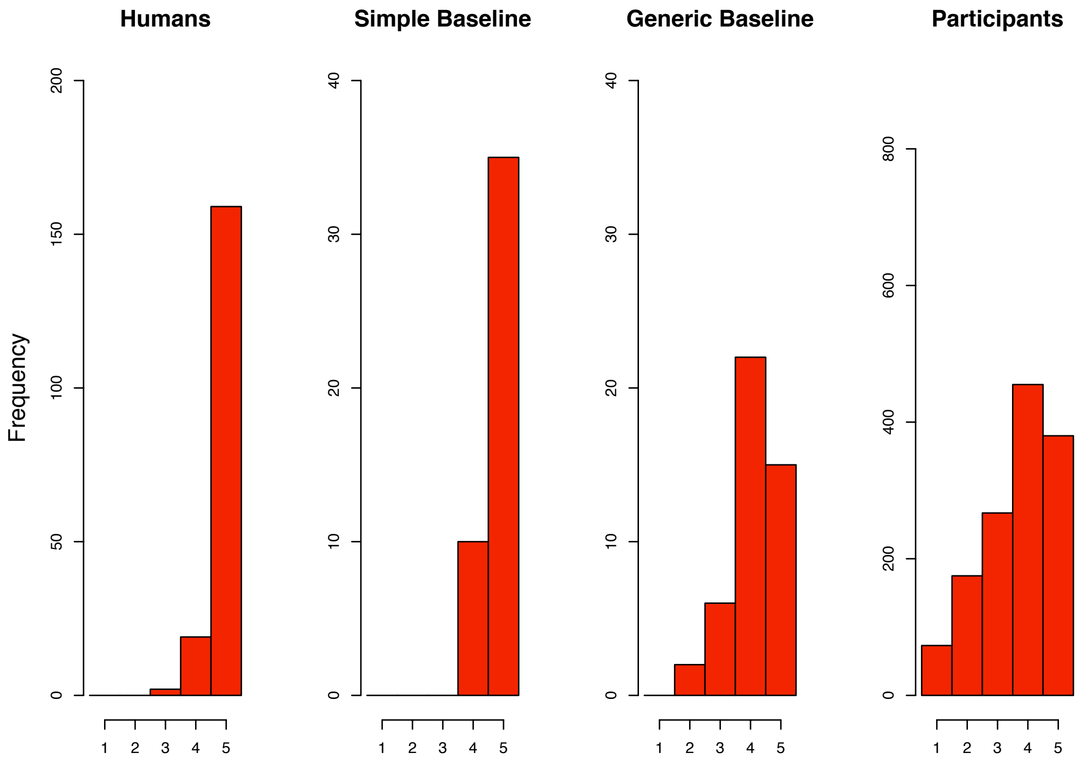


The summary should have no datelines, system-internal formatting, capitalization errors or obviously ungrammatical sentences (e.g., fragments, missing components) that make the text difficult to read.



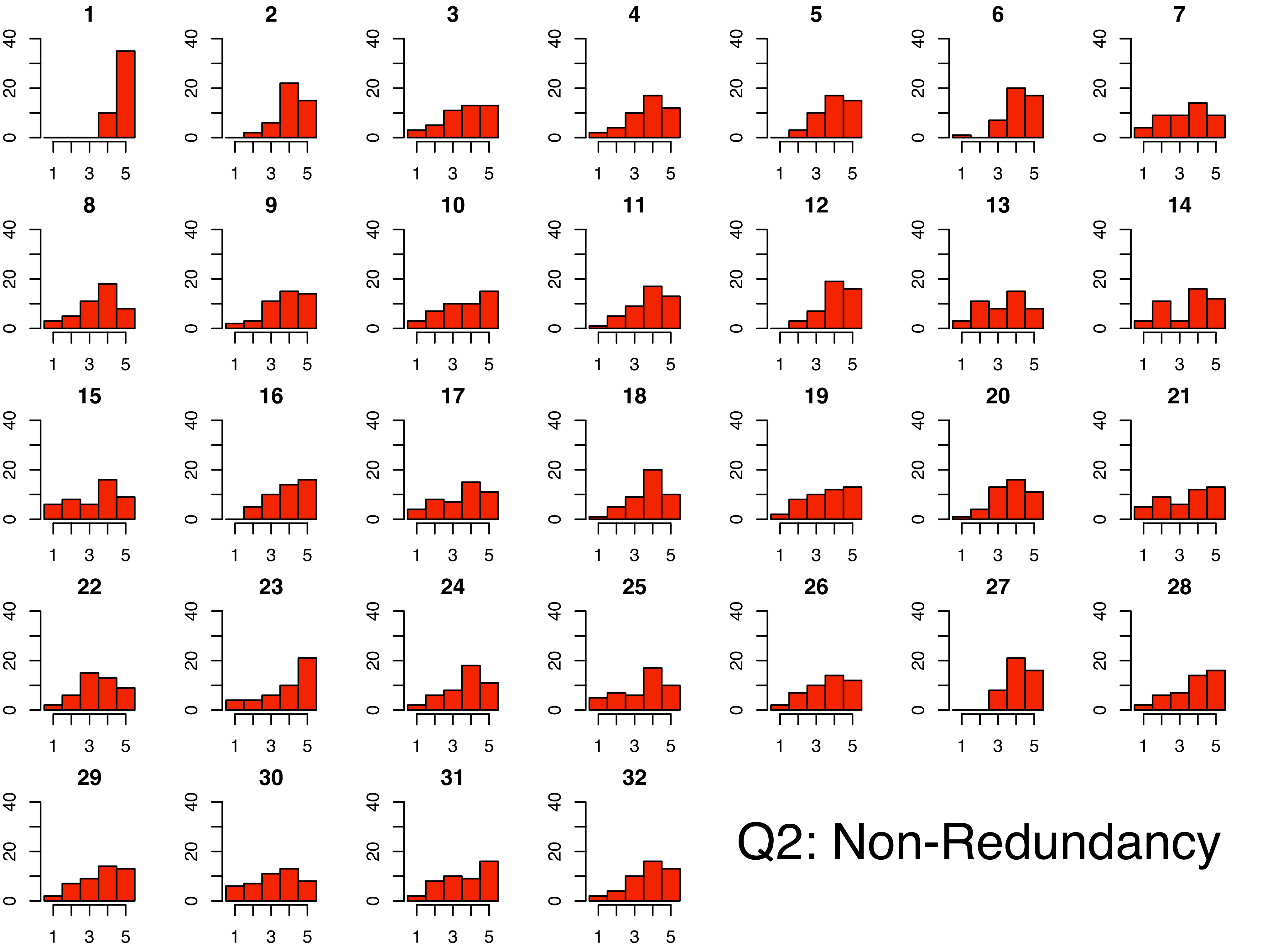
Q1: Grammaticality

Q2: Non-Redundancy

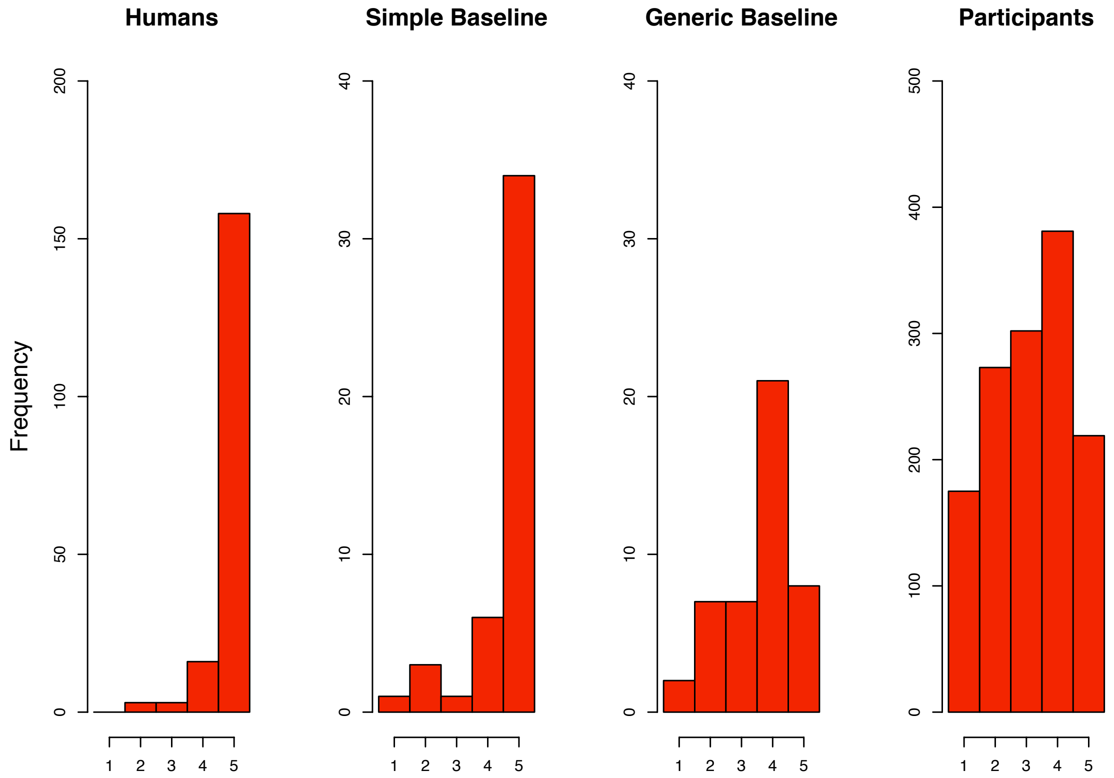


There should be no unnecessary repetition in the summary. Unnecessary repetition might take the form of whole sentences that are repeated, or repeated facts, or the repeated use of a noun or noun phrase (e.g., ``Bill Clinton") when a pronoun (``he") would suffice.

