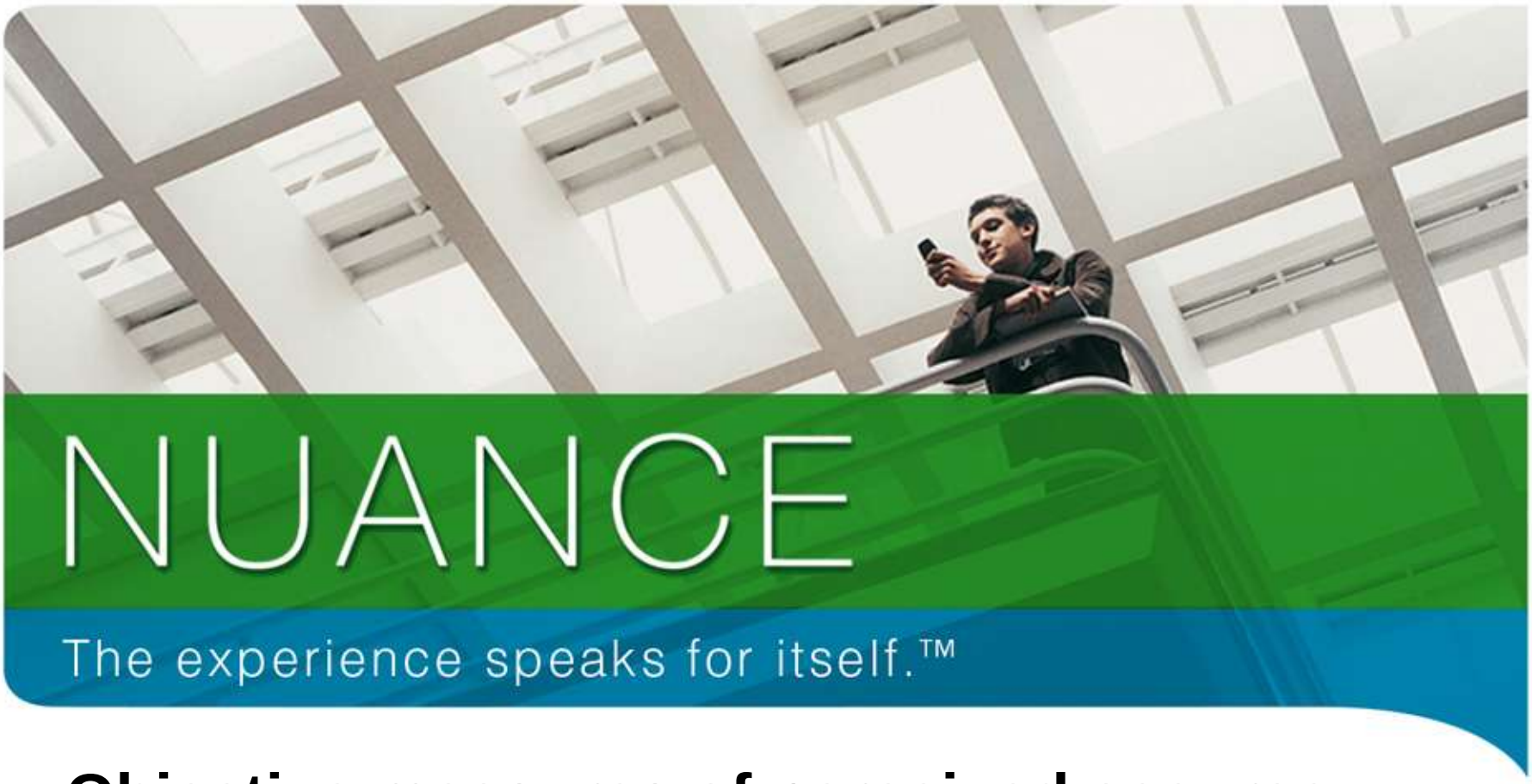


Objective measures of perceived accuracy

work by **David Thomson**

presentation and color commentary by **George Zavaliagkos**

Background

- Domain: Voicemail transcription
- The Question: what affects user perceived accuracy
- Methodology
  - Two studies try to quantify perceived accuracy and correlate it with measurable quantities
- Objective:
  - Create a cost function for use in optimizing recognition performance in areas that matter most to users.

Study #1: Concept Survival Rate

- Small but intensively scrutinized annotation of concepts

- 4050 total ratings
  - 162 messages
  - 5 recognizers
  - 5 judges

Example message

Goldfeder ,

- It is Catucci calling you back .
- Checking in .
- Give me a holler .
- I'm at the office 9082343758 .
- It is Friday about 10:00 .

Hope all is well .

