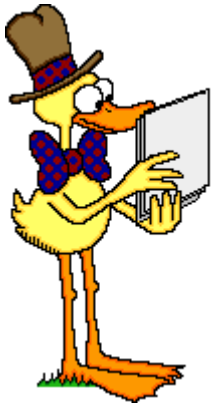


# An Introduction to DUC-2004

## Intrinsic Evaluation of Generic News Text Summarization Systems

Paul Over
Retrieval Group
Information Access Division

James Yen
Statistical Modeling and Analysis Group
Statistical Engineering Division

National Institute of Standards and Technology

Sponsored by DARPA and ARDA

# Document Understanding Conferences (DUC)

- Summarization has always been a TIDES component
- An evaluation roadmap created in 2000 after spring TIDES PI meeting
- Year 1 (DUC-2001 at SIGIR in September 2001)
  - Intrinsic evaluation of generic summaries,
    - of newswire/paper stories for single and multiple documents;
    - with fixed target lengths of 50, 100, 200, and 400 words
- Year 2 (DUC-2002 at ACL '02 in July 2002)
  - Abstracts of single documents and document sets
    - fixed lengths of 10, 50, 100, and 200 words
  - Extracts of document sets
    - fixed target lengths of 200 and 400 words
- Year 3 (DUC-2003 at HLT/NAACL in May 2003)
  - Abstracts of single documents and document sets
    - Target lengths of 10 and 100 words
    - multi-document summaries focused by
      - TDT event topics, Viewpoints, Question topics

# Goals of the talk

- Provide an overview of DUC 2004:
  - Data: documents, manual summaries, translations, questions
  - Tasks:
    - 1 - very short (<= 75 bytes) single document summaries
    - 2 - short (<= 665 bytes) single document summaries
    - 3 - very short
    - 4 - short
    - 5 - short
  - Evaluation: procedures, measures
- Introduce the results:
  - Basics of system performance on the measures
  - Sanity checking the results and measures
  - Exploration of various questions:
    - Performance of systems / baselines / humans on various measures
    - Relative performance among systems – significant differences?
    - Relationship of ROUGE scores to themselves and SEE coverage
- Invite  further exploration of the data …

# Data: Formation of test document sets

- 50 TDT English news clusters
  - 50 event topics chosen by NIST
  - ~ 10 documents / topic
  - NIST chose a subset of the documents the TDT annotator decided were “on topic”
- 24 TDT Arabic news clusters
  - 12 of the above topics with relevant docs in the Arabic source
  - 12 new topics in the same style, created by LDC
- 50 TREC English news clusters
  - NIST assessors explored the collection and created clusters
  - ~ 10 documents / cluster
  - Each cluster had to contain documents which contributed to answering a broad question “Who is X?”, where X was a person

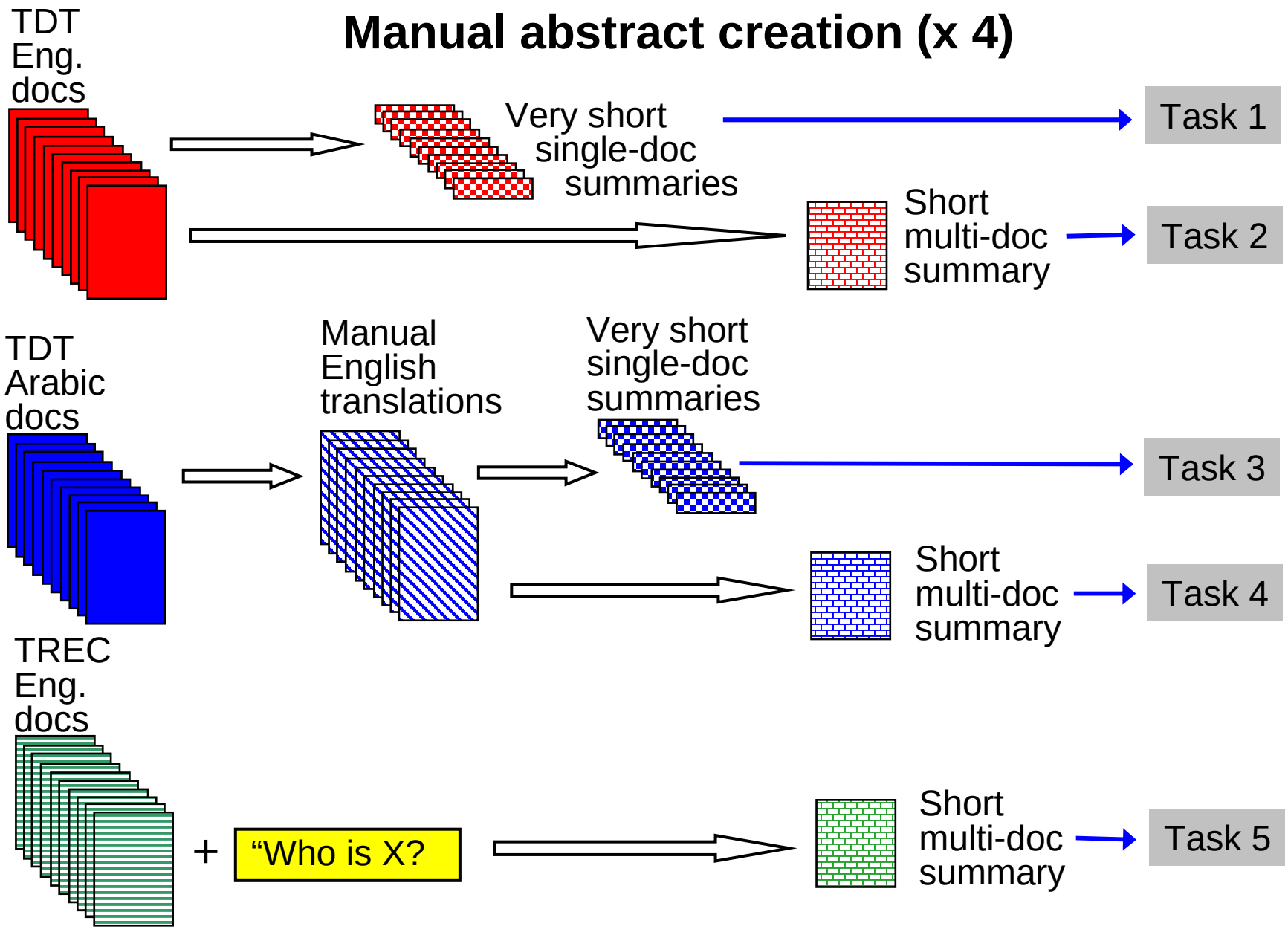
# Evaluation basics

- ROUGE (tasks 1 – 5) with 4 models max
- SEE Content coverage and linguistic quality (tasks 2, 5)
  - Intrinsic evaluation by humans using special rewritten version of SEE (thanks to Lei Ding, Chin-Yew Lin at ISI)
  - Compare:
    - a model summary - manual, authored by a human
    - a peer summary - system-created, baseline, or additional manual
  - Produce judgments of:
    - Peer linguistic quality (7 questions – thanks to Ani Nenkova et al)
    - Coverage of each model unit by the peer (recall)
    - Relevance of peer-only material (not covered here)
- Responsiveness (task 5):
  - Comparison together of all peer summaries for a given docset
  - Assignment of each summary to one of 5 bins

# ROUGE basics

- Recall-oriented, within-sentence word overlap with model(s)
- Developed by Chin-Yew Lin at ISI/USC
- Available from <http://www.isi.edu/~cyl/ROUGE>
- Models - no theoretical limit to number
  - compared system output to 4 models
  - compared manual summaries to 3 models
- ROUGE v1.2.1 measures for DUC 2004
  - ROUGE-1,2,3,4: N-gram matching where N = 1,2,3,4
  - ROUGE-LCS: Longest common substring
  - ROUGE-W-1.2 Favors LCS with least intervening material

# Manual abstract creation (x 4)



# ROUGE – runtime arguments for DUC 2004

rouge –a -c 95 -b 75 -m -n 4 -w 1.2

- a Evaluate all systems
- c 95 Calculate 95% confidence intervals
- b 75 Truncate model and peer at 75 (or 665) bytes
- m Stem (Porter) models and peers
- n 4 Calculate ROUGE-1..4
- w 1.2 Use 1.2 as the weighting factor for LCS-W
- Do not drop stop words

# SEE basics: Models

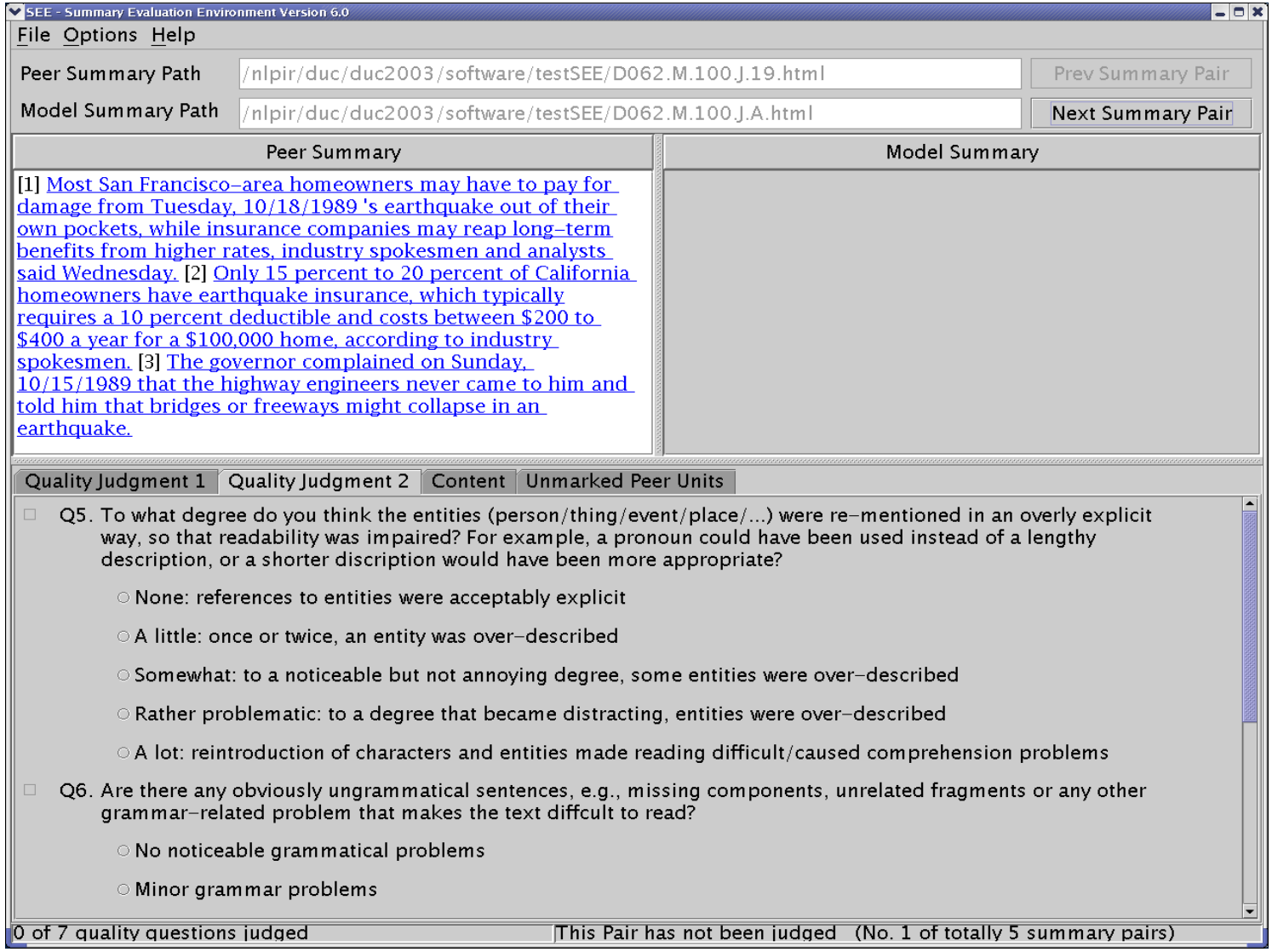
- Source:
  - Authored by a human
  - For 2004, the assessor is always the model's author
- Formatting:
  - Divided into model units (MUs) and more
    - (MUs == EDUs - thanks to Radu Soricut at ISI)
    - Surprise: some tokens changed by chunking process
      - Won't Ł will n't
  - Lightly edited by authors to integrate uninterpretable fragments
    - George Bush's selection of Dan Quayle
    - as his running mate surprised many
    - many political observers thought him a lightweight with baggage
    - to carry
  - Flowed together with HTML tags for SEE

# SEE basics: Peers

- Formatting:
  - Divided into peer units (PUs) –
    - simple automatically determined sentences
    - tuned slightly to documents and submissions
      - Abbreviations list
      - List of proper nouns
    - Flowed together with HTML tags for SEE

- 4 Sources:
  1. Automatically generated by baseline algorithms: 1 – 5
  2. Automatically generated by research systems: 6 – 151
  3. Authored by a human (the assessor): A – H

# SEE: overall peer quality



# **SEE: Overall peer quality**
## 7 Questions

1 Does the summary build from sentence to sentence to a coherent body of information about the topic?

**A**. Very coherently
**B**. Somewhat coherently
**C**. Neutral as to coherence
**D**. Not so coherently
**E**. Incoherent

2 If you were editing the summary to make it more concise and to the point, how much useless, confusing or repetitive text would you remove from the existing summary?

**A**. None
**B**. A little
**C**. Some
**D**. A lot
**E**. Most of the text

3 To what degree does the summary say the same thing over again?

**A**. None; the summary has no repeated information
**B**. Minor repetitions
**C**. Some repetition
**D**. More than half of the text is repetitive
**E**. Quite a lot; most sentences are repetitive

# SEE: Overall peer quality

4 How much trouble did you have identifying the referents of noun phrases in this summary? Are there nouns, pronouns or personal names that are not well-specified? For example, a person is mentioned and it is not clear what his role in the story is, or any other entity that is referenced but its identity and relation with the story remains unclear

**A**. No problems; it is clear who/what is being referred to throughout.

**B**. Slight problems, mostly cosmetic/stylistic

**C**. Somewhat problematic; some minor events/things/people/places are unclear, or a very few major ones, but overall the who and what are clear.

**D**. Rather problematic; enough events/things/people/places are unclear that parts of the summary are hard to understand

**E**. Severe problems; main events, characters or places are not well-specified and/or it's difficult to say how they relate to the topic

5 To what degree do you think the entities (person/thing/event/place/...) were re-mentioned in an overly explicit way, so that readability was impaired? For example, a pronoun could have been used instead of a lengthy description, or a shorter description would have been more appropriate?

**A**. None: references to entities were acceptably explicit

**B**. A little: once or twice, an entity was over-described

**C**. Somewhat: to a noticeable but not annoying degree, some entities were over-described

**D**. Rather problematic: to a degree that became distracting, entities were over-described

**E**. A lot: reintroduction of characters and entities made reading difficult/caused comprehension problems

# SEE: Overall peer quality

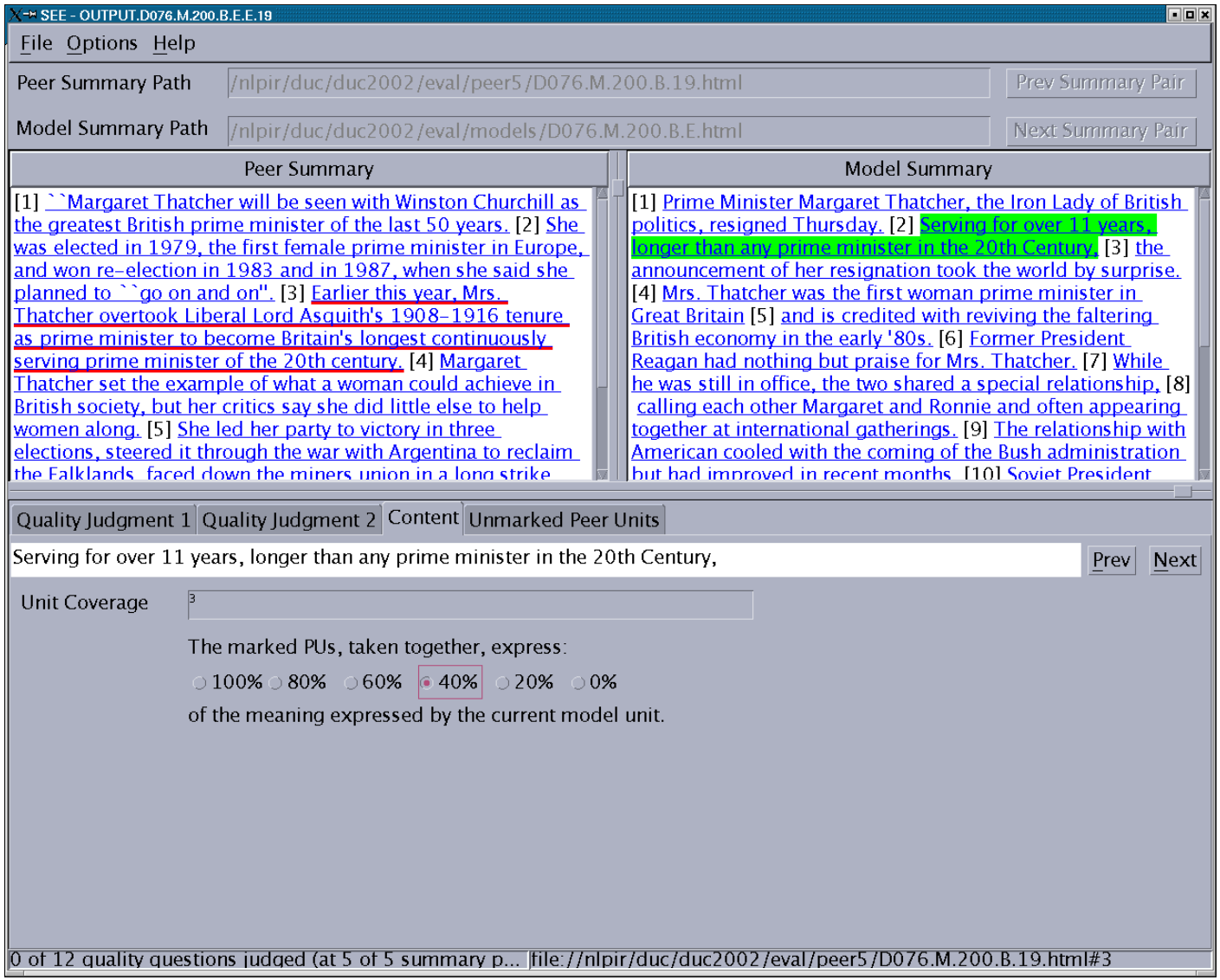
6 Are there any obviously ungrammatical sentences, e.g., missing components, unrelated fragments or any other grammar-related problem that makes the text difficult to read.

**A**. No noticeable grammatical problems
**B**. Minor grammar problems
**C**. Some problems, but overall acceptable
**D**. A fair amount of grammatical errors
**E**. Too many problems, the summary is impossible to read

7 Are there any datelines, system-internal formatting or capitalization errors that can make the reading of the summary difficult?