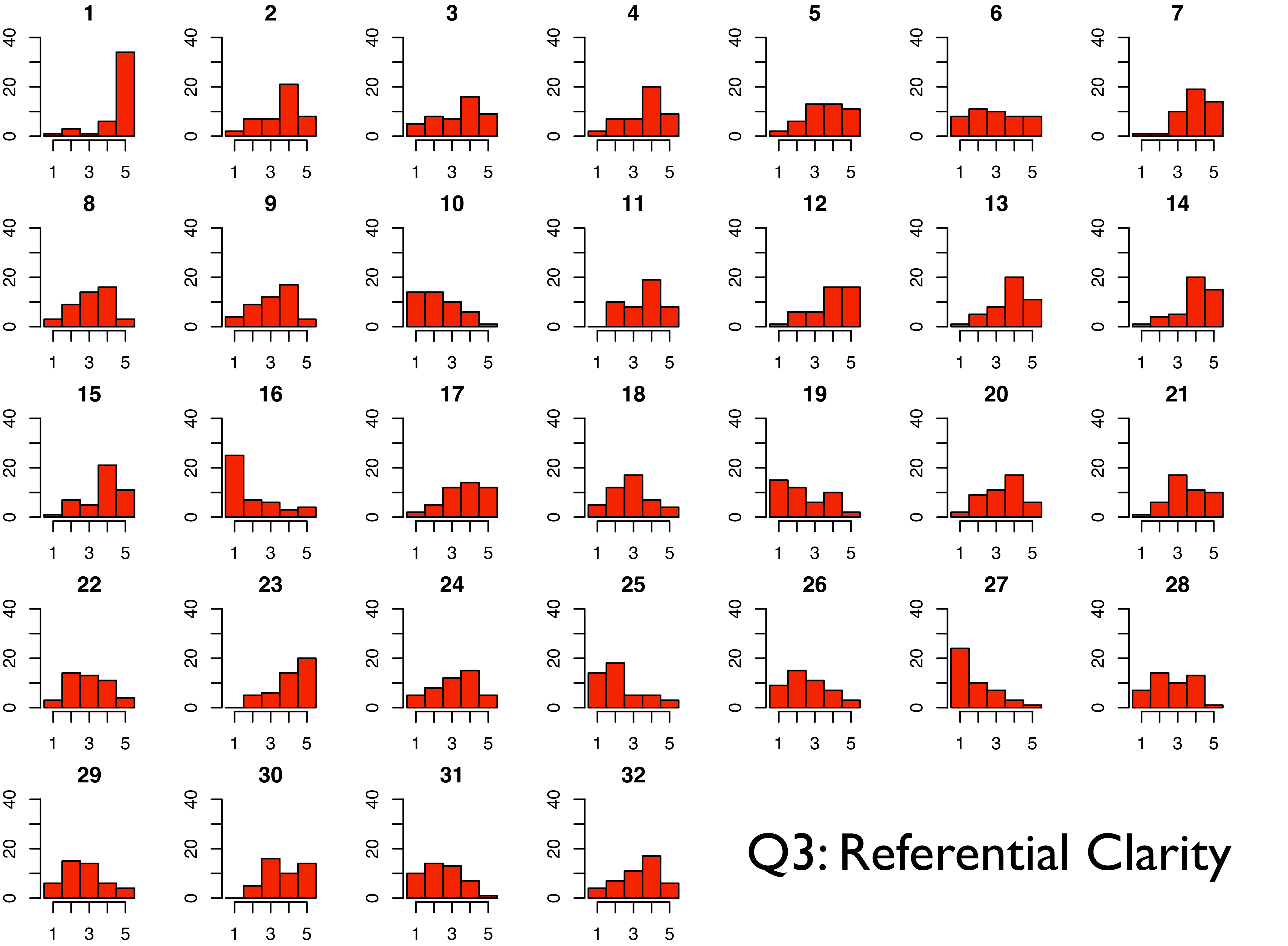
Q2: Non-Redundancy



Q3: Referential Clarity

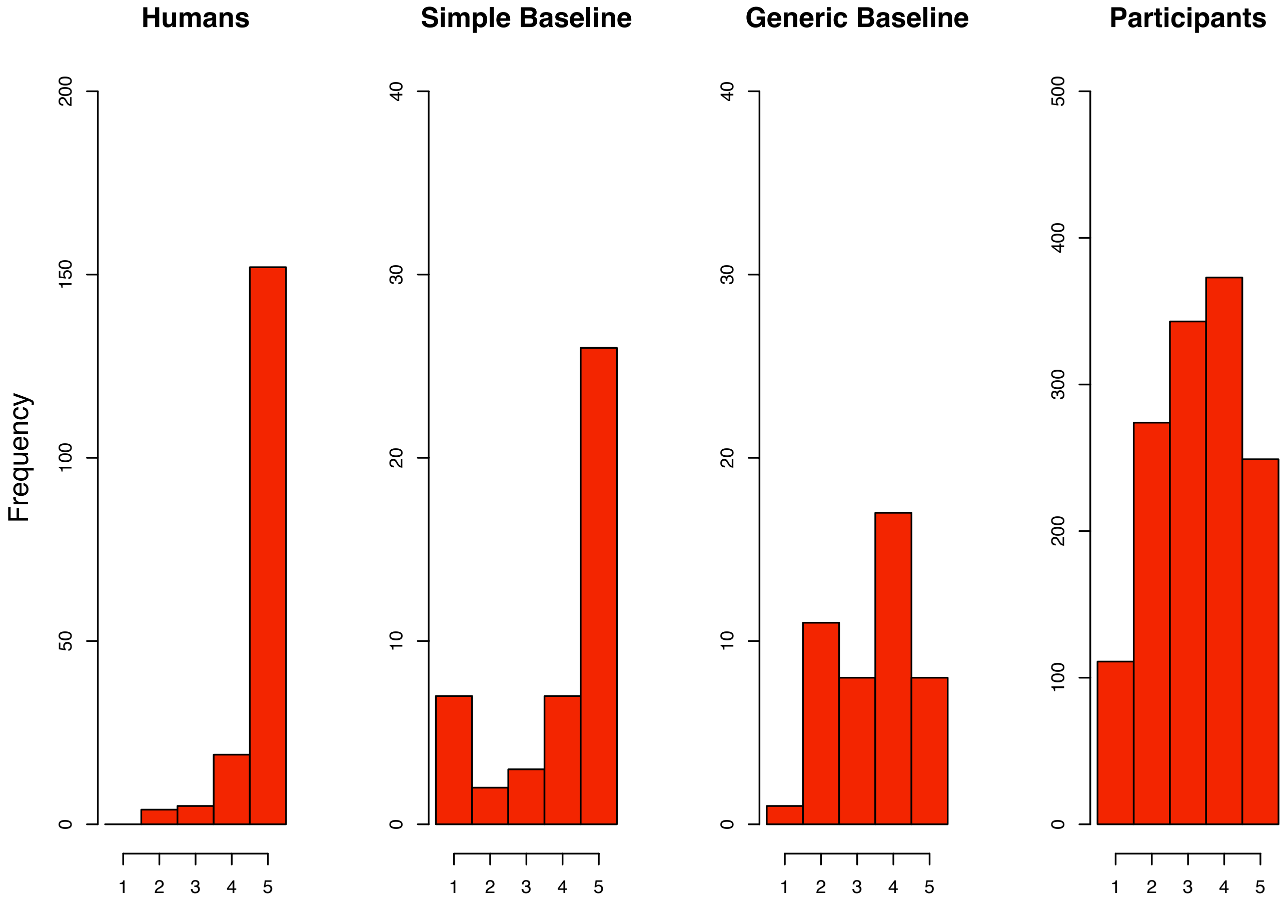


It should be easy to identify who or what the pronouns and noun phrases in the summary are referring to. If a person or other entity is mentioned, it should be clear what their role in the story is. So, a reference would be unclear if an entity is referenced but its identity or relation to the story remains unclear.



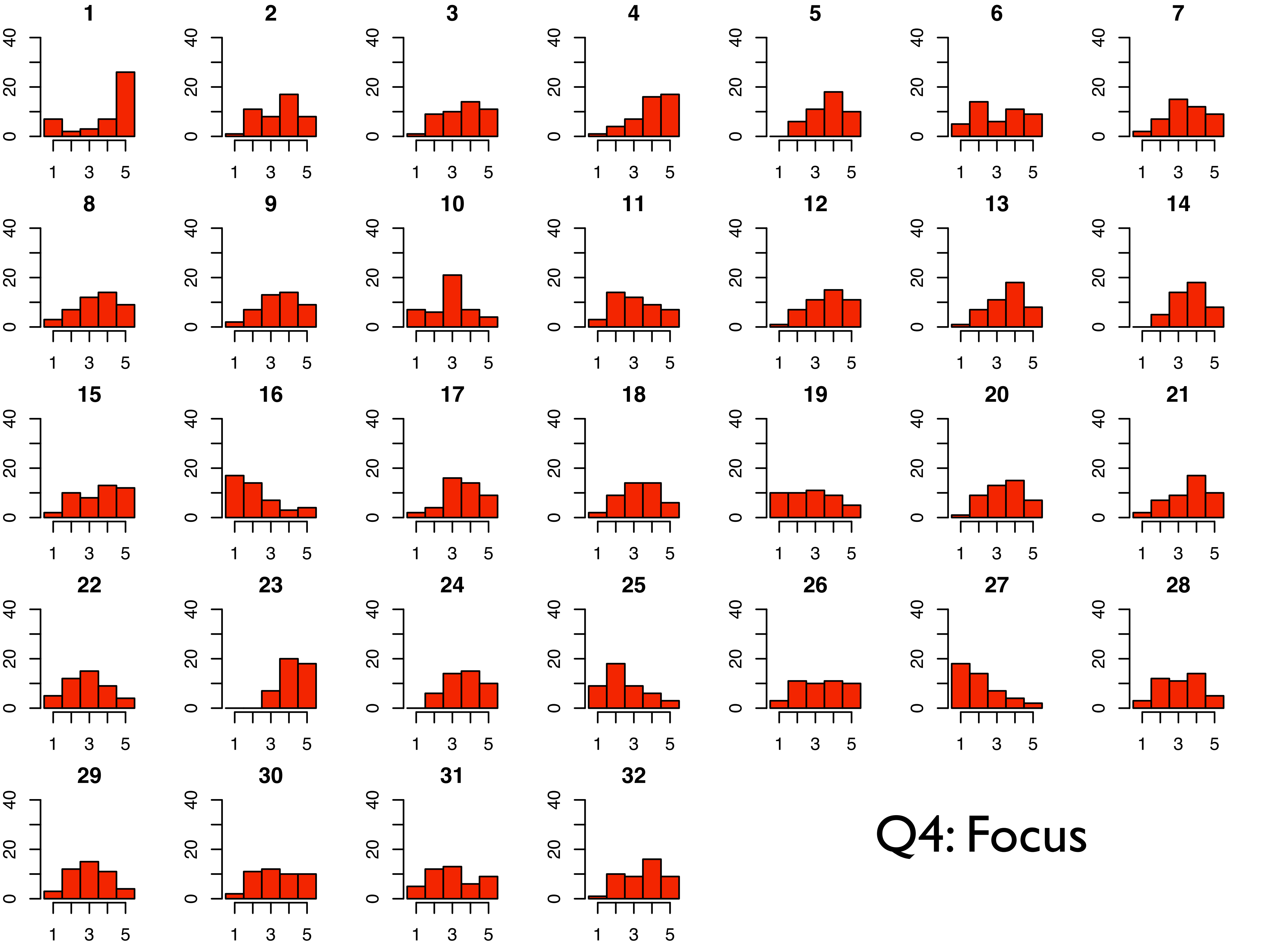
Q3: Referential Clarity

Q4: Focus

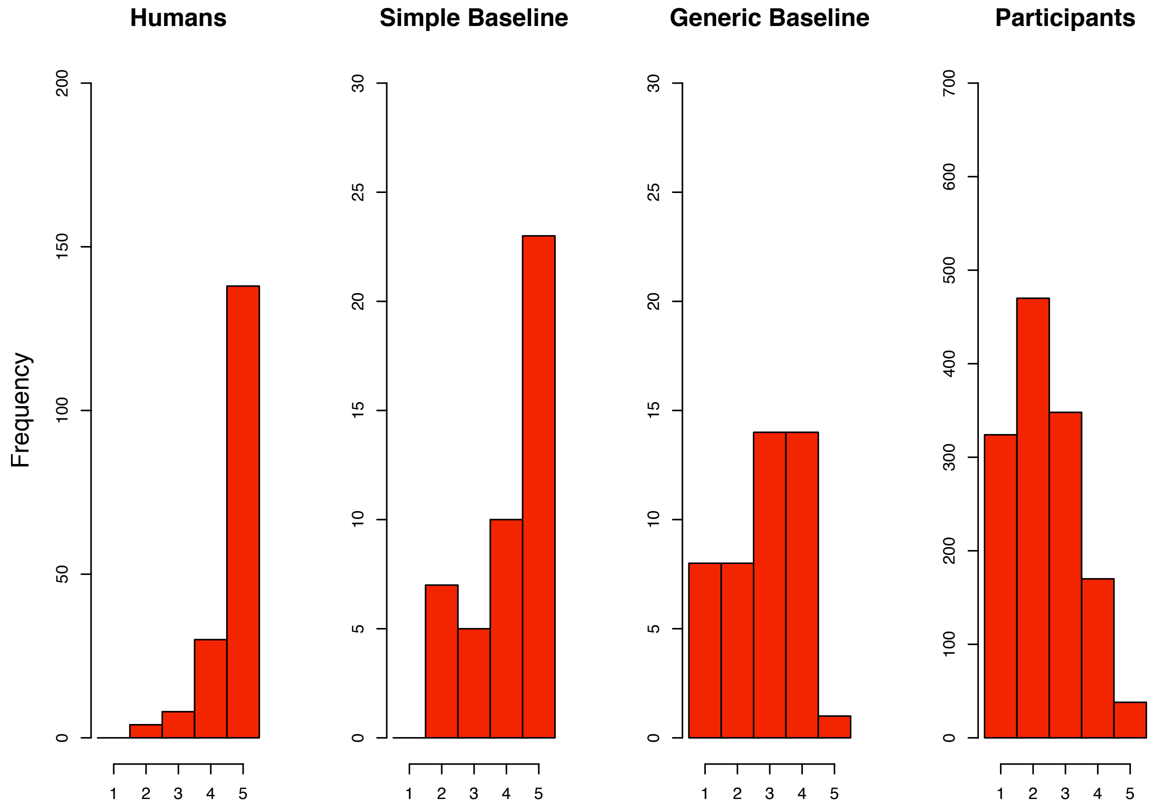


The summary should have a focus; sentences should only contain information that is related to the rest of the summary.

