What is Quality?

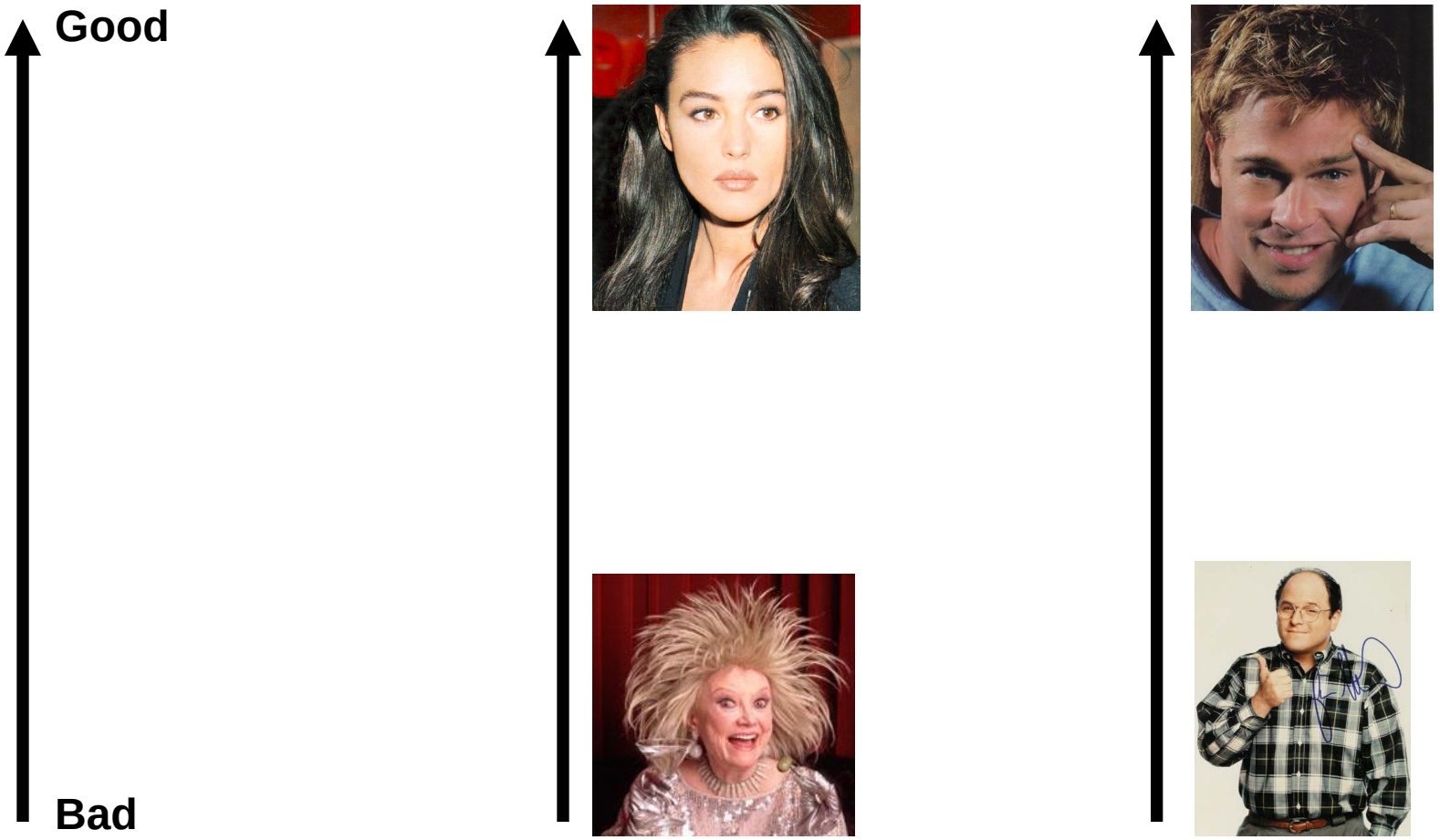
Workshop on Quality Assurance and Quality Measurement for Language and Speech Resources

Christopher Cieri
Linguistic Data Consortium

{ccieri}@ldc.upenn.edu

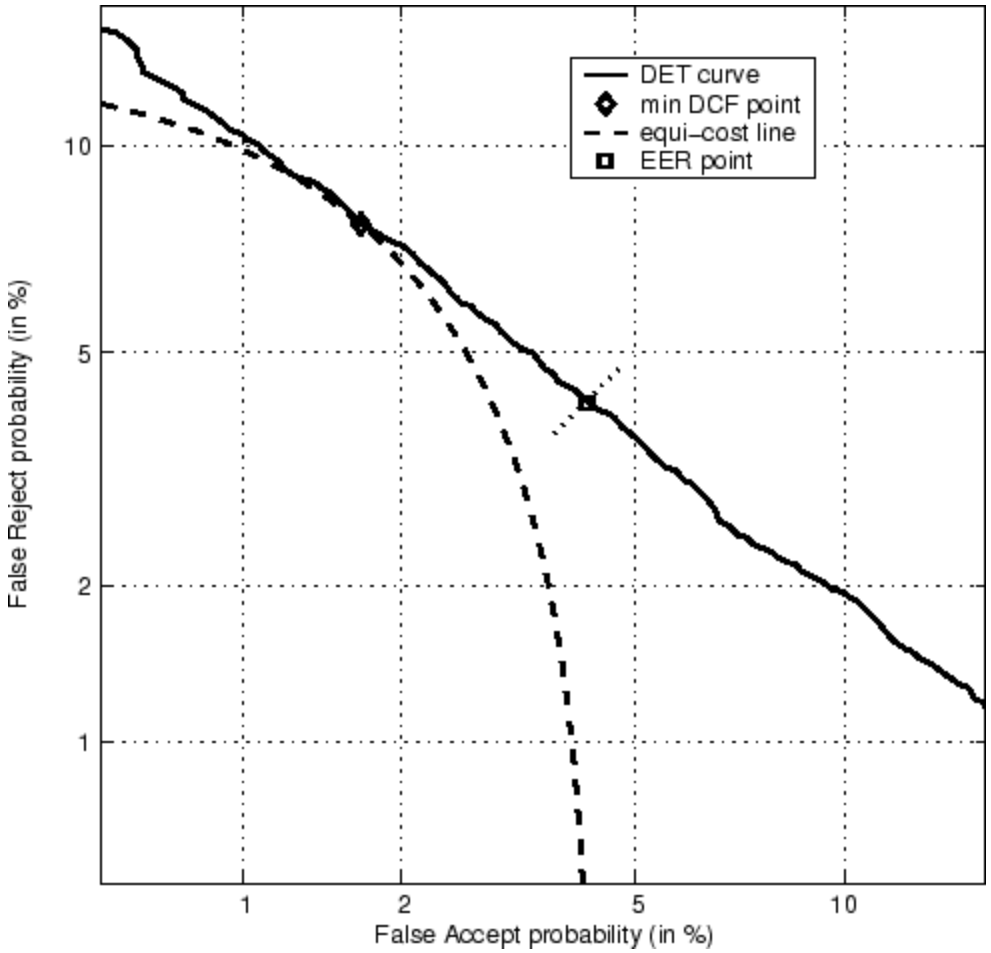
Common Quality Model

- A single dimension, a line that ranges from bad to good
  - goal is to locate ones data, software on the line and
  - move it toward better in a straight line.



- Appropriate as a tool for motivating improvements in quality
- But not the only model available and not accurate in many cases

Dimensions of IR Evaluation



- **Detection Error Trade-off (DET) curves.**
  - **describe system performance**
- **Equal Error Rate (EER) criterion**
  - **where false accept = false reject rate on DET**
  - **one-dimensional error figure**
  - **does not describe actual performance of realistic applications**
    - » **do not necessarily operate at EER point**
    - » **some require low false reject, others low false accept**
    - » **no a priori threshold setting; determined only after all access attempts processed (a posteriori)**

from ispeak.nl

- **Of course, human annotators are not IR systems**
  - **Human miss and false alarms rates are probably independent.**
- **However, project cost/timeline are generally fixed.**
  - **effort, funds devoted to some task are not available for some other**
- **Thus there are similar tradeoffs in corpus creation**

