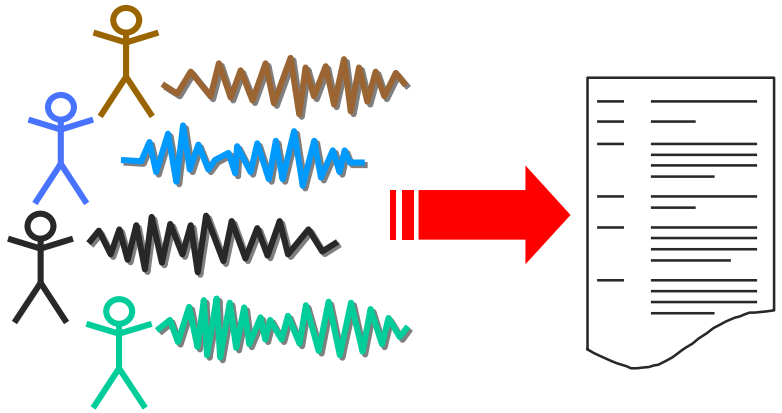
2004 Fall Rich Transcription Speech-to-Text Evaluation



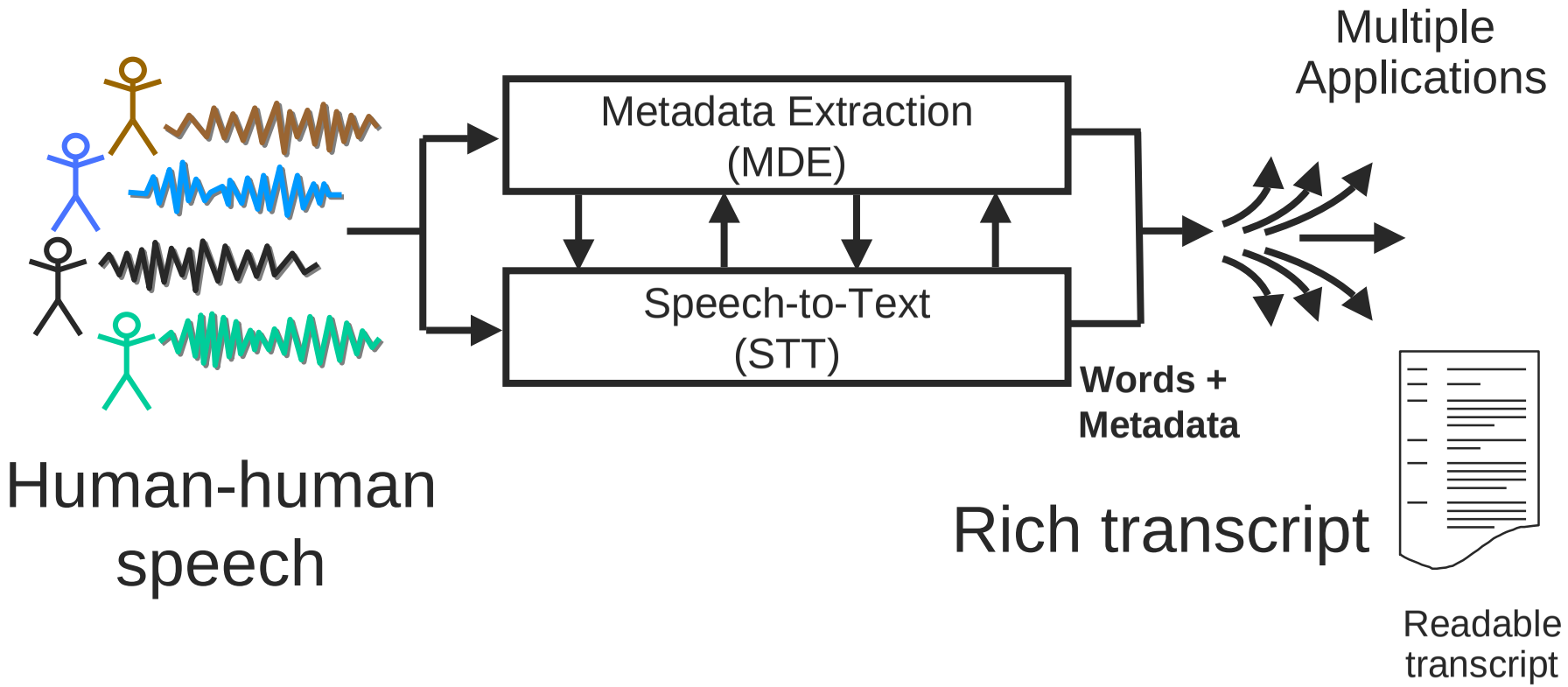
Audrey Le for the NIST Gang

RT-04F Workshop
November 7-10, 2004
Palisades, NY

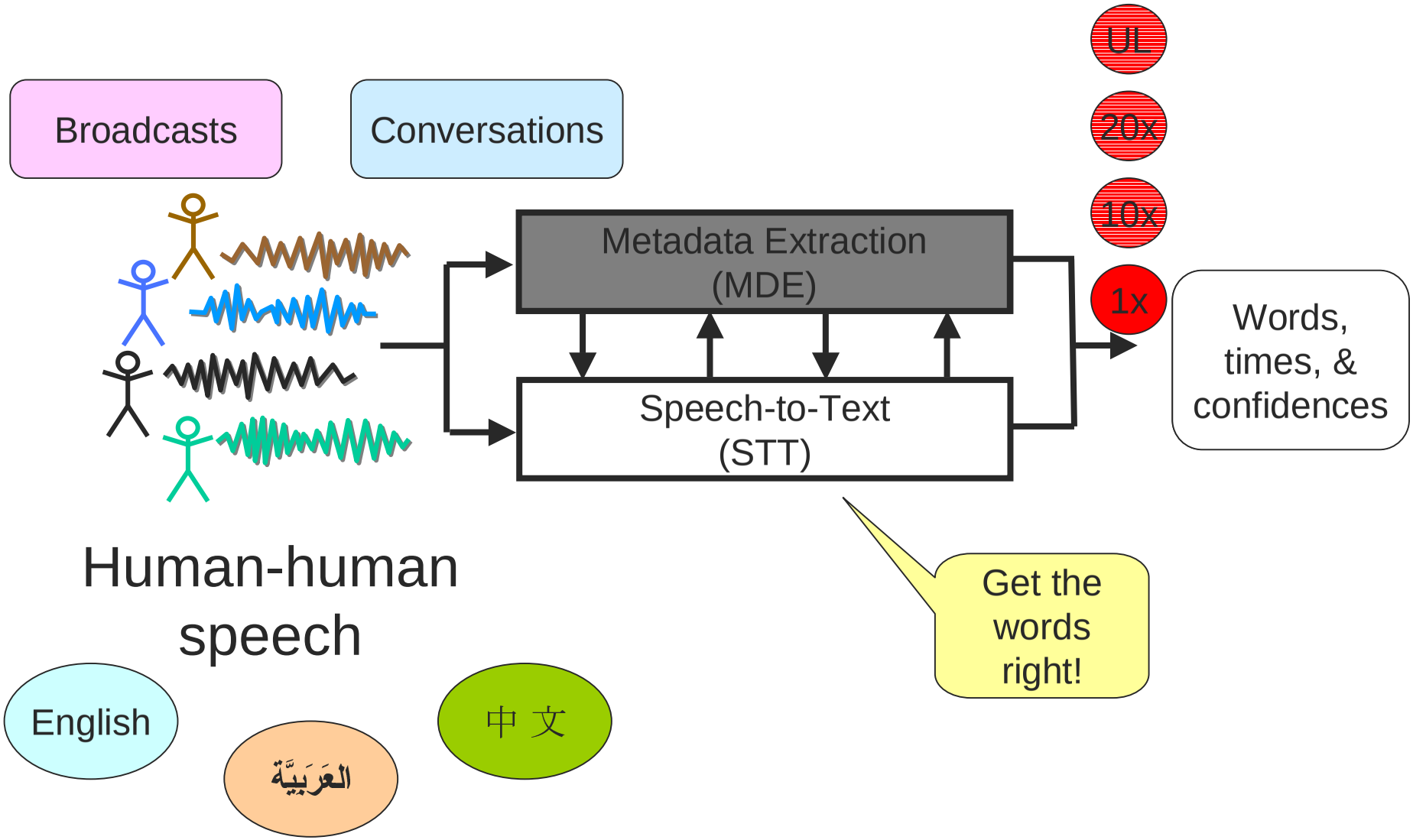
Speech-to-Text (STT) Talk Outline

- STT Task
- Test Sets
- Scoring Procedures
- STT Participants
- Evaluation Results

Rich Transcription



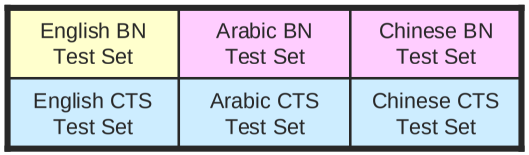
Speech-to-Text Task



Speech-to-Text Test Sets

English Broadcast News Test Set	Arabic Broadcast News Test Set	Chinese Broadcast News Test Set
English Conversational Telephone Speech Test Set	Arabic Conversational Telephone Speech Test Set	Chinese Conversational Telephone Speech Test Set

English Broadcast News Test Set



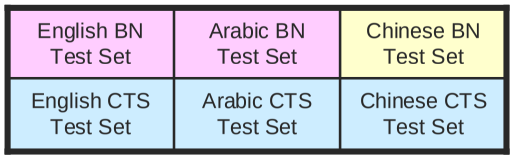
- English EARS 2003 Collection
- 12 shows: 30 minute excerpts
  - Consist of network news and local news shows
  - Broadcast from the US for a US audience
  - Cover international news, domestic and local US news

Arabic Broadcast News Test Set

English BN Test Set	Arabic BN Test Set	Chinese BN Test Set
English CTS Test Set	Arabic CTS Test Set	Chinese CTS Test Set

- Arabic EARS 2003 Collection
- 3 shows: 20 minute excerpts
  - Dubai TV, Al-Jazeera (2 shows, different dates)
    - Target Middle East and world-wide audience
    - Cover international and Middle East related news
- Main dialect - Modern Standard Arabic
- Format similar to US network news

Chinese Broadcast News Test Set



- Mandarin EARS 2004 Collection
- 3 shows: 20 minute excerpts
  - China Central Television (CCTV-4) *Financial and Economic Reports*
    - Broadcasts from China for overseas Chinese
    - Covers international and Chinese financial and economic news