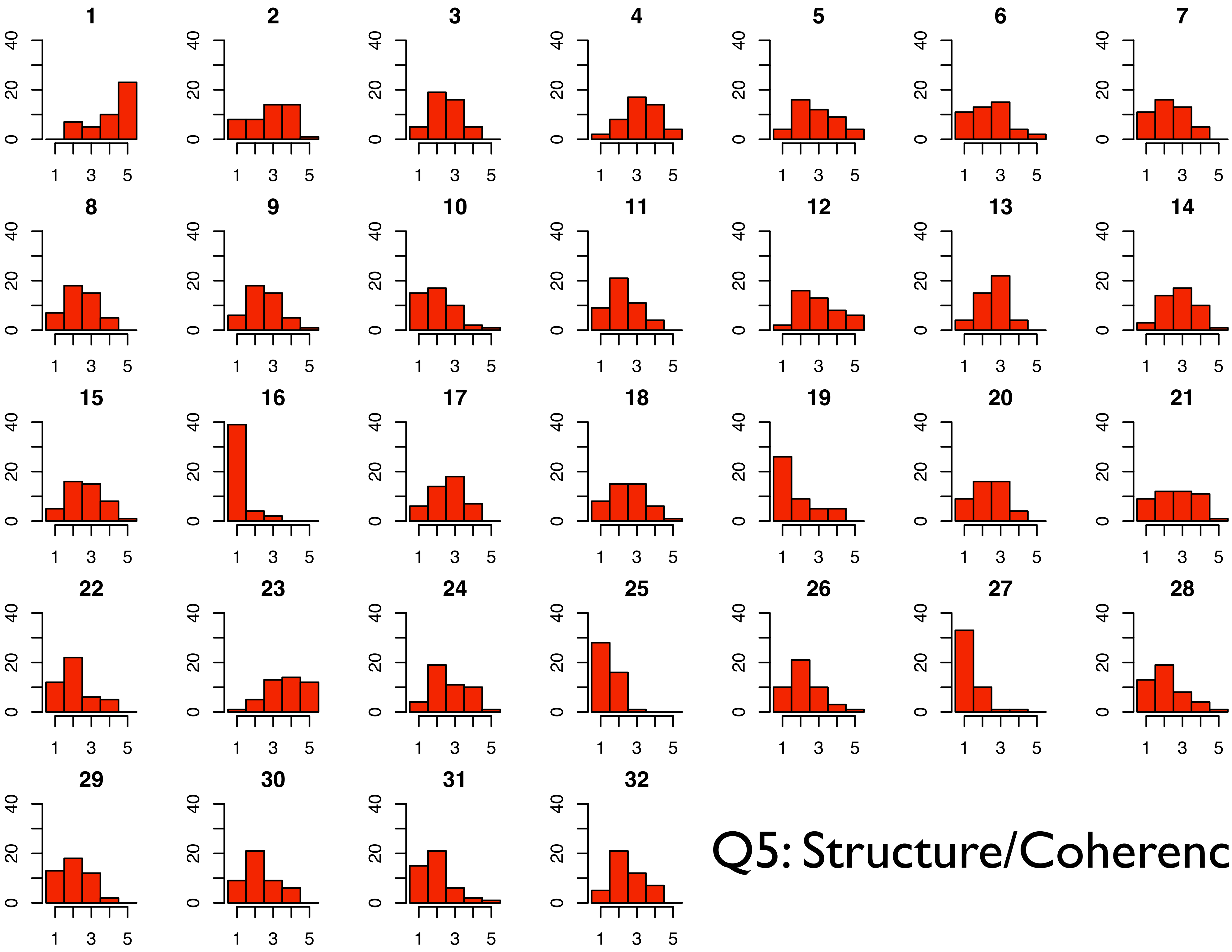
Q4: Focus


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
0
20
40
1
3
5

Q5: Structure and Coherence

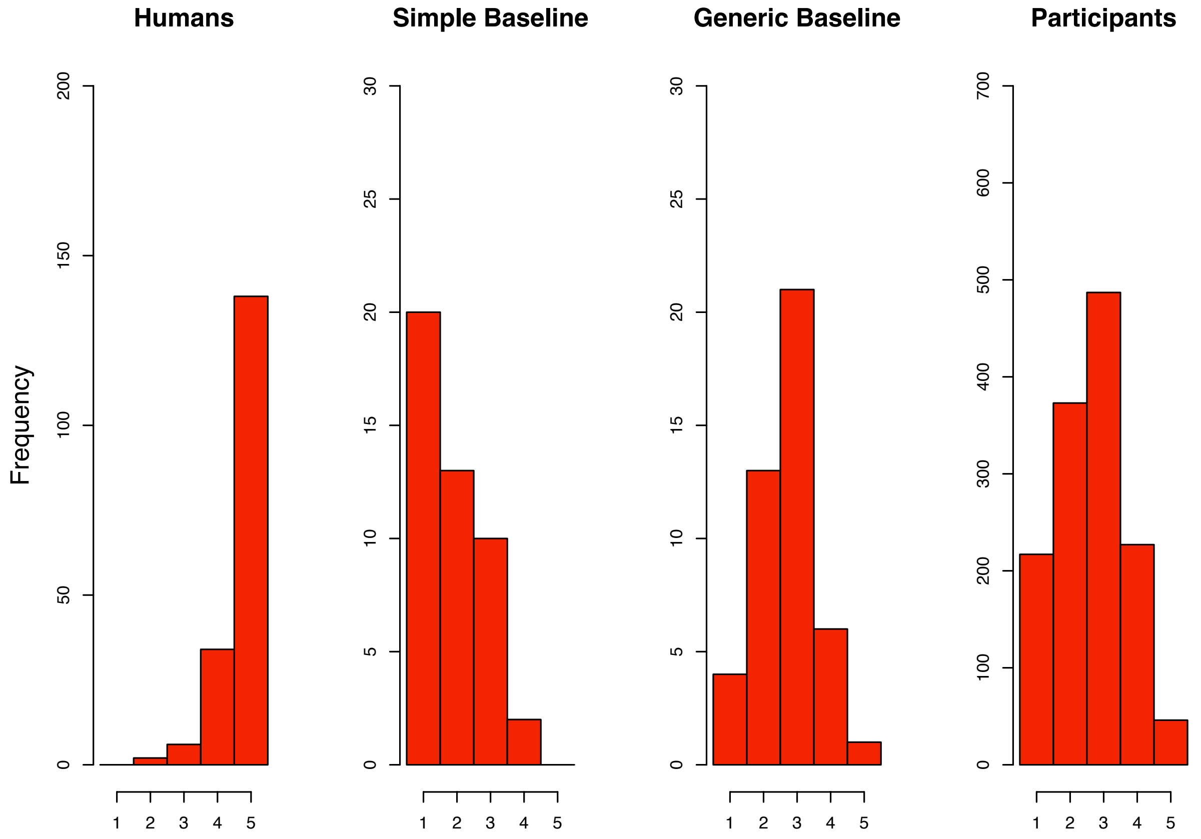


The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to sentence to a coherent body of information about a topic.

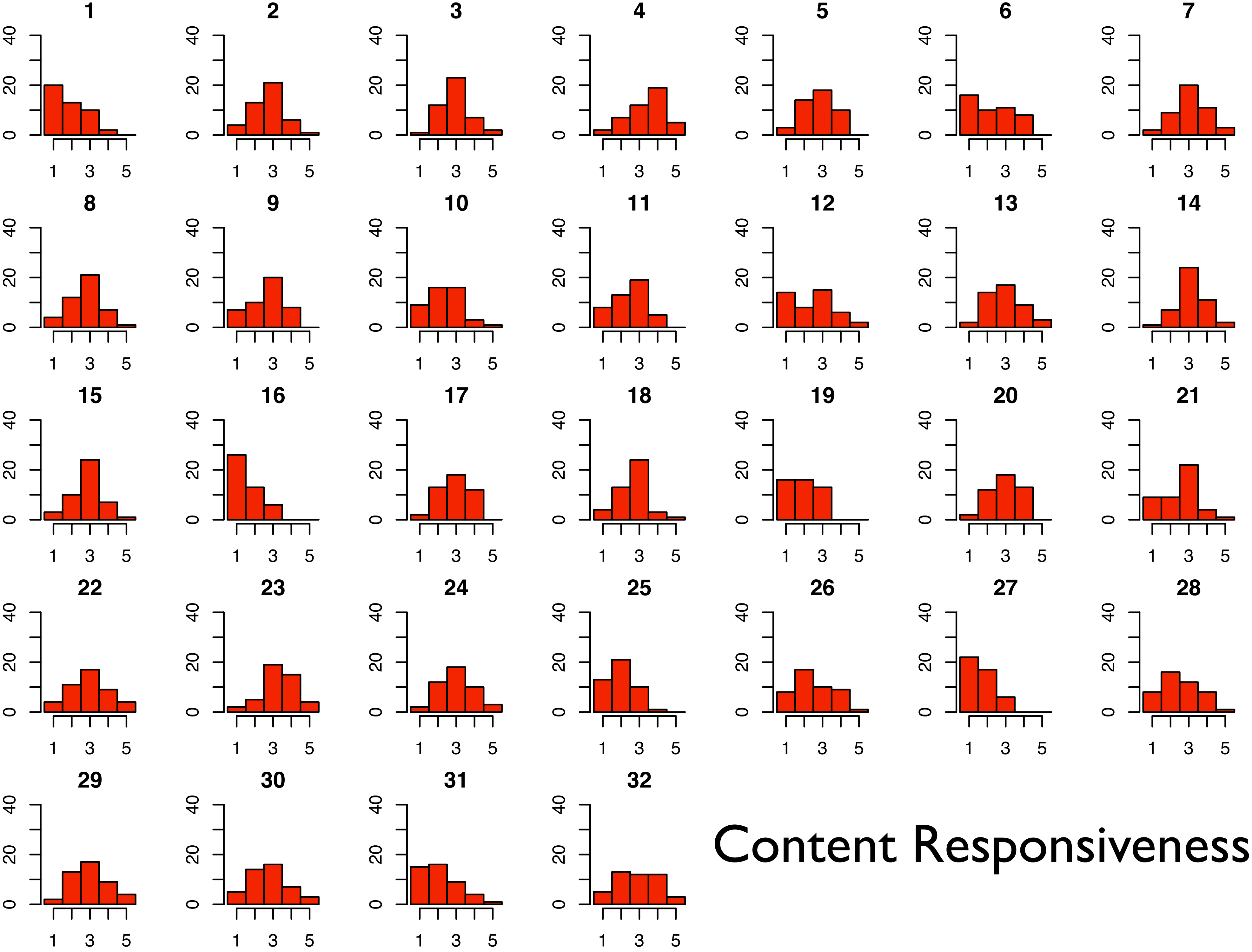


Q5: Structure/Coherence

Content Responsiveness



Based on amount of information in summary that contributes to meeting the information need expressed in the topic statement