Collection Quality

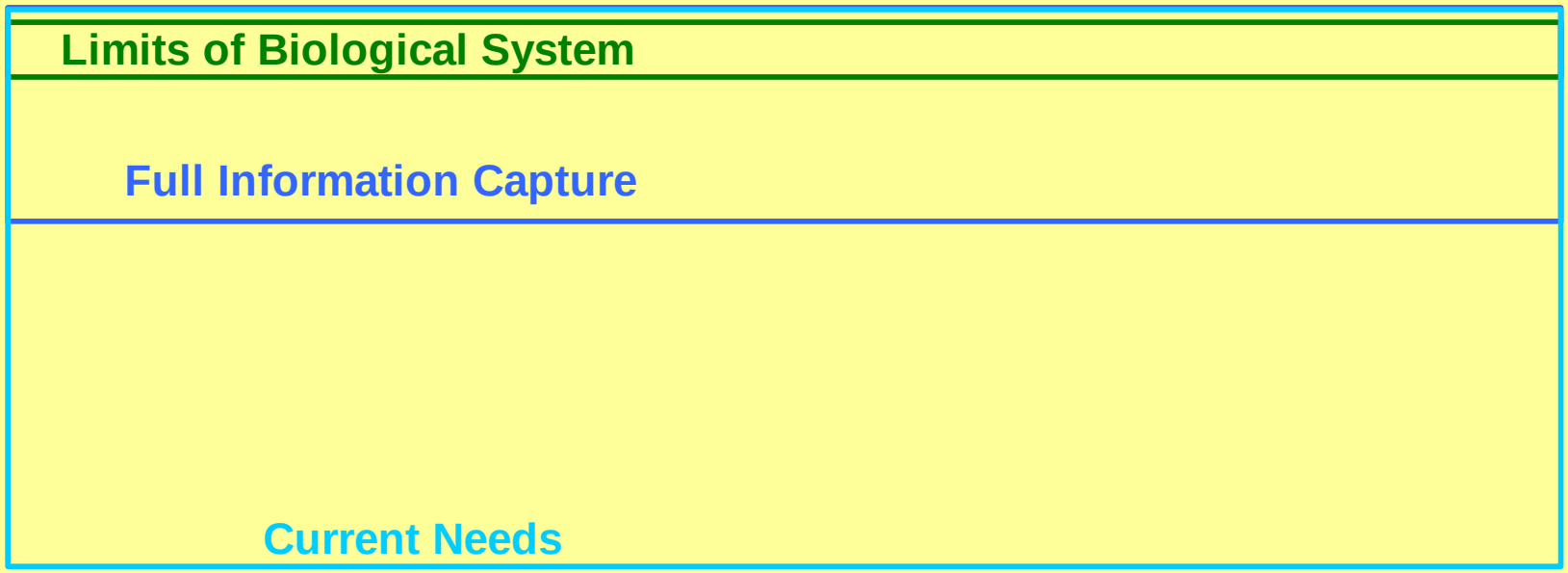
Limits of Biological System

Collection Quality

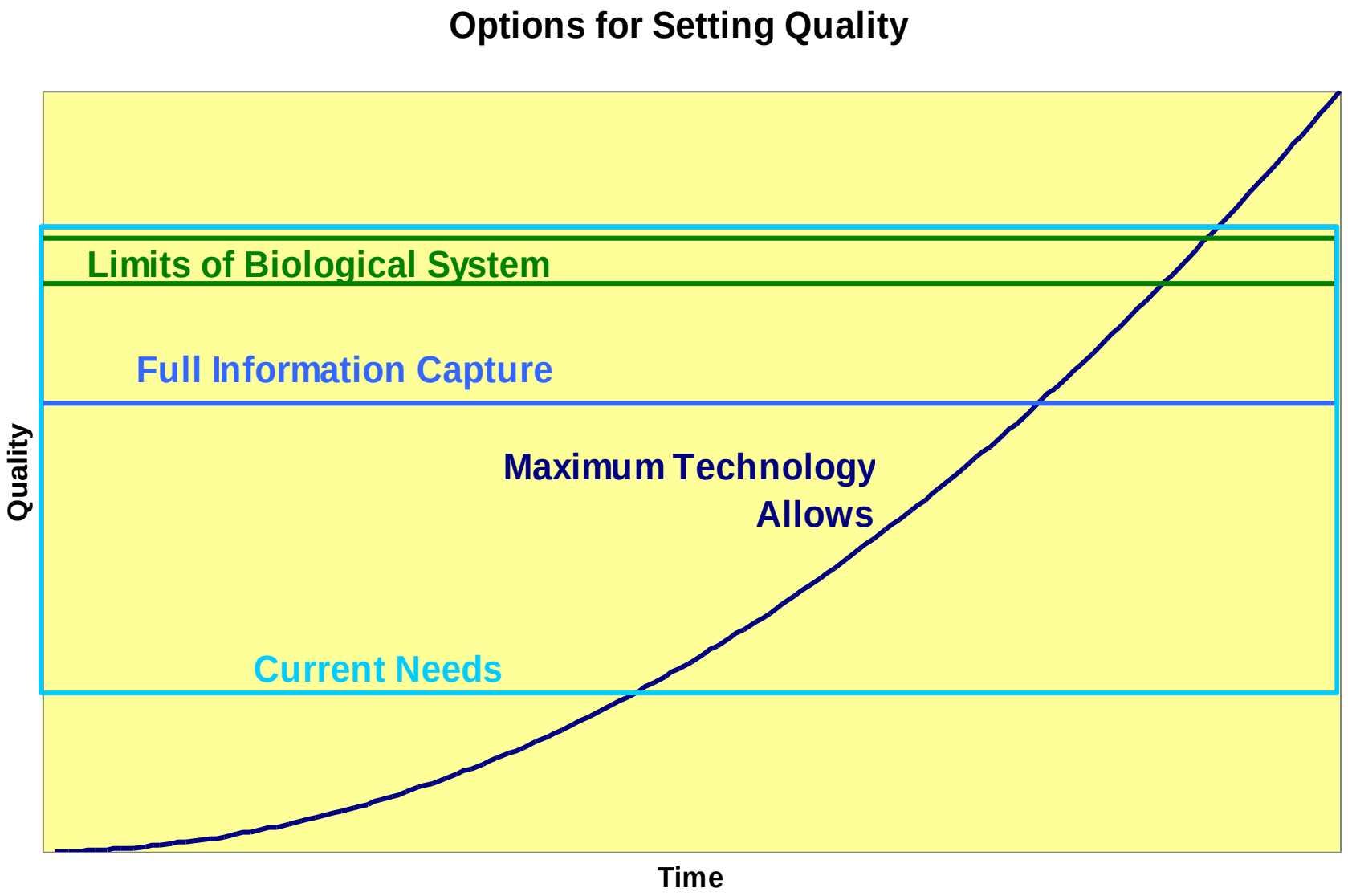
Limits of Biological System

Full Information Capture

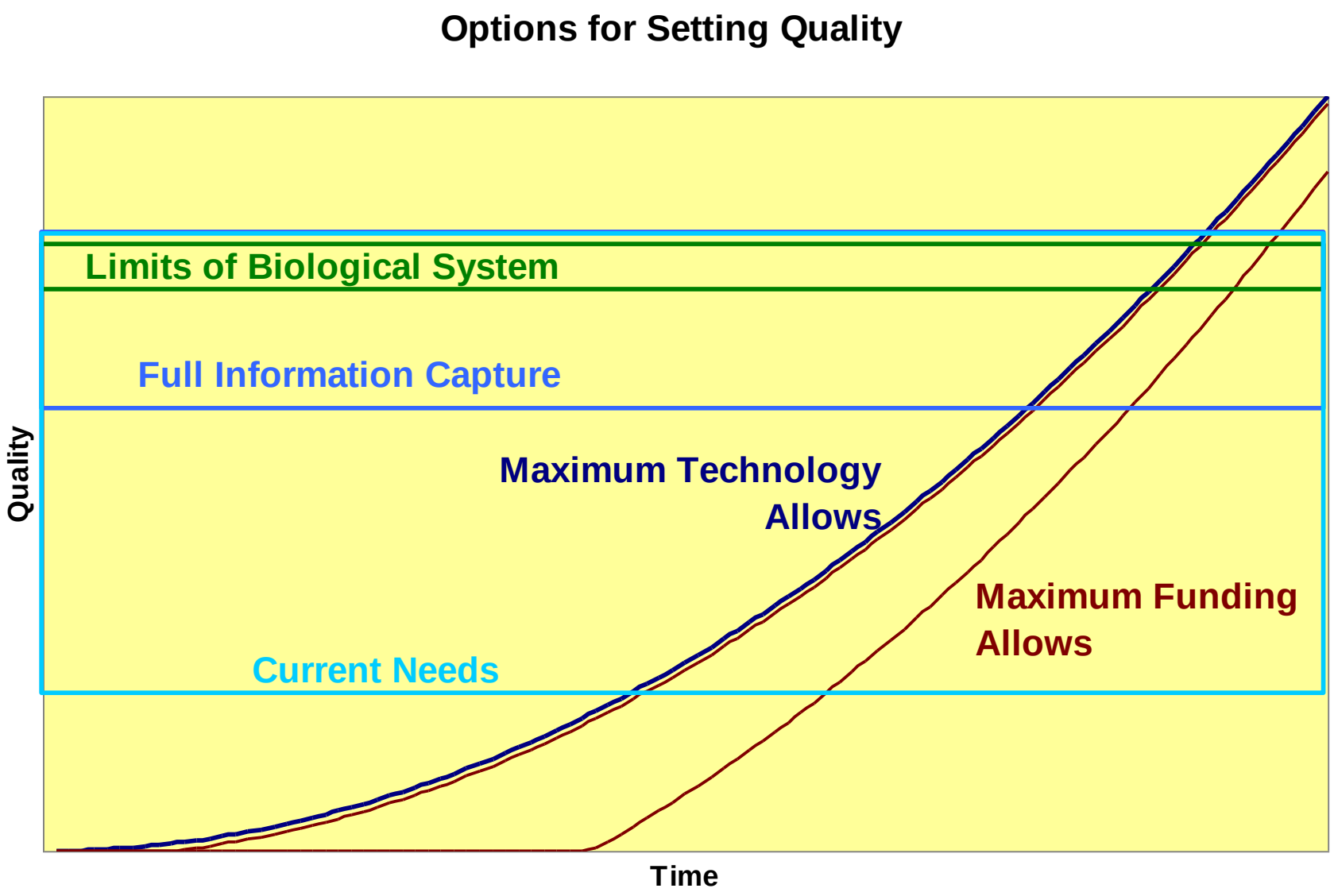
Collection Quality



Collection Quality

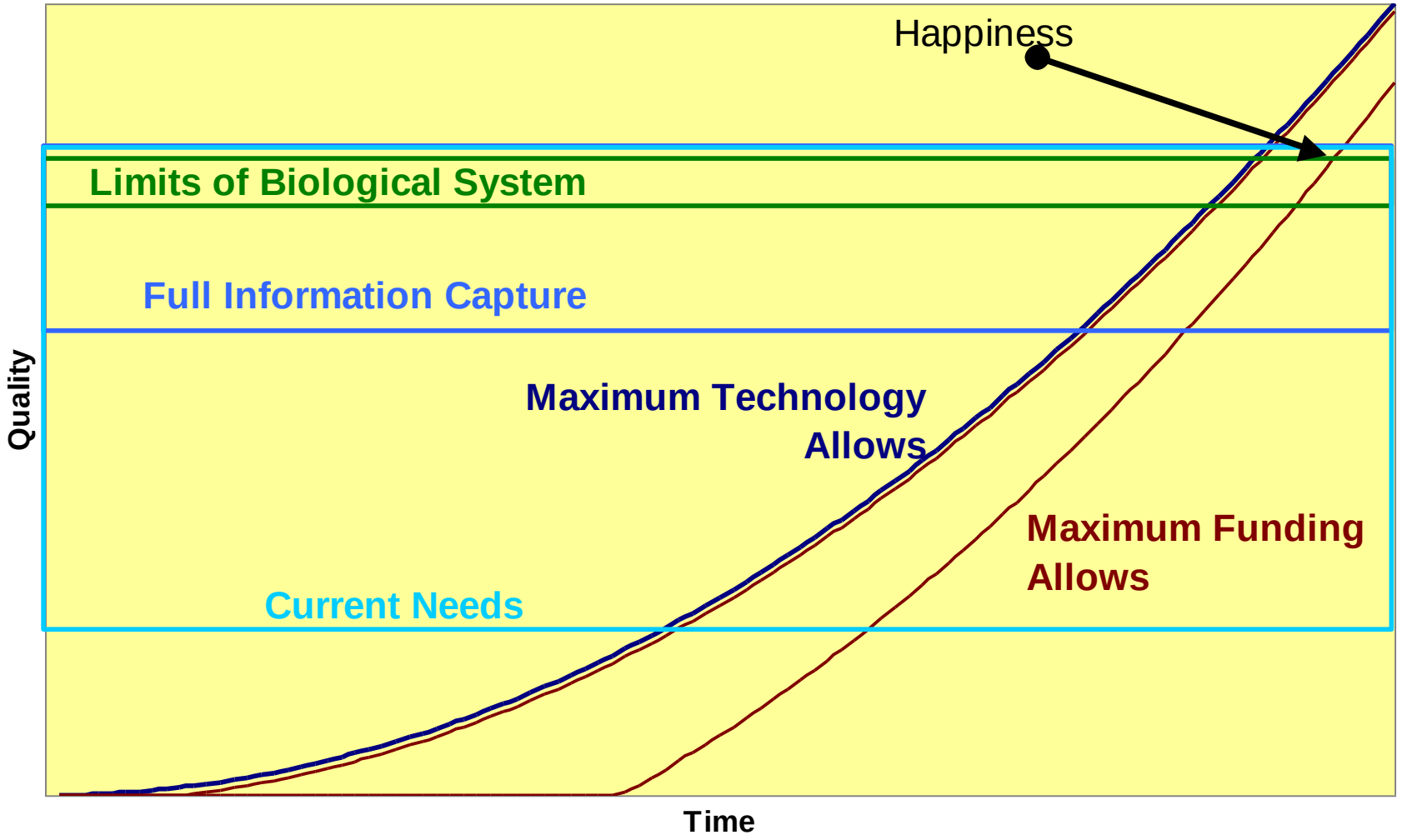


Collection Quality



Collection Quality

Options for Setting Quality



Components of Quality

- **Suitability: of design to need**
  - corpora created for specific purpose but frequently re-used
  - raw data is large enough, appropriate
  - annotation specification are adequately rich
  - publication formats are appropriate to user community
- **Fidelity: of implementation to design**
- **Internal Consistency:**
  - collection, annotation
  - decisions and practice
- **Granularity**
- **Realism**
- **Timeliness**
- **Cost Effectiveness**

Quality in Real World Data

- **Gigaword News Corpora**
  - **large subset of LDC's archive of news text**
  - **checked for language of the article**
  - **contain duplicates and near duplicates**

- **Systems that hope to process real world data must be robust against multiple languages in an archive or also against duplicate or near duplicates**

- **However, language models are skewed by document duplication**

Types of Annotation

- **Sparse or Exhaustive**
  - **Only some documents in a corpus are topic relevant**
  - **Only some words are named entities**
  - **All words in a corpus may be POS tagged**
- **Expert or Intuitive**
  - **Expert: there are right and wrong ways to annotate; the annotators goal is to learn the right way and annotate consistently**
  - **Intuitive: there are no right or wrong answers; the goal is to observe and then model human behavior or judgment**
- **Binary or Nary**
  - **A story is either relevant to a topic or it isn't**
  - **A word can have any of a number of MPG tags**

Annotation Quality

- **Miss/False Alarm and Insertion/Deletion/Substitution can be generalized and applied to human annotation.**

- **Actual phenomena are observed**
  - **failures are misses, deletions**
- **Observed phenomena are actual**
  - **failures are false alarms, insertions**
- **Observed phenomena are correctly categorized**
  - **failures are substitutions**

QA Procedures

- **Precision**
  - attempt to find incorrect assignments of an annotation
  - 100%
- **Recall**
  - attempt to find failed assignments of an annotation
  - 10-20%
- **Discrepancy**
  - resolve disagreements among annotators
  - 100%