**A**. No noticeable formatting problems
**B**. Minor formatting problems
**C**. Some, but they do not create any major difficulties
**D**. A fair amount of formatting problems
**E**. Many, to an extent that reading is difficult

# SEE: per-unit content



# Per-unit content: evaluation details

- “First, find all the peer units which tell you at least some of what the current model unit tells you, i.e., peer units which express at least some of the same facts as the current model unit. When you find such a PU, click on it to mark it.
- “When you have marked all such PUs for the current MU, then think about the whole set of marked PUs and answer the question:”
- “The marked PUs, taken together, express about
  0% 20% 40% 60% 80% 100%
  of the meaning expressed by the current model unit”
- Mean coverage:
  - average of the per-MU completeness judgments [0, 20, 40, 60, 80,100]% for a peer summary

# Tasks 1 & 2

# Task 2: Short summary of a TDT document set

- System task:
  - Use the 50 English TDT clusters
    - ~ 10 documents/cluster
  - Given:
    - each document cluster
  - Create a short summary (<=665 bytes) of the cluster
- Coverage baseline 2:
  - Take the first 665 bytes of the TEXT of the most recent document
  - Note: for linguistic quality this is really not a baseline – it is contiguous human-authored text.
- Evaluation:
  - SEE (unplanned – done to provide more info on meaning of ROUGE)
    - Linguistic quality
    - Coverage
    - Extra material
  - ROUGE

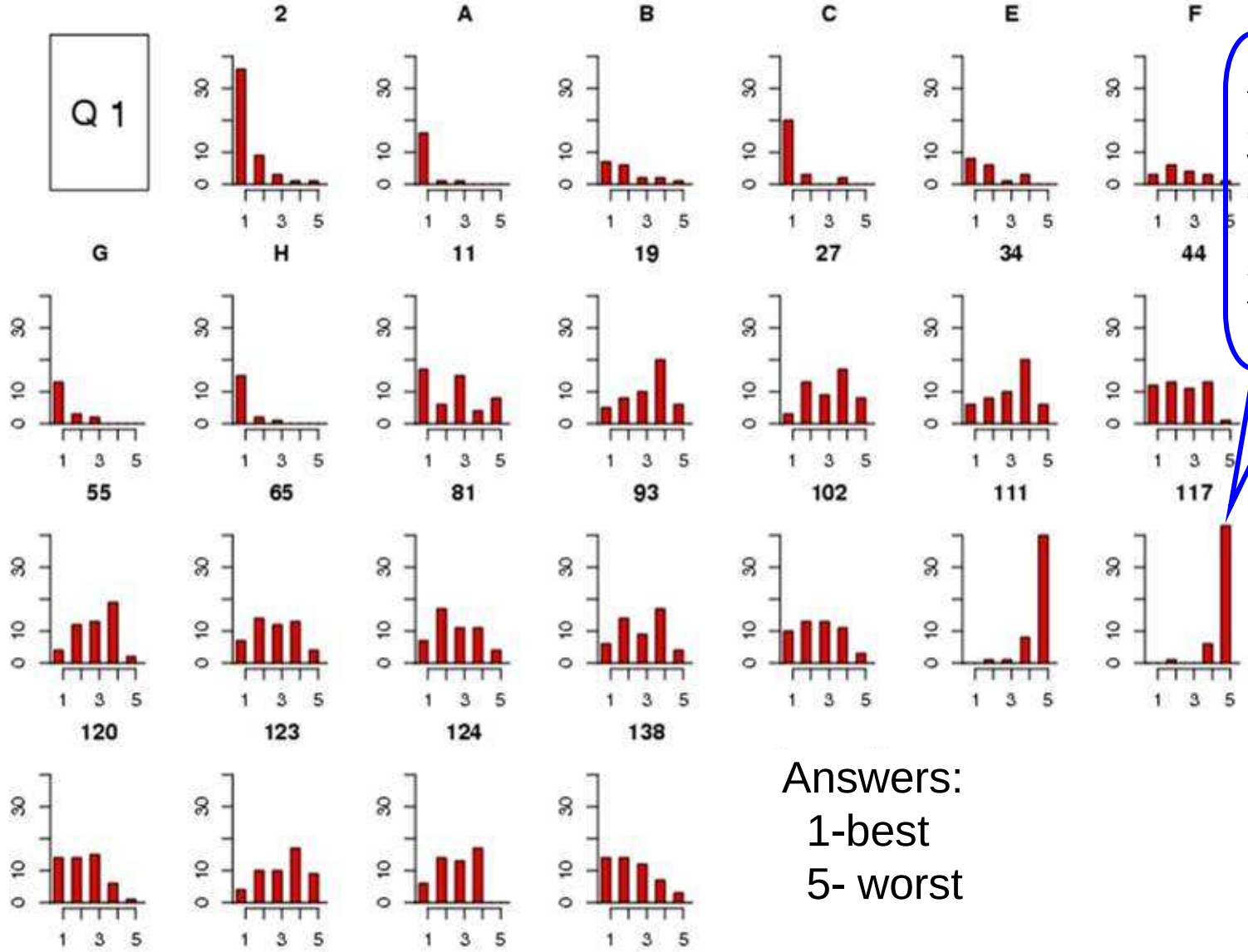
# Task 2: Participants and runs

| Sysid | Priority | Run | Group |
|---|---|---|---|
| CL | 1 | 11 | CL Research |
| LARIS.2004 | 1 | 19 | Laris Labs |
| ULeth2004 | 1 | 27 | U. Lethbridge |
| ULeth2004 | 2 | 28 | |
| ULeth2004 | 3 | 29 | |
| MEDLAB_Fudan | 1 | 34 | Fudan U. |
| MEDLAB_Fudan | 2 | 35 | |
| MEDLAB_Fudan | 3 | 36 | |
| columbia1 | 1 | 44 | Columbia U. |
| columbia1 | 2 | 45 | |
| CLaCDUCTape2 | 1 | 55 | Concordia U. |
| CLaCDUCTape2 | 2 | 56 | |
| CLaCDUCTape2 | 3 | 57 | |
| CCSNSA04 | 1 | 65 | NSA |
| CCSNSA04 | 2 | 66 | |
| CCSNSA04 | 3 | 67 | |
| webcl2004 | 1 | 81 | ISI |

| Sysid | Priority | Run | Group |
|---|---|---|---|
| kul.2004 | 1 | 93 | KU Leuven |
| kul.2004 | 2 | 94 | |
| kul.2004 | 3 | 95 | |
| lcc.duc04 | 1 | 102 | LCC |
| lcc.duc04 | 2 | 103 | |
| lcc.duc04 | 3 | 104 | |
| uofo | 1 | 111 | U. Ottawa |
| msr-nlp.duc2004 | 1 | 117 | Microsoft |
| msr-nlp.duc2004 | 2 | 118 | |
| msr-nlp.duc2004 | 3 | 119 | |
| crl_nyu.duc04 | 1 | 120 | CRL/NYU |
| crl_nyu.duc04 | 2 | 121 | |
| nttcslab.duc2004 | 1 | 123 | NTT |
| shef2004.saggion | 1 | 124 | U. Sheffield |
| UofM-MEAD | 1 | 138 | U. Michigan |
| UofM-MEAD | 2 | 139 | |
| UofM-MEAD | 3 | 140 | |

# Task 2: Linguistic quality

Question 1 – builds to a coherent body of information?




Answers:

1-best

5- worst

# Task 2: Linguistic quality

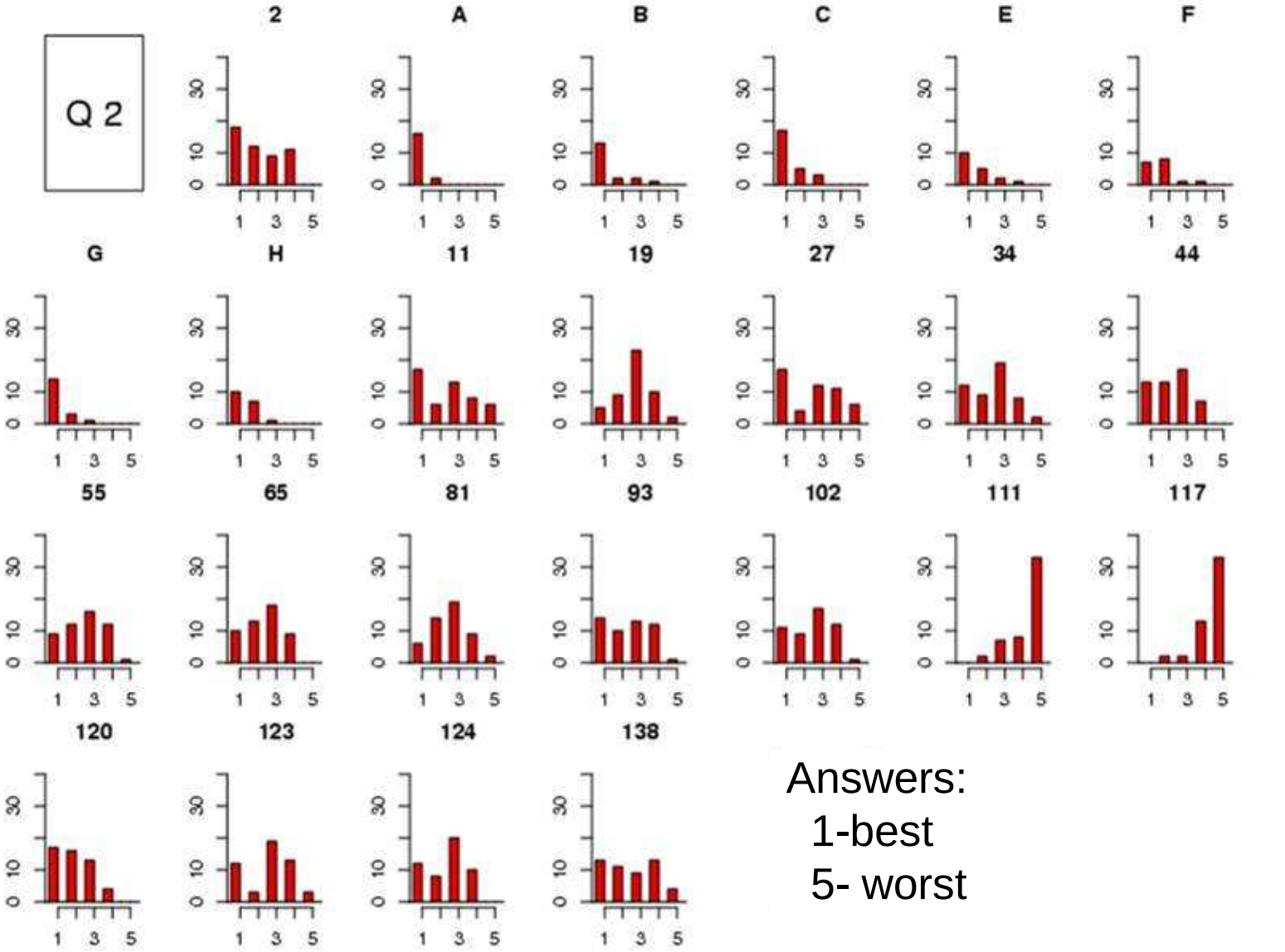
Question 1 – builds to a coherent body of information?

Example.

117: The senior opposition FUNCINPEC party refused to form a new one Party and the two party opposition had called on the monarch to lead top level talks Hun Sen holds bilateral talks "The CPP would like to launch an appeal, consider this draft resolution, and give justice to the CPP, Hun Sen, and Sihanouk's Cambodian people by not approving it." In Nov. 13, uncompromising enemies agreed a few months Hun Sen and FUNCINPEC President Prince Norodom Ranariddh ago to form a coalition government, with the party, at a summit convened by Sihanouk. Ranariddh and Sam Rainsy, in Cambodia, have remained outside the country since the convening of parliament in Sept. 24."No

# Task 2: Linguistic quality

Question 2 – useless, confusing, repetitive text ?



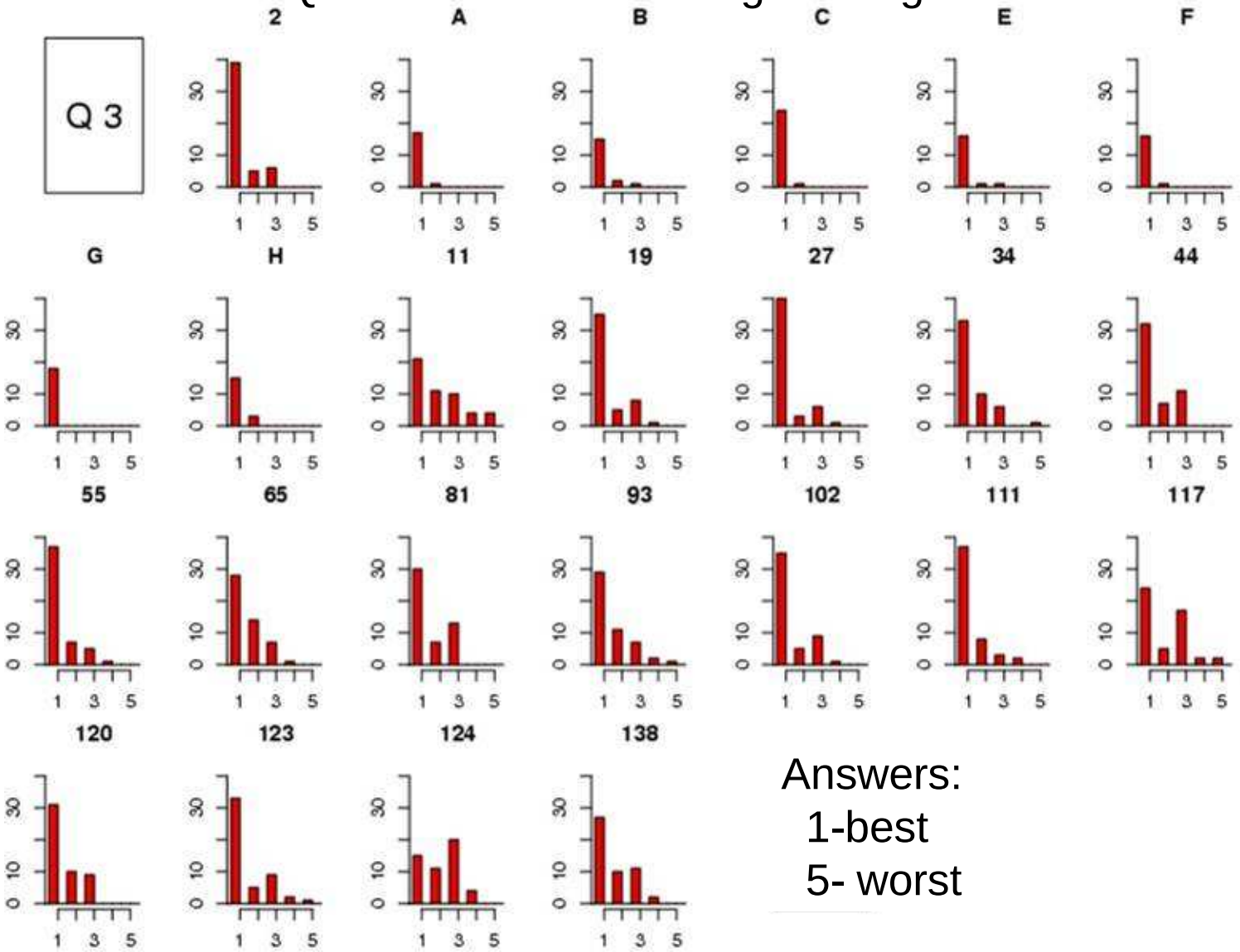

Answers:

1-best

5- worst

# Task 2: Linguistic quality

Question 3 – same thing over again ?



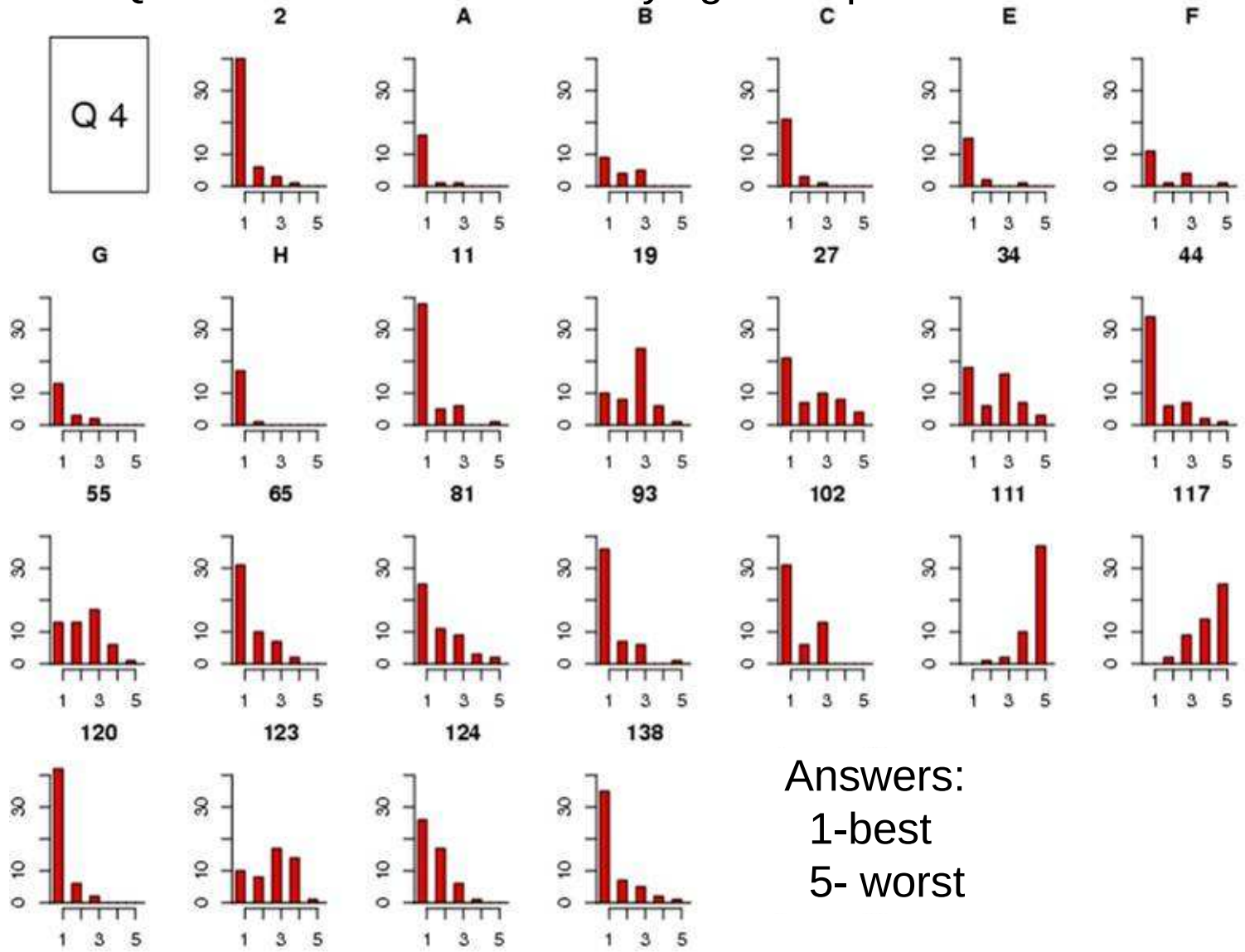

Answers:

1-best

5- worst

# Task 2: Linguistic quality

Question 4 – trouble identifying noun phrase referents?




Answers:

1-best

5- worst

# Task 2: Linguistic quality

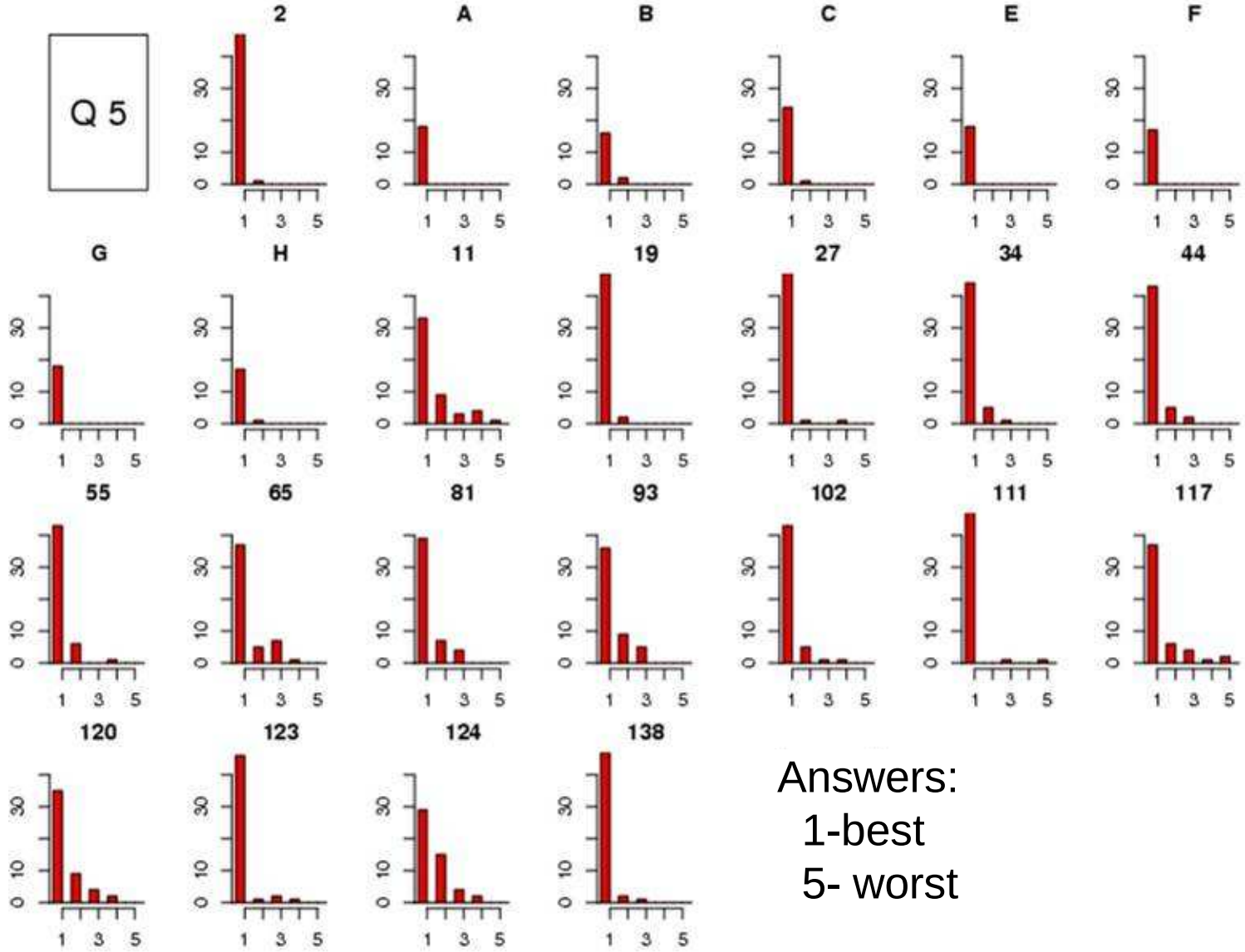
Question 4 – trouble identifying noun phrase referents?
Example

111: Furthermore, such a proposal is unconstitutional," the faxed statement said". At least four demonstrators were killed by police, but the discovery of more than 20 bodies in the aftermath has prompted = speculation that the death tally could be much higher. They have demanded a thorough = investigation into their election complaints as a precondition for their cooperation in getting the national assembly moving and a new government formed. The prince's party, in a statement dated Friday and seen Saturday, said such a scenario was unconstitutional. A copy of the resolution has since been submitted to the U.S. Senate Committee on Foreign Relations.