Content Responsiveness

ANOVA, multiple comparison of systems

Responsiveness

```
4       3.4000   A
23      3.3111   A B
14      3.1333   A B C
7       3.0889   A B C D
29      3.0000   A B C D E
24      3.0000   A B C D E
22      2.9556   A B C D E
3       2.9333   A B C D E F
20      2.9333   A B C D E F
13      2.9333   A B C D E F
32      2.8889   A B C D E F
17      2.8889   A B C D E F
15      2.8444   A B C D E F
5       2.7778     B C D E F G
8       2.7556     B C D E F G
30      2.7556     B C D E F G
2       2.7111       C D E F G
9       2.6444       C D E F G H
18      2.6444       C D E F G H
21      2.5333         D E F G H I
28      2.5111         D E F G H I J
26      2.5111         D E F G H I J
11      2.4667           E F G H I J
12      2.4222           E F G H I J K
10      2.3556             F G H I J K
6       2.2444               G H I J K
31      2.1111                 H I J K L
25      1.9778                   I J K L
19      1.9333                     J K L
1       1.8667                       K L
27      1.6444                         L
16      1.5556                         L
```

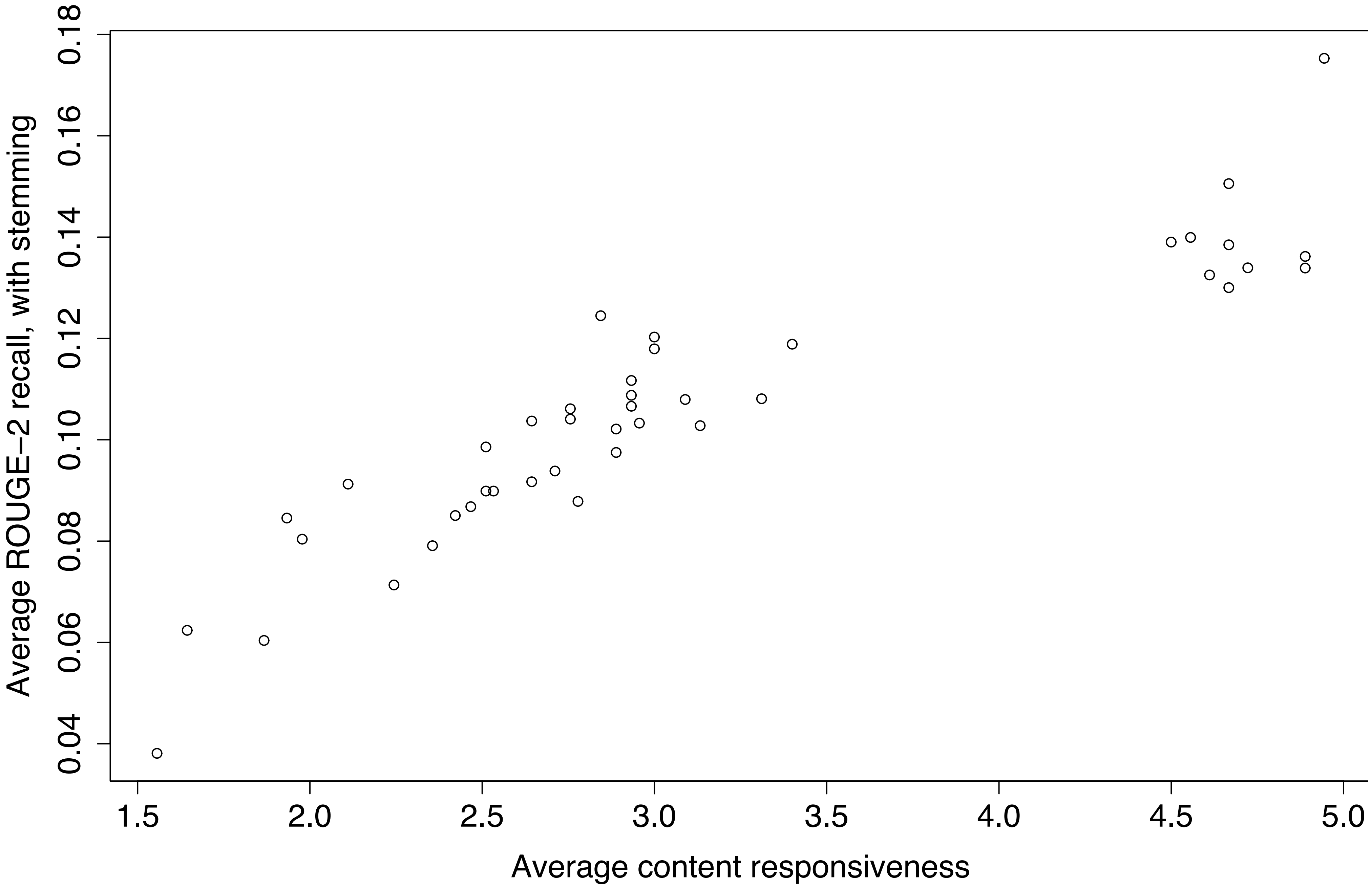
ROUGE-2

```
15      0.1245   A
29      0.1203   A B
4       0.1189   A B C
24      0.1180   A B C D
13      0.1118   A B C D E
20      0.1088   A B C D E F
23      0.1081     B C D E F
7       0.1079     B C D E F
3       0.1066     B C D E F G
30      0.1061     B C D E F G
8       0.1041       C D E F G H
9       0.1037       C D E F G H I
22      0.1033       C D E F G H I
14      0.1028         D E F G H I J
17      0.1022         D E F G H I J
28      0.0987           E F G H I J K
32      0.0975           E F G H I J K
2       0.0938             F G H I J K L
18      0.0917               G H I J K L
31      0.0912               G H I J K L
26      0.0900                 H I J K L
21      0.0899                 H I J K L
5       0.0878                   I J K L
11      0.0868                     J K L M
12      0.0850                       K L M
19      0.0846                       K L M
25      0.0805                         L M
10      0.0791                         L M
6       0.0714                           M N
27      0.0624                             N
1       0.0604                             N
16      0.0381                               O
```

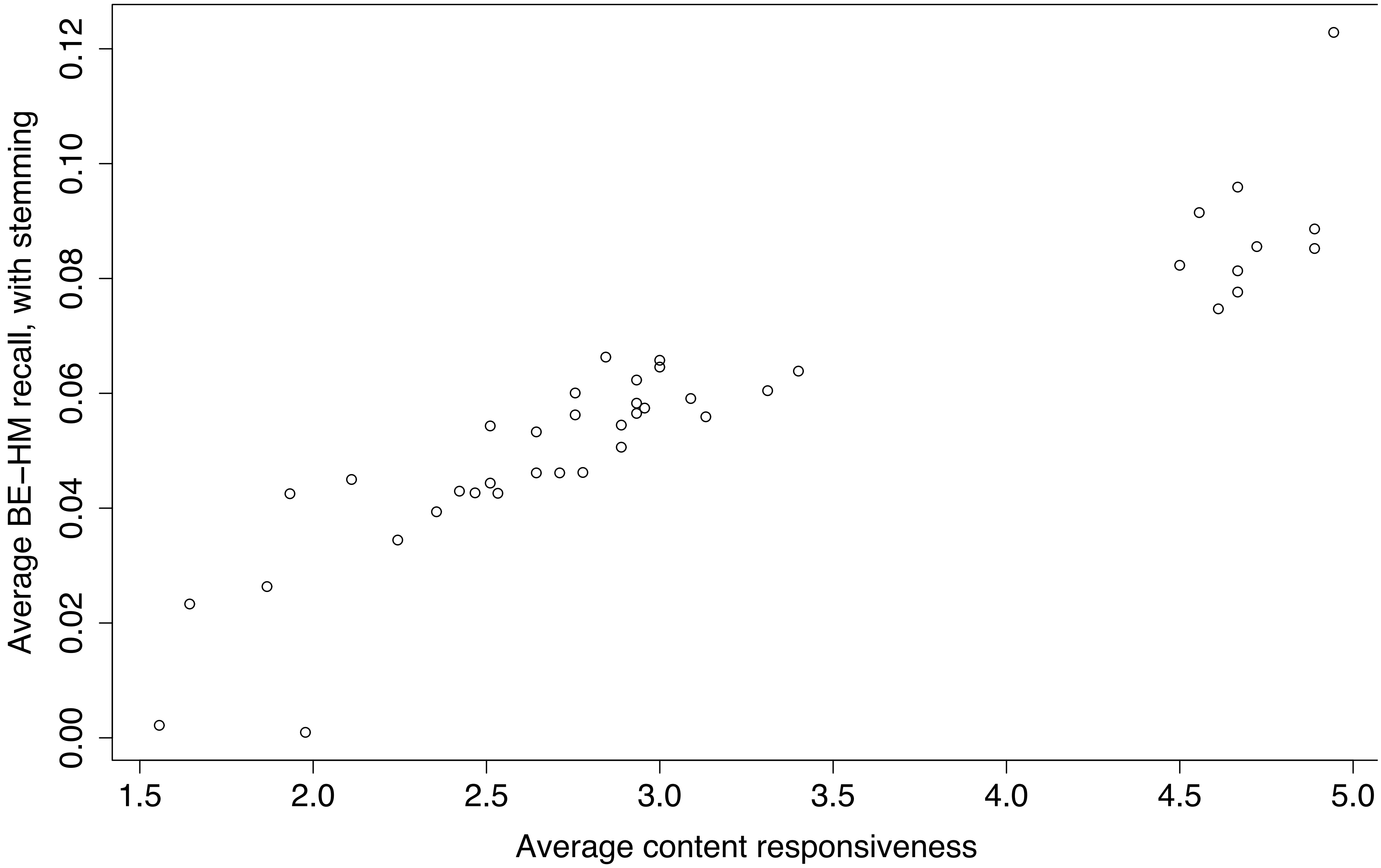
Multiple Comparison Test

- Conservative test, probability of erroneously declaring two systems to be different is <=5% over all comparisons of 32 systems
- Simple Baseline extremely easy to outperform
- Generic Baseline significantly worse than topic-dependent Systems 4 and 23
- *Topic focus matters*