- **Structural**
  - identify, better yet, prevent impossible combinations of annotations

Dual Annotation

- **Inter-annotator Agreement != Accuracy**
  - **studies of inter-annotator agreement indicate task difficulty or**
  - **overall agreement in the subject population as well as**
  - **project internal consistency**
  - **tension between these two uses**
    - » **As annotation team becomes more internally consistent it ceases to be useful for modeling task difficulty.**

- **Results from dual annotation used for**
  - **scoring inter-annotator agreement**
  - **adjudication**
  - **training**
  - **developing gold standard**

- **Quality of expert annotation may be judged by**
  - **comparison with another annotator of known quality**
  - **comparison to gold standard**

Limits of Human Annotation

- **Linguistic resources used to train and evaluate HLTs**
  - **as training they provide behavior for systems to emulate**
  - **as evaluation material they provide gold standards**
- **But, human are not perfect and don't always agree.**
- **Human errors, inconsistencies in LR creation provide inappropriate models and depress system scores**
  - **especially relevant as system performance approaches human performance**
- **HLT community needs to**
  - **understand limits of human performance in different annotation tasks**
  - **recognize/compensate for potential human errors in training**
  - **evaluate system performance in the context of human performance**
- **Example: STT R&D and Careful Transcription in DARPA EARS**
  - **EARS 2007 Go/No-Go requirement was WER 5.6%**

Transcription Process

Regular workflow:

Annotator 1	SEG: segmentation
Annotator 2	1P: verbatim transcript
Annotator 3	2P: check 1P transcript, add markup
Lead Annotator	QC: quality check, post-process

30+ hours labor/hour audio

Dual annotation workflow:

Annotator 1	SEG	SEG	Annotator 2
Annotator 1	1P	1P	Annotator 2
Annotator 1	2P	2P	Annotator 2

Lead Annotator:
Resolve discrepancies, QC & post-process

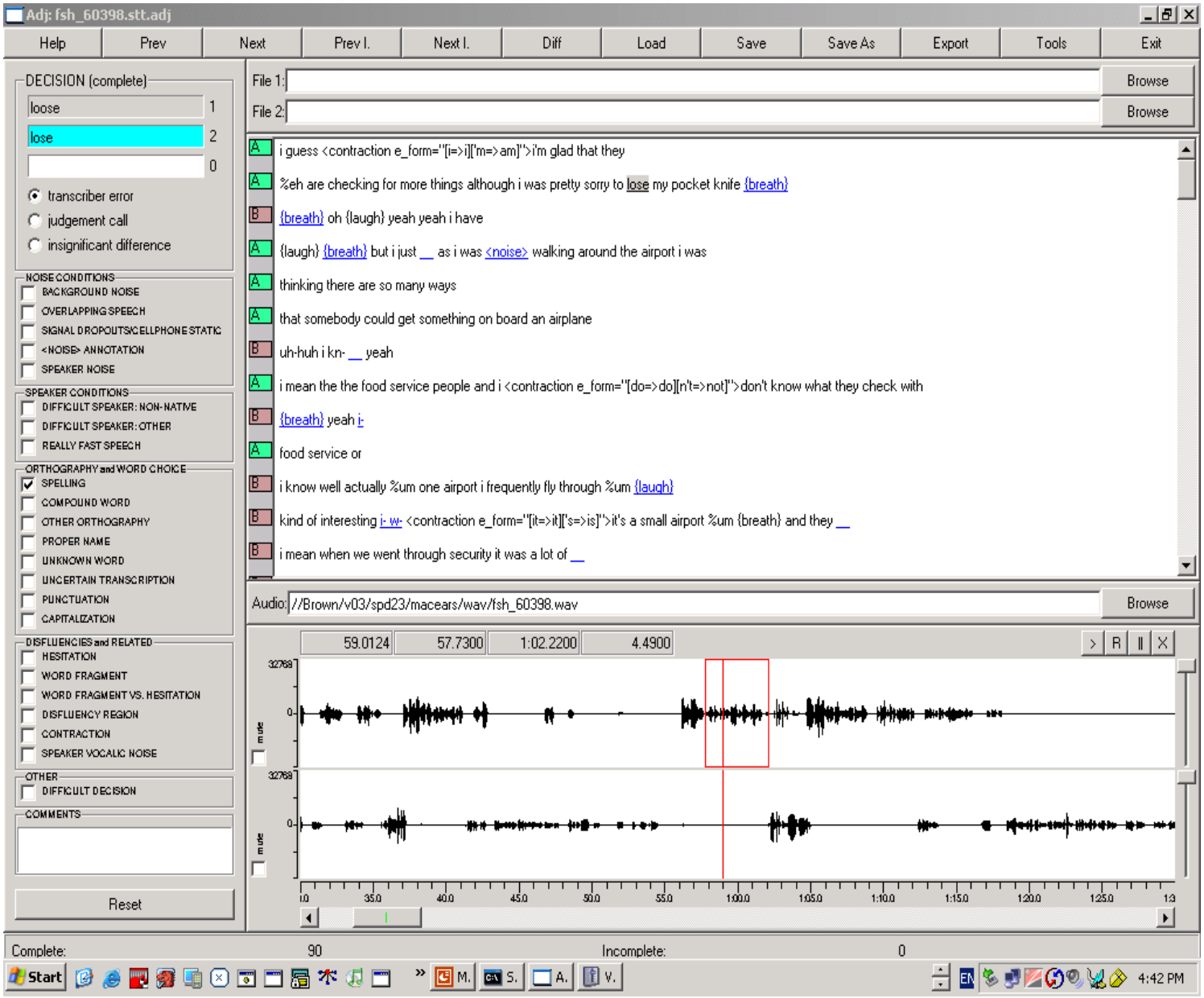
Results

- **EARS 2007 goal was WER 5.6%**

	LDC 1	*LDC 2*
LDC Careful Transcription 1	0	4.1
LDC Careful Transcription 2	4.5	0
WordWave Transcription	6.3	6.6
LDC Quick Transcription	6.5	6.2
LDC 2, Pass 1	5.3	
LDC 2, Pass 2	5.6	

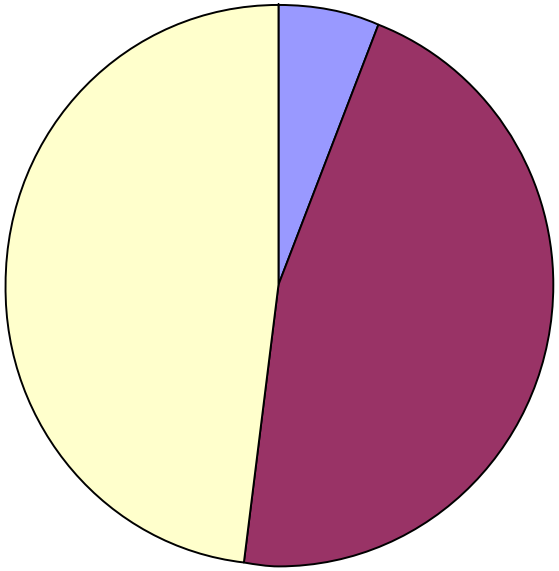
- **Best Human WER 4.1%**
- **Excluding fragments, filled pauses reduces WER by 1.5% absolute.**
- **Scoring against 5 independent transcripts reduces WER by 2.3%.**
- **Need to improve quality of human transcription!!!**