120: Worried that party colleagues still face arrest for their politics, opposition leader Sam Rainsy sought further clarification Friday of security guarantees promised by strongman Hun Sen. Sam Rainsy wrote in a letter to King Norodom Sihanouk that he was eager to attend the first session of the new National Assembly on Nov. 25, but complained that Hun Sen's assurances were not strong enough to ease concerns his party members may be arrested upon their return to Cambodia.

# Task 2: Linguistic quality

Question 5 – entities re-mentioned ?



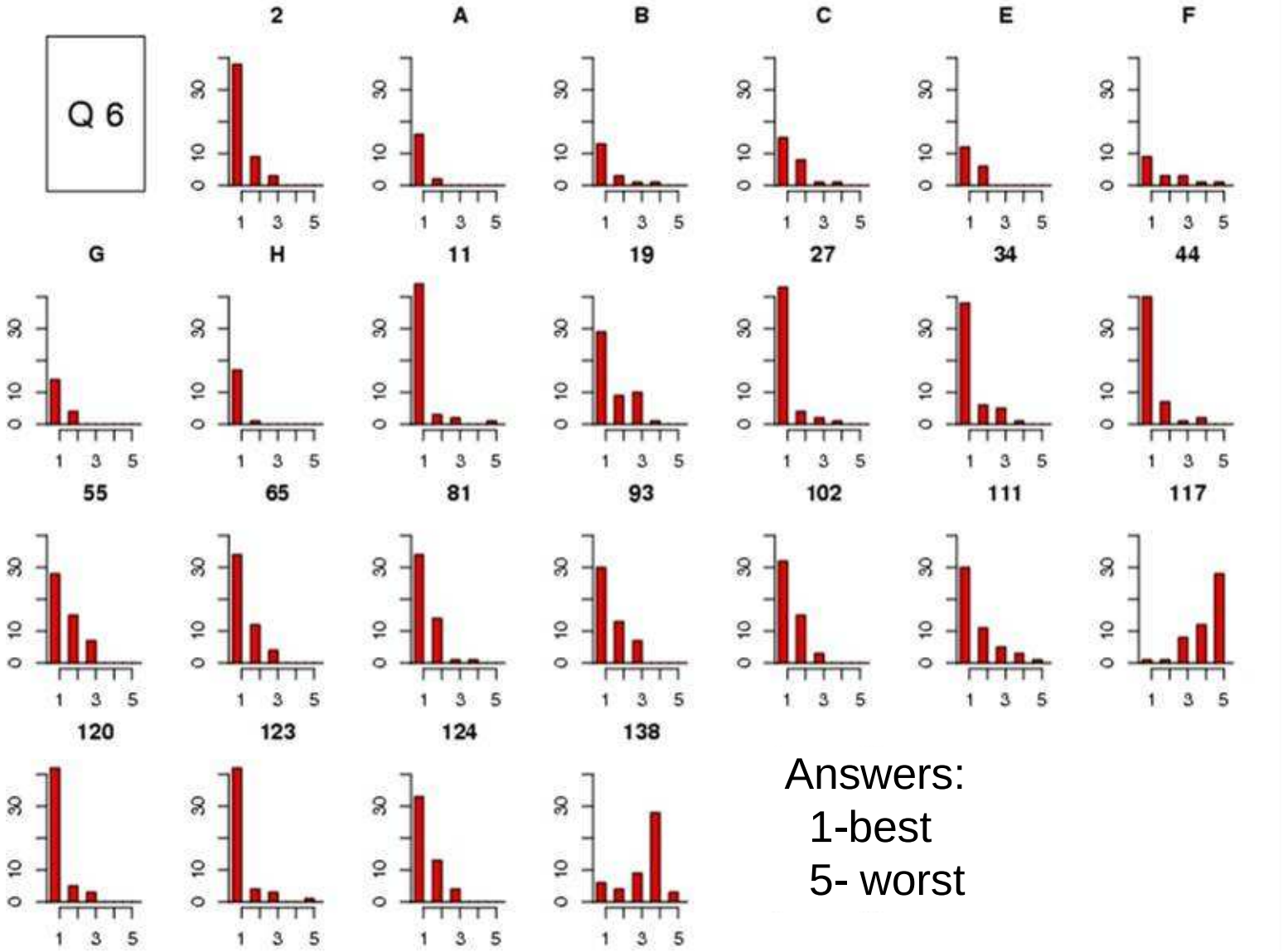

Answers:

1-best

5- worst

# Task 2: Linguistic quality

Question 6 – ungrammatical sentences?



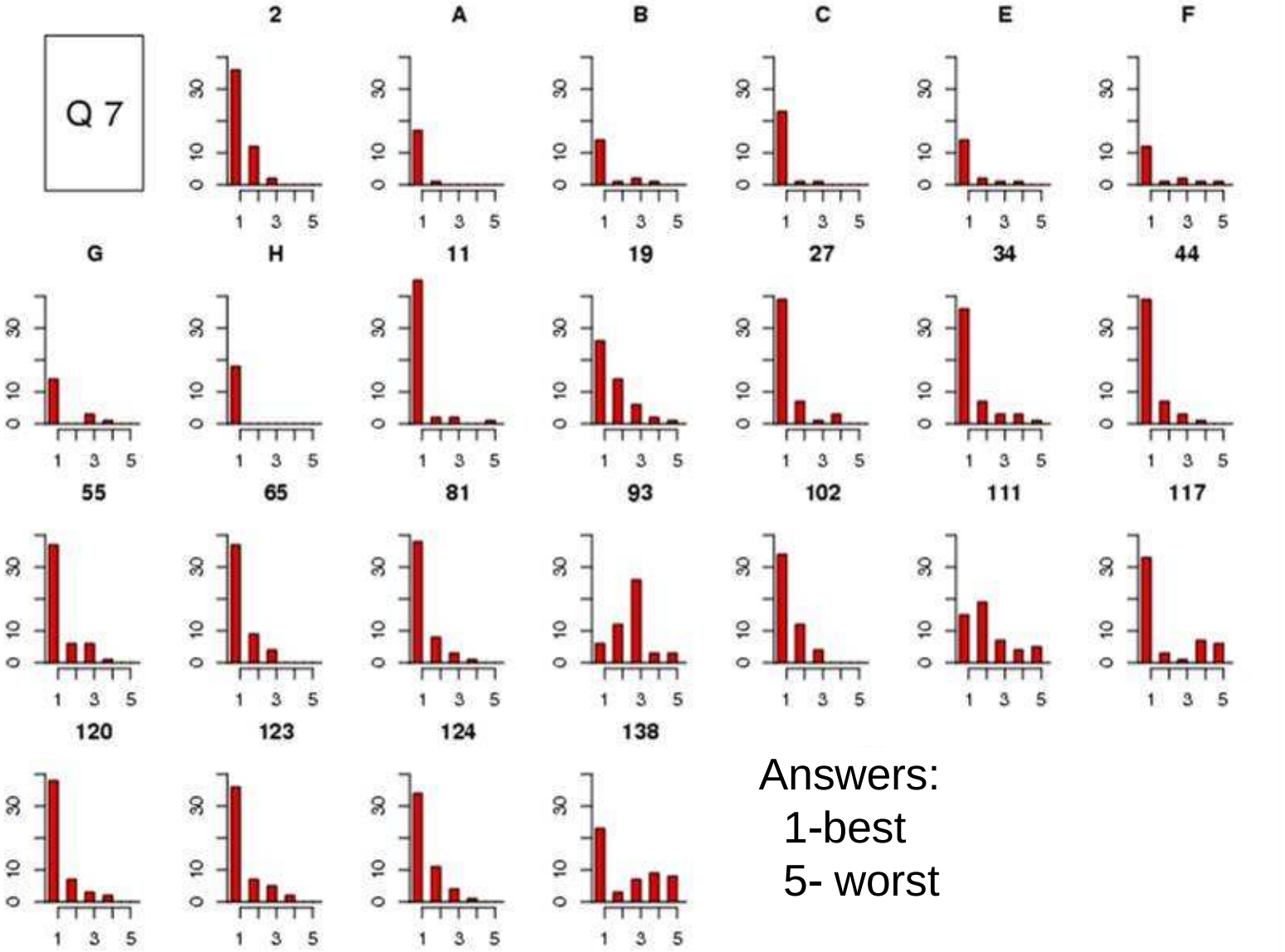

Answers:

1-best

5- worst

# Task 2: Linguistic quality

Question 7 – datelines, formatting, capitalization?




Answers:

1-best

5- worst

# Overall peer quality

## Task 2 – multiple comparisons (best on top)

**q1**

| | | | | |
|---|---|---|---|---|
| 120 | A | | | |
| 138 | A | B | | |
| 44 | A | B | | |
| 102 | A | B | D | |
| 11 | A | B | D | |
| 81 | A | B | D | |
| 65 | A | B | D | |
| 124 | A | B | D | |
| 93 | A | B | D | |
| 55 | A | B | D | |
| 34 | | B | D | |
| 123 | | | D | |
| 19 | | | D | |
| 27 | | | D | |
| 111 | | | | O |
| 117 | | | | O |

**q3**

| | | | | |
|---|---|---|---|---|
| 27 | A | | | |
| 111 | A | | | |
| 55 | A | C | | |
| 19 | A | C | | |
| 102 | A | C | | |
| 34 | A | C | | |
| 44 | A | C | | |
| 120 | A | C | | |
| 123 | A | C | | |
| 65 | A | C | | |
| 81 | A | C | L | |
| 93 | A | C | L | |
| 138 | A | C | L | |
| 117 | | C | L | |
| 11 | | | L | |
| 124 | | | L | |

**q5**

| | | | | |
|---|---|---|---|---|
| 19 | A | | | |
| 27 | A | B | | |
| 111 | A | B | | |
| 138 | A | B | | |
| 123 | A | B | | |
| 34 | A | B | F | |
| 44 | A | B | F | |
| 102 | A | B | F | |
| 55 | A | B | F | |
| 81 | A | B | F | J |
| 65 | A | B | F | J |
| 93 | A | B | F | J |
| 117 | A | B | F | J |
| 120 | | B | F | J |
| 11 | | | F | J |
| 124 | | | | J |

**q7**

| | | | | |
|---|---|---|---|---|
| 11 | A | | | |
| 81 | A | B | | |
| 27 | A | B | | |
| 65 | A | B | | |
| 120 | A | B | | |
| 44 | A | B | | |
| 102 | A | B | | |
| 55 | A | B | | |
| 124 | A | B | | |
| 123 | A | B | | |
| 34 | A | B | | |
| 117 | | B | L | |
| 19 | | B | L | |
| 138 | | | L | N |
| 111 | | | L | N |
| 93 | | | | N |

**q2**

| | | | | |
|---|---|---|---|---|
| 120 | A | | | |
| 44 | A | B | | |
| 65 | A | B | | |
| 93 | A | B | | |
| 124 | A | B | | |
| 34 | A | B | | |
| 11 | A | B | | |
| 102 | A | B | | |
| 55 | A | B | | |
| 138 | A | B | | |
| 81 | A | B | | |
| 27 | A | B | | |
| 123 | A | B | | |
| 19 | | B | | |
| 111 | | | O | |
| 117 | | | O | |

**q4**

| | | | | | | |
|---|---|---|---|---|---|---|
| 120 | A | | | | | |
| 11 | A | B | | | | |
| 93 | A | B | | | | |
| 138 | A | B | | | | |
| 44 | A | B | | | | |
| 65 | A | B | F | | | |
| 124 | A | B | F | | | |
| 102 | A | B | F | | | |
| 81 | A | B | F | I | | |
| 27 | | B | F | I | J | |
| 34 | | | F | I | J | |
| 55 | | | F | I | J | |
| 19 | | | | I | J | |
| 123 | | | | | J | |
| 117 | | | | | | O |
| 111 | | | | | | O |

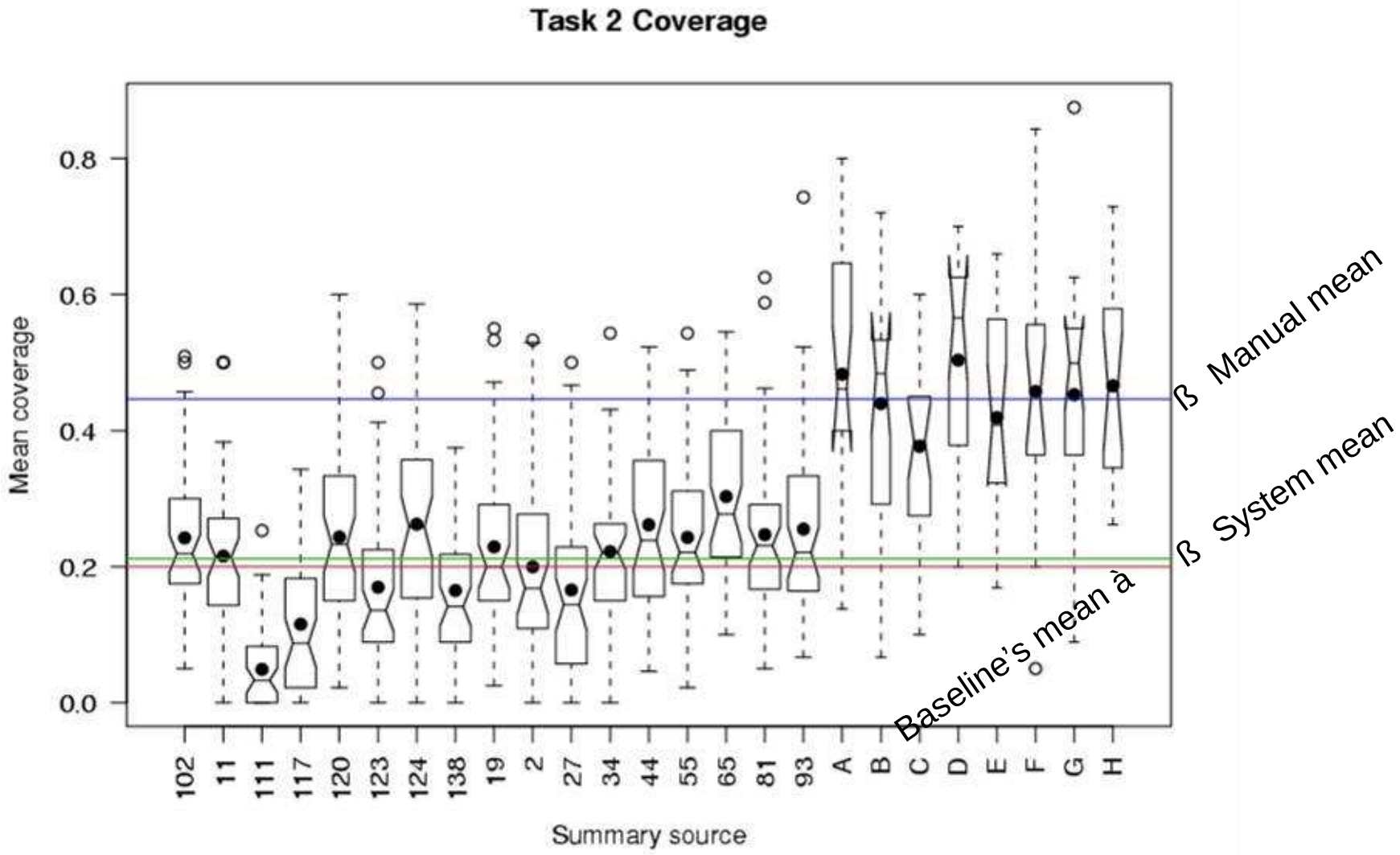
**q6**

| | | |
|---|---|---|
| 11 | A | |
| 27 | A | |
| 120 | A | |
| 123 | A | |
| 44 | A | |
| 34 | A | |
| 65 | A | |
| 81 | A | |
| 102 | A | |
| 124 | A | |
| 93 | A | |
| 111 | A | |
| 19 | A | |
| 55 | A | |
| 138 | | O |
| 117 | | O |

Means with the same letter are not significantly different.