ROUGE-2 vs. Content Responsiveness



BE-HM vs. Content Responsiveness



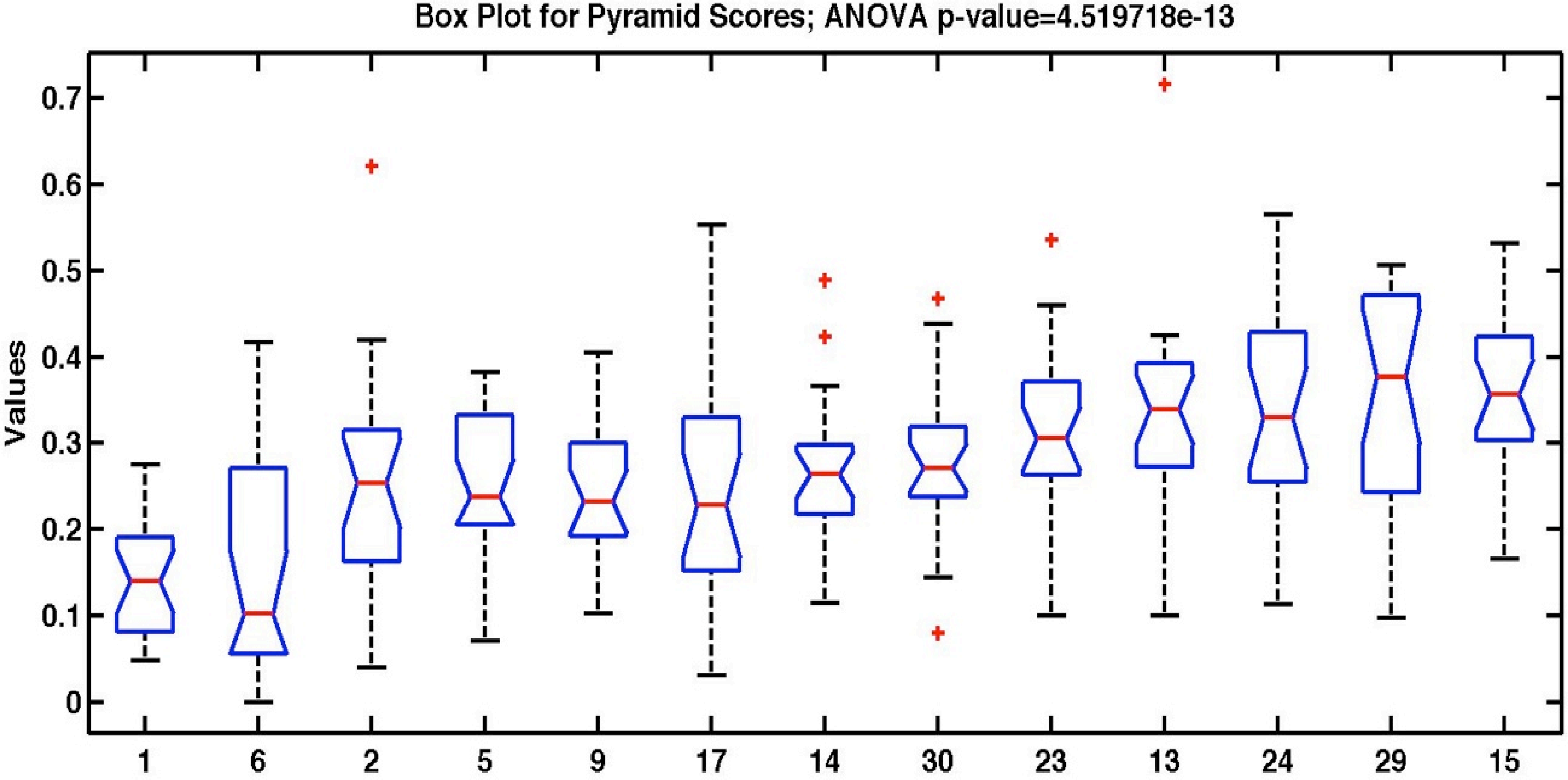
Correlation with Content Responsiveness

Automatic Peers

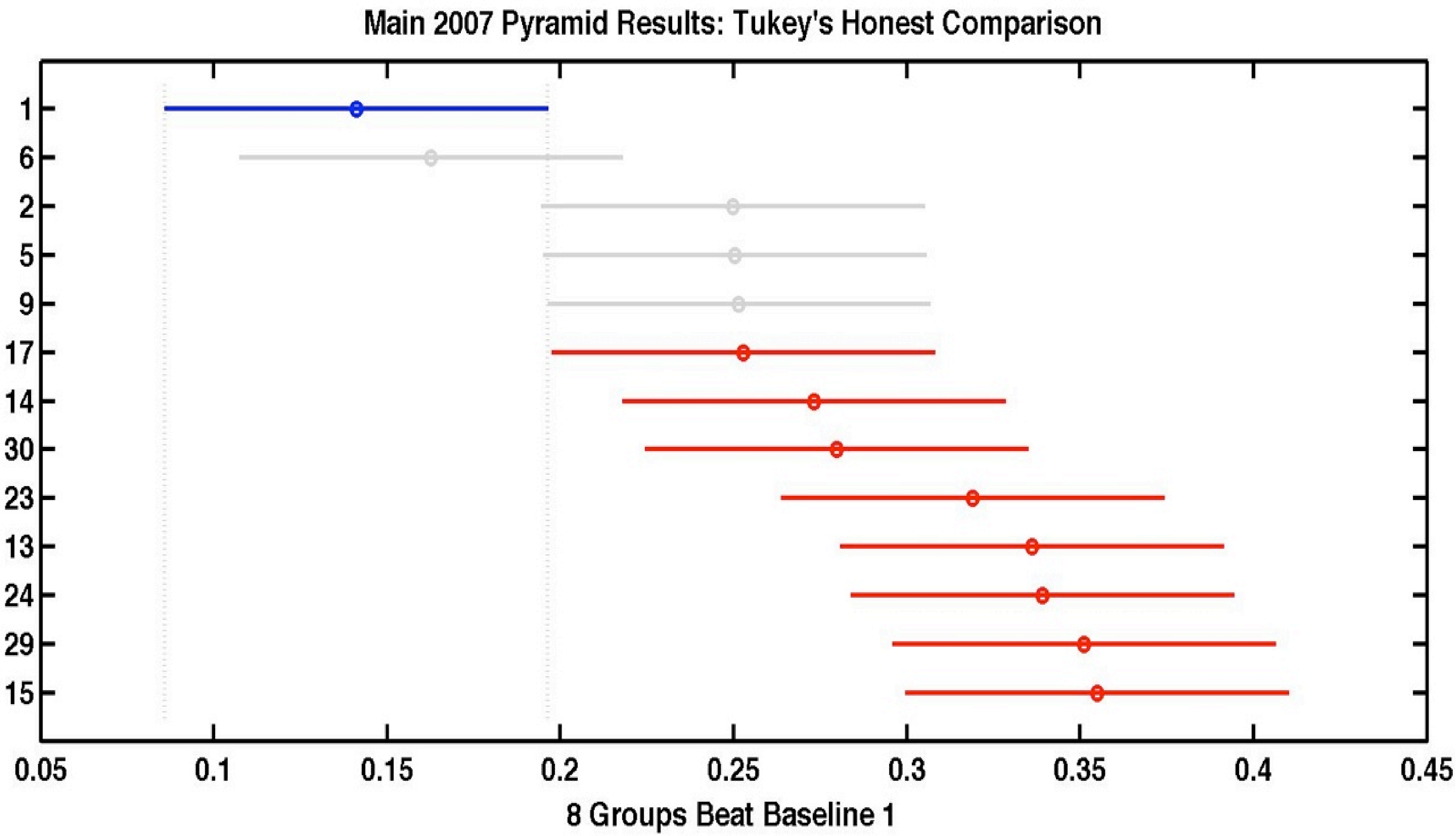
	Spearman	Pearson
ROUGE-2	0.869	0.878 [0.786,1.00]
ROUGE-SU4	0.827	0.831 [0.709, 1.00]
BE-HM	0.885	0.861 [0.759,1.00]

Pyramid Evaluation

- 23 topics selected from main task
- Topics had been rated for clarity by assessors who wrote summaries for topic; topics with highest clarity were selected
- 13 automatic peers: 11 task participants, 2 baselines
- 5 additional volunteers
- Organized by Lucy Vanderwende at Microsoft



(Courtesy, John Conroy)



(Courtesy, John Conroy)

Combined overall manual score

- No manual “overall responsiveness” assessment in 2007
- Estimate overall score using DUC 2006 multiple linear regression model
- Approximate weights

Quality	Weight
Grammaticality	0.05
Non-Redundancy	0.01
Referential Clarity	0.07
Focus	0.02
Structure and Coherence	0.20
Content Responsiveness	0.65

Example summary (23)

France and Germany on Thursday gave U.N. officials paperwork showing they have ratified a treaty banning anti-personnel land mines. Burkino Faso became the 40th country to ratify an international treaty to ban anti-personnel land mines Wednesday. Namibia has become the 24th country to ratify the Ottawa Convention banning land mines. Kenya will next year ratify the Ottawa Treaty banning the use of land mines. Jordan signed a global land mine treaty Saturday, joining 127 other countries that have endorsed the pact, which prohibits the use, production and stockpiling of the weapon. South Africa is to join more than 100 countries in Canada this week to sign a treaty banning the use or possession of anti-personnel mines. With a land mine treaty ratified in record time, they [UNITED NATION] want every nation to sign it and the millions of land mines that continue to kill, maim and sow terror around the world removed. Noor said it was encouraging that major producers and exporters including France, Germany, Britain and Hungary had already ratified the treaty.

The U.S. did not sign the Ottawa treaty and is therefore not obliged to destroy its own mines, but anti-mine campaigners have been pressuring signatories to destroy all mines within their borders.

Despite a treaty signed by 135 countries to ban their use, production and stockpiling, anti-personnel land mines appear to be as popular as ever in fighting wars these days.

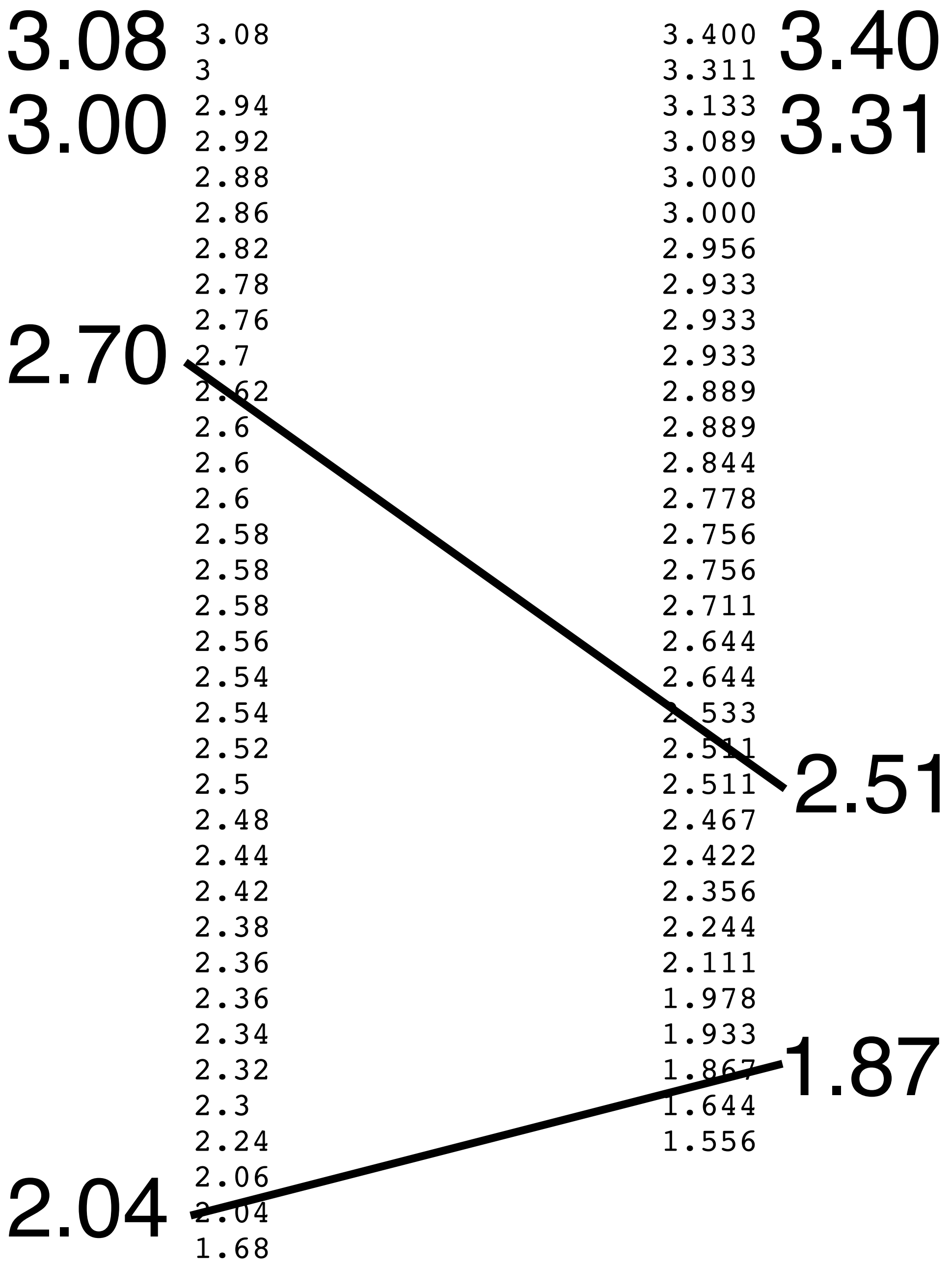
Q1-Q5={5,3,4,4,3}, content=4: overall=3.84 (91.7 percentile)