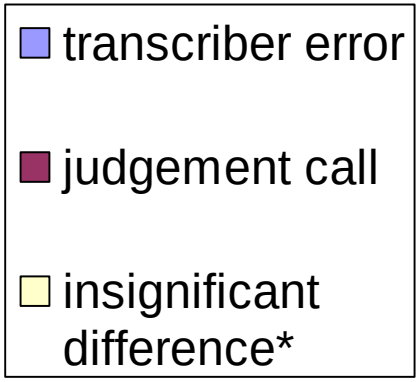
Transcript Adjudication



CTS Consistency

Word Disagreement Rate (WER)		
System	Orig RT-03	Retrans RT-03
Orig RT-03	0%	4.1%
Retrans RT-03	4.5%	0%

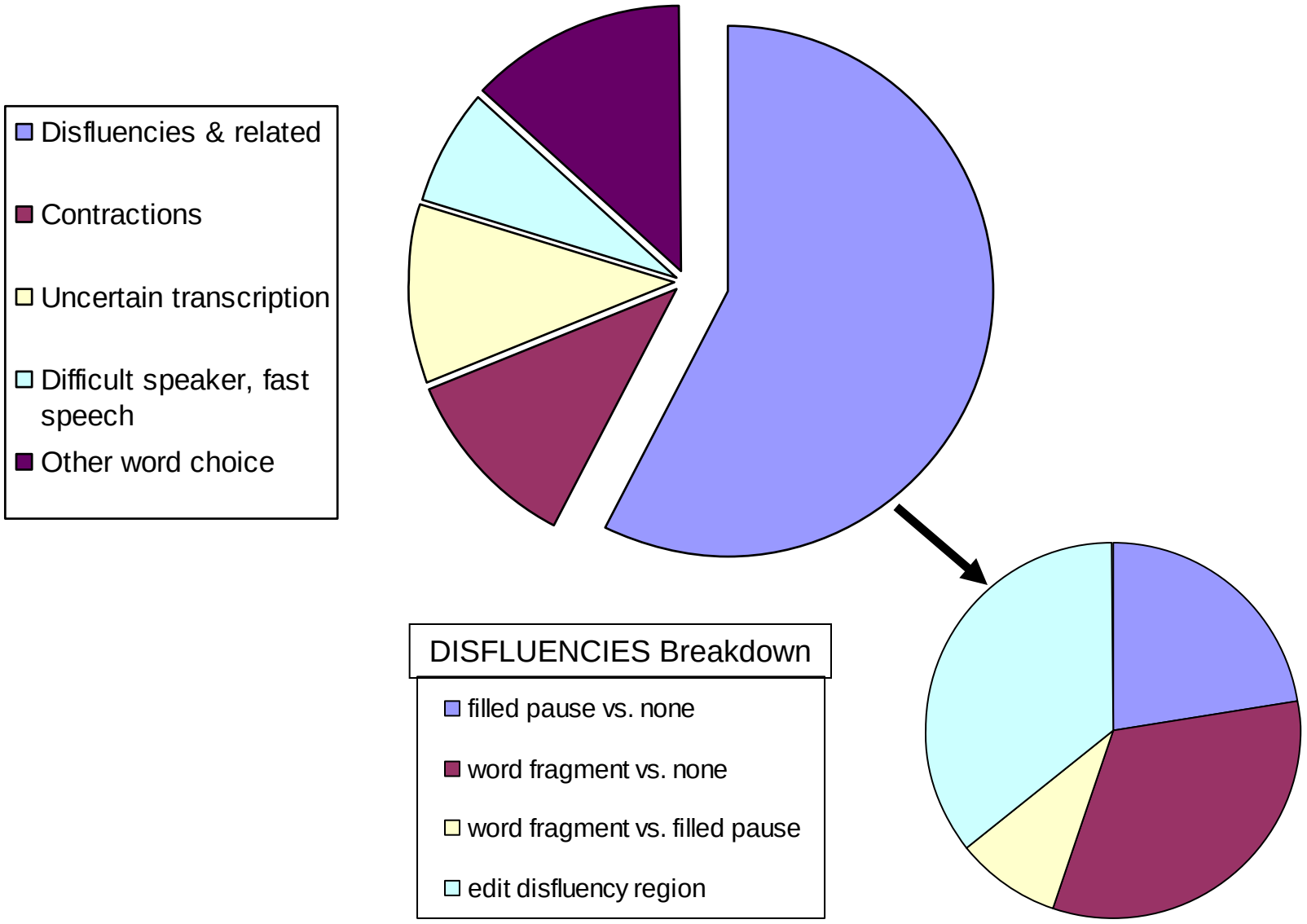




*most, but not all, insignificant differences are removed from scoring

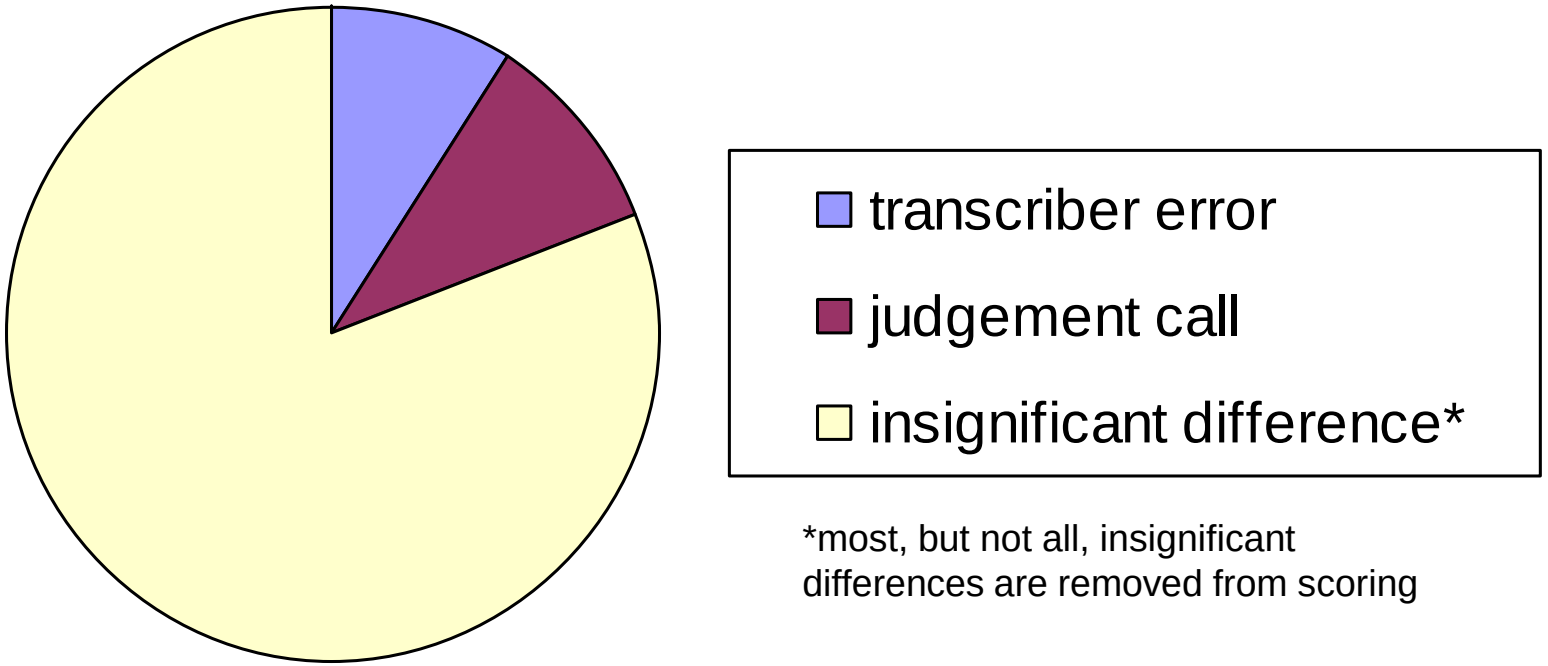
WER based on Fisher data from RT-03 Current Eval Set (36 calls)
Preliminary analysis based on subset of 6 calls; 552 total discrepancies analyzed

CTS Judgment Calls



BN Consistency

Word disagreement rate (equiv. to WER)		
Basic	RT-03 GLM	RT-04 GLM
1.3%	1.1%	0.9%



*most, but not all, insignificant differences are removed from scoring

WER based on BN data from RT-03 Current Eval Set (6 programs)
Analysis based on all files; 2503 total discrepancies analyzed

Conclusions

- **Many scorable annotator discrepancies involve disfluencies that have no clear target**
- **Cost to "get it right" high relative to benefit**
- **Proposal**
  - **Fully transcribe clear cases**
  - **Mark unclear as such and ignore**
    - » **In further annotation**
    - » **In scoring**

Head Room

- **TDT Goal was a system to monitor news performing automatic transcription & translation, division of the broadcast into stories and categorization of the stories by topic.**
- **Data is transcribed, translated broadcast news sessions from multiple media, languages that are segmented into stories and then categorized by topic.**

	Months	Hours	English	Topics	Decisions
TDT-2	6	800	72000	100	7.2M
TDT-3	3	600	51000	120	6.1M
TDT-4	4	615	57000	60	3.4M

Story Segmentation

- **Listen to audio file, view waveform & transcript**
- **Segment**
  - **Review story boundaries inserted during transcription; add, delete, modify boundaries as needed**
  - **Classify sections as *news*, *not news* (miscellaneous), teaser or *un(der)transcribed***
  - **Set and confirm timestamps for all story boundaries**
- **Every file receives a single pass by LDC annotators**
  - **Independent second pass optional**
  - **Quality control through annotator training, spot checking**
- **Evaluation text is bereft of segments; they are encoded in stand-off file.**

Story Segmentation and QC

- Additional QA on segmented material
  - ratio of text words to audio duration for each section
  - sections with unusual ratios re-examined

- 5% of files dually segmented/second-passed by independent annotators; results reconciled by team leaders

- Results of QC showed high rates of consistency among annotators relative to the scores of systems – head room
  - total cost of story boundary detection:
    - Human Cseg: 0.036
    - System Cseg: 0.319-0.873

- But, what about other uses of story boundaries???