  - Radio Free Asia (RFA) *Asia-Pacific Headline News* and *Asia-Pacific Reports*
    - Broadcasts from the US for Chinese who live in China
    - Covers news about China and China-related international news
  - New Tang Dynasty TV (NTDTV) *Hourly News*
    - Broadcasts from the US for a Chinese audience
    - Covers international news
- Main dialect - mainland China standard Mandarin
- Format similar to US network news

English Conversational Telephone Speech Test Set

English BN Test Set	Arabic BN Test Set	Chinese BN Test Set
English CTS Test Set	Arabic CTS Test Set	Chinese CTS Test Set

- English Fisher Collection
- 36 conversations: 72 speakers – 5 minute excerpts
  - Equal number of male and female speakers
  - All speakers are native English speakers living in the US
    - Speakers came from four US dialect regions
      - North, South, Midland, West
  - Numbers of standard, cordless, and cellular phones used are ~ equally distributed
  - All conversations were recorded in the US
  - Topics from current events and social issues

Arabic Conversational Telephone Speech Test Set

English BN Test Set	Arabic BN Test Set	Chinese BN Test Set
English CTS Test Set	Arabic CTS Test Set	Chinese CTS Test Set

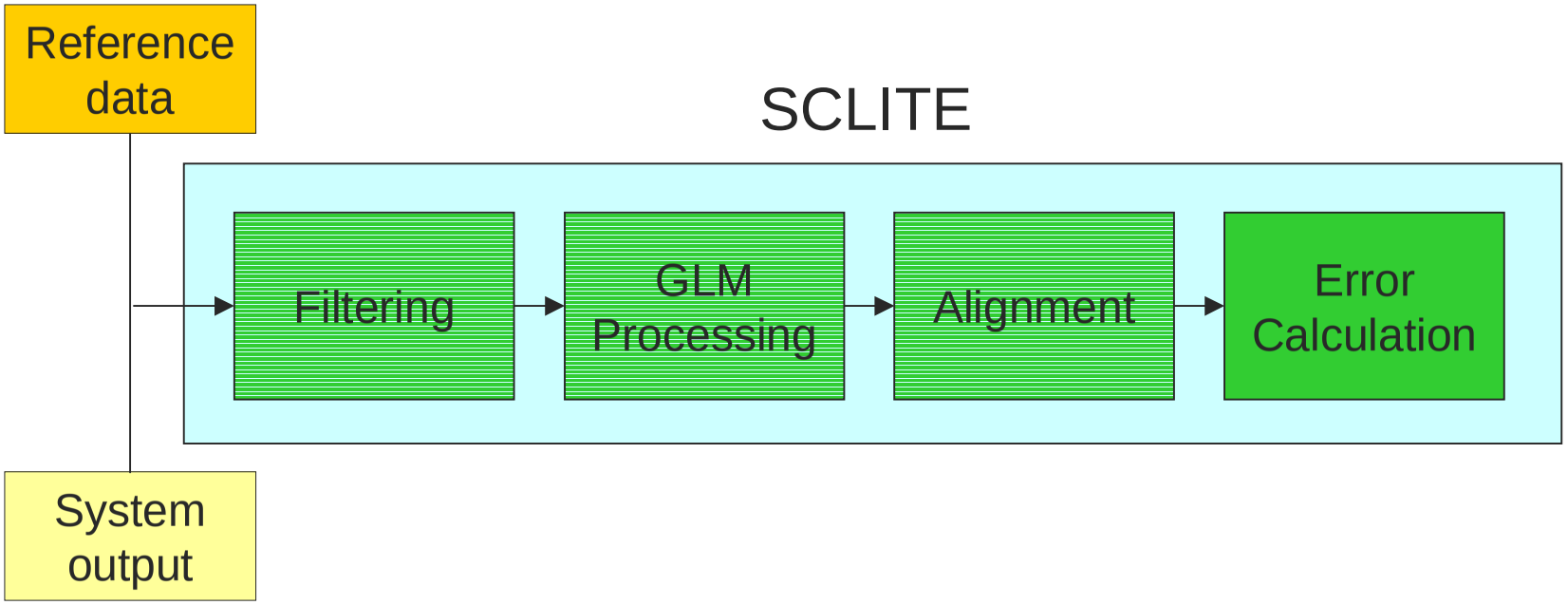
- Levantine Dialect Fisher Collection
- 12 conversations: 24 speakers – 5 minute excerpts
  - ~ Equal number of male and female speakers
  - All speakers are native Arabic speakers with majority living in the Middle East and a few in the US
    - Majority of the speakers was raised in Jordan
  - Majority of the phones used was landlines
  - All conversations were recorded in the US
  - Topics from current events and social issues