Example summary (4)

A three-day conference on the land mines ban is being held in Ottawa, Canada, and most of the participating countries are expected to sign the treaty. South Africa is to join more than 100 countries in Canada this week to sign a treaty banning the use or possession of anti-personnel mines. France and Germany on Thursday gave UN officials paperwork showing they have ratified a treaty banning anti-personnel land mines. Burkino Faso became the 40th country to ratify an international treaty to ban anti-personnel land mines Wednesday, meaning the treaty will go into effect in six months, the United Nations announced. In a statement, the ICBL expressed ``grave concern about reports of the continued laying of mines in a number of countries that have signed but not ratified the treaty, such as Angola, Cambodia, Senegal and Sudan. The United States has refused to ratify the treaty, arguing that such weapons are needed on the Korean peninsula to deter an invasion by North Korea of South Korea. Jordan signed a global land mine treaty Saturday, joining 127 other countries that have endorsed the pact, which prohibits the use, production and stockpiling of the weapon. Kenya will next year ratify the Ottawa Treaty banning the use of land mines, according to a senior Kenyan government official. Despite a treaty signed by 135 countries to ban their use, production and stockpiling, anti-personnel land mines appear to be as popular as ever in fighting wars these days.

Q1-Q5={5,4,4,4,4}, content=4: overall=4.05 (94.8 percentile)

Average Content Responsiveness



Humans 2006	Humans 2007		Systems 2006	Systems 2007	
4.9	4.944	3.08	3.08	3.400	3.40
4.9	4.889		3	3.311	
4.9	4.889	3.00	2.94	3.133	3.31
4.85	4.722		2.92	3.089	
4.75	4.667		2.88	3.000	
4.7	4.667		2.86	3.000	
4.65	4.667		2.82	2.956	
4.65	4.611		2.78	2.933	
4.6	4.556		2.76	2.933	
4.6	4.500	2.70	2.7	2.933	
			2.62	2.889	
			2.6	2.889	
			2.6	2.844	
			2.6	2.778	
			2.58	2.756	
			2.58	2.756	
			2.58	2.711	
			2.56	2.644	
			2.54	2.644	
			2.54	2.533	
			2.52	2.511	
			2.5	2.511	2.51
			2.48	2.467	
			2.44	2.422	
			2.42	2.356	
			2.38	2.244	
			2.36	2.111	
			2.36	1.978	
			2.34	1.933	
			2.32	1.867	1.87
			2.3	1.644	
			2.24	1.556	
			2.06		
		2.04	2.04		
			1.68		

DUC 2007 Update Task (Pilot)

Update Task: Topics and documents

- 10 topics selected from main task, each developed by a different assessor
- topics selected based on whether it was likely to contain new information over time in the period covered by the documents.
- Documents partitioned into 3 sets, A-C, ordered by date: Date(A) < Date(B) < Date(C)
- ~10 docs in A, ~8 in B, ~7 in C

Update Task

- Given a topic and its 3 clusters of documents, A-C, create three brief (<=100 words), fluent summaries that contribute to satisfying the information need expressed in the topic statement:
  - Summary A: summary of cluster A
  - Summary B: summary of cluster B, assuming reader has read cluster A
  - Summary C: summary of cluster C, assuming reader has read clusters A and B

Update summaries

- 10 assessors; each topic assigned to 4 different assessors, including topic developer
- 22 participants (ID = 36-57)
- Simple Baseline (ID = 35)
- Generic Baseline (ID = 58)
  - Generic A: straight application of CLASSY04
  - Generic B: signature terms from docsets A and B
  - Generic C: signature terms from docsets A-C

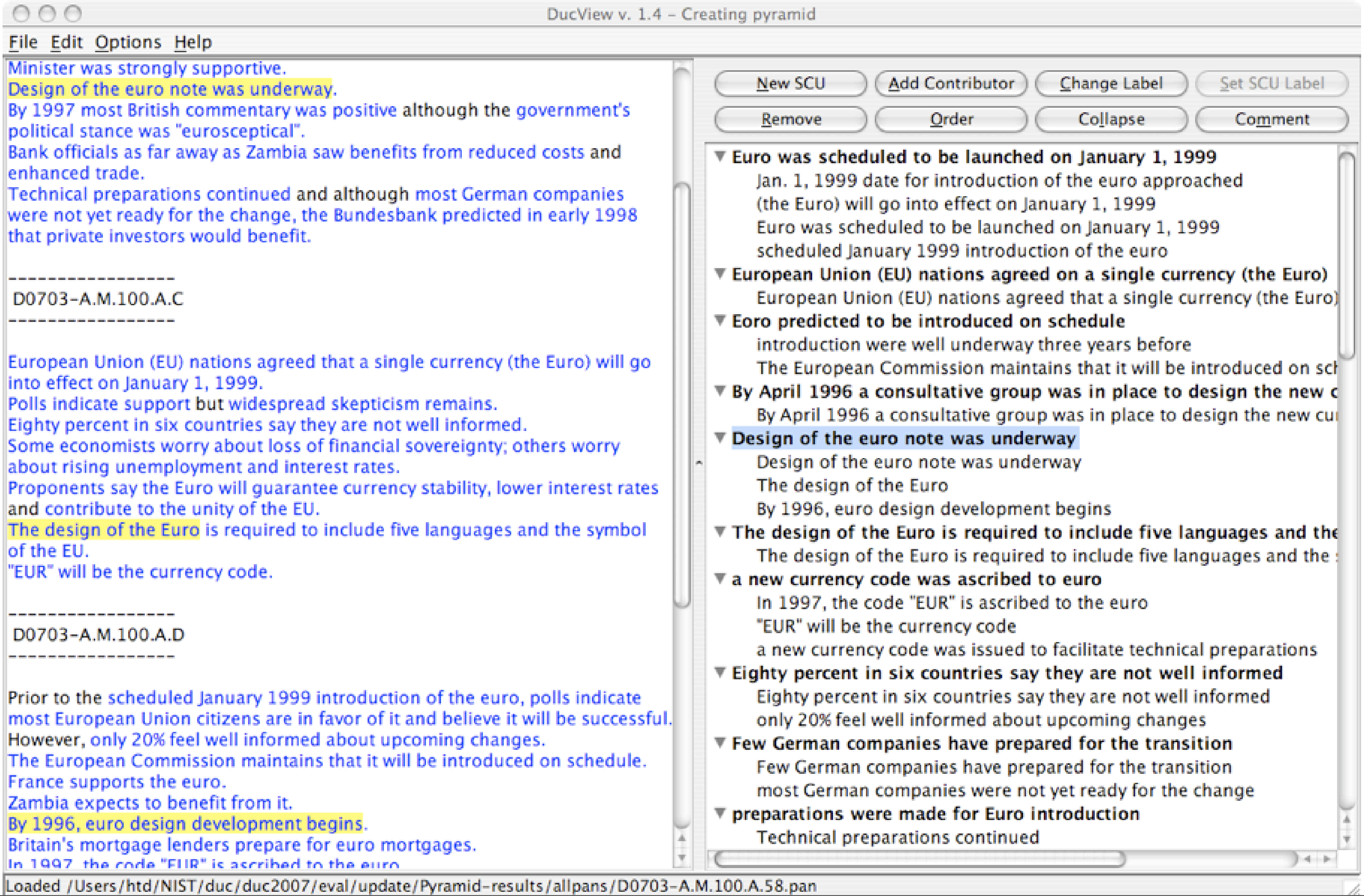
Evaluation

- 9 Assessors (usually same as topic developer, always a summarizer for topic)
- Single assessor for each topic:
  - Content Responsiveness: same as for main task, except discount relevant information in Summary B that is already in Cluster A; discount information in Summary C that is already in Clusters A and B
  - Pyramid Evaluation
    - Pyramid creation
    - Peer annotation

Content Responsiveness ANOVA

- ANOVA, multiple comparison of 10 humans and 24 automatic peers using Tukey's HSD criterion:
  - All 30 doc clusters (10 topics x 3 clusters/topic):
    - All Humans better than all systems, but worst human "close" to best system (3.8 vs. 3.0 average responsiveness)
  - By summary type (A, B, or C): 10 doc clusters not enough to distinguish humans from systems
- Small number of topics; topic variation hides any differences in peers

Pyramid Creation



Minister was strongly supportive.
Design of the euro note was underway.
By 1997 most British commentary was positive although the government's political stance was "eurosceptical".
Bank officials as far away as Zambia saw benefits from reduced costs and enhanced trade.
Technical preparations continued and although most German companies were not yet ready for the change, the Bundesbank predicted in early 1998 that private investors would benefit.

D0703-A.M.100.A.C

European Union (EU) nations agreed that a single currency (the Euro) will go into effect on January 1, 1999.
Polls indicate support but widespread skepticism remains.
Eighty percent in six countries say they are not well informed.
Some economists worry about loss of financial sovereignty; others worry about rising unemployment and interest rates.
Proponents say the Euro will guarantee currency stability, lower interest rates and contribute to the unity of the EU.
The design of the Euro is required to include five languages and the symbol of the EU.
"EUR" will be the currency code.