Topic Detection and Tracking

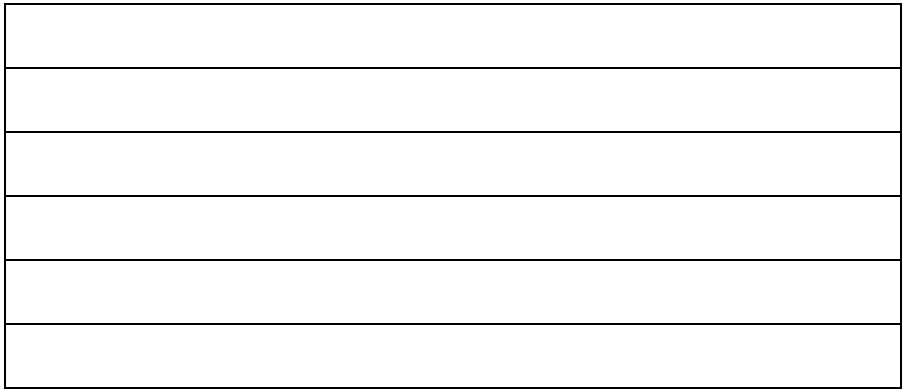
- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**

TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

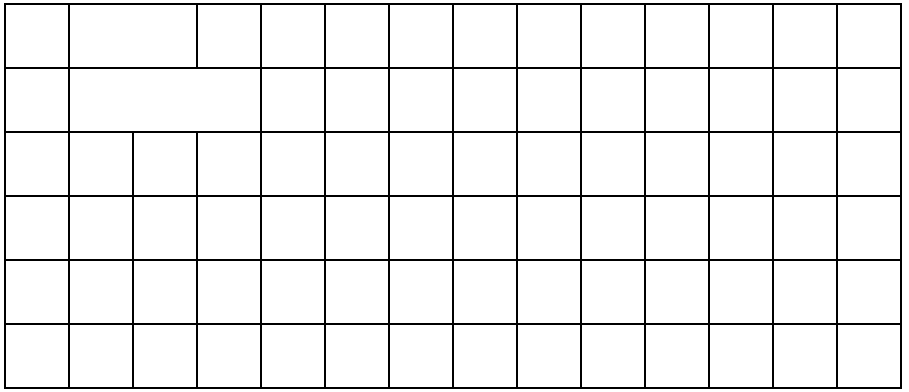
- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

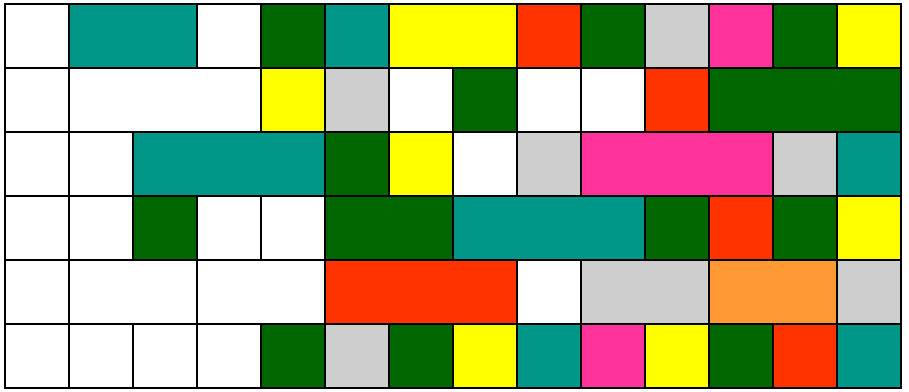
- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

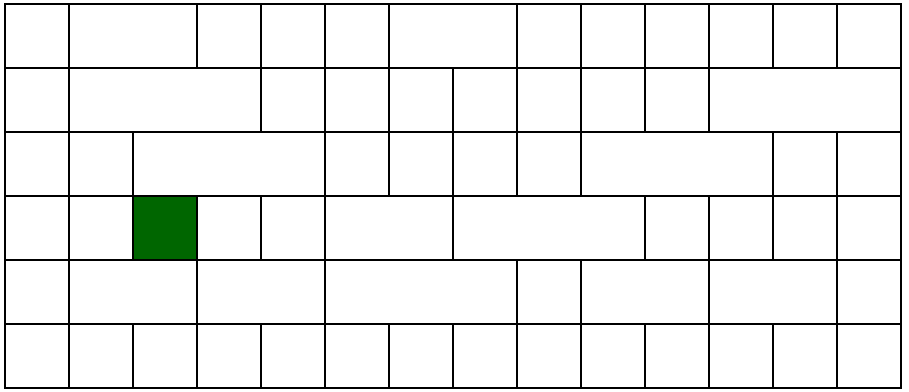
- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

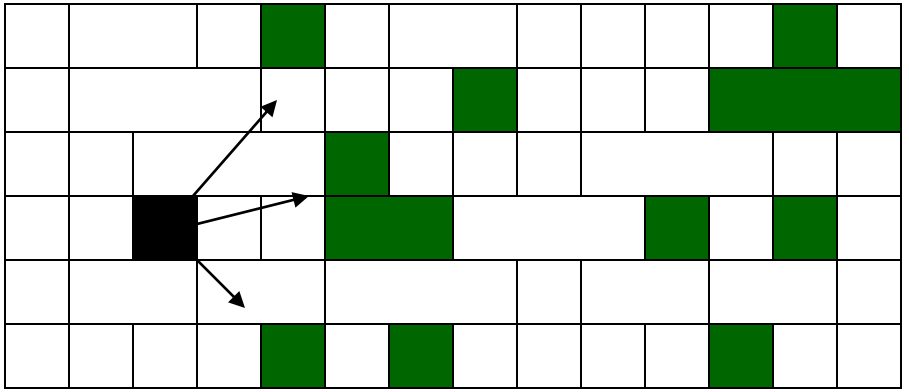
- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

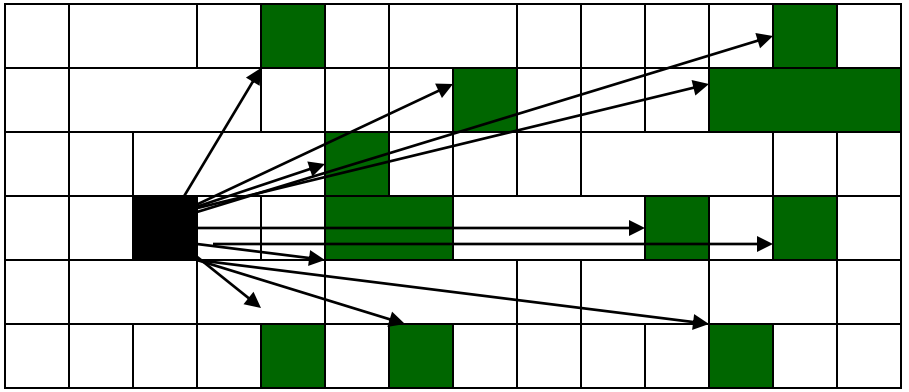
- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

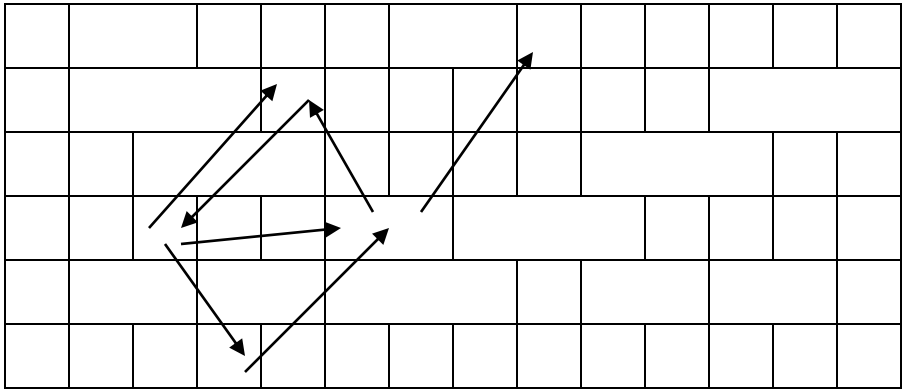
- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



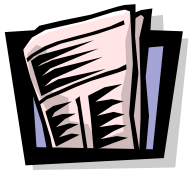
TDT Overview

- **US sponsored, common task program**
- **Manage information in archives of broadcast news and news text.**

- **Tasks**
  - » **segmentation**
  - » **topic detection**
  - » **first story detection**
  - » **topic tracking**
  - » **story link detection**



TDT Process



Transcription

Automatic Formatting

Machine Translation

Segmentation

SGML-encoded Text Files

Topic Labelling

Translated Files

Boundary Table

Relevance Table

TDT Corpus

Conclusion

- **Story boundaries have significant effect on other tasks, in particular detection.**

- **Additional effort on segmentation warranted.**

When is Less More?

- **DARPA EARS researchers needed 2000 hours of transcribed speech to reach programs aggressive go/no-go criteria.**
- **At 35-50xRT program could not afford careful transcription used previously.**
- **How to create the required transcripts within budget?**

- **Solution: Lower Quality**
  - **Larger quantity of lower quality data sooner will provide better results that smaller quantity of higher quality data later.**

Experiment

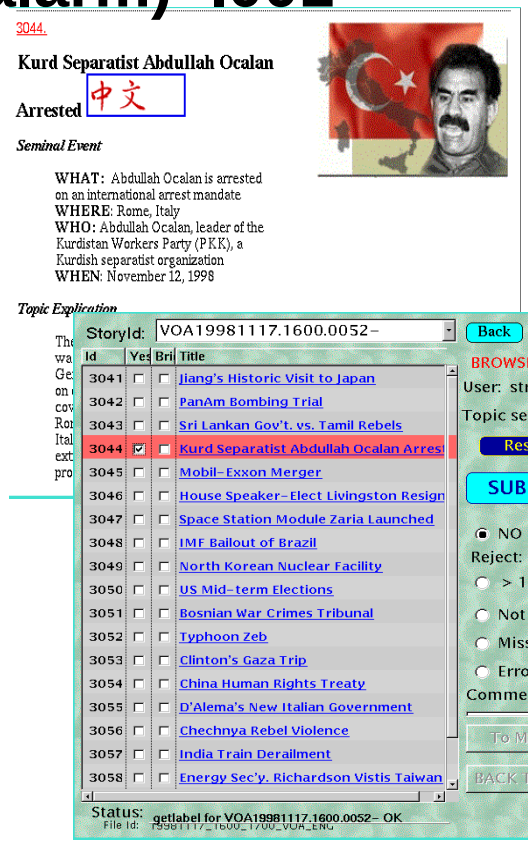
- **Select 20 hours of Switchboard audio for which careful transcripts existed from MSU.**
- **Transcribe them using quick transcription (QTR) specification.**
- **Train fresh systems on either 20 hour training set.**
- **Test against current evaluation corpus.**

	Training Hrs	%WER
MSU	23.4	38.0
LDC QTR	17.9	39.4
WordWave	19.6	38.8

- **Systems trained on 20 hours of QTR perform as well as systems trained on equal amounts of carefully transcribed data.**
- **And they cost much less**
- **So volume was increased to 2700 hours in Year 1.**