Tukey-Kramer criterion (.05) on average ranks from Friedman's test

# Task 2: Mean coverage by summary source



# Tasks 2: ANOVA on coverage

| Source | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| Model | 99 | 13.38023203 | 0.13515386 | 10.35 |
| Error | 849 | 11.08426032 | 0.01305567 | |
| Corrected Total | 948 | 24.46449235 | | |

| Source | Pr > F |
|---|---|
| Model | <.0001 |
| Error | |
| Corrected Total | |

| R-Square | Coeff Var | Root MSE | meanco Mean |
|---|---|---|---|
| 0.546925 | 45.91959 | 0.114261 | 0.248829 |

| Source | DF | Type I SS | Mean Square | F Value |
|---|---|---|---|---|
| docset | 49 | 5.15955811 | 0.10529710 | 8.07 |
| group | 1 | 6.94165185 | 6.94165185 | 531.70 |
| docset*group | 49 | 1.27902207 | 0.02610249 | 2.00 |

| Source | Pr > F |
|---|---|
| docset | <.0001 |
| group | <.0001 |
| docset*group | <.0001 |

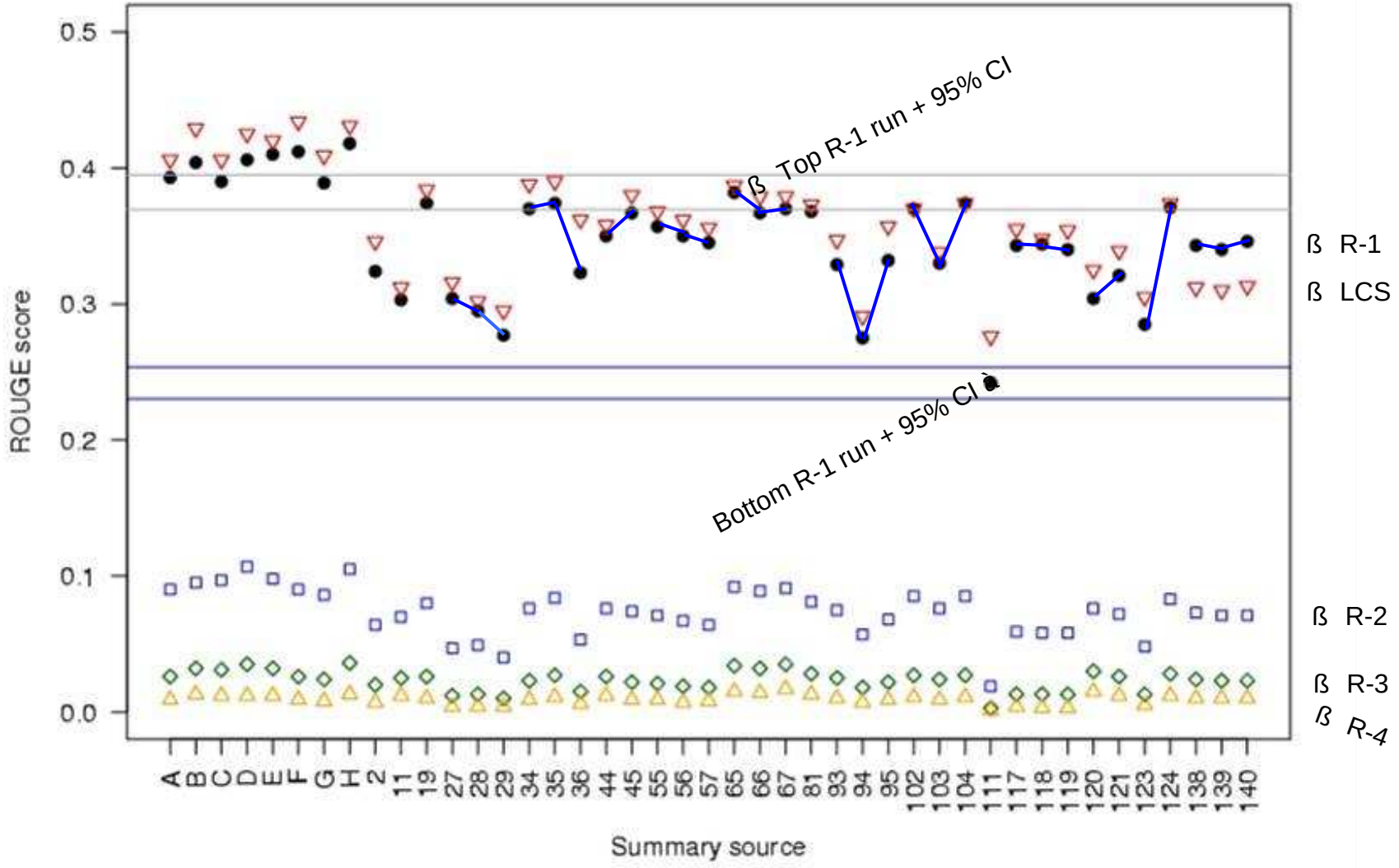
# Task 2: Multiple comparisons on mean coverage

| REGWQ Grouping | | Mean | N | peer |
|---|---|---|---|---|
| | A | 0.30304 | 50 | 65 |
| B | A | 0.26228 | 50 | 124 |
| B | A | 0.26152 | 50 | 44 |
| B | A | 0.25512 | 50 | 93 |
| B | A | 0.24704 | 50 | 81 |
| B | A | 0.24346 | 50 | 120 |
| B | A | 0.24284 | 50 | 55 |
| B | A | 0.24220 | 50 | 102 |
| B | | 0.22906 | 49 | 19 |
| B | C | 0.22198 | 50 | 34 |
| B | C | 0.21552 | 50 | 11 |
| D | C | 0.16968 | 50 | 123 |
| D | C | 0.16566 | 50 | 27 |
| D | C | 0.16492 | 50 | 138 |
| D | | 0.11536 | 50 | 117 |
| | E | 0.04900 | 50 | 111 |

Means with the same letter are not significantly different.

# Task 2: ROUGE scores by summary source

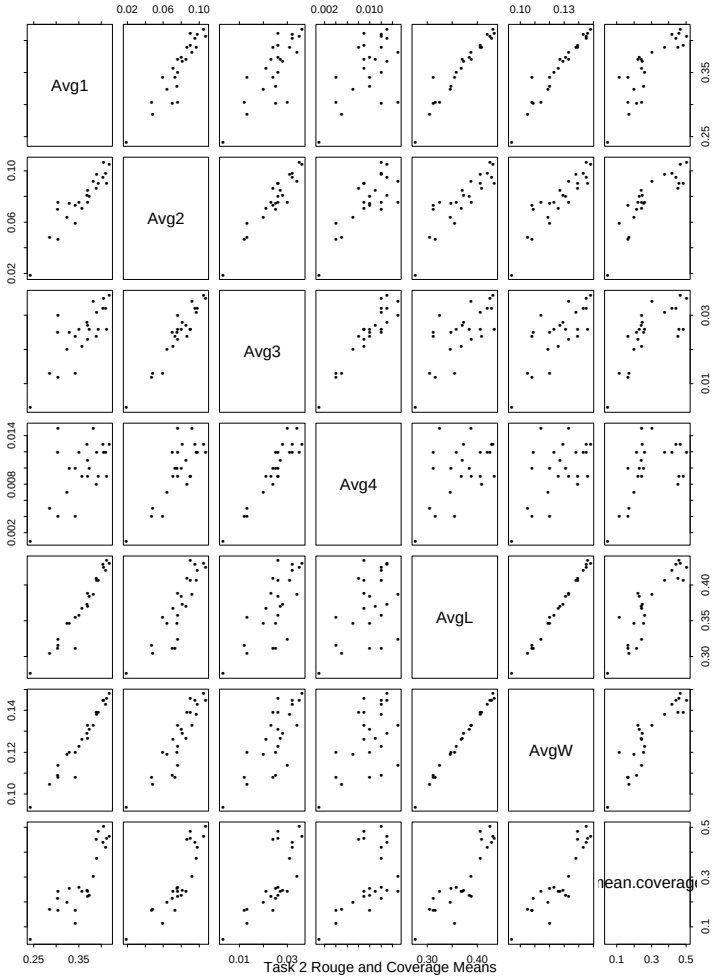
(blue lines connect runs (priority 1à 2à 3) from same group)



# SEE coverage and ROUGE scoring

- Are any of the ROUGE scores redundant?
  - ROUGE-1,2,3,4
  - ROUGE-LCS
  - ROUGE-W-1.2

- How well do the ROUGE scores predict human coverage judgments (SEE)?
  - DUC 2004
    - Task 2
    - Task 5
  - DUC 2003
    - Tasks 1-4



Task 2 Rouge and Coverage Means

# SEE coverage and ROUGE scoring

Correlations* of means for priority-1 2004 runs

**Task 2**

| | R-2 | R-3 | R-4 | LCS | LCS-W | Mean coverage |
|---|---|---|---|---|---|---|
| R-1 | 0.917480 | 0.780080 | 0.561491 | 0.965666 | 0.968852 | 0.843490 |
| R-2 | | 0.951575 | 0.797900 | 0.873155 | 0.897242 | 0.858969 |
| R-3 | | | 0.933759 | 0.725824 | 0.762940 | 0.747137 |
| R-4 | | | | 0.490218 | 0.536869 | 0.525070 |
| LCS | | | | | 0.997705 | 0.885321 |
| LCS-W1.2 | | | | | | 0.892690 |

**Task 5**

| | R-2 | R-3 | R-4 | LCS | LCS-W | Mean Coverage |
|---|---|---|---|---|---|---|
| R-1 | 0.975219 | 0.924497 | 0.846288 | 0.991905 | 0.991800 | 0.954240 |
| R-2 | | 0.977023 | 0.915436 | 0.967596 | 0.976287 | 0.962451 |
| R-3 | | | 0.977227 | 0.919919 | 0.937268 | 0.896223 |
| R-4 | | | | 0.847506 | 0.870654 | 0.801645 |
| LCS | | | | | 0.998633 | 0.947881 |
| LCS-W1.2 | | | | | | 0.948312 |

* Pearson's product moment

# Task 1: Very short summary of a TDT document

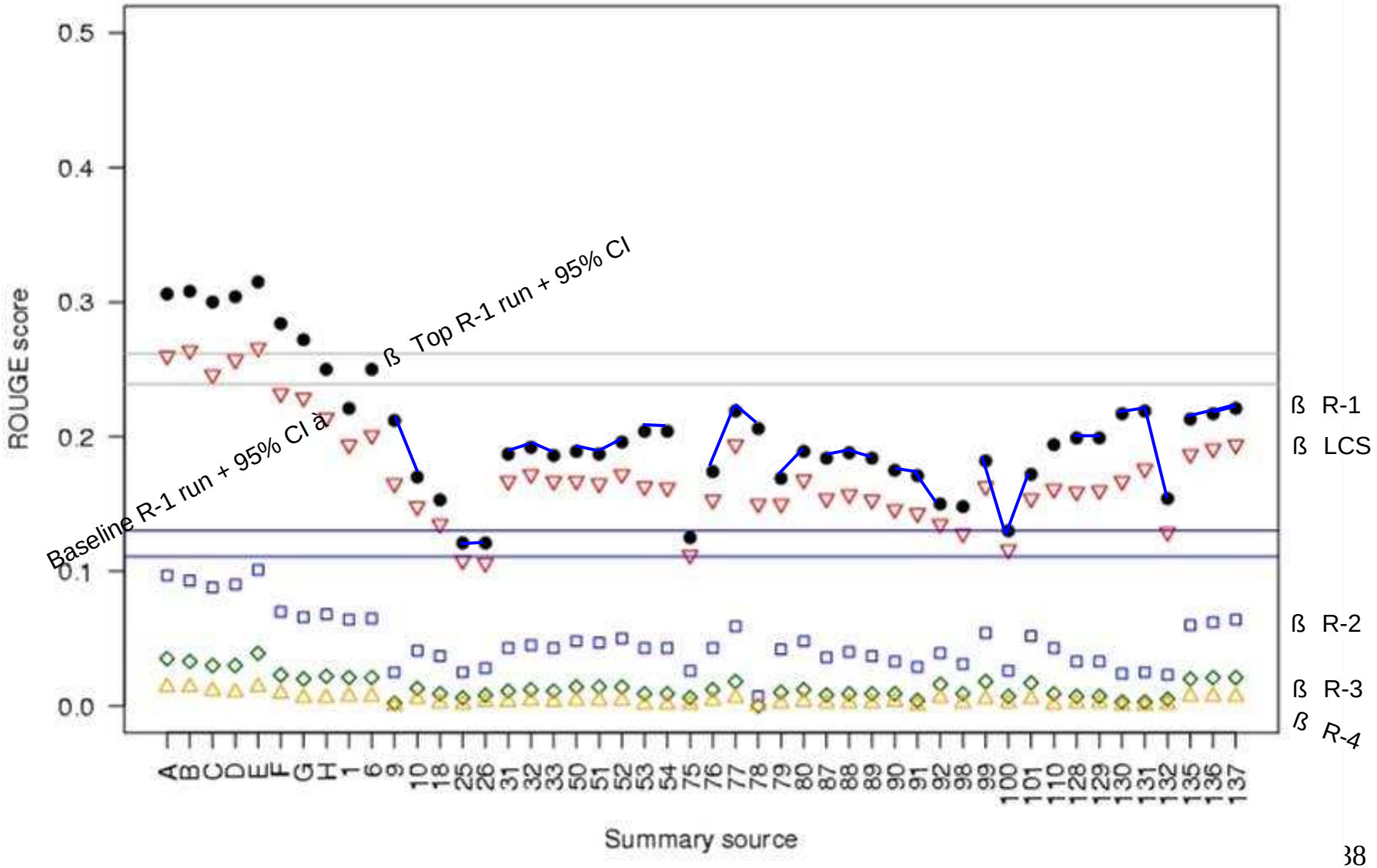
- System task:
  - Use the 50 English TDT clusters
    - ~ 10 documents/cluster
  - Given:
    - each document cluster
  - Create a short summary (<=75 bytes) of each document in the cluster (no formatting).
- Coverage baseline 1:
  - Take the first 75 bytes of the TEXT of the document
- Evaluation:
  - ROUGE

# Task 1: Participants and runs

| Sysid | Priority | Run | Group |
|---|---|---|---|
| UMD.BBN.Trimmer | 1 | 6 | U.Md/BBN |
| CL | 1 | 9 | CL Research |
| CL | 2 | 10 | |
| LARIS.2004 | 1 | 18 | Laris Labs |
| ULeth2004 | 1 | 25 | U. Lethbridge |
| ULeth2004 | 2 | 26 | |
| MEDLAB_Fudan | 1 | 31 | Fudan U. |
| MEDLAB_Fudan | 2 | 32 | |
| MEDLAB_Fudan | 3 | 33 | |
| SummariserPort.UniS | 1 | 50 | U. Nijmegen |
| SummariserPort.UniS | 2 | 51 | |
| SummariserPort.UniS | 3 | 52 | |
| CLaCDUCTape2 | 1 | 53 | Concordia U. |
| CLaCDUCTape2 | 2 | 54 | |
| ISI.ReWrite | 1 | 75 | ISI/USC |
| Fennie-summariser | 1 | 76 | U. Sunderland |
| Fennie-summariser | 2 | 77 | |
| Fennie-summariser | 3 | 78 | |
| webcl2004 | 1 | 79 | ISI/USC |
| webcl2004 | 2 | 80 | |
| irstduc041 | 1 | 87 | IRST |
| irstduc041 | 2 | 88 | |
| irstduc041 | 3 | 89 | |
| kul.2004 | 1 | 90 | KU Leuven |
| kul.2004 | 2 | 91 | |
| kul.2004 | 3 | 92 | |
| usheffield.gotoh | 1 | 98 | U. Sheffield |
| lcc.duc04 | 1 | 99 | LCC |
| lcc.duc04 | 2 | 100 | |
| lcc.duc04 | 3 | 101 | |
| uofo | 1 | 110 | U. Ottawa |
| uam.duc2004.v3 | 1 | 128 | U. Madrid |
| uam.duc2004.v3 | 2 | 129 | |
| ie_ucd_iirg | 1 | 130 | U. College Dublin |
| ie_ucd_iirg | 2 | 131 | |
| ie_ucd_iirg | 3 | 132 | |
| UofM-MEAD | 1 | 135 | U. Michigan |
| UofM-MEAD | 2 | 136 | |
| UofM-MEAD | 3 | 137 | |

# Task 1: ROUGE scores by summary source

(blue lines connect runs (priority 1à 2à 3) from same group)



# Tasks 1 & 2: Recap

- Linguistic quality questions
  - Pass some sanity checks; seem to provide lots of detailed feedback
  - Mixed per-system results
  - Multiple comparisons finds differences only between extremes
- SEE coverage
  - Manual summaries' coverage more than twice that of others
  - Systems' mean indistinguishable from baseline's
  - Multiple comparisons finds differences in systems at extremes
- ROUGE
  - LCS and R-1 track each other; likewise R-3,4
  - LCS < R-1 in task 1 but R-1 > LCS in task 2 – due to bug in ROUGE
  - LCS/R-1 in task 1 is about ½ the value in task 2
  - LCS/R-1 >> R-2,3,4
  - Correlation of SEE coverage and ROUGE means range from .747 (R-3) to .893 (LCS-W-1.2)

# Tasks 3 & 4

# Task 3: Very short summary of English translation of TDT Arabic document

- System task:
  - Use the 24 TDT clusters of Arabic documents in:
    1. (Required) automatic English translations
       1. IBM and ISI MT output
       2. 10 best translations and single best for each sentence
    2. (Required) manual English translations
    3. (Optional) automatic English translations

       \+ additional relevant English documents from about same time period
  - Given
    - each document cluster in translation
    - for subtask 3: additional relevant English documents
  - Create a very short summary (<=75 bytes) of each document in the cluster (no formatting)

# Task 3: Very short summary of English translation of TDT Arabic document

- Coverage baseline 3:
  - Use the ISI translations for half the docsets and the IBM translation for the rest.
  - Use the best translation for each sentence as determined by the MT system.
  - Take the first 75 bytes of the TEXT of the best translation of the document from the assigned MT system

- Evaluation:
  - ROUGE

# Task 3: Participants and runs

| Sysid | Priority | Run | Group |
|---|---|---|---|
| UMD.BBN.Trimmer | 1 | 7 | U.Md/BBN |
| UMD.BBN.Trimmer | 2 | 8 | |
| CL | 1 | 12 | CL Research |
| CL | 2 | 13 | |
| LARIS.2004 | 1 | 20 | Laris Labs |
| LARIS.2004 | 2 | 21 | |
| MEDLAB_Fudan | 1 | 37 | Fudan U. |
| MEDLAB_Fudan | 2 | 38 | |
| MEDLAB_Fudan | 3 | 39 | |
| CLaCDUCTape2 | 1 | 58 | Concordia U. |
| CLaCDUCTape2 | 2 | 59 | |
| Lakhas0001 | 1 | 74 | U. Montreal |
| webcl2004 | 1 | 82 | ISI/USC |

| Sysid | Priority | Run | Group |
|---|---|---|---|
| lcc.duc04 | 1 | 105 | LCC |
| lcc.duc04 | 2 | 106 | |
| uofo | 1 | 112 | U. Ottawa |
| uofo | 2 | 113 | |
| ie_ucd_iirg | 1 | 133 | U. College Dublin |
| ie_ucd_iirg | 2 | 134 | |
| UofM-MEAD | 1 | 141 | U. Michigan |
| UofM-MEAD | 2 | 142 | |
| UofM-MEAD | 3 | 143 | |
| webcl2004 | 2 | 151 | ISI/USC |

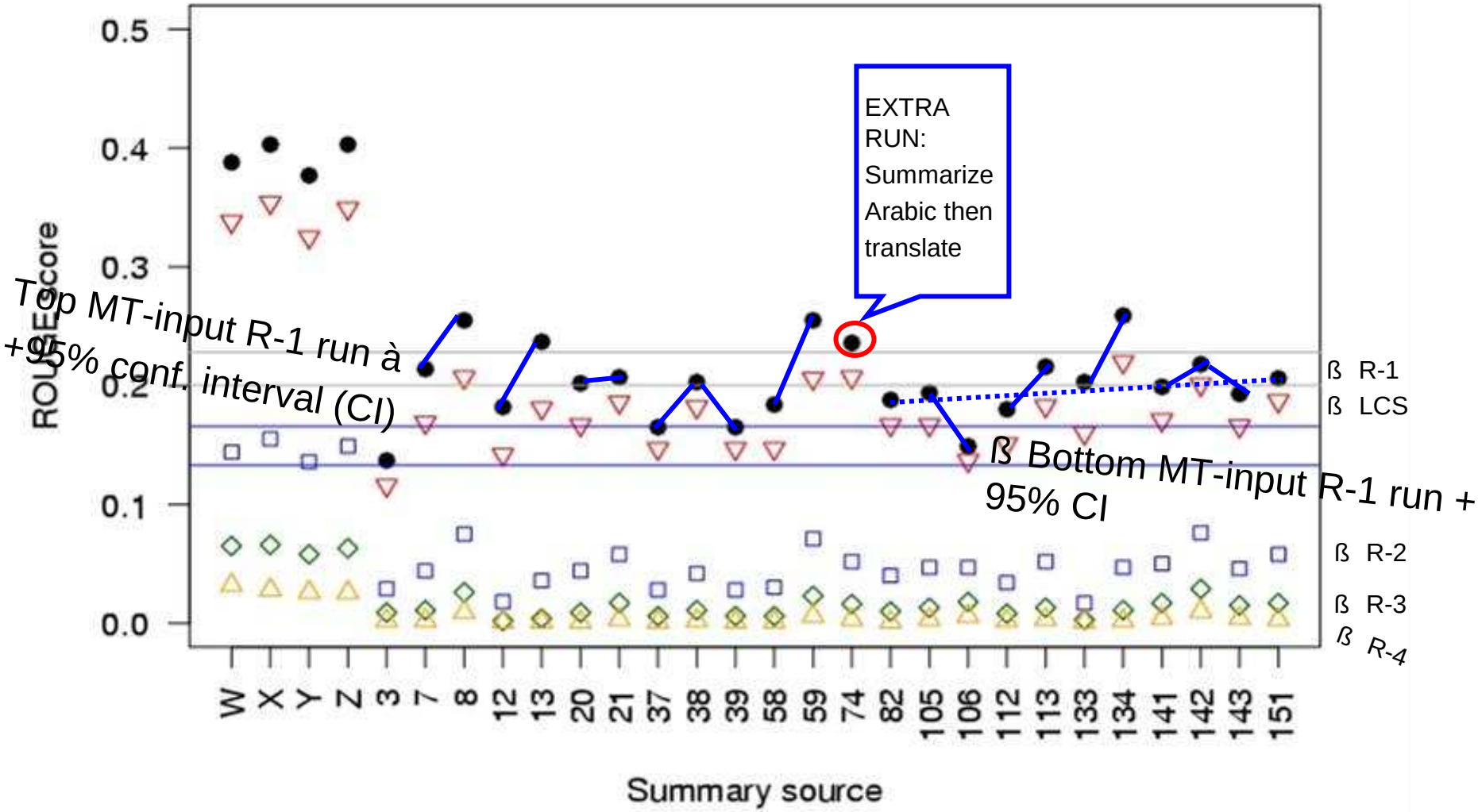
Priority 1 (required): input = IBM/ISI automatic translations