Chinese Conversational Telephone Speech Test Set

English BN Test Set	Arabic BN Test Set	Chinese BN Test Set
English CTS Test Set	Arabic CTS Test Set	Chinese CTS Test Set

- Hong Kong Univ. of Science & Tech. Collection
- 12 conversations: 24 speakers – 5 minute excerpts
  - ~ Equal number of male and female speakers
  - All speakers are native Mandarin speakers and live in China
    - All speakers are in their twenties
  - All phones used were landlines
  - All conversations were recorded in China
  - Topics from current events and social issues

Scoring Procedures



$$\text{WER} = \frac{(\#\text{ Deletions} + \#\text{ Insertions} + \#\text{ Substitutions})}{\#\text{ Reference Words}}$$

Scoring Procedures: Variations for Arabic

- Transcription of colloquial Arabic is a challenge
  - Modern Standard Arabic (MSA) orthographic conventions do not easily apply to dialectal Arabic speech
  - New transcription conventions and methods were devised by LDC with community input
- New Arabic text normalization steps
  - Word-initial alif+hamza normalization
    - **Rule**: All word-initial alif+hamza characters are translated to bare alif
    - The word-initial “alif” can be written omitting a “hamza” or with any of three vocalic variants
    - Neither form changes the meaning and transcribers rarely agree
    - Unresolved Issue: normalization was not applied to inflected forms
  - Tanween character removal
    - **Rule**: All tanween characters removed
    - The three vowel diacritics, fathah tanween, kasrah tanween, and dhammah tanween are inconsistently transcribed
    - Omitting or including the vowel does not affect the word’s meaning

Scoring Procedures: Variations for Chinese

- Pause filler characters
  - List of pause filler characters expanded to 5 as an allowance for CTS training data
  - Sometimes they occur as non-pause filler characters in broadcast news
- No lexical normalizations
- Character Error Rate (CER)
  - No “word” analog in the native orthography

Reference STT Words

- Ordinary words
  - Scored as is without case
- Speaker-attributed noises (e.g., cough, laugh)
  - Stripped out
- Non-vocal noises (e.g., door slam)
  - Stripped out
- Word fragments
  - Counted as optionally deletable
  - Counted as correct if substring of system output matches
- Uncertain or foreign words
  - Counted as optionally deletable
- Pause-fillers
  - Counted as optionally deletable
- Contractions
  - Expanded to the most likely correct form

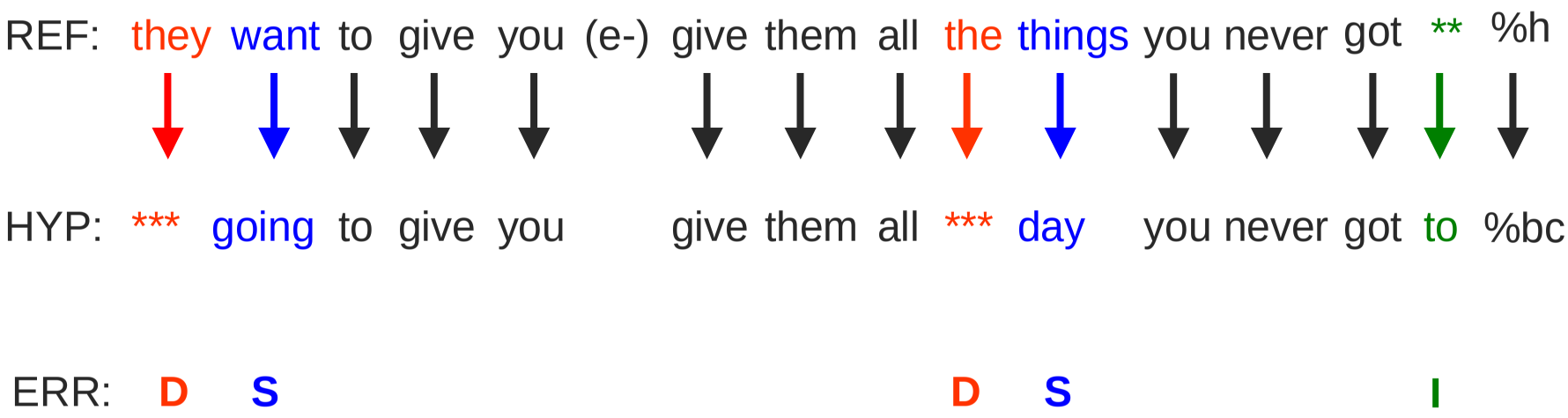
System Output STT Words

- Ordinary words
  - Scored as is without case
- Speaker-attributed noises (e.g., cough, laugh)
  - Stripped out
- Non-vocal noises (e.g., door slam)
  - Stripped out
- Word fragments
  - Stripped out
- Uncertain or foreign words (as marked by the system)
  - Stripped out
- Pause-fillers
  - Stripped out if identified as pause-fillers
  - Transformed to a generic pause-fillers if not identified as such

Global Mapping (GLM) Processing

- Allowed equivalent system output to be scored as correct
- Contained equivalence rules to transform data to a canonical form
  - Rules to expand contractions (only applied to the system output)
    - she's => she is, she has, she was
  - Rules to split hyphenated words
    - processing-speed task => processing speed task
  - Rules to split compounds with ambiguous representations
    - servicemen => service men
    - halftime => would not be split
  - Rules to map backchannels and hesitations
    - uh-huh => %BCACK
  - Rules to allow legitimate alternative spellings
    - Mr. => Mister