Topic Annotation

Exhaustive annotation; read each story, indicate topic relevance. TDT2 encoded 5.8M decisions. TDT3 corpus encodes 2.6M decisions. Quality: p(miss)=.04, p(false-alarm)=.001



article "VOA19981130.1600.1970"

<DOC>
<DOCNO> VOA19981130.1600.1970 </DOCNO>
<DOCTYPE>
<DATE_TIM
<BODY>

Internation
Urjalon, co
economic t
minister. In
internation

</BODY>
<END_TIME
</DOC>

article "APW19981128.0528"

<DOC>
<DOCNO> APW19981128.0528 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIM
<HEADER>
w2936 &Cx
r i &Cx13
</HEADER>
<BODY>
<SLUG> BC
<HEADLINE
Turkey pr
</HEADLIN
By CANDIC
Associate

ROME (A
to hand ov
alleged cri

Italy has re
Turkey. De
would be n

He said the
rebel leade
and videos
news agen

article "XIN19981120.0131"

<DOC>
<DOCNO> XIN19981120.0131 </DOCNO>
<DOCTYPE> NEWS STORY </DOCTYPE>
<DATE_TIME> 1998-11-20 23:32:26 </DATE_TIME>
<BODY>
<DOCOLDNO> CB415020.BFJ (275) </DOCOLDNO>
<HEADLINE> 德政府暂时放弃引渡奥贾兰 </HEADLINE>

新华社波恩１１月２０日电（记者刘钢）德国政府发言人海耶２０日在波恩说，德国政府决定暂时不要求引渡被扣押在意大利的土耳其库尔德工人党领导人奥贾兰。

１９９０年，德国法院曾因奥贾兰涉嫌参与发生在德国领土上的谋杀和纵火而发出了对他的逮捕令。本月１２日奥贾兰因持假护照进入意大利而在罗马机场被意警方拘留后，德国曾表示要求引渡奥贾兰到德国受审。

海耶在解释德国政府的决定时说，德国政府认为，土耳其政府已经提出了引渡奥贾兰的要求，而土耳其政府指控奥贾兰所犯的罪行更加严重。但他同时强调，德国法院对奥兰贾的逮捕令并未取消。（完）

</BODY>
</DOC>

Annotation Strategy

- **Overview**
  - **Search-guided complete annotation**
  - **Work with one topic at a time**
  - **Multiple stages for each topic**

- **Stage 1: Initial query**
  - **Submit seed story or keywords as query to search engine**
  - **Read through resulting relevance-ranked list**
  - **Label each story as YES/NO/BRIEF**
    - » ***BRIEF*: 10% or less of story discusses topic**
  - **Stop after finding 5-10 on-topic stories, or**
  - **After reaching "off-topic threshold"**
    - » **At least 2 off-topic stories for every 1 OT read AND**
    - » **The last 10 consecutive stories are off-topic**

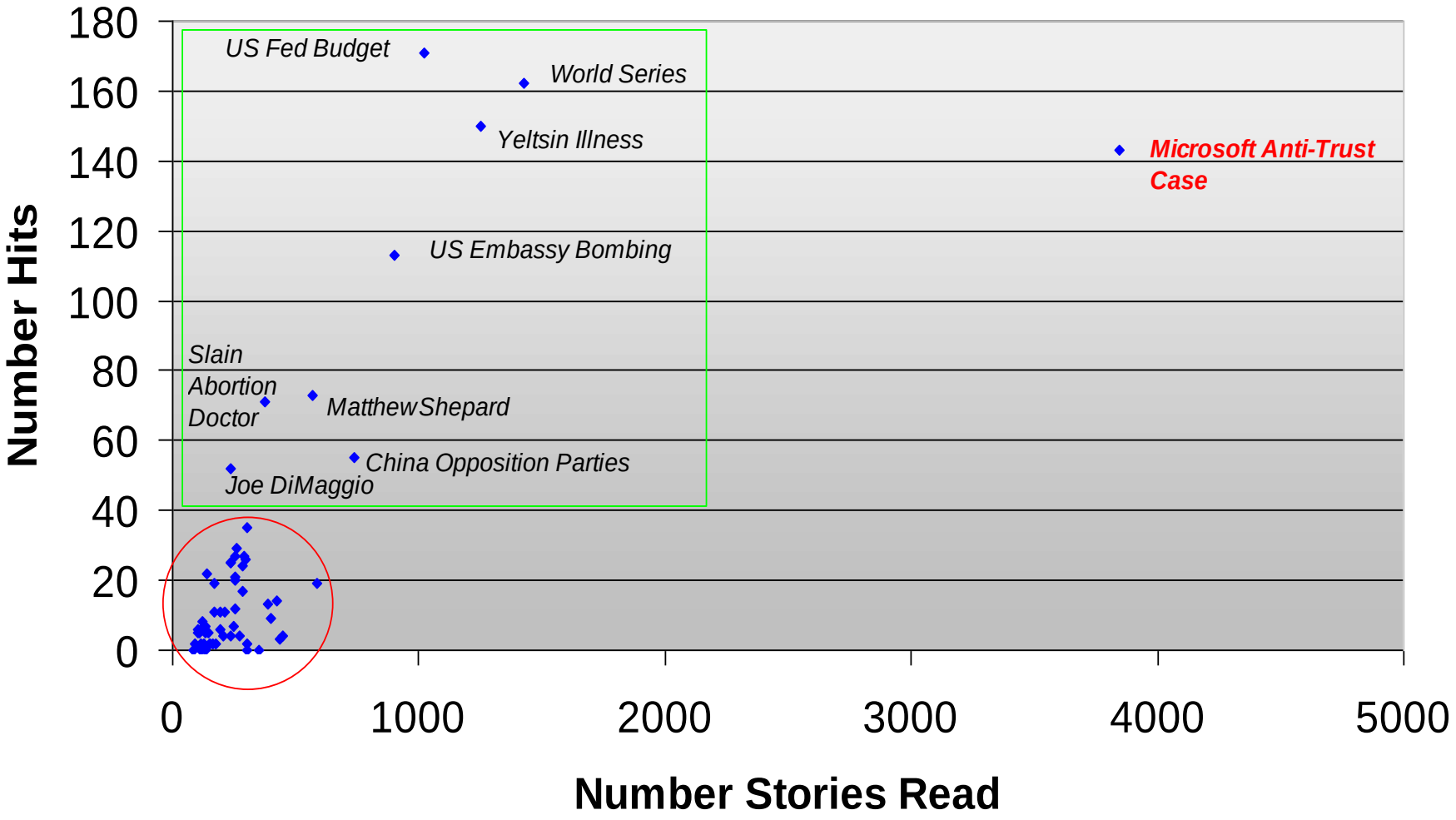
Annotation Strategy

- **Stage 2: Improved query using OT stories from Stage 1**
  - **Issue new query using concatenation of all known OT stories**
  - **Read and annotate stories in resulting relevance-ranked list until reaching off-topic threshold**
- **Stage 3: Text-based queries**
  - **Issue new query drawn from topic research & topic definition documents plus any additional relevant text**
  - **Read and annotate stories in resulting relevance-ranked list until reaching off-topic threshold**
- **Stage 4: Creative searching**
  - **Annotators instructed to use specialized knowledge, think creatively to find novel ways to identify additional OT stories**

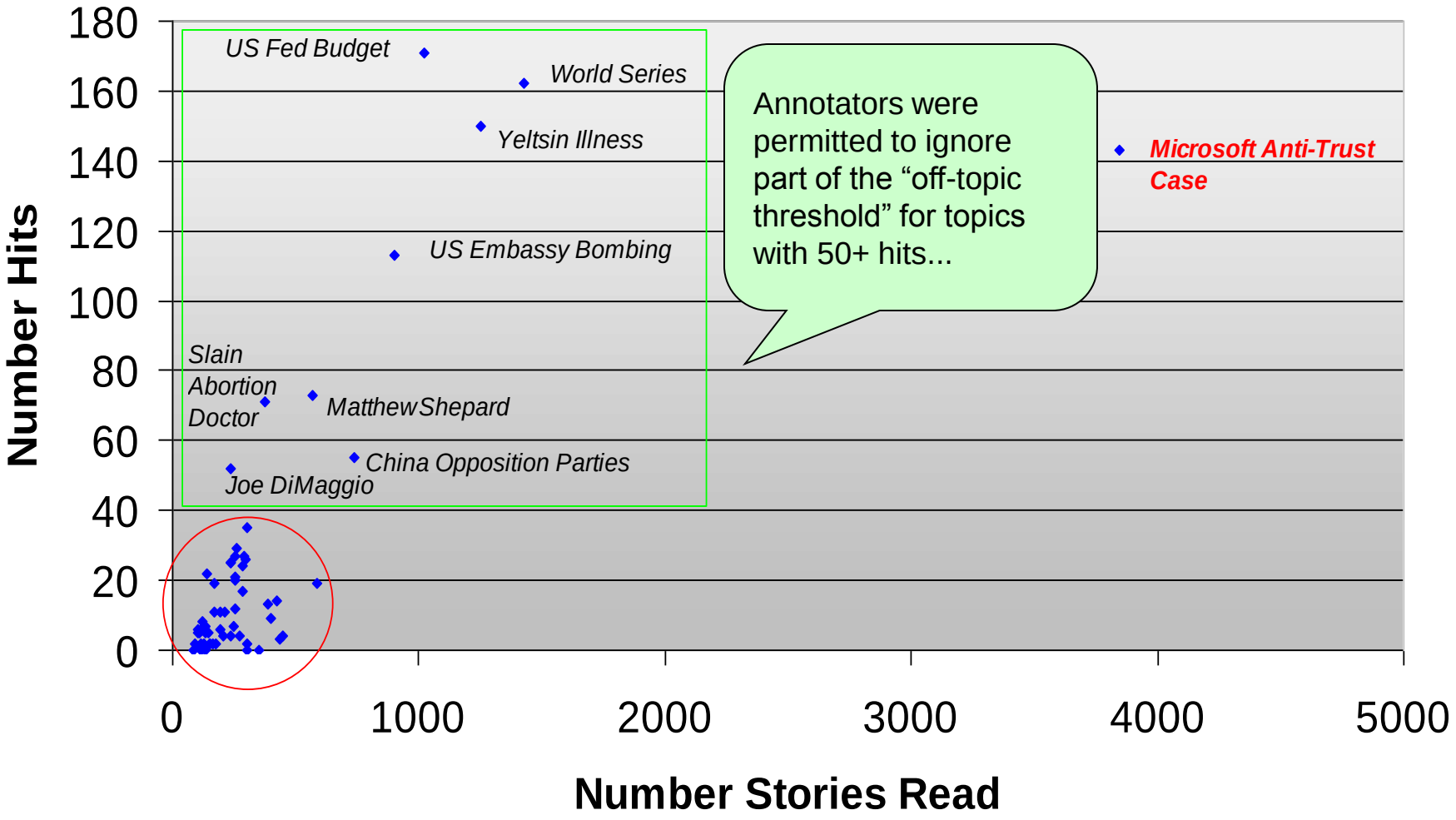
Annotation QC Measures

- **Precision**
  - **All on-topic (YES) stories reviewed by senior annotator to identify false alarms**
- **Recall**
  - **Search stories marked off topic looking for misses.**
- **Adjudication**
  - **Review sites' results and adjudicate cases where majority of sites disagree with annotators' judgments**
- **Dual annotation**
  - **10% of topics entirely re-annotated by independent annotators**
    - » **Impossible to re-annotate 10% of *stories* due to annotation approach**
  - **Compare YES/BRIEF judgments for both sets of results to establish some measure of inter-annotator agreement**