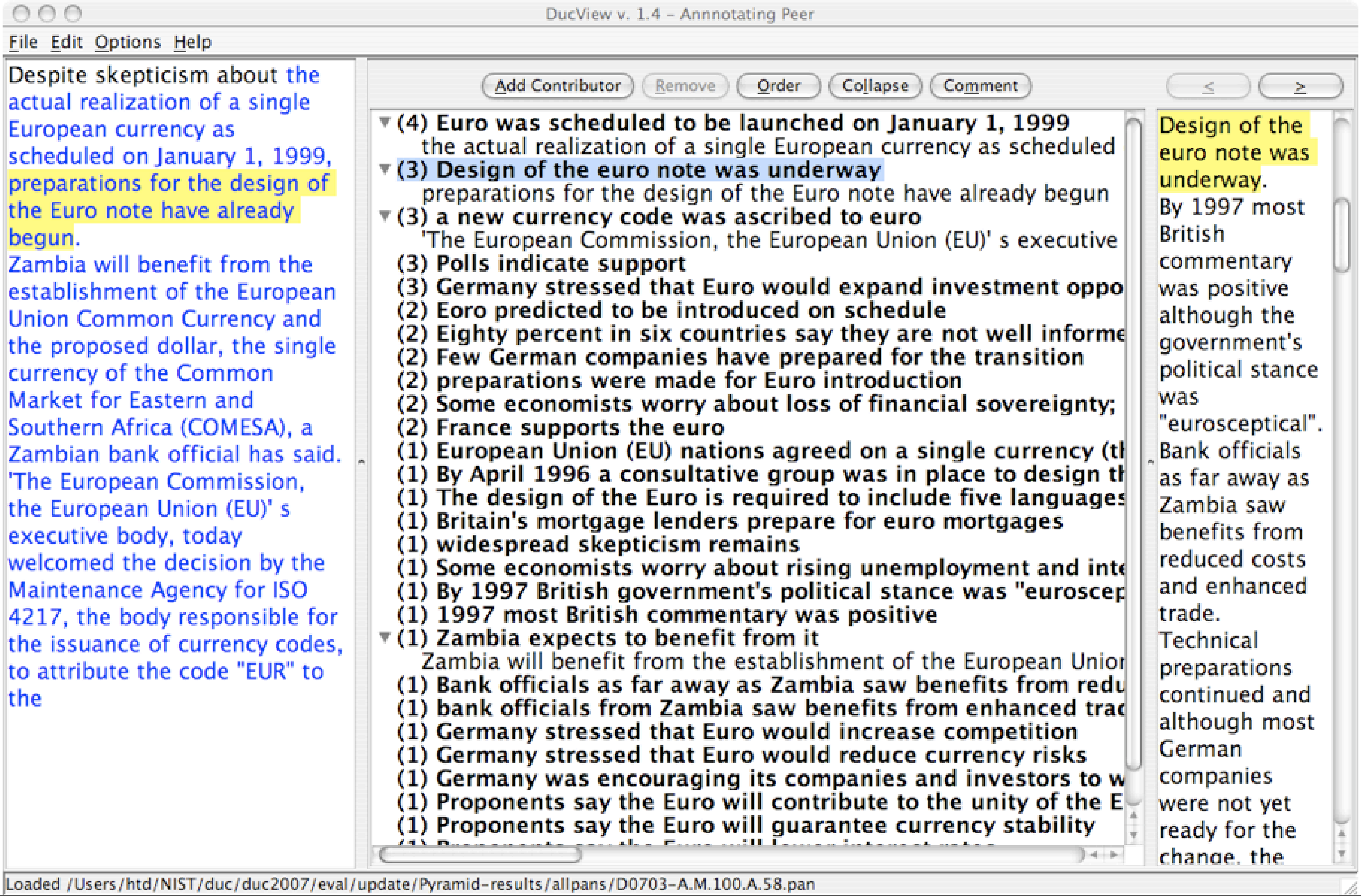
D0703-A.M.100.A.D

Prior to the scheduled January 1999 introduction of the euro, polls indicate most European Union citizens are in favor of it and believe it will be successful.
However, only 20% feel well informed about upcoming changes.
The European Commission maintains that it will be introduced on schedule.
France supports the euro.
Zambia expects to benefit from it.
By 1996, euro design development begins.
Britain's mortgage lenders prepare for euro mortgages.
In 1997, the code "EUR" is ascribed to the euro

- ▼ **Euro was scheduled to be launched on January 1, 1999**
  - Jan. 1, 1999 date for introduction of the euro approached
  - (the Euro) will go into effect on January 1, 1999
  - Euro was scheduled to be launched on January 1, 1999
  - scheduled January 1999 introduction of the euro
- ▼ **European Union (EU) nations agreed on a single currency (the Euro)**
  - European Union (EU) nations agreed that a single currency (the Euro)
- ▼ **Eoro predicted to be introduced on schedule**
  - introduction were well underway three years before
  - The European Commission maintains that it will be introduced on sch
- ▼ **By April 1996 a consultative group was in place to design the new c**
  - By April 1996 a consultative group was in place to design the new cu
- ▼ **Design of the euro note was underway**
  - Design of the euro note was underway
  - The design of the Euro
  - By 1996, euro design development begins
- ▼ **The design of the Euro is required to include five languages and the**
  - The design of the Euro is required to include five languages and the
- ▼ **a new currency code was ascribed to euro**
  - In 1997, the code "EUR" is ascribed to the euro
  - "EUR" will be the currency code
  - a new currency code was issued to facilitate technical preparations
- ▼ **Eighty percent in six countries say they are not well informed**
  - Eighty percent in six countries say they are not well informed
  - only 20% feel well informed about upcoming changes
- ▼ **Few German companies have prepared for the transition**
  - Few German companies have prepared for the transition
  - most German companies were not yet ready for the change
- ▼ **preparations were made for Euro introduction**
  - Technical preparations continued

Loaded /Users/htd/NIST/duc/duc2007/eval/update/Pyramid-results/allpans/D0703-A.M.100.A.58.pan

Peer Annotation



Despite skepticism about the actual realization of a single European currency as scheduled on January 1, 1999, preparations for the design of the Euro note have already begun. Zambia will benefit from the establishment of the European Union Common Currency and the proposed dollar, the single currency of the Common Market for Eastern and Southern Africa (COMESA), a Zambian bank official has said. 'The European Commission, the European Union (EU)' s executive body, today welcomed the decision by the Maintenance Agency for ISO 4217, the body responsible for the issuance of currency codes, to attribute the code "EUR" to the

- ▼ (4) Euro was scheduled to be launched on January 1, 1999
  - the actual realization of a single European currency as scheduled
- ▼ (3) Design of the euro note was underway
  - preparations for the design of the Euro note have already begun
- ▼ (3) a new currency code was ascribed to euro
  - 'The European Commission, the European Union (EU)' s executive
- (3) Polls indicate support
- (3) Germany stressed that Euro would expand investment oppo
- (2) Eoro predicted to be introduced on schedule
- (2) Eighty percent in six countries say they are not well informe
- (2) Few German companies have prepared for the transition
- (2) preparations were made for Euro introduction
- (2) Some economists worry about loss of financial sovereignty;
- (2) France supports the euro
- (1) European Union (EU) nations agreed on a single currency (tl
- (1) By April 1996 a consultative group was in place to design tl
- (1) The design of the Euro is required to include five languages
- (1) Britain's mortgage lenders prepare for euro mortgages
- (1) widespread skepticism remains
- (1) Some economists worry about rising unemployment and inte
- (1) By 1997 British government's political stance was "euroscep
- (1) 1997 most British commentary was positive
- ▼ (1) Zambia expects to benefit from it
  - Zambia will benefit from the establishment of the European Unior
- (1) Bank officials as far away as Zambia saw benefits from redu
- (1) bank officials from Zambia saw benefits from enhanced trac
- (1) Germany stressed that Euro would increase competition
- (1) Germany stressed that Euro would reduce currency risks
- (1) Germany was encouraging its companies and investors to w
- (1) Proponents say the Euro will contribute to the unity of the E
- (1) Proponents say the Euro will guarantee currency stability

Design of the euro note was underway. By 1997 most British commentary was positive although the government's political stance was "eurosceptical". Bank officials as far away as Zambia saw benefits from reduced costs and enhanced trade. Technical preparations continued and although most German companies were not yet ready for the change, the

Loaded /Users/htd/NIST/duc/duc2007/eval/update/Pyramid-results/allpans/D0703-A.M.100.A.58.pan

Modified Pyramid Score

- N = average number of SCUs in human summary
- W= sum of weights of SCUs in a summary containing the N most highly weighted SCUs
- D = sum of weights of all matched SCUs in peer
- Modified Pyramid Score (recall-based) = D/W

Pyramid ANOVA, all doc clusters

```
40      0.3403   A
46      0.3078   A B
44      0.2997   A B
55      0.2940   A B
47      0.2727   A B C
45      0.2684   A B C D
38      0.2659   A B C D
52      0.2617   A B C D
36      0.2616   A B C D
51      0.2578   A B C D
48      0.2500   A B C D
58      0.2446   A B C D E
49      0.2288   A B C D E
53      0.2283   A B C D E
37      0.2120     B C D E
43      0.1969     B C D E F
42      0.1923     B C D E F
56      0.1628       C D E F
54      0.1569       C D E F
39      0.1565       C D E F
50      0.1521       C D E F
41      0.1412         D E F
35      0.1217           E F
57      0.0740             F
```

Pyramid ANOVA, by update sequence

Summary A

40	0.4019	A					
51	0.3453	A	B				
55	0.3353	A	B				
36	0.3285	A	B				
44	0.3216	A	B				
48	0.3094	A	B	C			
52	0.3054	A	B	C	D		
49	0.3039	A	B	C	D		
47	0.3025	A	B	C	D		
38	0.2937	A	B	C	D		
53	0.2897	A	B	C	D		
58	0.2796	A	B	C	D	E	
43	0.2787	A	B	C	D	E	
46	0.2766	A	B	C	D	E	
42	0.2531	A	B	C	D	E	
41	0.2420	A	B	C	D	E	
45	0.2306	A	B	C	D	E	
37	0.2217		B	C	D	E	F
39	0.1794		B	C	D	E	F
56	0.1787		B	C	D	E	F
54	0.1412			C	D	E	F
35	0.1274				D	E	F
50	0.1020					E	F
57	0.0478						F

Summary B

44	0.2869	A	
40	0.2754	A	
58	0.2359	A	B
46	0.2333	A	B
52	0.2320	A	B
55	0.2239	A	B
51	0.2112	A	B
48	0.2073	A	B
45	0.1993	A	B
36	0.1954	A	B
53	0.1853	A	B
38	0.1852	A	B
47	0.1768	A	B
35	0.1573	A	B
37	0.1546	A	B
56	0.1421	A	B
42	0.1253	A	B
54	0.1220	A	B
43	0.1187	A	B
50	0.1011	A	B
49	0.1000	A	B
57	0.0978	A	B
39	0.0971	A	B
41	0.0760		B

Summary C

46	0.4135	A	
45	0.3755	A	
40	0.3436	A	B
47	0.3387	A	B
55	0.3229	A	B
38	0.3188	A	B
44	0.2906	A	B
49	0.2826	A	B
36	0.2608	A	B
37	0.2596	A	B
50	0.2531	A	B
52	0.2475	A	B
48	0.2333	A	B
58	0.2184	A	B
51	0.2169	A	B
53	0.2097	A	B
54	0.2075	A	B
42	0.1984	A	B
43	0.1932	A	B
39	0.1930	A	B
56	0.1675	A	B
41	0.1056		B
35	0.0806		B
57	0.0765		B

Pyramid ANOVA, by update sequence

Summary A

40	0.4019	A					
51	0.3453	A	B				
55	0.3353	A	B				
36	0.3285	A	B				
44	0.3216	A	B				
48	0.3094	A	B	C			
52	0.3054	A	B	C	D		
49	0.3039	A	B	C	D		
47	0.3025	A	B	C	D		
38	0.2937	A	B	C	D		
53	0.2897	A	B	C	D		
58	0.2796	A	B	C	D	E	
43	0.2787	A	B	C	D	E	
46	0.2766	A	B	C	D	E	
42	0.2531	A	B	C	D	E	
41	0.2420	A	B	C	D	E	
45	0.2306	A	B	C	D	E	
37	0.2217		B	C	D	E	F
39	0.1794		B	C	D	E	F
56	0.1787		B	C	D	E	F
54	0.1412			C	D	E	F
35	0.1274				D	E	F
50	0.1020					E	F
57	0.0478						F