2 (required): input = Manual translations

3 (optional): input = automatic translations + relevant Eng. documents

# Task 3: ROUGE scores by summary source

(blue lines connect runs (subtasks MTà Manà MT+) from same group)



# Task 3: ROUGE-1 scores

MT input vs Manual translation input

| Group | MT | Man. Trans | Diff | %Diff |
|---|---|---|---|---|
| UMd/BBN | **0.21412** | **0.25492** | **0.0408** | **19.05** |
| CL Research | **0.18235** | **0.23743** | **0.05508** | **30.20** |
| Laris Labs | **0.20166** | **0.20713** | **0.00547** | **2.71** |
| Fudan U. | **0.16511** | **0.20266** | **0.03755** | **22.74** |
| Concordia | **0.18402** | **0.25548** | **0.07146** | **38.83** |
| ISI/USC (Lin) | **0.18794** | **0.20562** | **0.01768** | **9.41** |
| LCC | **0.19412** | **0.14930** | **-0.04482** | **-23.09** |
| U. Ottawa | **0.17961** | **0.21578** | **0.03617** | **20.14** |
| DCU | **0.20320** | **0.25866** | **0.05546** | **27.29** |
| U. Michigan | **0.19856** | **0.21751** | **0.01895** | **9.54** |

# Tasks 4: Short summary of English translation of TDT Arabic document set

- System task:
  - Use the 24 TDT clusters of Arabic documents in:
    1. (Required) automatic English translations
    2. (Required) manual English translations
    3. (Optional) automatic English translations
       + additional relevant English documents
  - Given
    - each document cluster in translation
    - for subtask 3: additional relevant English documents
  - create a short summary (<=665 bytes) of each document cluster

# Tasks 4: Short summary of English translation of TDT Arabic document set

- Coverage baseline 4:
  - Use the ISI translations for half the docsets and the IBM translation for the rest.
  - Use the best translation for each sentence as determined by the MT system.
  - Take the first 665 bytes of the TEXT of the most recent document in the best translation by the assigned MT system

- Evaluation:
  - ROUGE

# Task 4: Participants and runs

| Sysid | Priority | Run | Group |
|---|---|---|---|
| CL | 1 | 14 | CL Research |
| CL | 2 | 15 | |
| LARIS.2004 | 1 | 22 | Laris Labs |
| LARIS.2004 | 2 | 23 | |
| MEDLAB_Fudan | 1 | 40 | Fudan U. |
| MEDLAB_Fudan | 2 | 41 | |
| MEDLAB_Fudan | 3 | 42 | |
| columbia1 | 1 | 46 | Columbia U. |
| columbia1 | 2 | 47 | |
| columbia1 | 3 | 48 | |
| CLaCDUCTape2 | 1 | 60 | Concordia U. |
| CLaCDUCTape2 | 2 | 61 | |
| CCSNSA04 | 1 | 68 | NSA |
| CCSNSA04 | 2 | 69 | |
| CCSNSA04 | 3 | 70 | |

| Sysid | Priority | Run | Group |
|---|---|---|---|
| webcl2004 | 1 | 83 | ISI/USC |
| webcl2004 | 2 | 84 | |
| webcl2004 | 3 | 85 | |
| lcc.duc04 | 1 | 107 | LCC |
| lcc.duc04 | 2 | 108 | |
| uofo | 1 | 114 | U. Ottawa |
| uofo | 2 | 115 | |
| UofM-MEAD | 1 | 144 | U. Michigan |
| UofM-MEAD | 2 | 145 | |
| UofM-MEAD | 3 | 146 | |
| columbia2 | 1 | 148 | Columbia U. |
| columbia2 | 2 | 149 | |
| columbia2 | 3 | 150 | |

Priority 1 (required): input = IBM/ISI automatic translations

2 (required): input = Manual translations

3 (optional): input = automatic translations + relevant Eng. documents

# Task 4: Examples (MEAD on D1001)

**Manual summary, group Y** -- In response to Iraq's noncompliance with UN weapons inspectors, Britain and the U.S. bombed Iraq just before Ramadan from December 16 to December 19, 1998. The attack included off site cruise missile launches which targeted transport, air defense sites, and command and control facilities and hit several Iraqi hospitals. Although a "concerted effort was made to minimize civilian casualties", 68 civilians were dead Saturday. By the end of the bombing thousands were said to have been wounded. Funeral goers denounced Americans as "enemies of God". The Russia Foreign Ministry criticized solutions that "do not contribute to establishing suitable environments".

**MEAD, run 145, using the manual translations** -- The humane organization Caritas announced today Friday in Germany quoting its branch in Iraq that several Iraqi hospitals especially Saddam Hospital which is considered the biggest hospital in Iraq were hit in the American-British bombing of the Iraqi capital An Iraq official reported today Saturday that 68 Iraqi civilians were killed as a result of the American and British bombing on Iraq and that their funerals were held today in Baghdad The American television network CNN reported that officials at the American Defense Department The Pentagon will recommend to President Bill Clinton this evening Saturday halting the bombing on Iraq

**MEAD, run 144, using the machine translations** -- Caritas humanitarian organization announced today in Germany quoting the section in Iraq that several Iraqi hospitals especially hospital Saddam which is the largest hospital in Iraq was hit in the American bombardment British Iraqi capital An Iraqi official reported on Saturday that 68 Iraqi civilians were killed as a result of the American and British strikes on Iraq has been the funeral corpses today in Baghdad it was announced by the representative of Iraq in the United Nations Nizar Hamdoon on Sunday that thousands of people were killed or wounded during the four days of the bombing of Iraq

# Task 4: Examples (MEAD on D1043)

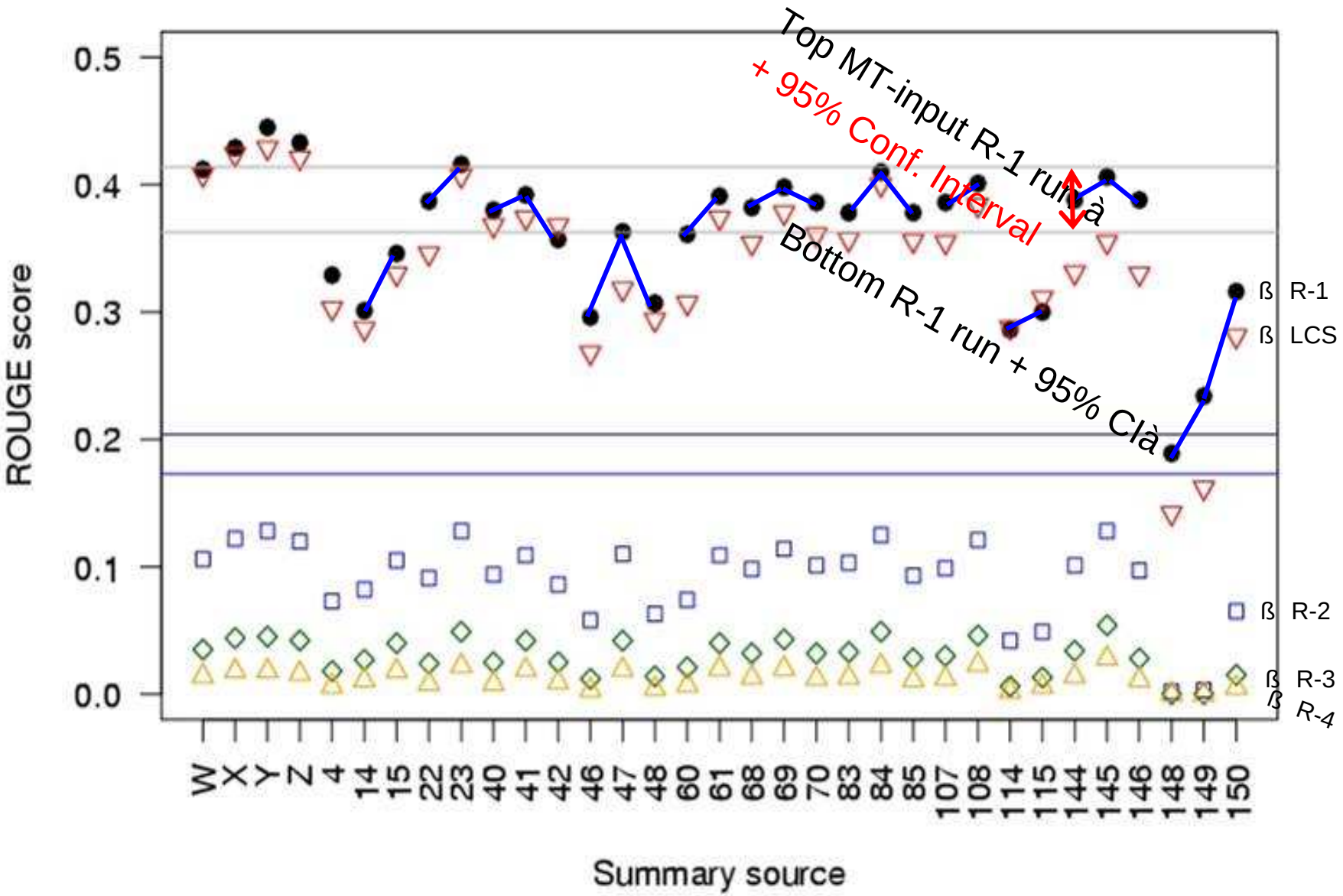
**Manual summary, group Y** -- From November 3-5, 1998 US defense secretary William Cohen visited Egypt, Saudi Arabia, Jordan, United Arab Emirates, Qatar, and other Gulf states, attempting to gauge the reaction of Arab countries if the US were to bomb Baghdad. The previous Saturday, Iraq decided to stop cooperating with UN Special Commission regarding weapons of mass destruction. Cohen warned Iraq "all options [were] open" to force it to revoke this decision. Reportedly, Cohen and President Clinton were confident of allies' support of "a diplomatic solution" in the crisis between Iraq and the UN. Every Gulf country opposed military force against Iraq.

**MEAD, run 145, using the manual translations** -- The American Defense Department said today Tuesday that Secretary William Cohen is confident of Saudi Arabia s support in the crisis between the United Nations and Iraq The American Embassy in Doha announced that the American Secretary of Defense William Cohen arrived in Qatar today Wednesday as part of his Gulf tour to hold consultations with the United States allies on the crisis with Iraq The Embassy said that Cohen arrived in Doha coming from Bahrain where he held talks with Crown Prince Sheikh Hamad Bin Issa Al Khalifa and Prime Minister Sheikh Khalifa Bin Salman Al Khalifa dealing with Iraq s decision to stop cooperation with United Nations disarmamen

**MEAD, run 144, using the machine translations** -- The Pentagon said today that the minister William Cohen is confident of support for Saudi Arabia in the crisis between the United Nations and Iraq American minister was received by the Saudi Crown Prince in Riyadh and talked about the crisis between Iraq and United Nations experts in the field of disarmament The American embassy in Doha announced that the American Secretary of Defense William Cohen arrived today to Qatar in the framework of the tour Gulf allies to hold consultations with the United states on the crisis with Iraq Embassy said that Cohen arrived in Doha coming from Bahrain where he held talks with Crown Prince Sheikh Hamad bin Isa al-Khalifa

# Task 4: ROUGE scores by summary source

(blue lines connect runs (subtasks MTà Manà MT+) from same group)



# Task 4: ROUGE-1 scores

MT input vs Manual translation input

| Group | MT | Man. Trans. | Diff | %Diff |
|---|---|---|---|---|
| CL Research | **0.30072** | **0.34628** | **0.04556** | **15.15** |
| Laris Labs | **0.38654** | **0.41577** | **0.02923** | **7.56** |
| Fudan U. | **0.37960** | **0.39223** | **0.01263** | **3.33** |
| Columbia U. (1) | **0.29599** | **0.36269** | **0.0667** | **22.53** |
| Concordia | **0.36055** | **0.39125** | **0.0307** | **8.51** |
| NSA | **0.38156** | **0.39844** | **0.01688** | **4.42** |
| ISI/USC (Lin) | **0.37812** | **0.41012** | **0.032** | **8.46** |
| LCC | **0.38615** | **0.40059** | **0.01444** | **3.74** |
| U. Ottawa | **0.28589** | **0.30037** | **0.01448** | **5.06** |
| U. Michigan | **0.38827** | **0.40602** | **0.01775** | **4.57** |
| Columbia U. (2) | **0.18853** | **0.23381** | **0.04528** | **24.02** |

# Tasks 3 & 4: Recap

- LCS and R-1 track each other; also R-3,4 (as in task 1,2)
- LCS/R-1 >> R-2,3,4 (as in task 1,2)
- Using manual translations repeatedly yield better R-1/LCS
  - Within-system difference does not appear to be significant
  - Cross-system pattern seems unlikely to be due to chance
- Adding relevant English documents doesn't seem to help scores
- CI around top MT input run includes runs using manual and automatic translations

# Tasks 3 & 4: Questions

- Did groups do something special to handle translations as input?
- What choices did groups make about:
  - IBM vs ISI translations
  - All variants vs best only
- What did groups do with the extra, relevant English documents?
- ?
- ?

# Task 5

# Task 5: short summary of a document set focused by a “Who is X?” question

- Questions and document sets created by NIST assessors

- Instructions:
  - Create questions of the form "Who is X?", where X is the name of a person or group of people
  - Find sets of 10 documents each, such that each document in a given set, contributes to answering the associated question.
  - Different documents may contribute different amounts of material.
  - There may be some repetition within a set.

# Task 5: Example questions/people

| | |
|---|---|
| d132d | Robert Rubin |
| d133c | Stephen Hawking |
| d134h | Desmond Tutu |
| d135g | Brian Jones |
| d136c | Gene Autry |
| d137c | Harry A. Blackmun |
| d139b | Joerg Haider |
| d141d | Sir John Gielgud |
| d144c | Jon Postel |
| d147d | Mel Carnahan |
| d148g | Carole Sund |
| d149d | Louis J. Freeh |
| d151h | Alan Greenspan |
| d153h | Kofi Annan |
| d154c | Wilt Chamberlain |
| d155c | JFK, Jr. |
| d156b | Wen Ho Lee |
| d157d | John C. Danforth |
| d159a | Theodore John Kaczynski |
| d161f | Karl Rove |
| d164g | Mia Hamm |
| d165h | Jimmy Carter |
| d166d | Jesse Helms |
| d168d | Helmut Kohl |
| d169a | Dr. Jack Kevorkian |
| d170e | Hugo Chavez |
| …. | |

- Great variety in
  - Density of information
  - Level of detail
  - Organization
- Examples:
  - (Multiple) obituaries
  - Short mention (among other subjects)
  - Second-hand accounts of
    - Appearances
    - Accomplishments
  - Interviews

# Task 5: short summary of a document set focused by a “Who is X?” question

- System task:
  - Use the 50 TREC question clusters
    - ~10 documents/cluster
  - Given:
    - The question
    - Each document cluster
  - Create a short summary of the cluster that contributes to answering the question
    - <= 665 bytes

- Coverage baseline 5:
  - Take the first 665 bytes of the most recent document

# Task 5: short summary of a document set focused by a “Who is X?” question

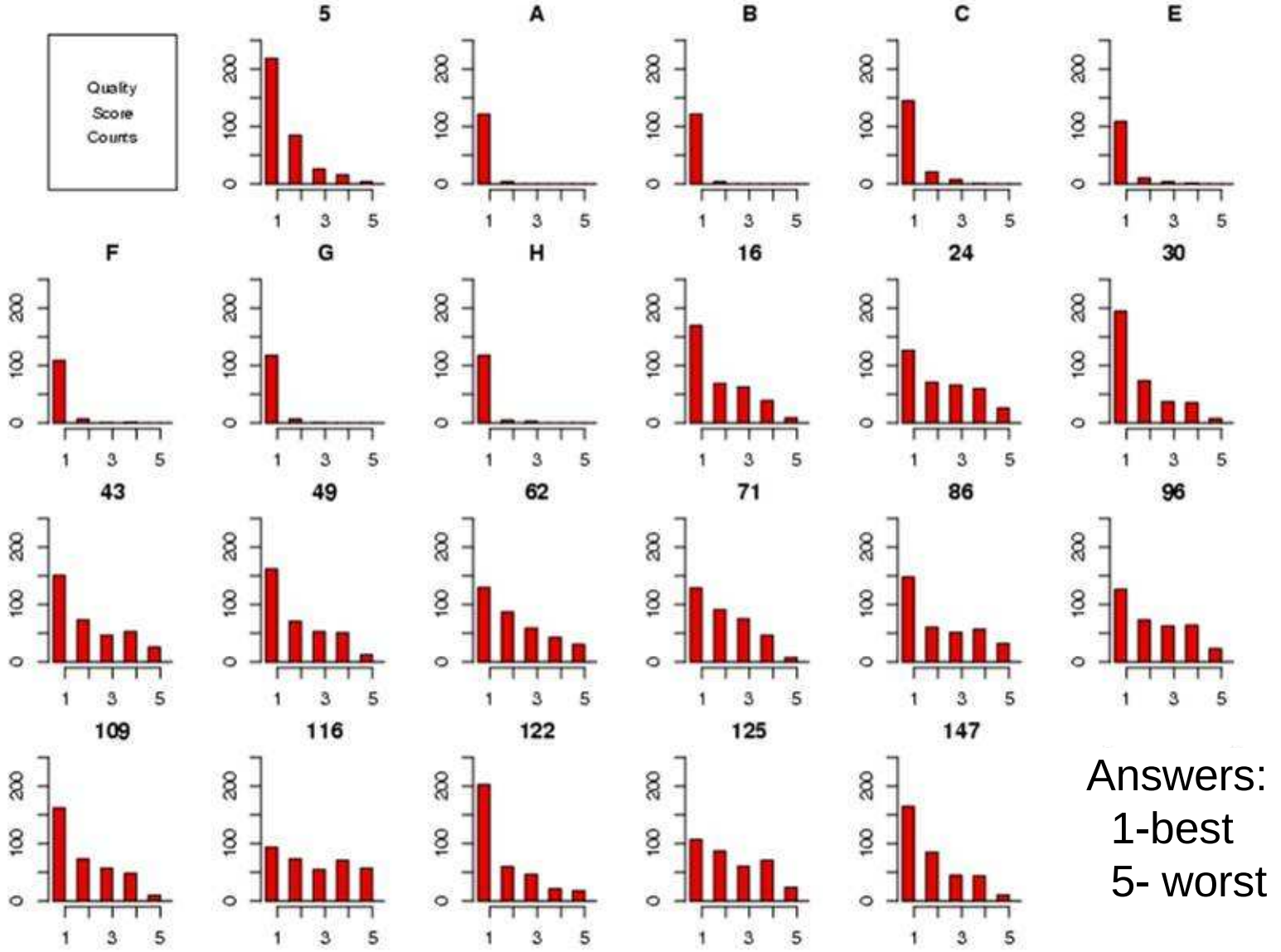
- Evaluation:
  - SEE
    - Linguistic quality
    - Coverage
    - Extra material
  - Responsiveness
  - ROUGE

# Task 5: Participants and runs

| | | | |
|---|---|---|---|
| CL | 1 | 16 | CL Research |
| CL | 2 | 17 | |
| LARIS.2004 | 1 | 24 | Laris Labs |
| ULeth2004 | 1 | 30 | U. Lethbridge |
| MEDLAB_Fudan | 1 | 43 | Fudan U. |
| columbia1 | 1 | 49 | Columbia U. |
| CLaCDUCTape2 | 1 | 62 | Concordia U. |
| CLaCDUCTape2 | 2 | 63 | |
| CLaCDUCTape2 | 3 | 64 | |
| CCSNSA04 | 1 | 71 | NSA |
| CCSNSA04 | 2 | 72 | |
| CCSNSA04 | 3 | 73 | |
| webcl2004 | 1 | 86 | ISI/USC |
| kul.2004 | 1 | 96 | KU Leuven |
| kul.2004 | 2 | 97 | |
| lcc.duc04 | 1 | 109 | LCC |
| uofo | 1 | 116 | U. Ottawa |
| crl_nyu.duc04 | 1 | 122 | CRL/NYU |
| shef2004.saggion | 1 | 125 | U. Sheffield |
| shef2004.saggion | 2 | 126 | |
| shef2004.saggion | 3 | 127 | |
| UofM-MEAD | 1 | 147 | U. Michigan |