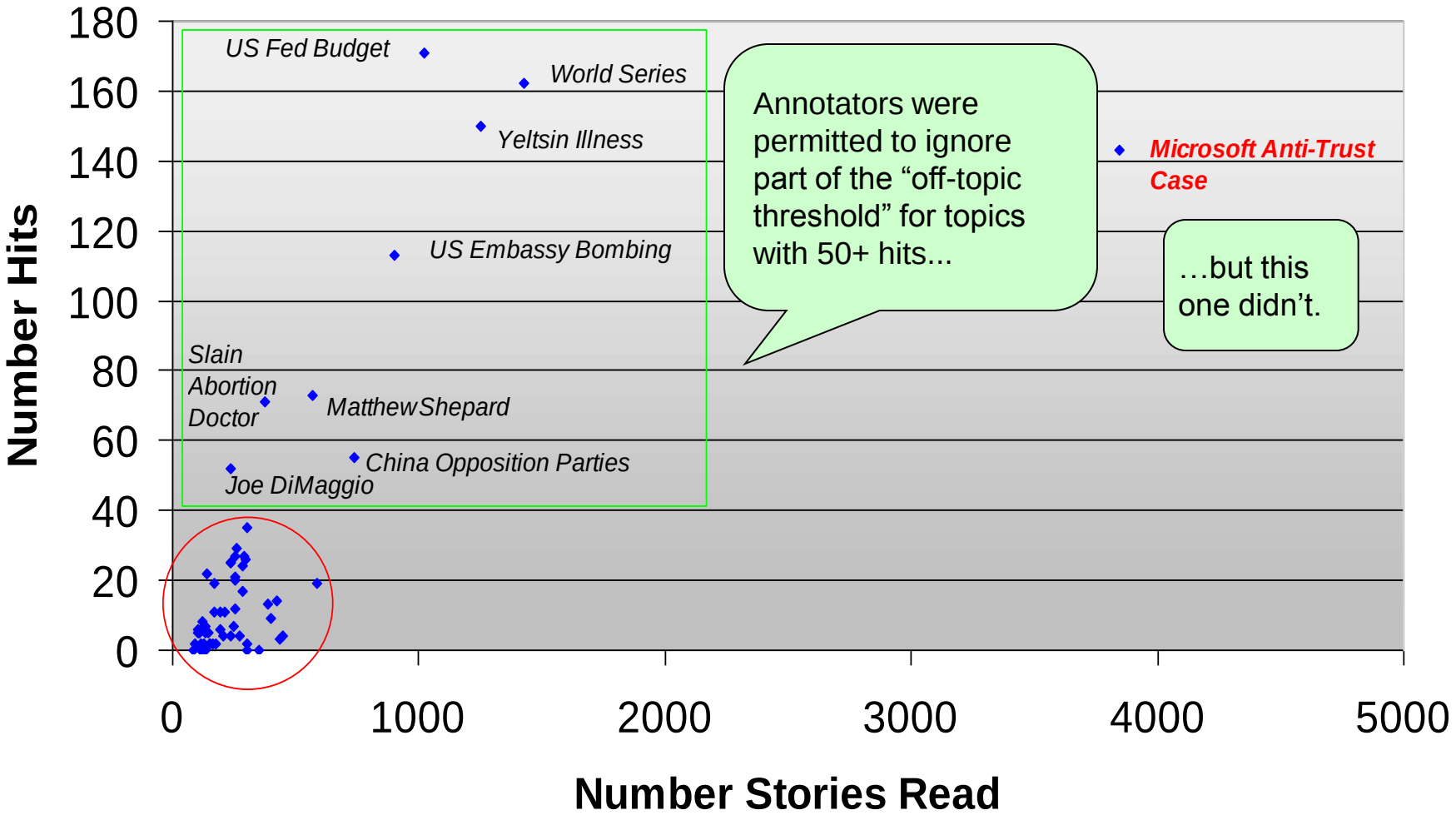
English Hits vs. Stories Read



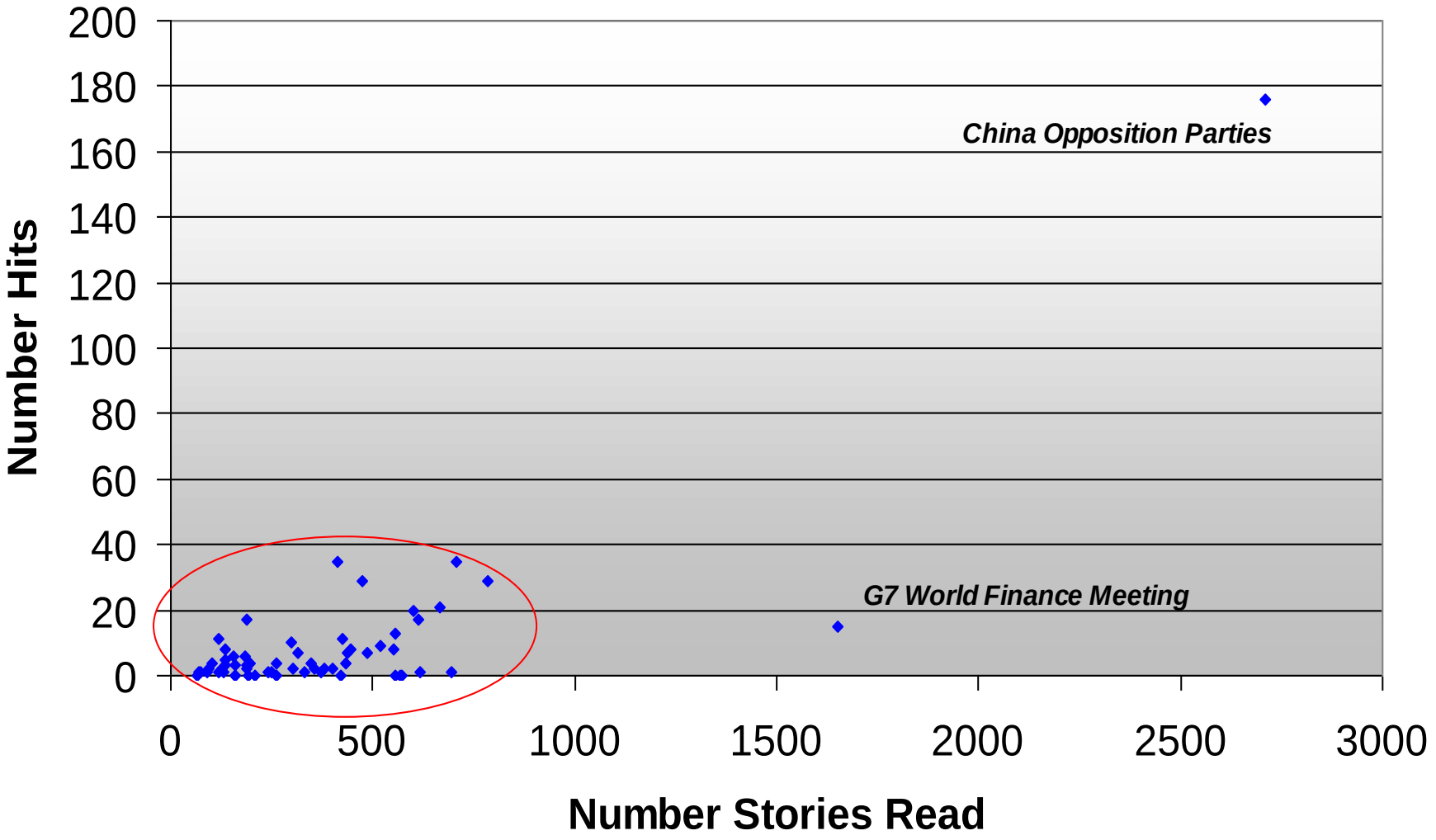
English Hits vs. Stories Read



English Hits vs. Stories Read



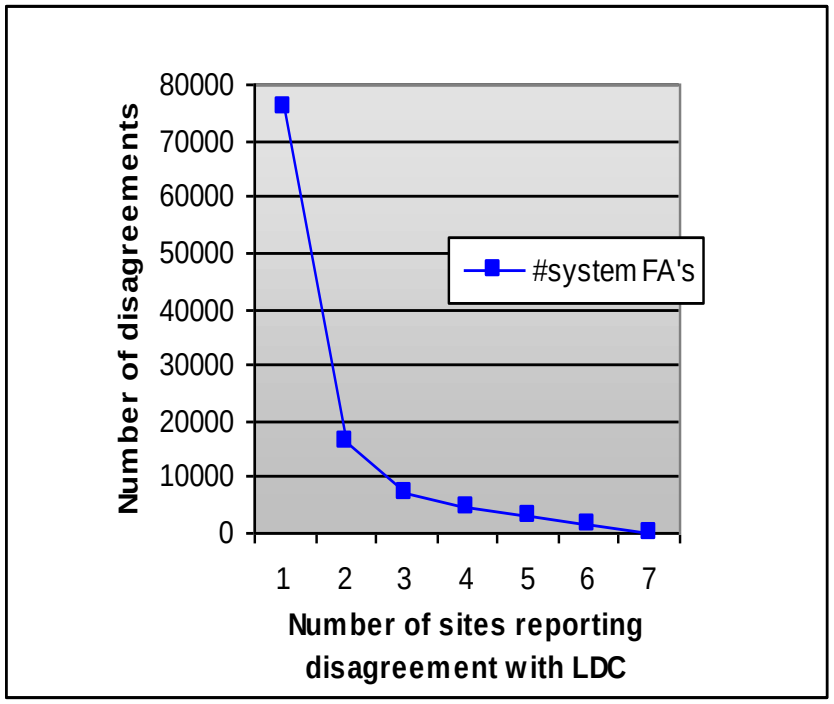
Mandarin Hits vs. Stories Read

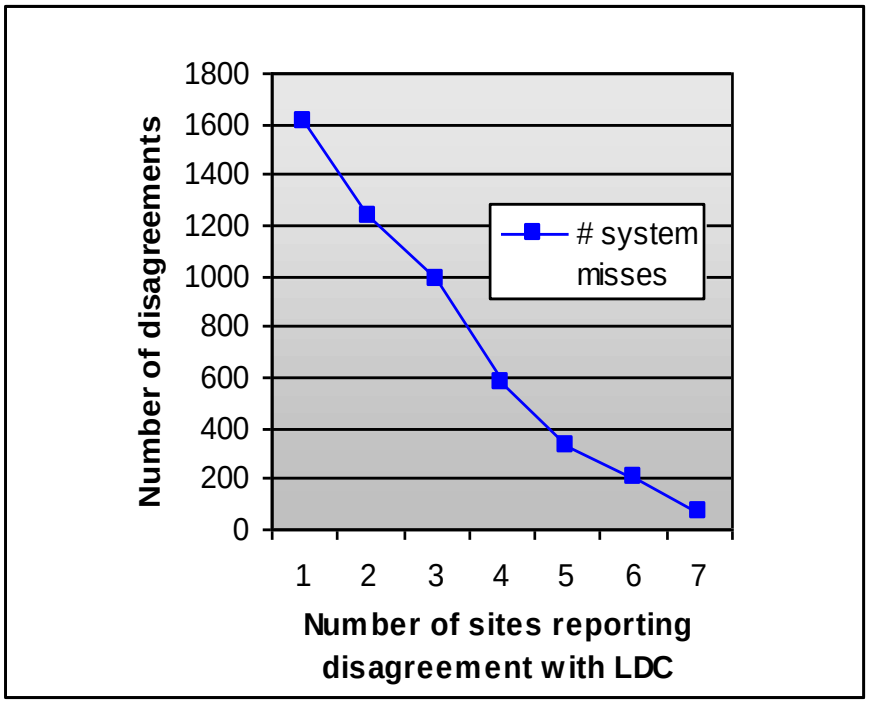


Topic-Story QC

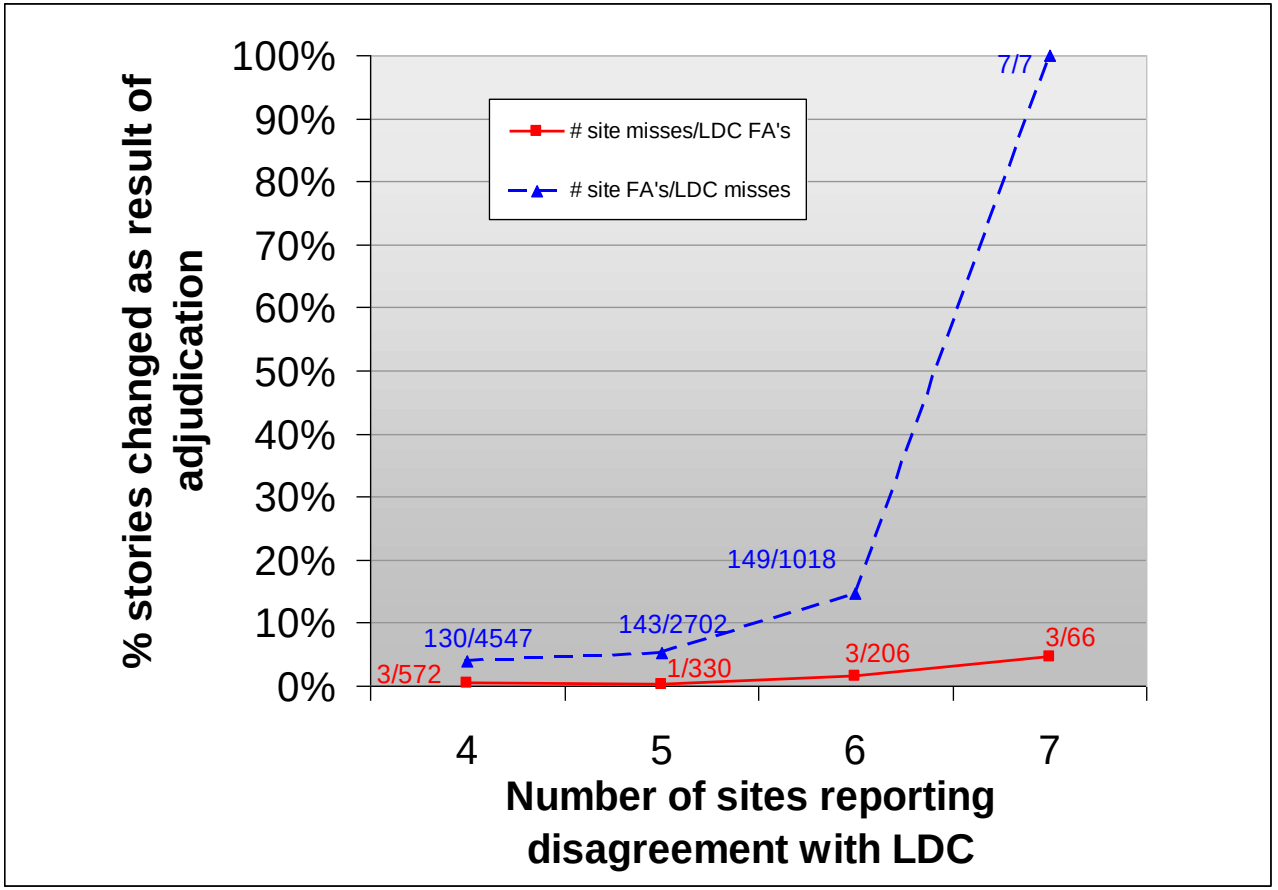
- Review rejects
  - **all rejection judgements reviewed and confirmed or vetoed**
  - **corrections made where possible and stories returned to pipeline or discarded**
- Dual Annotation & Discrepancy
  - **8% of Mandarin & English files receive 2 separate annotations**
  - **double-blind file assignment part of automated work distribution**
  - **inter-annotator consistency is good (compares favorably with TDT2 kappas)**
    - **Topic List 2 ~ kappa 0.8648106**
    - **Topic List 3 ~ kappa 0.777349**
    - **Topic List 4 ~ kappa 0.7248981**