Alignment Example



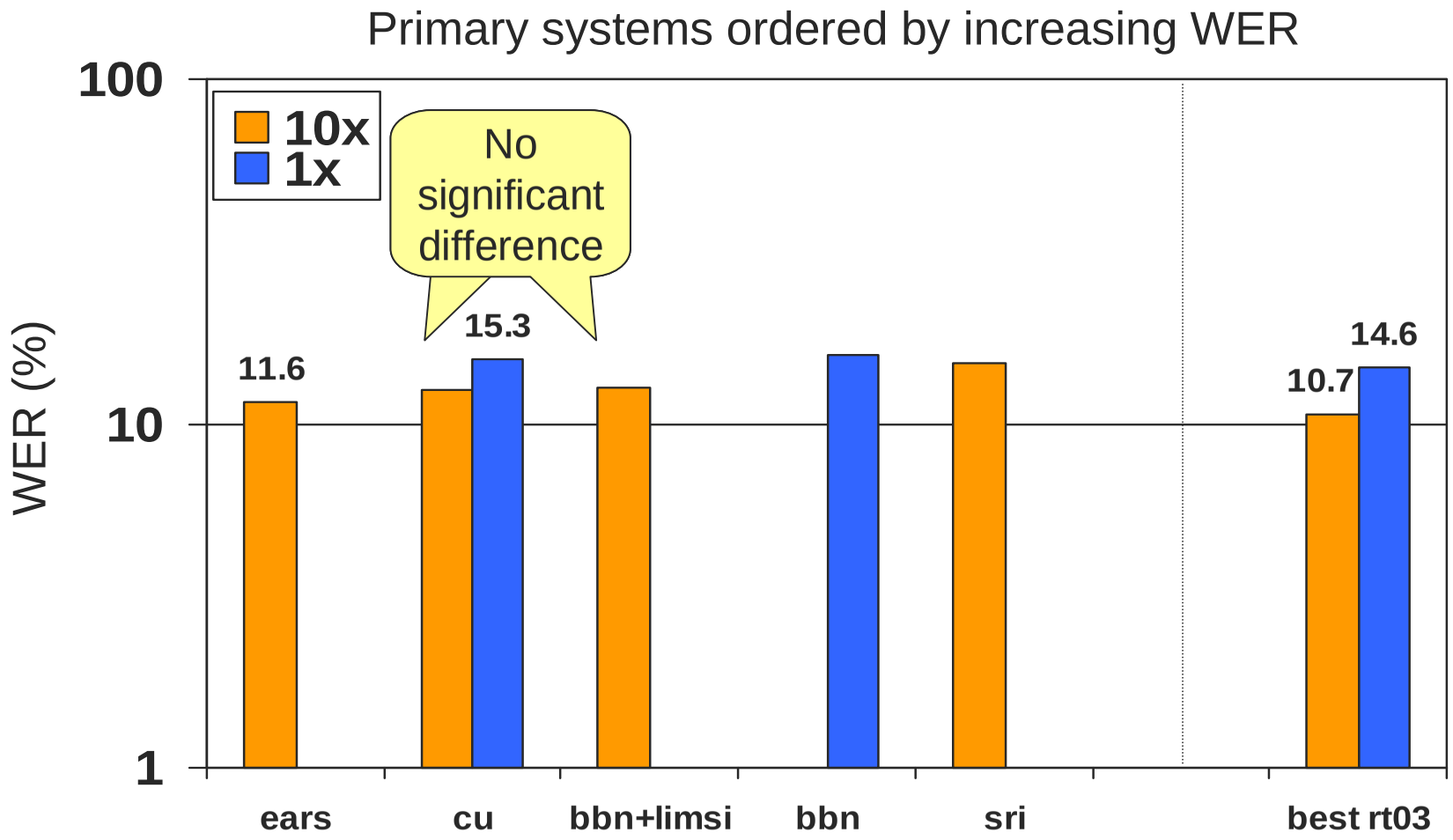
Summary:

two deletions, two substitutions, one insertion

STT Participants

- BBN
- BBN+LIMSI
- Cambridge Univ.
- IBM
- IBM+SRI
- ISL
- LIMSI
- SRI
- SRI + Univ. of Washington
- Super EARS
  (Collaborative effort involving BBN+CU+LIMSI+SRI)

English BN STT Results



Why Twelve Shows in the English BN Test Set?

- Cambridge Univ. observed that the LDC-provided RT-04 dev set had a higher WER than RT-03 eval set
  - RT-04 dev set is different from RT-03 eval set
- NIST decided to enlarge the RT-04 eval set to provide
  - Comparable subset of RT-04 with RT-03
  - Better sampling of broadcast news shows

English BN Test Set Selections

- LDC selections
  - News From CNN (anchor: Wolf Blitzer)
  - PBS News Hour (anchor: Gwen Ifill)
  - CNN Headline News (anchors: Steven Fraser, Sophia Choi)
  - CNBC The News (anchor: Tom Costello)
  - ABC6 WPVI Action News (12/17)
  - CSPAN Book TV: Texas Book Festival
- NIST selections
  - CNN Daybreak (anchor: Carol Costello)
  - CNBC The News (anchor: Brian Williams)
  - ABC World News Tonight (anchor: Peter Jennings)
  - PBS BBC News (anchor: Alistair Yates)
  - ABC6 WPVI Action News (12/03)
  - WB17 News WPHL