# Task 5: Linguistic quality

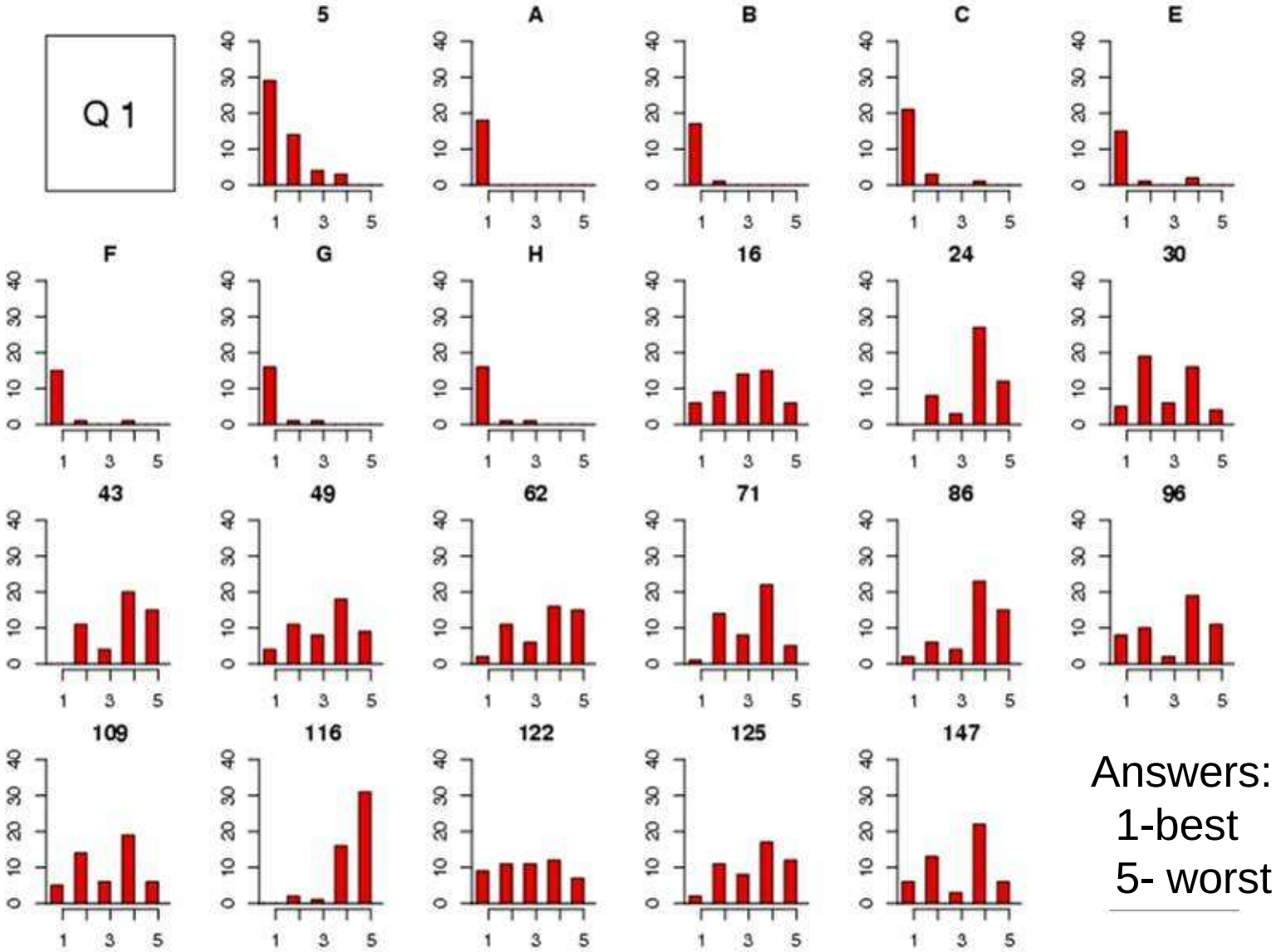
Counts of answers(1-5) by summary source




Answers:
1-best
5- worst

# Task 5: Linguistic quality

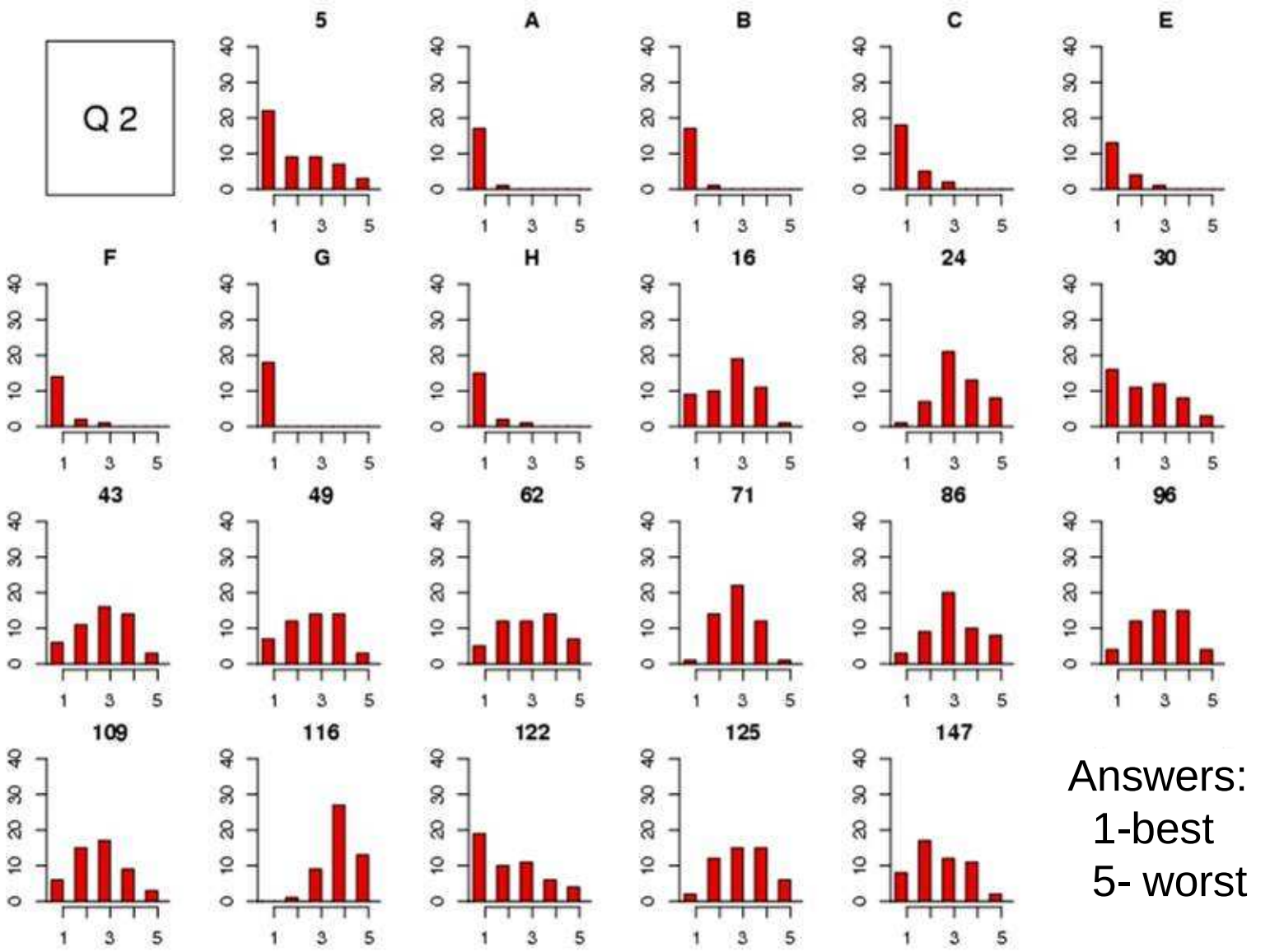
Question 1 - builds to coherent body of information?




Answers:
1-best
5- worst

# Task 5: Linguistic quality

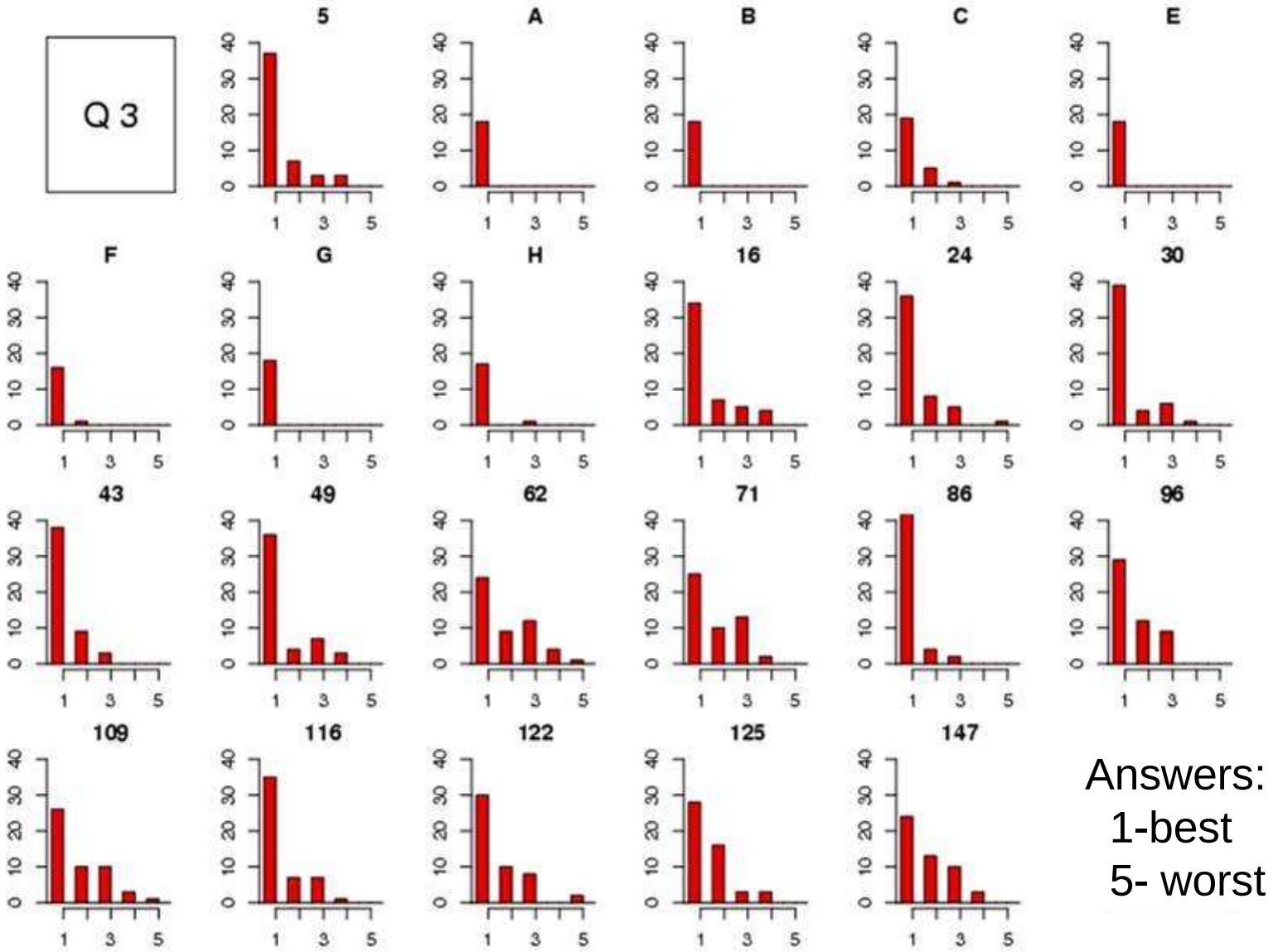
Question 2 – useless, confusing, repetitive text?




Answers:
1-best
5- worst

# Task 5: Linguistic quality

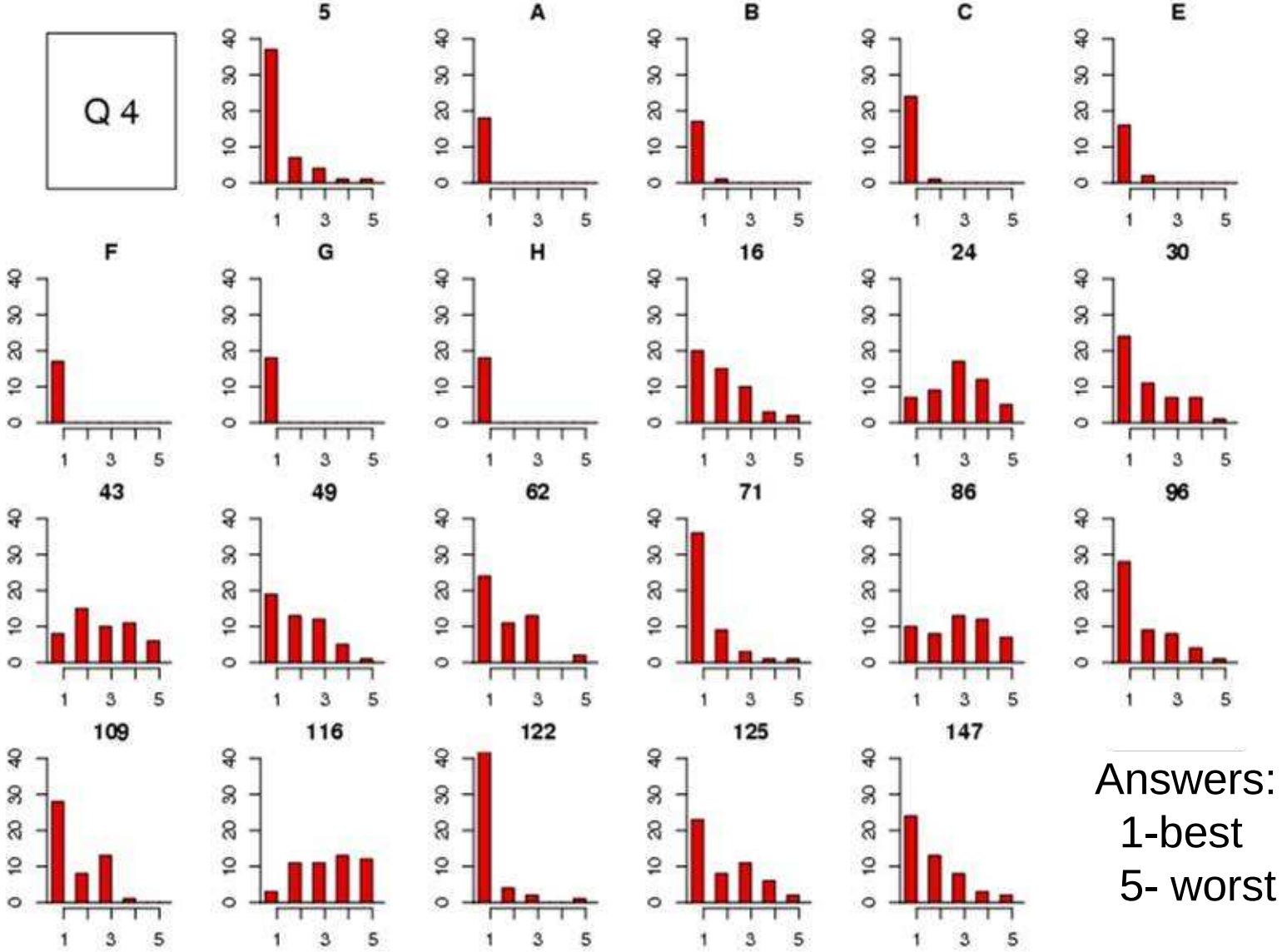
Question 3 – same thing over again?




Answers:
1-best
5- worst

# Task 5: Linguistic quality

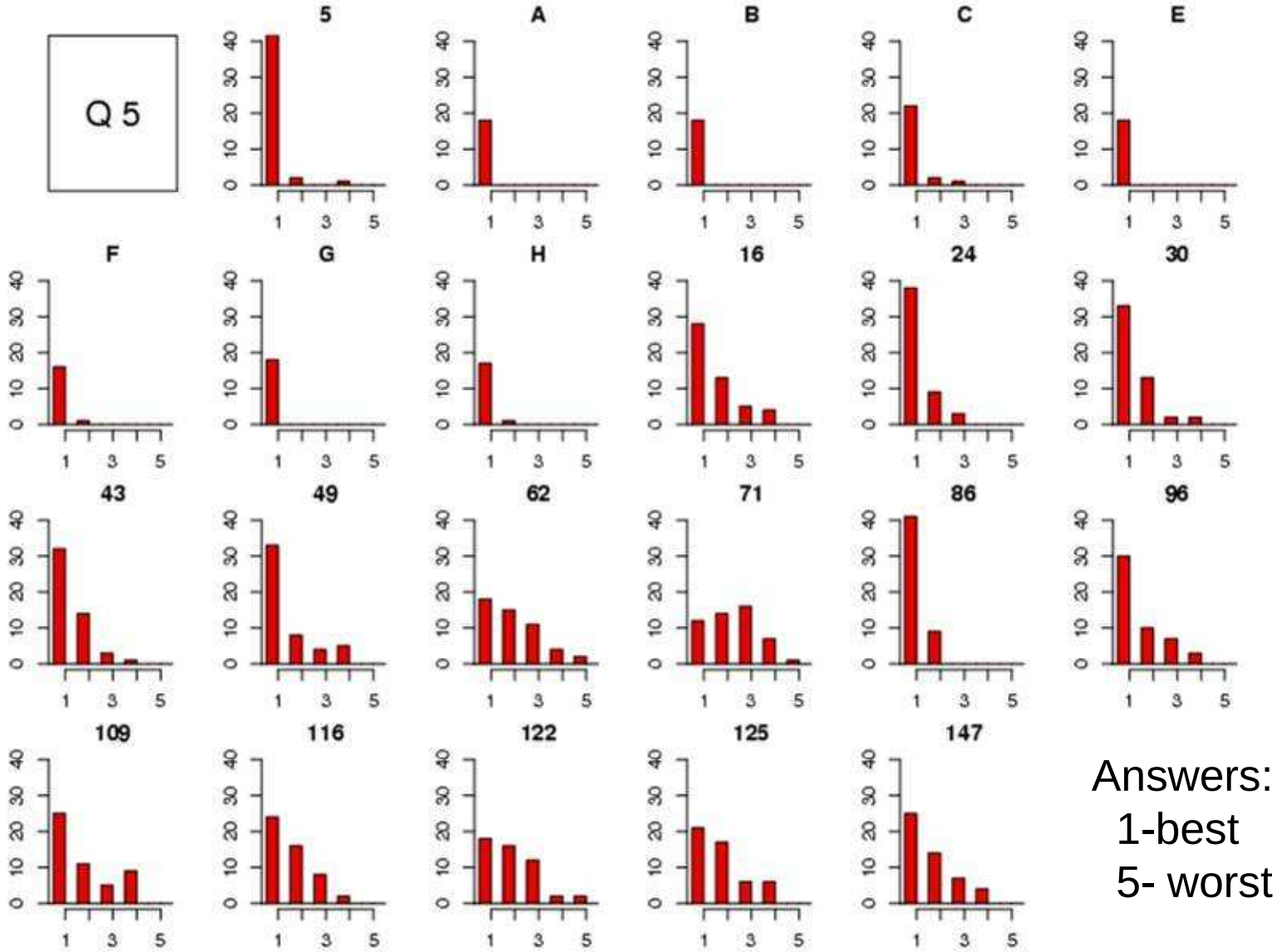
Question 4 – trouble identifying noun phrase referents?