Summary B

44	0.2869	A	
40	0.2754	A	
58	0.2359	A	B
46	0.2333	A	B
52	0.2320	A	B
55	0.2239	A	B
51	0.2112	A	B
48	0.2073	A	B
45	0.1993	A	B
36	0.1954	A	B
53	0.1853	A	B
38	0.1852	A	B
47	0.1768	A	B
35	0.1573	A	B
37	0.1546	A	B
56	0.1421	A	B
42	0.1253	A	B
54	0.1220	A	B
43	0.1187	A	B
50	0.1011	A	B
49	0.1000	A	B
57	0.0978	A	B
39	0.0971	A	B
41	0.0760		B

Summary C

46	0.4135	A	
45	0.3755	A	
40	0.3436	A	B
47	0.3387	A	B
55	0.3229	A	B
38	0.3188	A	B
44	0.2906	A	B
49	0.2826	A	B
36	0.2608	A	B
37	0.2596	A	B
50	0.2531	A	B
52	0.2475	A	B
48	0.2333	A	B
58	0.2184	A	B
51	0.2169	A	B
53	0.2097	A	B
54	0.2075	A	B
42	0.1984	A	B
43	0.1932	A	B
39	0.1930	A	B
56	0.1675	A	B
41	0.1056		B
35	0.0806		B
57	0.0765		B

Pyramid ANOVA, by update sequence

Summary A

40	0.4019	A
51	0.3453	A B
55	0.3353	A B
36	0.3285	A B
44	0.3216	A B
48	0.3094	A B C
52	0.3054	A B C D
49	0.3039	A B C D
47	0.3025	A B C D
38	0.2937	A B C D
53	0.2897	A B C D
58	0.2796	A B C D E
43	0.2787	A B C D E
46	0.2766	A B C D E
42	0.2531	A B C D E
41	0.2420	A B C D E
45	0.2306	A B C D E
37	0.2217	B C D E F
39	0.1794	B C D E F
56	0.1787	B C D E F
54	0.1412	C D E F
35	0.1274	D E F
50	0.1020	E F
57	0.0478	F

Summary B

44	0.2869	A
40	0.2754	A
58	0.2359	A B
46	0.2333	A B
52	0.2320	A B
55	0.2239	A B
51	0.2112	A B
48	0.2073	A B
45	0.1993	A B
36	0.1954	A B
53	0.1853	A B
38	0.1852	A B
47	0.1768	A B
35	0.1573	A B
37	0.1546	A B
56	0.1421	A B
42	0.1253	A B
54	0.1220	A B
43	0.1187	A B
50	0.1011	A B
49	0.1000	A B
57	0.0978	A B
39	0.0971	A B
41	0.0760	B

Summary C

46	0.4135	A
45	0.3755	A
40	0.3436	A B
47	0.3387	A B
55	0.3229	A B
38	0.3188	A B
44	0.2906	A B
49	0.2826	A B
36	0.2608	A B
37	0.2596	A B
50	0.2531	A B
52	0.2475	A B
48	0.2333	A B
58	0.2184	A B
51	0.2169	A B
53	0.2097	A B
54	0.2075	A B
42	0.1984	A B
43	0.1932	A B
39	0.1930	A B
56	0.1675	A B
41	0.1056	B
35	0.0806	B
57	0.0765	B

Pyramid ANOVA, by update sequence

Summary A

40	0.4019	A
51	0.3453	A B
55	0.3353	A B
36	0.3285	A B
44	0.3216	A B
48	0.3094	A B C
52	0.3054	A B C D
49	0.3039	A B C D
47	0.3025	A B C D
38	0.2937	A B C D
53	0.2897	A B C D
58	0.2796	A B C D E
43	0.2787	A B C D E
46	0.2766	A B C D E
42	0.2531	A B C D E
41	0.2420	A B C D E
45	0.2306	A B C D E
37	0.2217	B C D E F
39	0.1794	B C D E F
56	0.1787	B C D E F
54	0.1412	C D E F
35	0.1274	D E F
50	0.1020	E F
57	0.0478	F

Summary B

44	0.2869	A
40	0.2754	A
58	0.2359	A B
46	0.2333	A B
52	0.2320	A B
55	0.2239	A B
51	0.2112	A B
48	0.2073	A B
45	0.1993	A B
36	0.1954	A B
53	0.1853	A B
38	0.1852	A B
47	0.1768	A B
35	0.1573	A B
37	0.1546	A B
56	0.1421	A B
42	0.1253	A B
54	0.1220	A B
43	0.1187	A B
50	0.1011	A B
49	0.1000	A B
57	0.0978	A B
39	0.0971	A B
41	0.0760	B

Summary C

46	0.4135	A
45	0.3755	A
40	0.3436	A B
47	0.3387	A B
55	0.3229	A B
38	0.3188	A B
44	0.2906	A B
49	0.2826	A B
36	0.2608	A B
37	0.2596	A B
50	0.2531	A B
52	0.2475	A B
48	0.2333	A B
58	0.2184	A B
51	0.2169	A B
53	0.2097	A B
54	0.2075	A B
42	0.1984	A B
43	0.1932	A B
39	0.1930	A B
56	0.1675	A B
41	0.1056	B
35	0.0806	B
57	0.0765	B

Pyramid ANOVA, by update sequence

Summary A

40	0.4019	A					
51	0.3453	A	B				
55	0.3353	A	B				
36	0.3285	A	B				
44	0.3216	A	B				
48	0.3094	A	B	C			
52	0.3054	A	B	C	D		
49	0.3039	A	B	C	D		
47	0.3025	A	B	C	D		
38	0.2937	A	B	C	D		
53	0.2897	A	B	C	D		
58	0.2796	A	B	C	D	E	
43	0.2787	A	B	C	D	E	
46	0.2766	A	B	C	D	E	
42	0.2531	A	B	C	D	E	
41	0.2420	A	B	C	D	E	
45	0.2306	A	B	C	D	E	
37	0.2217		B	C	D	E	F
39	0.1794		B	C	D	E	F
56	0.1787		B	C	D	E	F
54	0.1412			C	D	E	F
35	0.1274				D	E	F
50	0.1020					E	F
57	0.0478						F

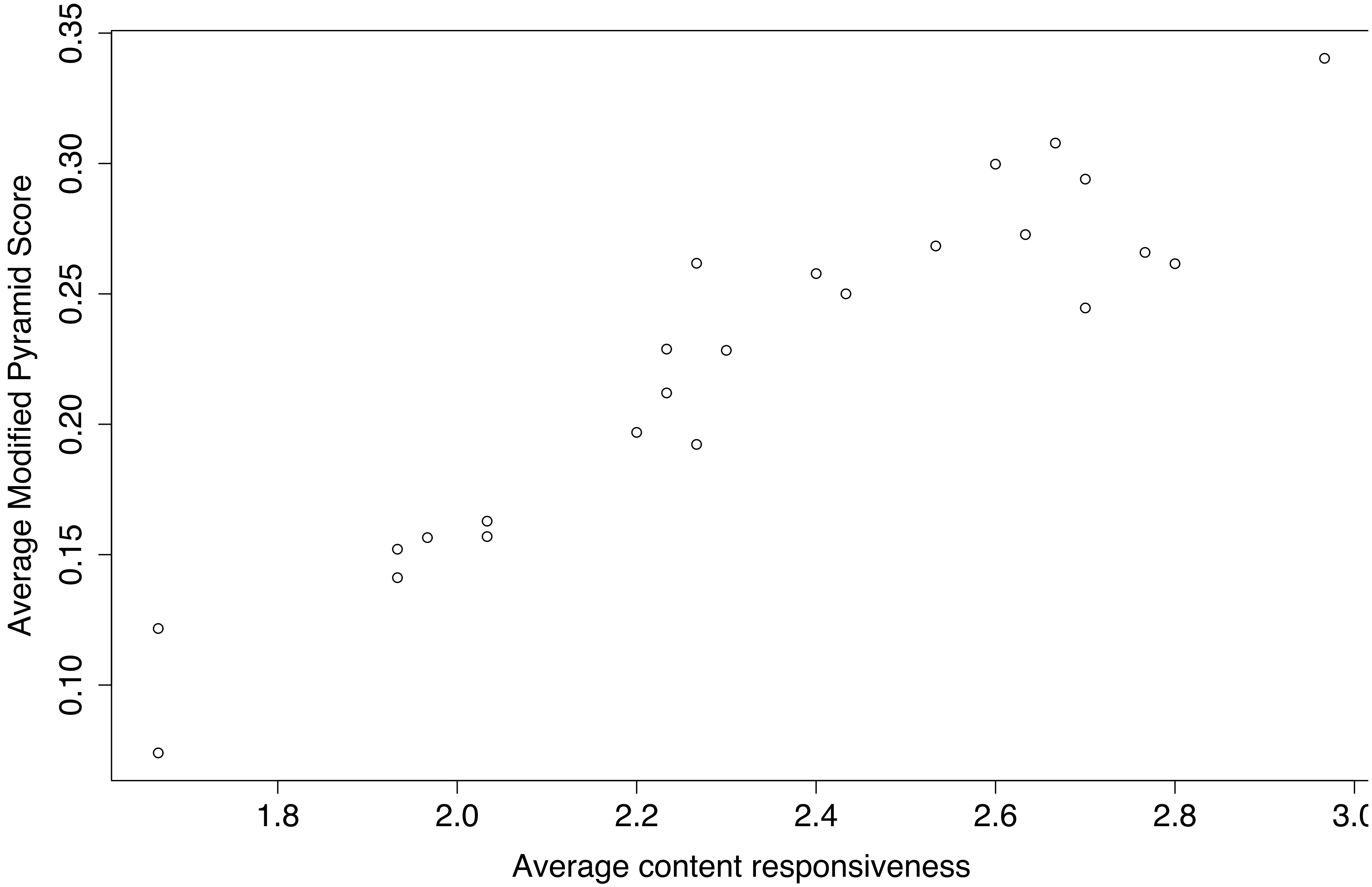
Summary B

44	0.2869	A	
40	0.2754	A	
58	0.2359	A	B
46	0.2333	A	B
52	0.2320	A	B
55	0.2239	A	B
51	0.2112	A	B
48	0.2073	A	B
45	0.1993	A	B
36	0.1954	A	B
53	0.1853	A	B
38	0.1852	A	B
47	0.1768	A	B
35	0.1573	A	B
37	0.1546	A	B
56	0.1421	A	B
42	0.1253	A	B
54	0.1220	A	B
43	0.1187	A	B
50	0.1011	A	B
49	0.1000	A	B
57	0.0978	A	B
39	0.0971	A	B
41	0.0760		B

Summary C

46	0.4135	A	
45	0.3755	A	
40	0.3436	A	B
47	0.3387	A	B
55	0.3229	A	B
38	0.3188	A	B
44	0.2906	A	B
49	0.2826	A	B
36	0.2608	A	B
37	0.2596	A	B
50	0.2531	A	B
52	0.2475	A	B
48	0.2333	A	B
58	0.2184	A	B
51	0.2169	A	B
53	0.2097	A	B
54	0.2075	A	B
42	0.1984	A	B
43	0.1932	A	B
39	0.1930	A	B
56	0.1675	A	B
41	0.1056		B
35	0.0806		B
57	0.0765		B

Responsiveness vs. Pyramid



Spearman: 0.899 Pearson: 0.937 [0.875, 1.00]

Conclusion

- Main Task:
  - Systems are getting better at task
  - Topic focus matters
- Update Pilot:
  - Straightforward representation of user knowledge
  - Good correlation between average responsiveness and pyramid scores (30 doc clusters x 24 systems)
- NIST assessors make good pyramid builders!