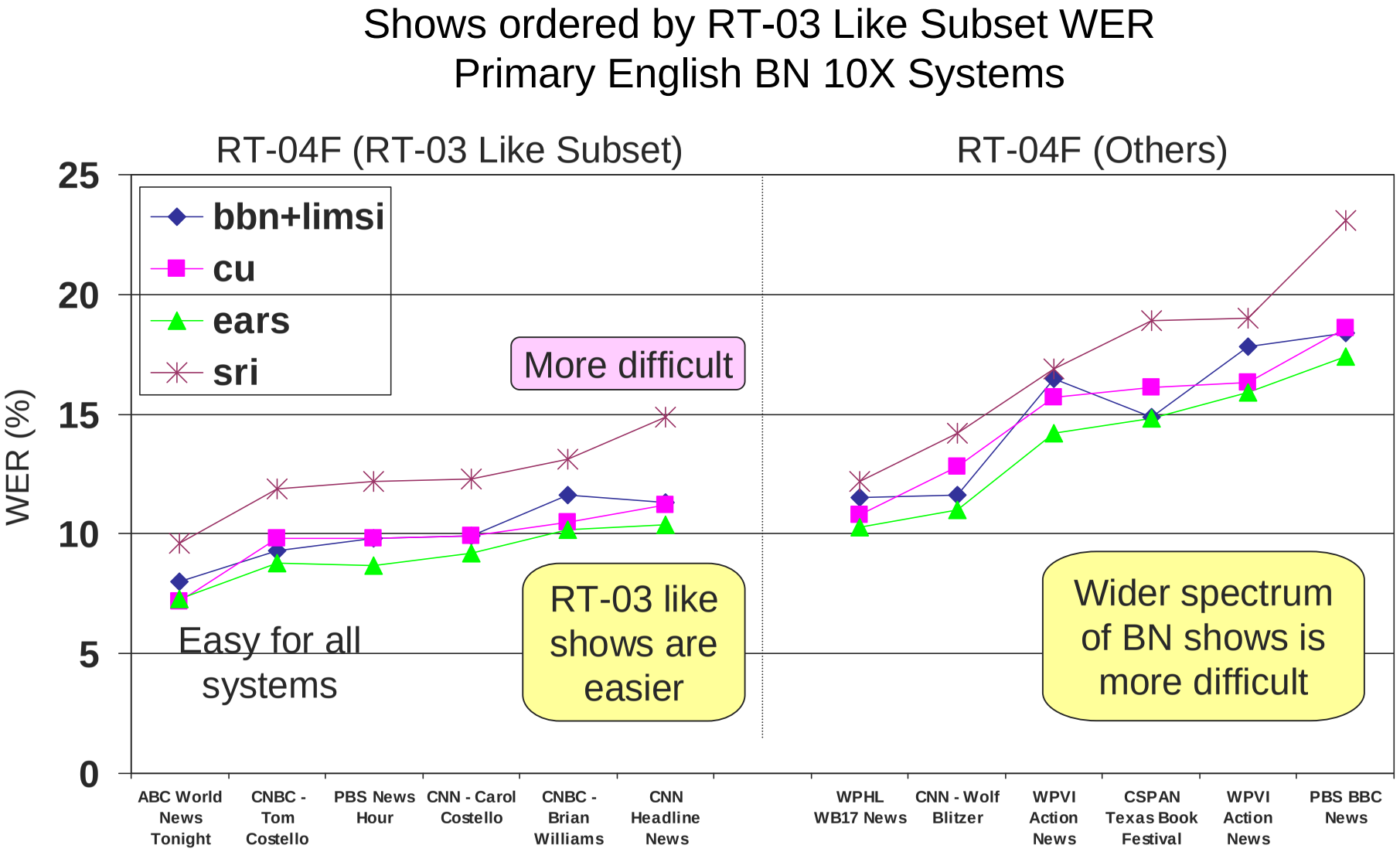
Comparability with RT-03

RT-03 Like Subset of RT-04 Shows	Characteristics	RT-03 Shows
ABC World News Tonight	Same show, different dates	ABC World News Tonight
CNBC The News (anchor: Tom Costello)	Typical network news with both having male speakers speaking most of the time	NBC Nightly News (anchor: Tom Brokaw)
PBS News Hour	Similar types of news content and speaking style	VOA News Now
CNN Daybreak	Similar anchor styles	PRI The World
CNBC The News (anchor: Brian Williams)	Both anchored by Brian Williams	MSNBC News (anchor: Brian Williams)
CNN Headline News	Similar types of news contents, different dates	CNN Headline News

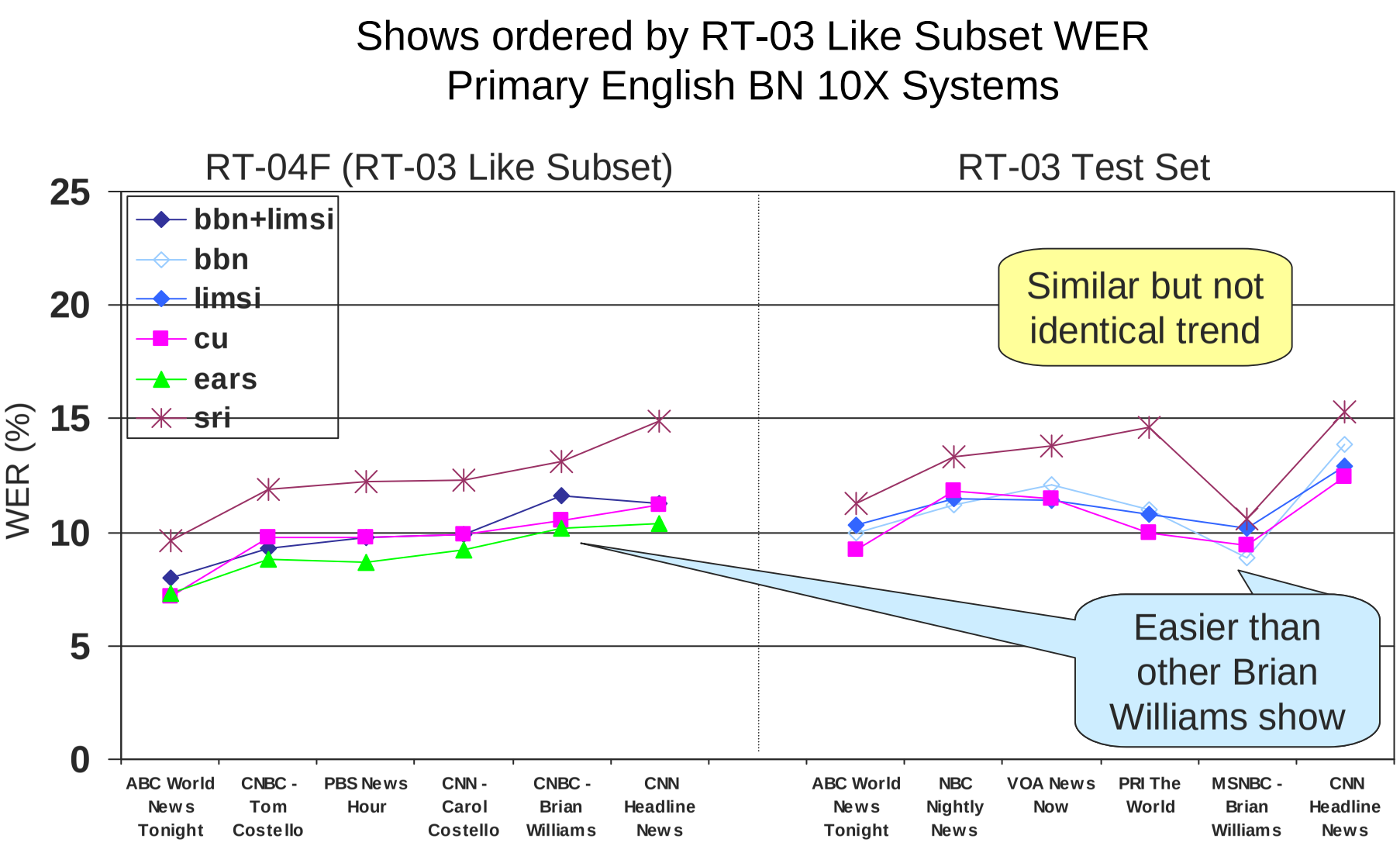
Reusability Requires Robustness

- Hypothesis: A robust system should have little variation in WER for different sources
- Observation: Dramatic sensitivity to test sets / subsets
- Implication: Current technology is not adequately robust?