Answers:
1-best
5- worst

# Task 5: Linguistic quality

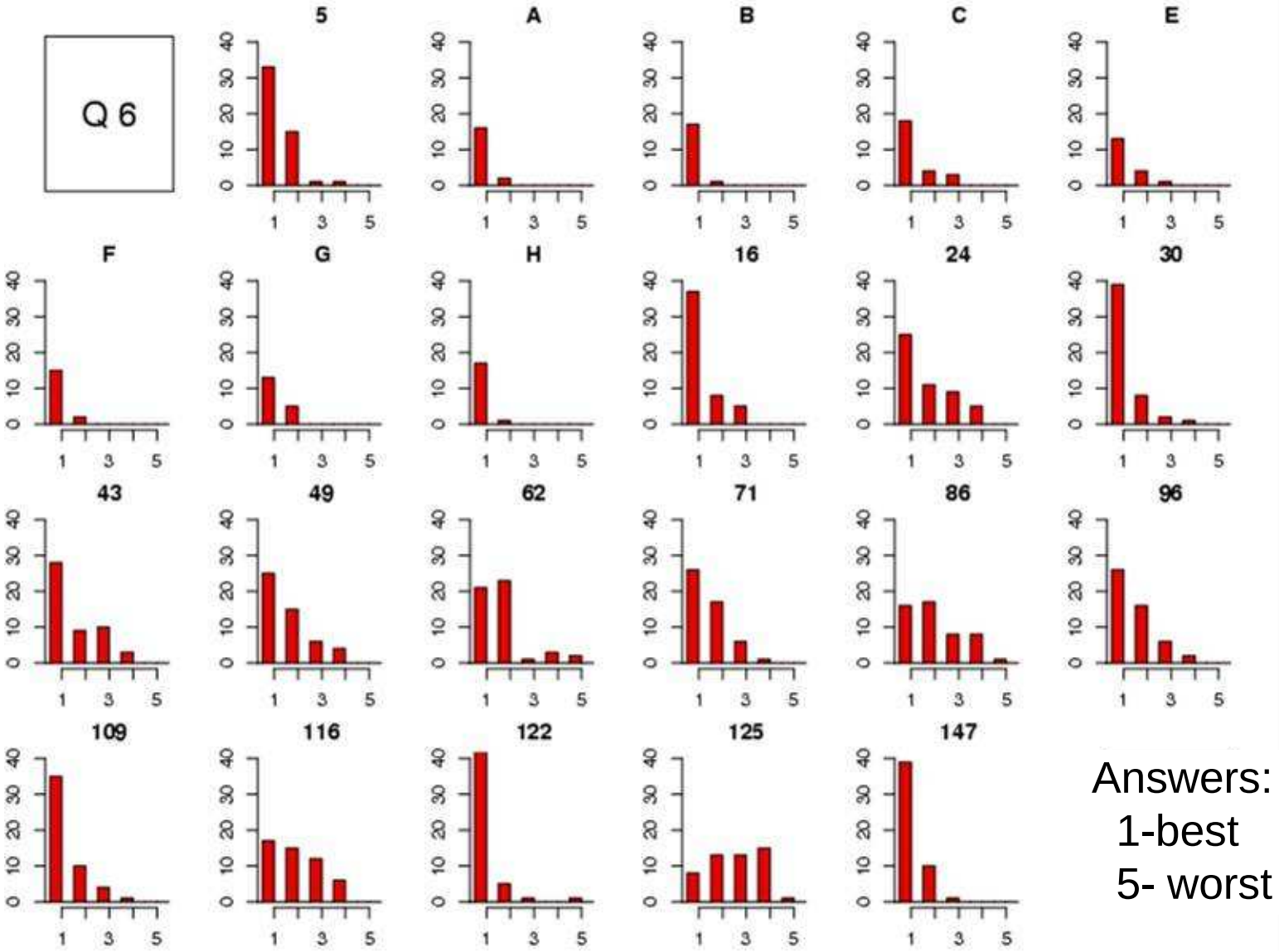
Question 5 – entities re-mentioned?




Answers:
1-best
5- worst

# Task 5: Linguistic quality

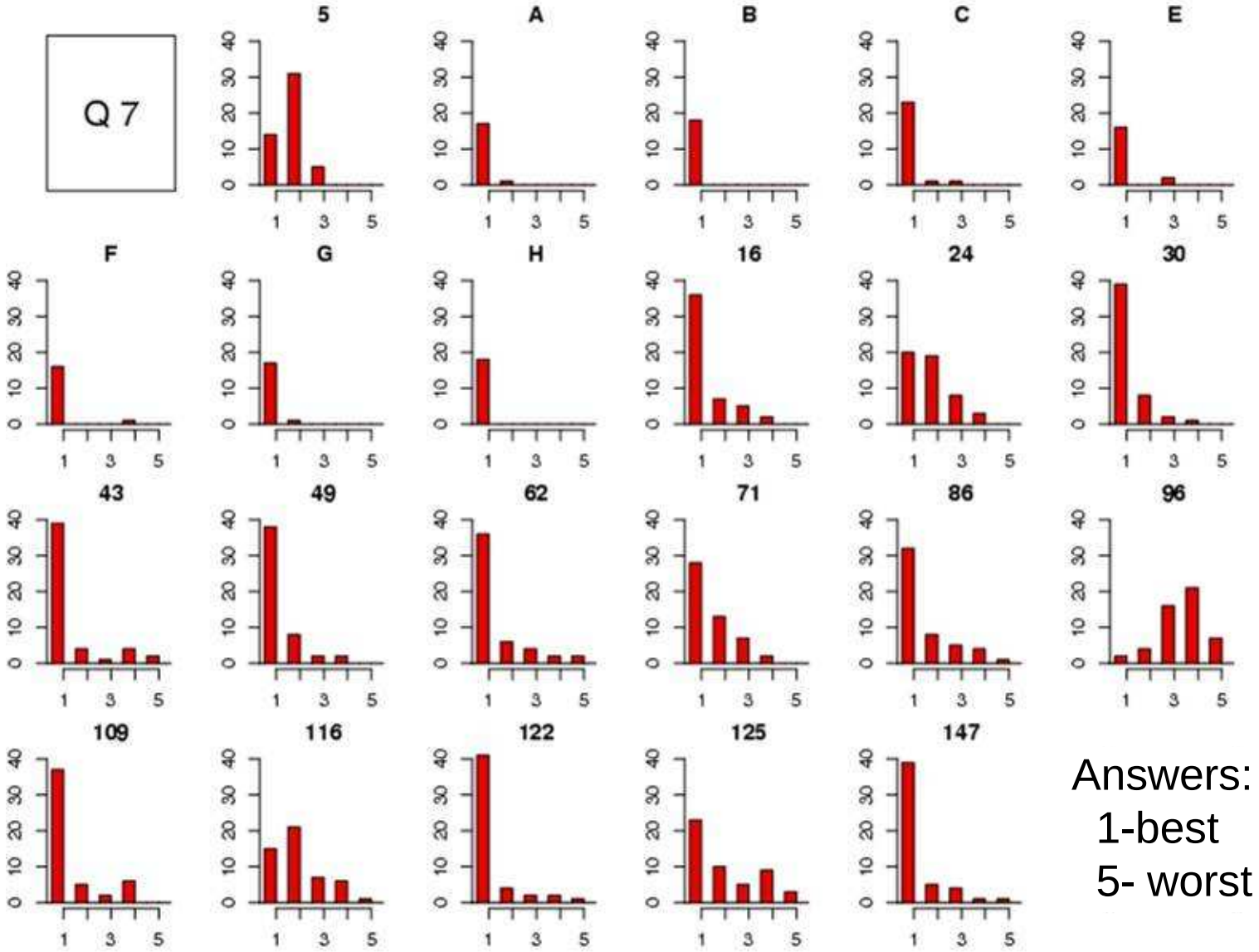
Question 6 – ungrammatical sentences?




Answers:
1-best
5- worst

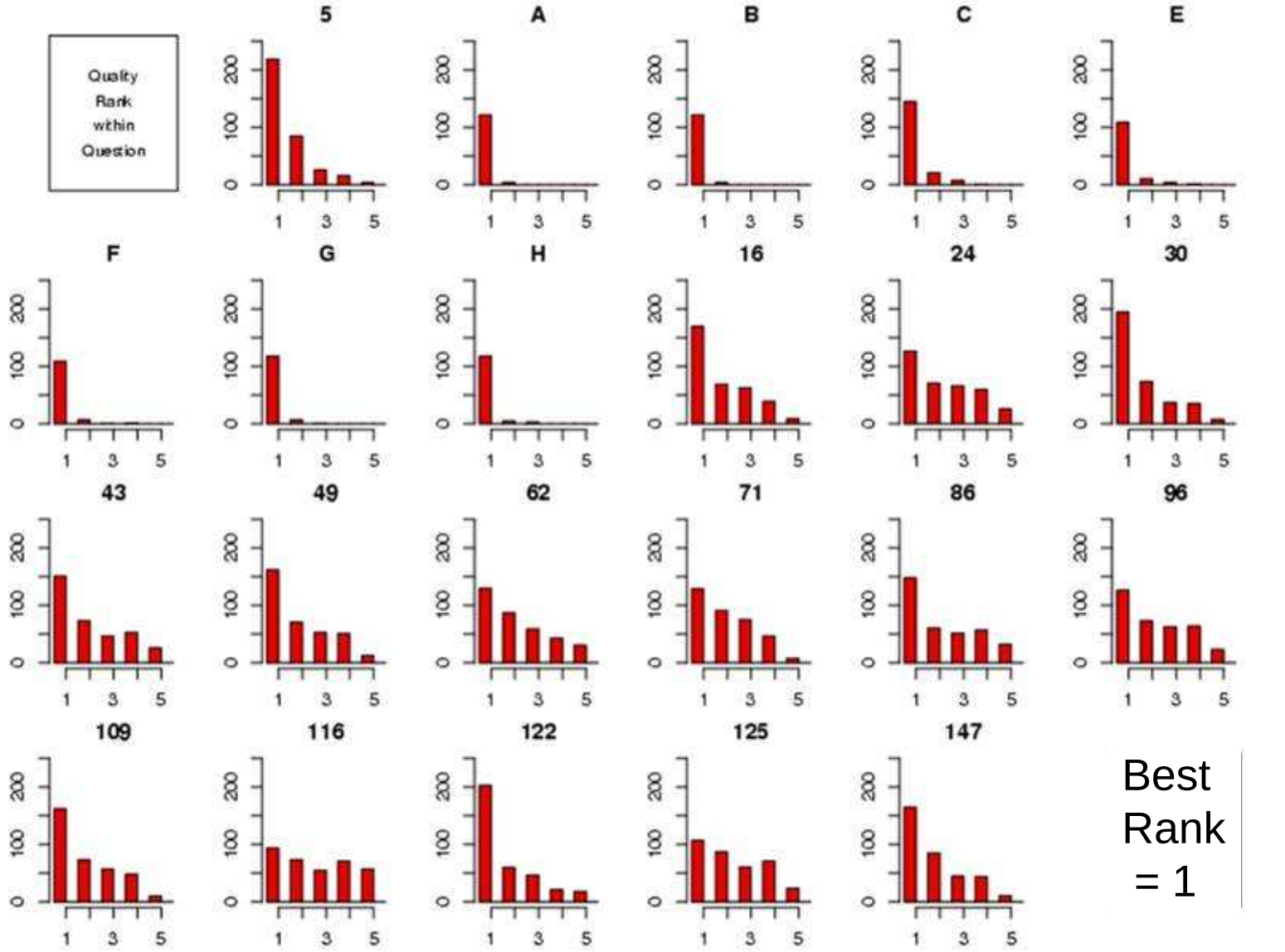
# Task 5: Linguistic quality

Question 7- datelines, formatting, capitalization problems?




Answers:
1-best
5- worst

# Task 5: Quality question group rank within docset by summary source



# Overall peer quality

## Task 5 – multiple comparisons (best on top)

**q1**

| | | | | | |
|---|---|---|---|---|---|
| 30 | A | | | | |
| 122 | A | | | | |
| 16 | A | C | | | |
| 147 | A | C | D | | |
| 109 | A | C | D | E | |
| 71 | A | C | D | E | |
| 96 | A | C | D | E | |
| 49 | A | C | D | E | |
| 125 | A | C | D | E | |
| 62 | A | C | D | E | |
| 43 | | C | D | E | G |
| 86 | | | D | E | G |
| 24 | | | | E | G |
| 116 | | | | | G |

**q3**

| | | | | |
|---|---|---|---|---|
| 86 | A | | | |
| 43 | A | B | | |
| 24 | A | B | | |
| 30 | A | B | | |
| 116 | A | B | | E |
| 49 | A | B | | E |
| 16 | A | B | | E |
| 125 | A | B | | E |
| 122 | A | B | | E |
| 96 | | B | | E |
| 71 | | B | | E |
| 147 | | B | | E |
| 109 | | B | | E |
| 62 | | | | E |

**q5**

| | | | | | |
|---|---|---|---|---|---|
| 86 | A | | | | |
| 24 | A | B | | | |
| 30 | A | B | C | | |
| 43 | A | B | C | | |
| 49 | A | B | C | E | |
| 96 | A | B | C | E | |
| 16 | A | B | C | E | |
| 147 | | B | C | E | H |
| 116 | | B | C | E | H |
| 109 | | B | C | E | H |
| 125 | | | C | E | H |
| 122 | | | C | E | H |
| 62 | | | | E | H |
| 71 | | | | | H |

**q7**

| | | | | | |
|---|---|---|---|---|---|
| 30 | A | | | | |
| 122 | A | | | | |
| 49 | A | | | | |
| 147 | A | | | | |
| 16 | A | E | | | |
| 43 | A | E | F | | |
| 62 | A | E | F | | |
| 109 | A | E | F | | |
| 86 | A | E | F | I | |
| 71 | A | E | F | I | |
| 24 | | E | F | I | |
| 125 | | | F | I | |
| 116 | | | | I | |
| 96 | | | | | N |

**q2**

| | | | | | |
|---|---|---|---|---|---|
| 122 | A | | | | |
| 30 | A | B | | | |
| 147 | A | B | | | |
| 16 | A | B | D | | |
| 109 | A | B | D | | |
| 49 | A | B | D | | |
| 43 | A | B | D | | |
| 71 | A | B | D | | |
| 96 | A | B | D | | |
| 62 | | B | D | | |
| 125 | | B | D | | |
| 86 | | B | D | | |
| 24 | | | D | | M |
| 116 | | | | | M |

**q4**

| | | | | | |
|---|---|---|---|---|---|
| 122 | A | | | | |
| 71 | A | B | | | |
| 96 | A | B | | | |
| 109 | A | B | | | |
| 62 | A | B | | | |
| 147 | A | B | | | |
| 30 | | B | | | |
| 125 | | B | H | | |
| 16 | | B | H | | |
| 49 | | B | H | | |
| 43 | | | H | J | |
| 86 | | | H | J | |
| 24 | | | | J | |
| 116 | | | | J | |

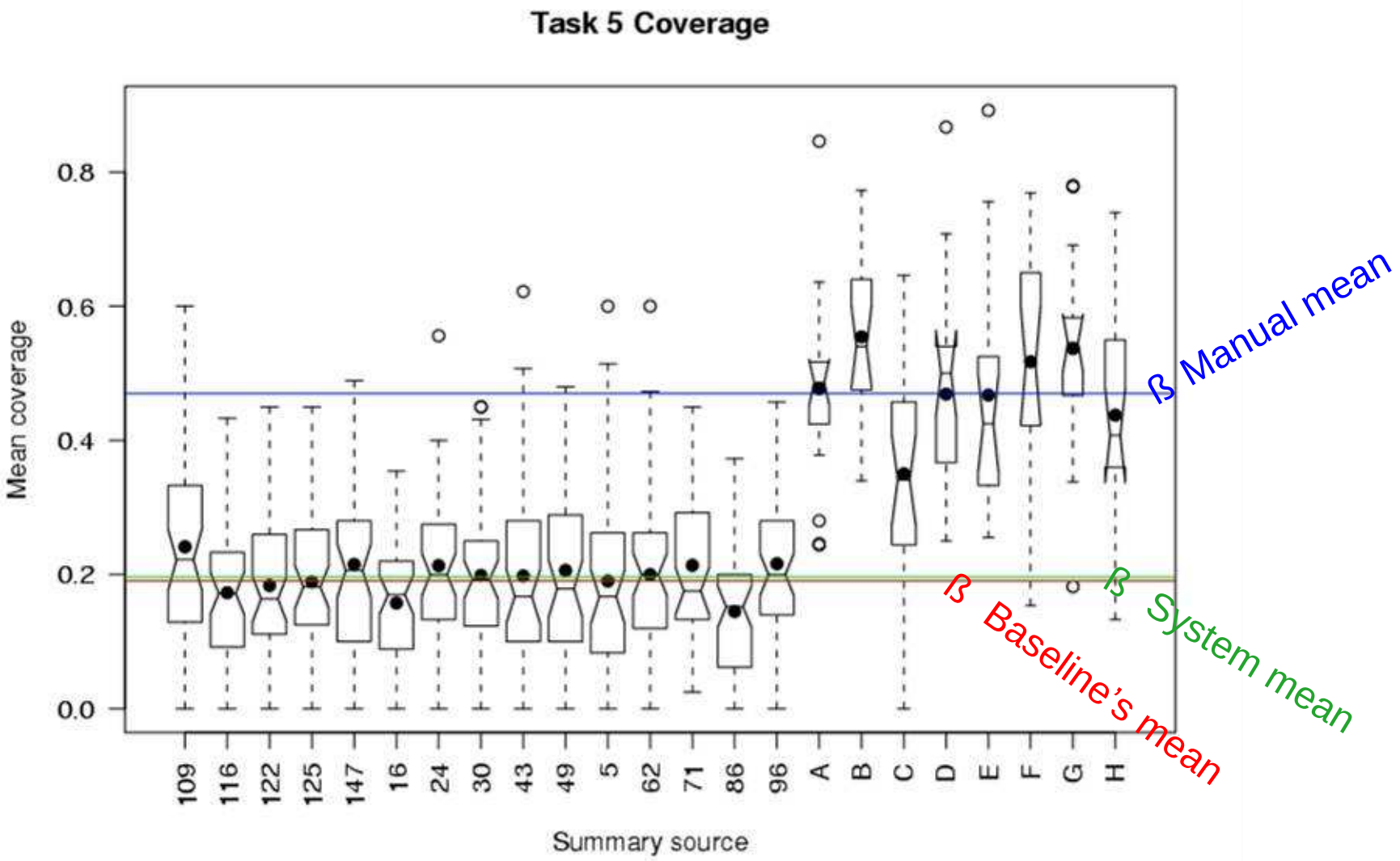
**q6**

| | | | | | |
|---|---|---|---|---|---|
| 122 | A | | | | |
| 147 | A | B | | | |
| 30 | A | B | C | | |
| 16 | A | B | C | | |
| 109 | A | B | C | | |
| 71 | A | B | C | F | |
| 96 | A | B | C | F | |
| 43 | A | B | C | F | |
| 49 | | B | C | F | |
| 24 | | | C | F | |
| 62 | | | C | F | |
| 116 | | | | F | L |
| 86 | | | | F | L |
| 125 | | | | | L |

Means with the same letter are not significantly different.

Tukey-Kramer criterion (.05) on average ranks from Friedman's test

# Task 5: Mean coverage by summary source



# Task 5: ANOVA on coverage

Number of observations 9922

The GLM Procedure

| R-Square | Coeff Var | Root MSE | Mean |
|---|---|---|---|
| 0.297547 | 67.80859 | 0.208265 | 0.307137 |

| Source | DF | Type I SS | Mean Square | F Value |
|---|---|---|---|---|
| docset | 59 | 42.1070990 | 0.7136796 | 16.45 |
| peer | 22 | 138.6796453 | 6.3036202 | 145.33 |

| Source | Pr > F |
|---|---|
| docset | <.0001 |
| peer | <.0001 |

# Task 5: Multiple comparisons on coverage (@ 0.05 confidence level)

| REGWQ Grouping | Mean | N | Summary source |
|---|---|---|---|
| A | 0.47009 | 150 | Manual |
| B | 0.19646 | 700 | System |
| B | 0.19038 | 50 | Baseline |

Means with the same letter are not significantly different.

# Task 5: Multiple comparisons on coverage (@ 0.05 confidence level)

| REGWQ Grouping | | | Mean | N | Summary source |
|---|---|---|---|---|---|
| | A | | 0.24144 | 50 | 109 |
| B | A | | 0.21604 | 50 | 96 |
| B | A | | 0.21500 | 50 | 147 |
| B | A | | 0.21370 | 50 | 71 |
| B | A | | 0.21328 | 50 | 24 |
| B | A | C | 0.20620 | 50 | 49 |
| B | A | C | 0.19988 | 50 | 62 |
| B | A | C | 0.19868 | 50 | 30 |
| B | A | C | 0.19792 | 50 | 43 |
| B | A | C | 0.18918 | 50 | 125 |
| B | A | C | 0.18370 | 50 | 122 |
| B | | C | 0.17294 | 50 | 116 |
| B | | C | 0.15752 | 50 | 16 |
| | | C | 0.14492 | 50 | 86 |

Means with the same letter are not significantly different.