- Precision
  - **all 'on topic' stories verified by senior annotators to identify false alarms**
  - **precision vetoed 2.5% of original judgments (213 of 8570 stories)**

- Adjudication of sites' hit lists from tracking task
  - **NIST delivered results containing ~1.5M topic-story tuples from 7 sites**
  - **LDC reviewed cases where a majority of systems (i.e. 4 or more) disagreed with original annotation**

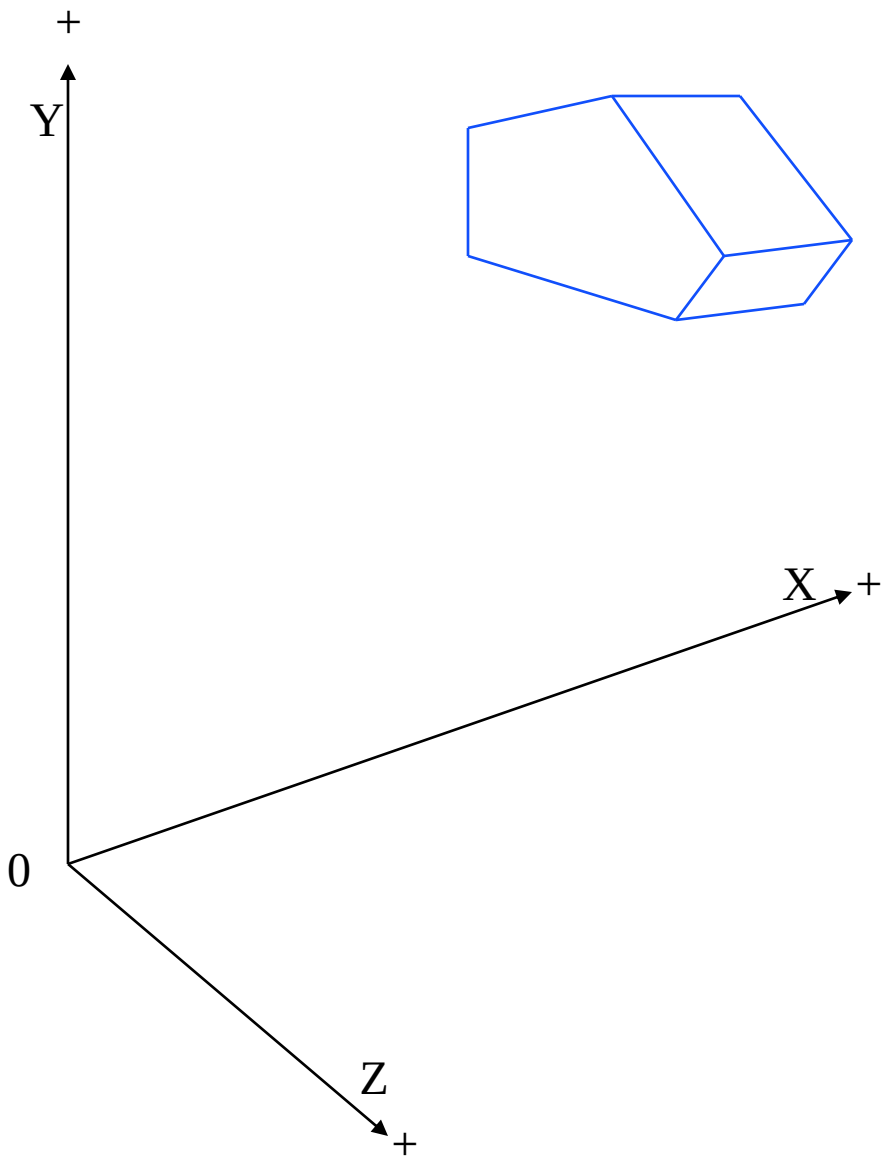




- Adjudication results
  - **rate of system miss leading to LDC false alarm very low (complete precision QC)**
  - **rate of system FA leading to LDC miss somewhat higher but still quite low (no recall on test set)**



Quality's Multiple Dimensions



Preliminary Conclusions

- **Quality is multidimensional**
- **Quality defined or evaluated with respect to needs**
- **Trade-offs with volume, cost, richness, appropriateness, timeliness, etc**