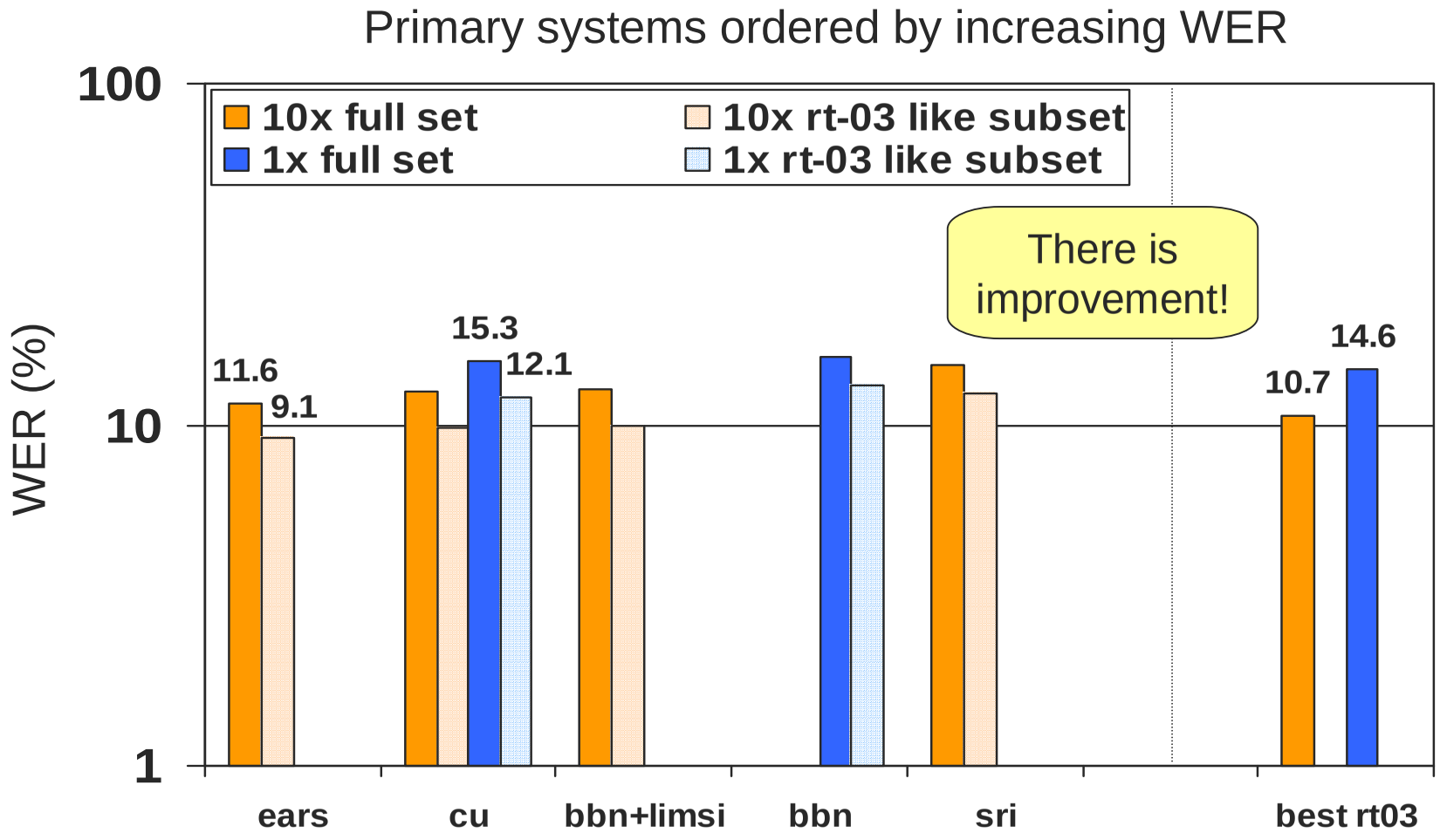
WER for Different English BN Shows



Comparison with RT-03 English BN Shows

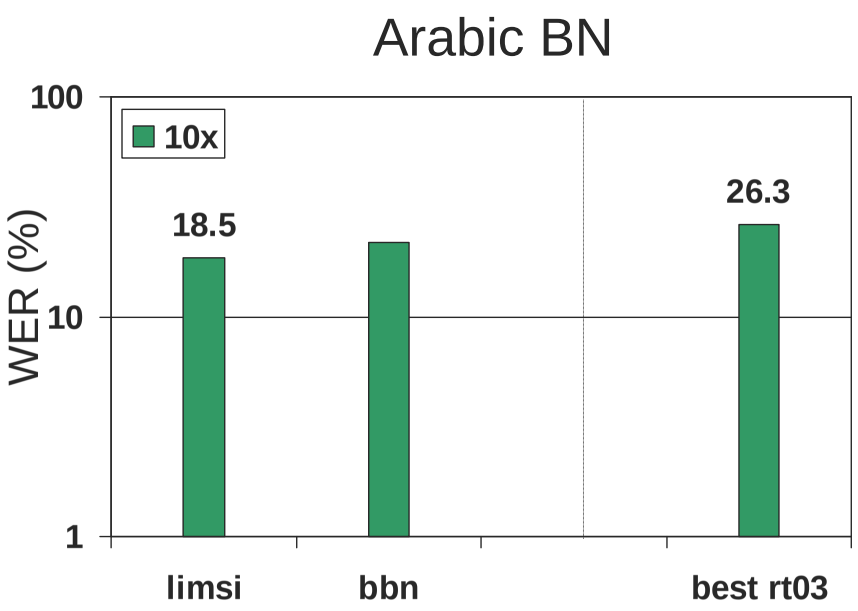


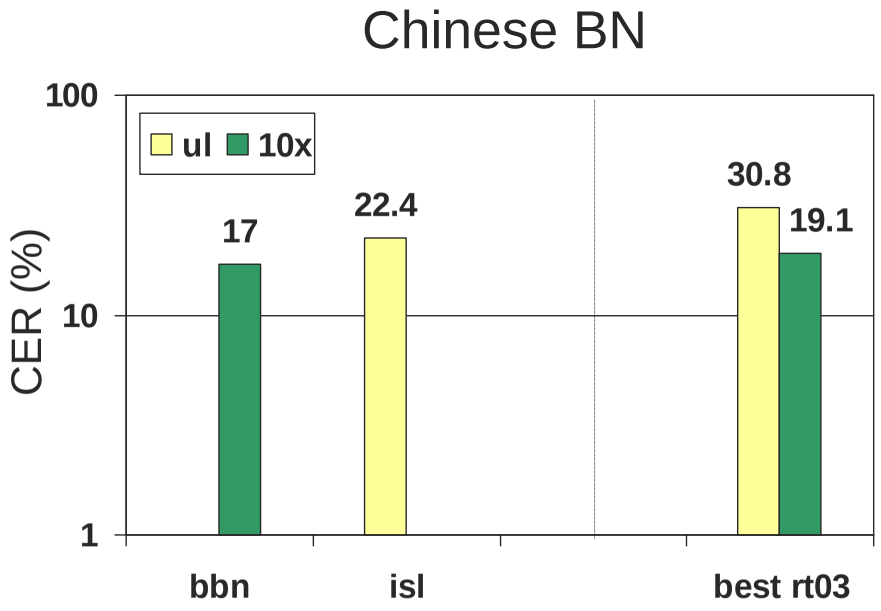
English BN STT Results - Again



Non-English BN STT Results

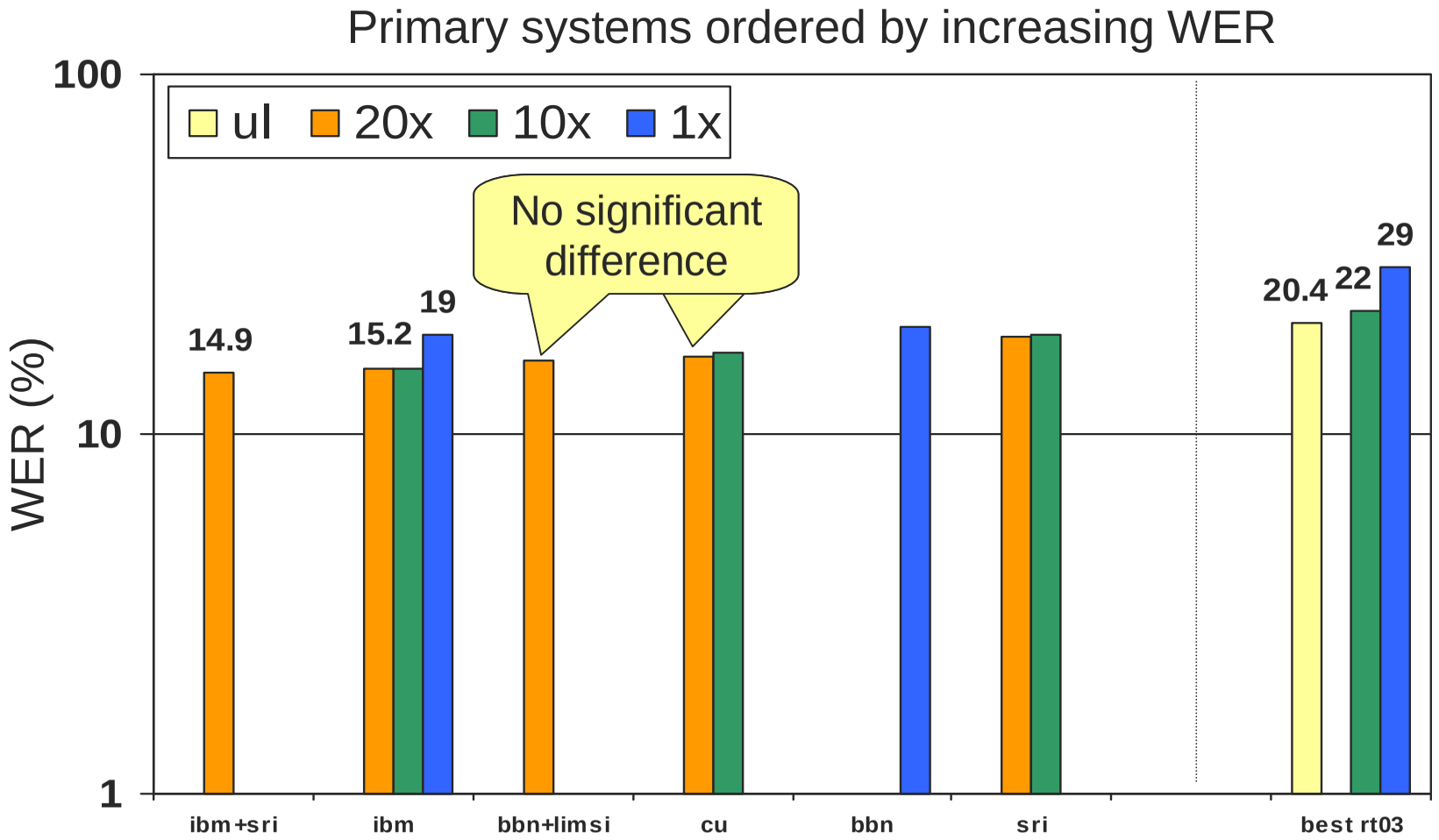
Primary systems ordered by increasing WER