# Task 5: Responsiveness

- Why?
  - Allows form as well as content
  - Compares all summaries to each other (not each to the model)

- Assessor sees
  - each document
  - all summaries of the document set

- Assessor asked to:
  - Read all the summaries.
  - Grade each summary according to how **relatively** responsive it is to the question – considering form and content

    0 (worst), 1, 2, 3, or 4 (best)

# Task 5: Responsiveness – Examples

**0** D202.M.100.E.F.F.86.Then when Wesley Clark was attending Oxford as a Rhodes scholar, one of the cousins, Molly Friedman from Cleveland, was visiting England and called Wesley at the university. An Associated Press report quoted several U.S. administration and military officials as saying that Clark had rifts with the Clinton administration during the 78-day campaign. Clark also has become fluent in the Russian language and in the past three years has delved into the family history. But in interviews, some of his relatives and friends say that Clark was inspired by the story of his grandfather 's persecution and escape from his native land, and that his determination to defeat

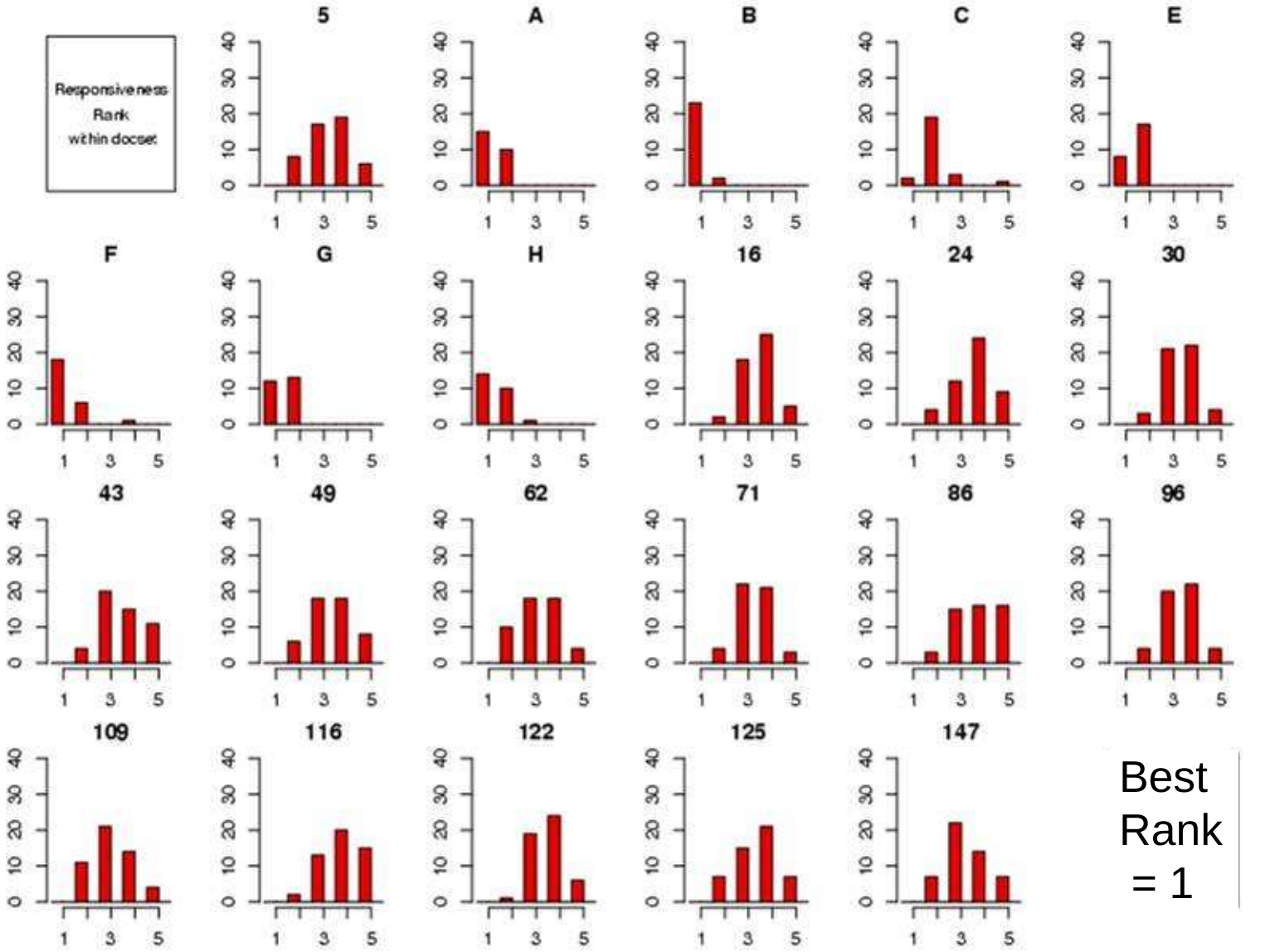
**1** D202.M.100.E.F.F.122 NATO's Defense Planning Committee adopted today a resolution appointing General Wesley K. Clark of the United States' army as the successor of General George Joulwan, who decided to resign next month from his post at Supreme Allied Forces Europe. U.S. Gen. Wesley Clark, NATO's supreme commander for Europe, discussed plans for NATO enlargement and the situation in the Balkans with his Slovene hosts on Monday. French President Jacques Chirac Friday met here with the NATO Supreme Commander in Europe, General Wesley Clark, who is on a visit in the French capital.

**2** D202.M.100.E.F.F.62 Wesley Clark, NATO's supreme commander for Europe, discussed plans for NATO enlargement and the situation in the Balkans with his Slovene hosts on Monday. NATO's Supreme Allied Commander in Europe, Gen. Wesley Clark, thanked Slovenia on Friday for hosting a military maneuver that the alliance held for the first time in this former Yugoslav country. Gen. Wesley Kanne Clark was raised as a Protestant in Little Rock, Ark., where he was rought up by his mother and stepfather, Victor Clark. He left behind his widow, Veneta Kanne, and a lone son, Wesley. General Clark, 52, was nominated by US President Bill Clinton on March 31 to succeed Joulwan who is also a US

**3** D202.M.100.E.F.F.24 As Wesley K. Clark graduated first in his high school class, then first in his class at West Point, the family heaped praise on Veneta for raising him so well and also avoided any direct contact with Wesley out of respect for Veneta's new life, according to cousin Barry Kanne. The White House denied Wednesday that an early outgoing of U.S. General Wesley Clark as NATO commander is because of the way he handles the air campaign against Yugoslavia. The Russian Defense Ministry on Thursday rejected NATO Supreme Allied Commander Europe Wesley Clark's criticism of its military doctrine. Slovenia is in NATO's Partnership for Peace program.

**4** D202.M.100.E.F.F.F Army 4-Star General Wesley Clark graduated 1st in his class at high school and West Point , was a Rhodes Scholar at Oxford , and is fluent in Russian . Wounded 4 times in Vietnam , he was appointed White House fellow after the war . He headed the US Southern Command and was the senior military member in the team brokering the 1995 Dayton peace accords that ended the war in Bosnia . In 1997 Clinton picked him to head NATO and US forces in Europe . In 2000 Clark argued to not rule out use of ground troops in Kosovo and was removed from command 3 months early . Clark grew up in Little Rock , Arkansas and discovered as an adult that he is the grand

# Task 5: Responsiveness group rank within docset by summary source



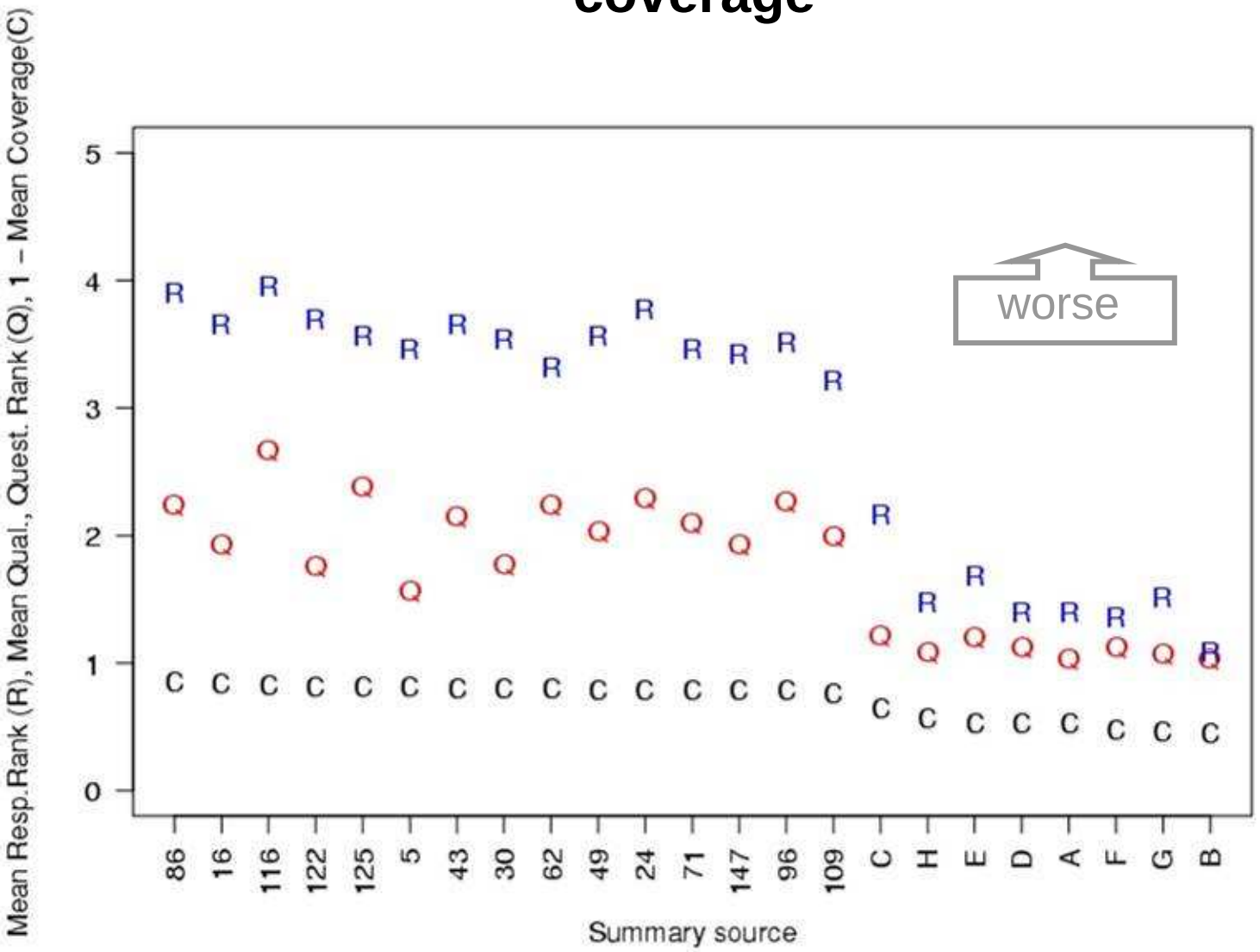
# Task 5: Responsiveness – multiple comparisons

| | |
|---|---|
| 116 | A |
| 86 | A B |
| 24 | A B C D |
| 16 | A B C D |
| 43 | A B C D |
| 125 | A B C D |
| 49 | A B C D |
| 96 | A B C D |
| 30 | A B C D |
| 71 | A B C D |
| 147 | B C D |
| 62 | C D |
| 109 | D |

- Comparing average ranks from Friedman's test (non-parametric)
- Using Tukey-Kramer criterion, at .05 significance level for group of all pair-wise comparisons.
- Lowest/worst score on top, highest/best on bottom
- One docset was omitted because of missing values (49 docsets).

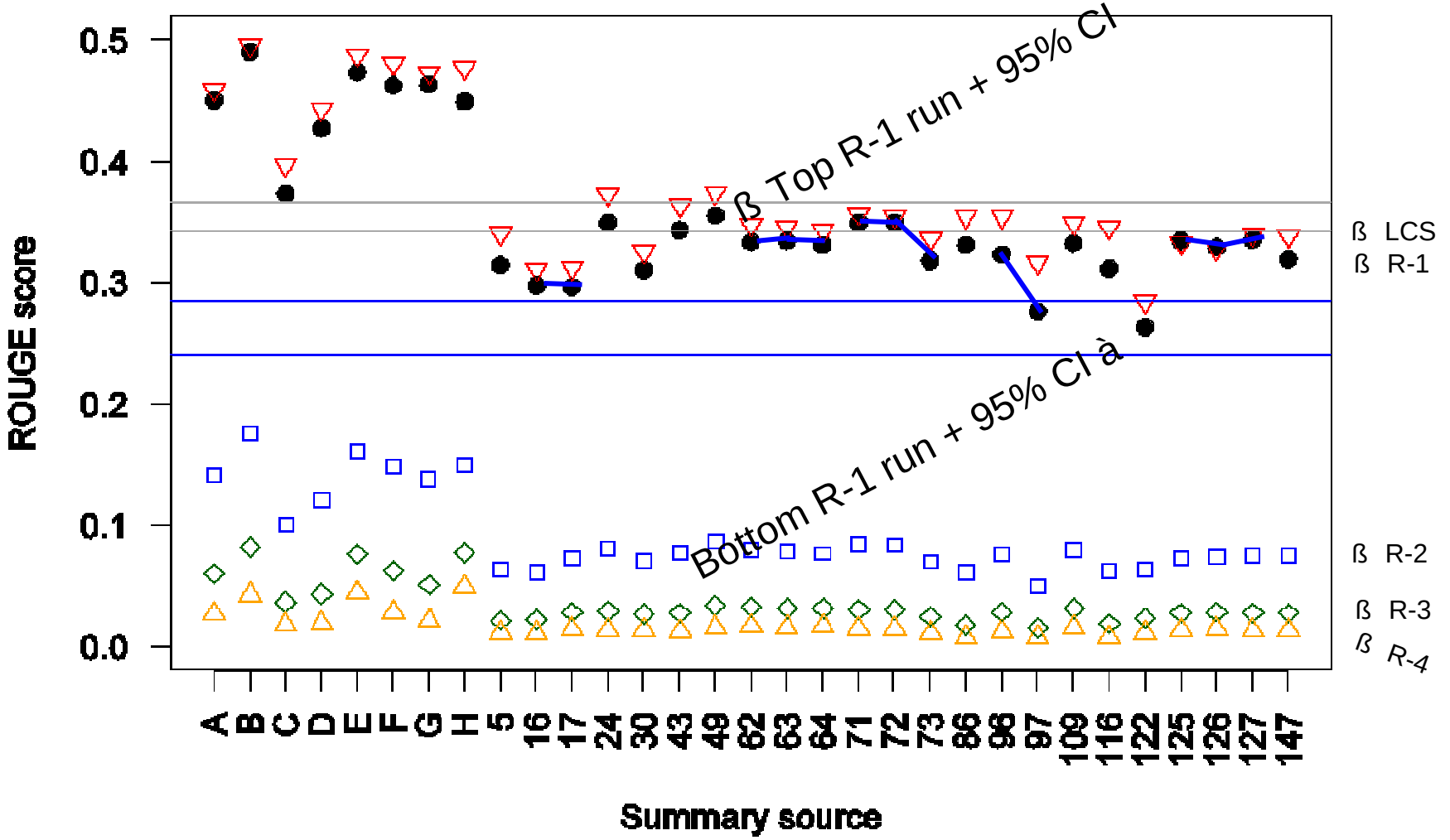
Runs with the same letter are not significantly different.

# Task 5: Responsiveness, linguistic quality, coverage



# Task 5: ROUGE scores by summary source

(blue lines connect runs (priority 1à 2à 3) from same group)



# Task 5: Coverage versus ROUGE

Correlations* of means for priority 1 runs

| | R-2 | R-3 | R-4 | LCS | LCS-W | Mean Coverage |
|---|---|---|---|---|---|---|
| R-1 | 0.975219 | 0.924497 | 0.846288 | 0.991905 | 0.991800 | 0.954240 |
| R-2 | | 0.977023 | 0.915436 | 0.967596 | 0.976287 | 0.962451 |
| R-3 | | | 0.977227 | 0.919919 | 0.937268 | 0.896223 |
| R-4 | | | | 0.847506 | 0.870654 | 0.801645 |
| LCS | | | | | 0.998633 | 0.947881 |
| LCS-W1.2 | | | | | | 0.948312 |

* Pearson's product moment

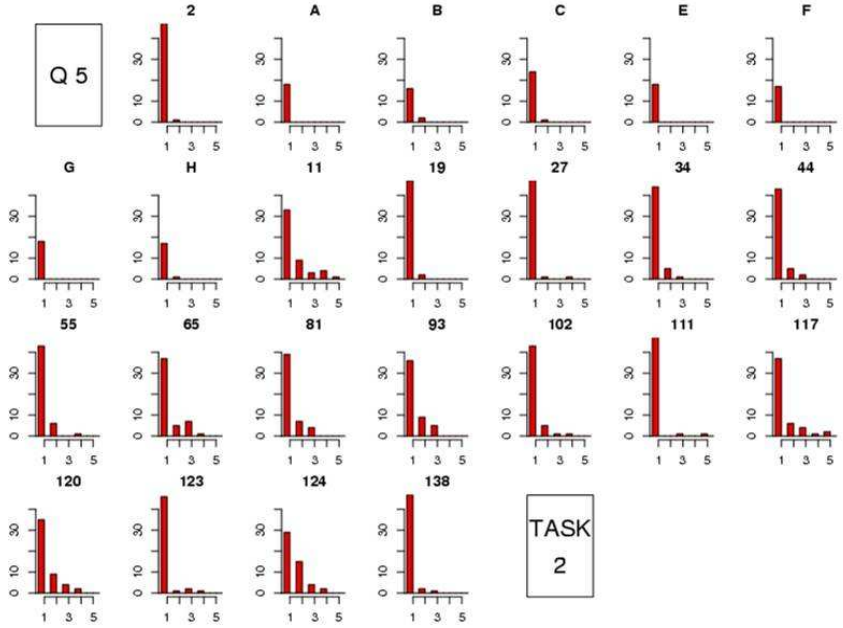
# Task 5: Recap

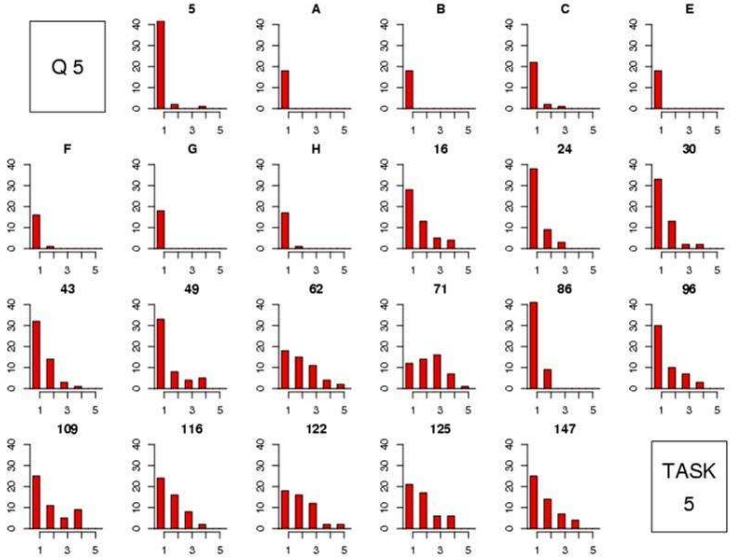
- Linguistic quality questions
  - Pass some sanity checks; seem to provide lots of detailed feedback
  - Mixed per-system results
  - Multiple comparisons finds differences only between extremes
- SEE coverage
  - Manual summaries' coverage more than twice that of others
  - Systems seem more alike than in other tasks
  - Systems' mean indistinguishable from baseline's
  - Multiple comparisons finds differences in systems at extremes
- Responsiveness
  - Afforded assessors different view/criteria but results resemble coverage
- ROUGE
  - LCS and R-1 track each other; likewise R-3,4
  - LCS/R-1 >> R-2,3,4
  - Correlation of SEE coverage and ROUGE means range from .802 (R-4) to .962 (R-1)

# Task 5: Questions

- Did groups do anything special because the task required the summary to be focused by a question / a person's name?

- Oddities: Why do the quality scores for Q5 (entities re-mentioned) seem to be worse than for task 2? – repetition of person X's name?



# Task 5: Questions

- Oddities: Why do the manual summaries in task 5 seem to have scored better than in task 2? – focus provided by question?