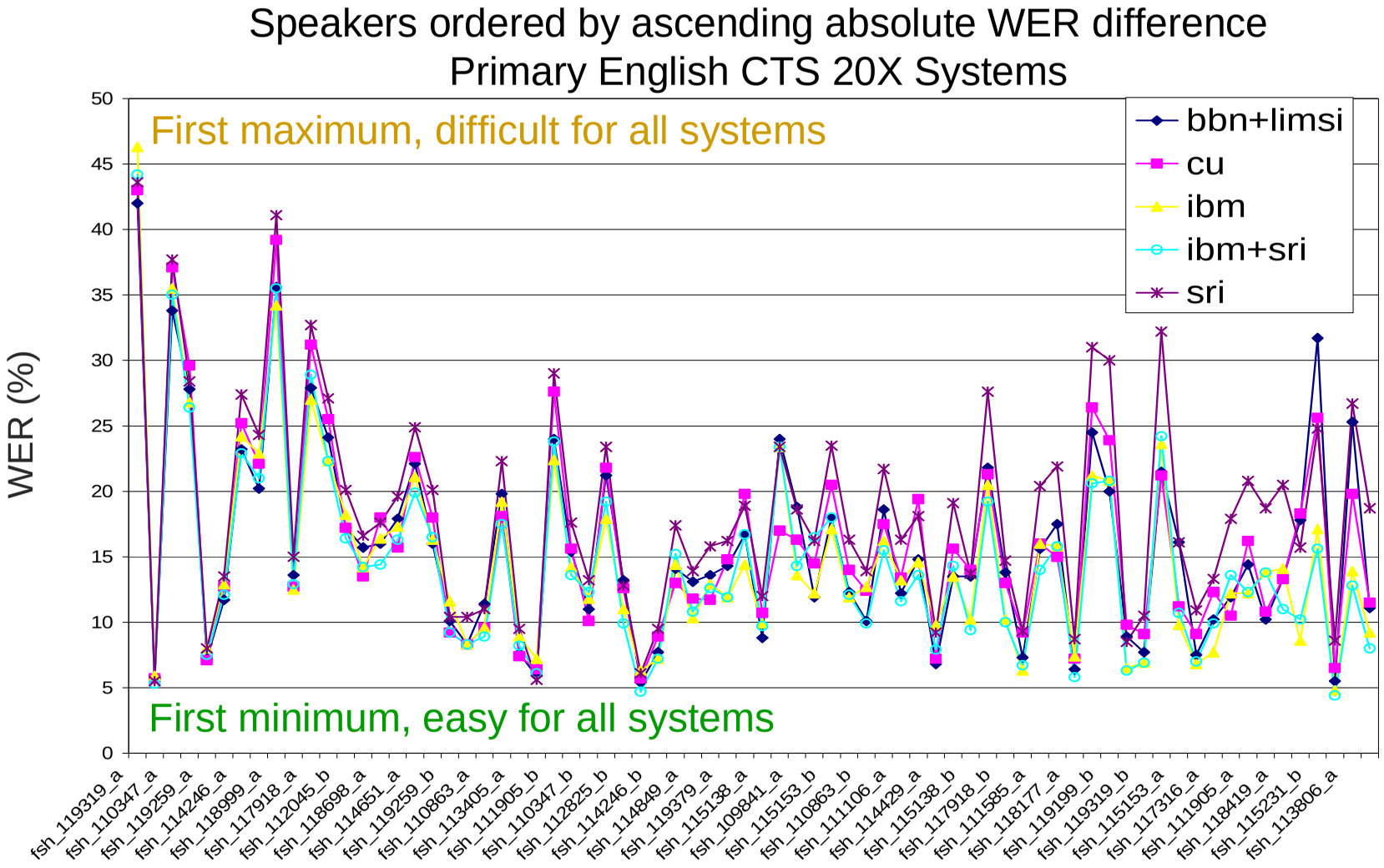


- All differences at a given speed are statistically significant

English CTS STT Results



WER for Different CTS Speakers

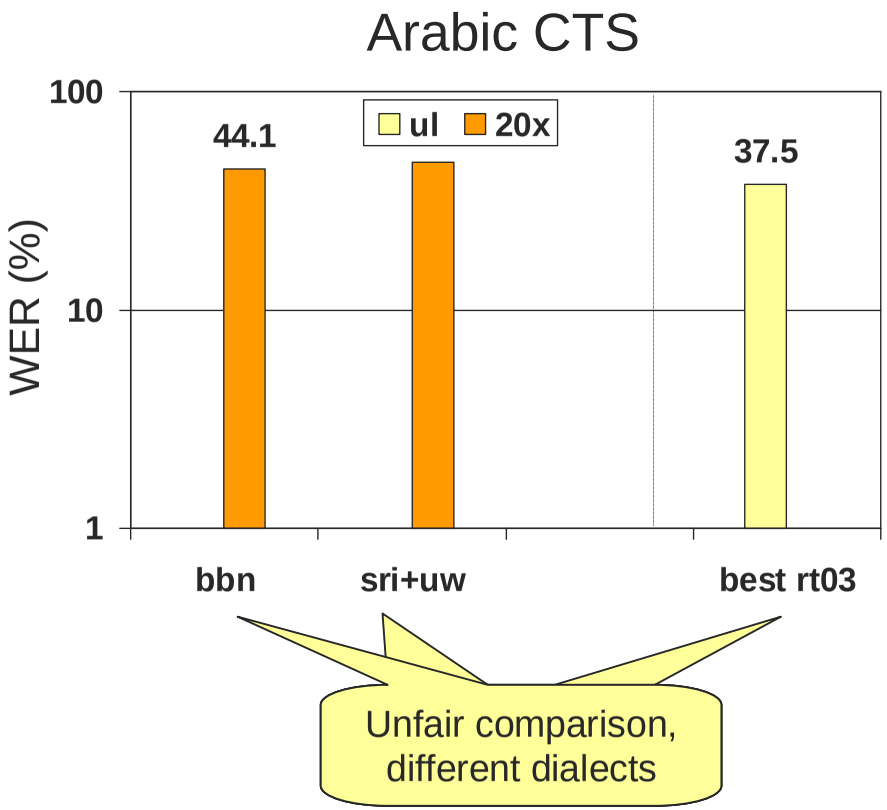


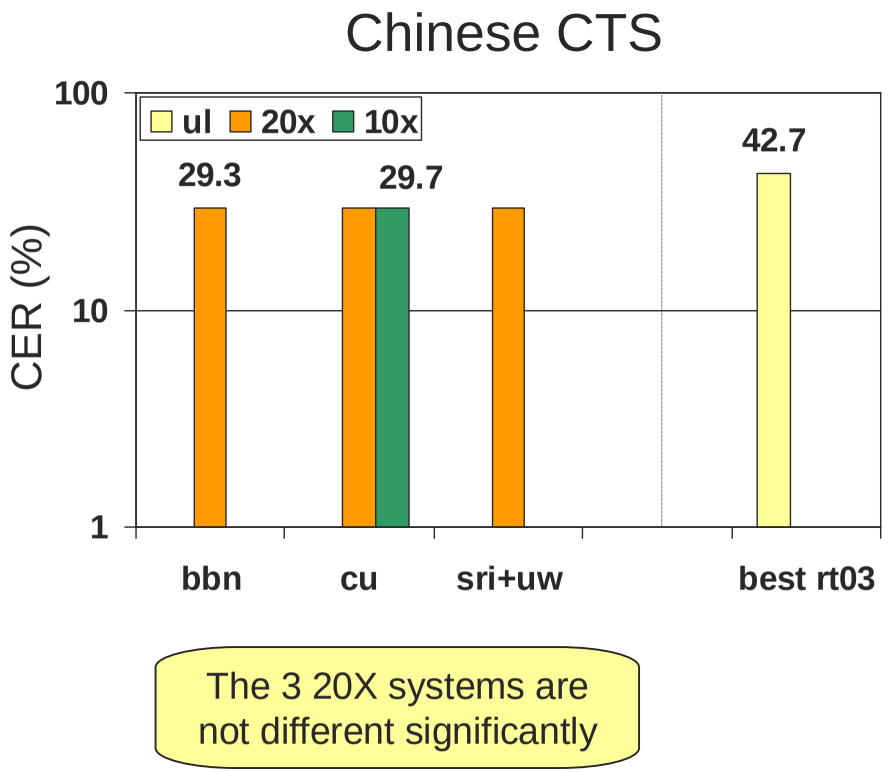
Different system outputs

- Systems don't do well on rapid and disfluent speech

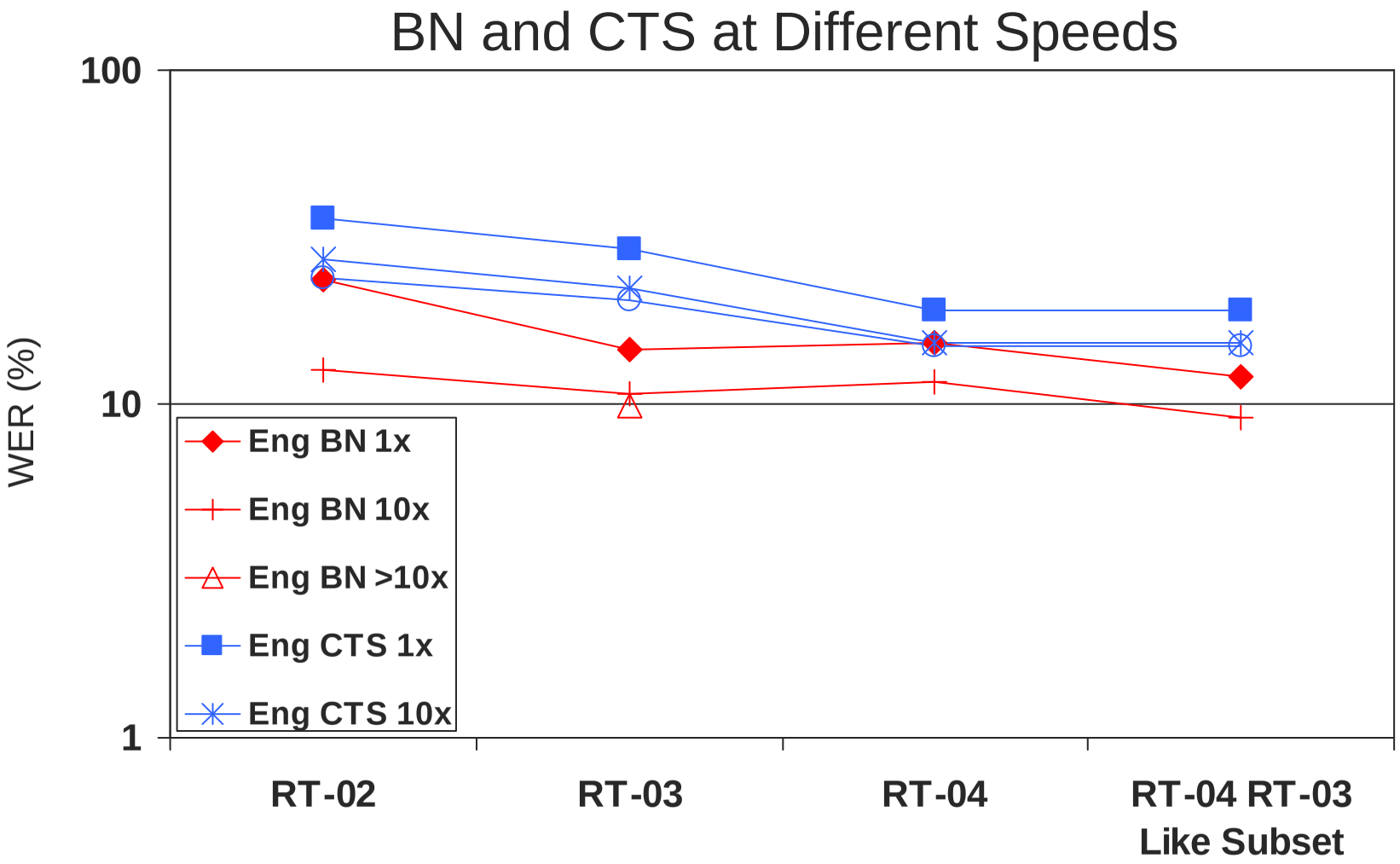
Non-English CTS STT Results

Primary systems ordered by increasing WER





Conclusion:
EARS RT Progress – English



Thank You

- Phil Woodland
- The LDC
- Our Arabic tutors
  - Mohamed Maamouri
  - John Makhoul
- The rest of the NIST Gang!!!