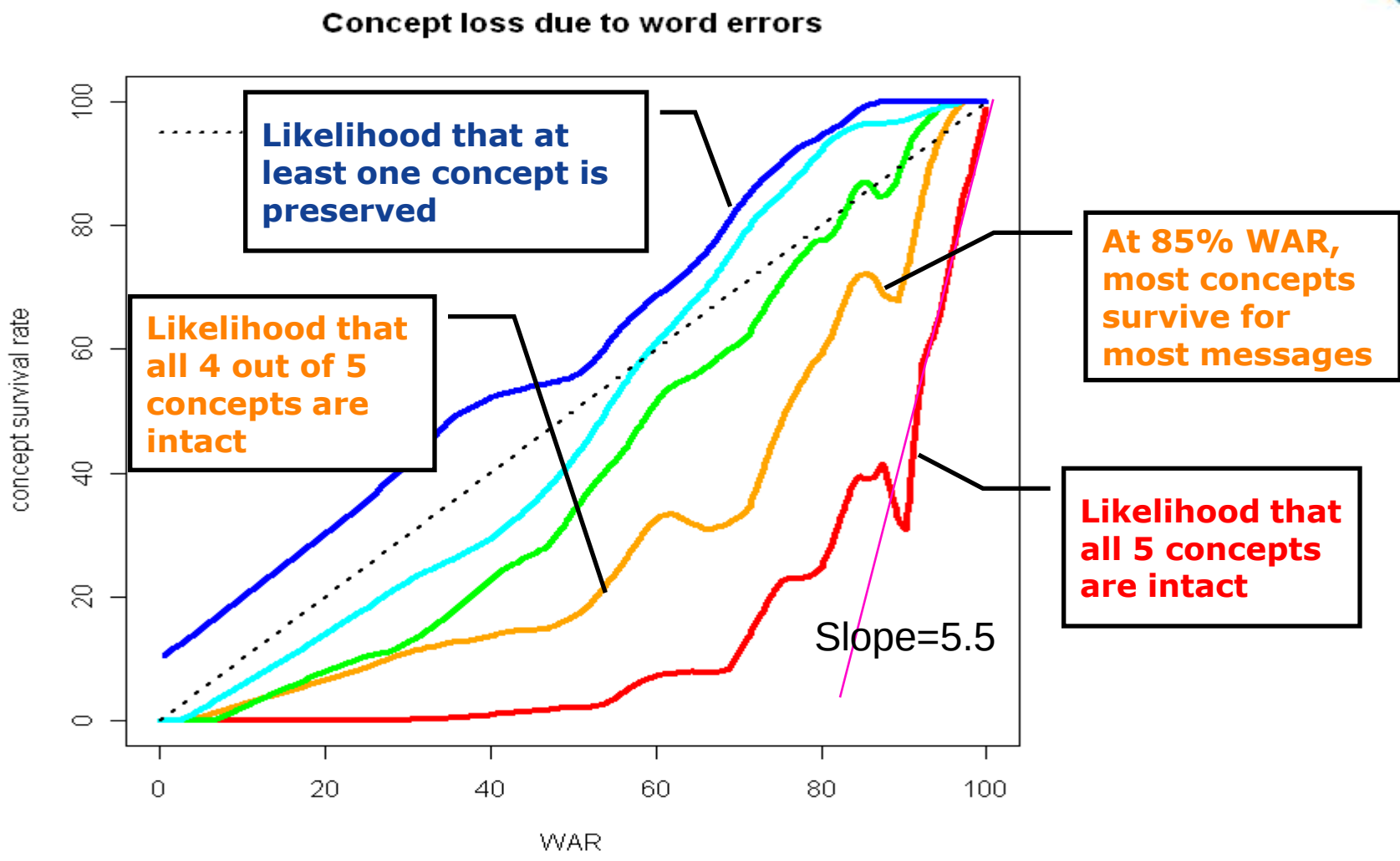
Bye Bye .

Concept marking – judge's score sheet

Original message						Recognizer #1 89% WAR 5/5 concepts						Recognizer #2 78% WAR 4/5 concepts						Recognizer #3 67% WAR 2/5 concepts
Goldfeder . (3. It is Catucci calling you back) . (2. Checking in) . (1. Give me a holler) . I'm (4. at the office 9082343758) . (5. It is Friday about 10:00) . Hope all is well . Bye Bye .	1	2	3	4	5	Gold Fedder . It is Catochi calling you back . Checking in . Give me a holler . I'm at the office 9082343758 . It is Friday about 10:00 . Hope all is well . Bye Bye .	1	2	x	4	5	cole better get these kids to keep calling you back checking in give me a holler i'm at the office 9082343758 it is friday about 10:00 hope all's well bye bye	1	2	x	x	x	go frederick it is to to g calling you back checking in give me a holler i'm at the office 90823437580 this for day about 10 o'clock hope all is well bye bye

Concept survival ratings

Individual concept survival



Perceived Accuracy is a function of expectations

If my expectations is

Perfect transcription

Actionable transcription

Satisfy Curiosity
without getting annoyed

Then

Most messages must achieve 95% WAR

Most messages must achieve 85% WAR

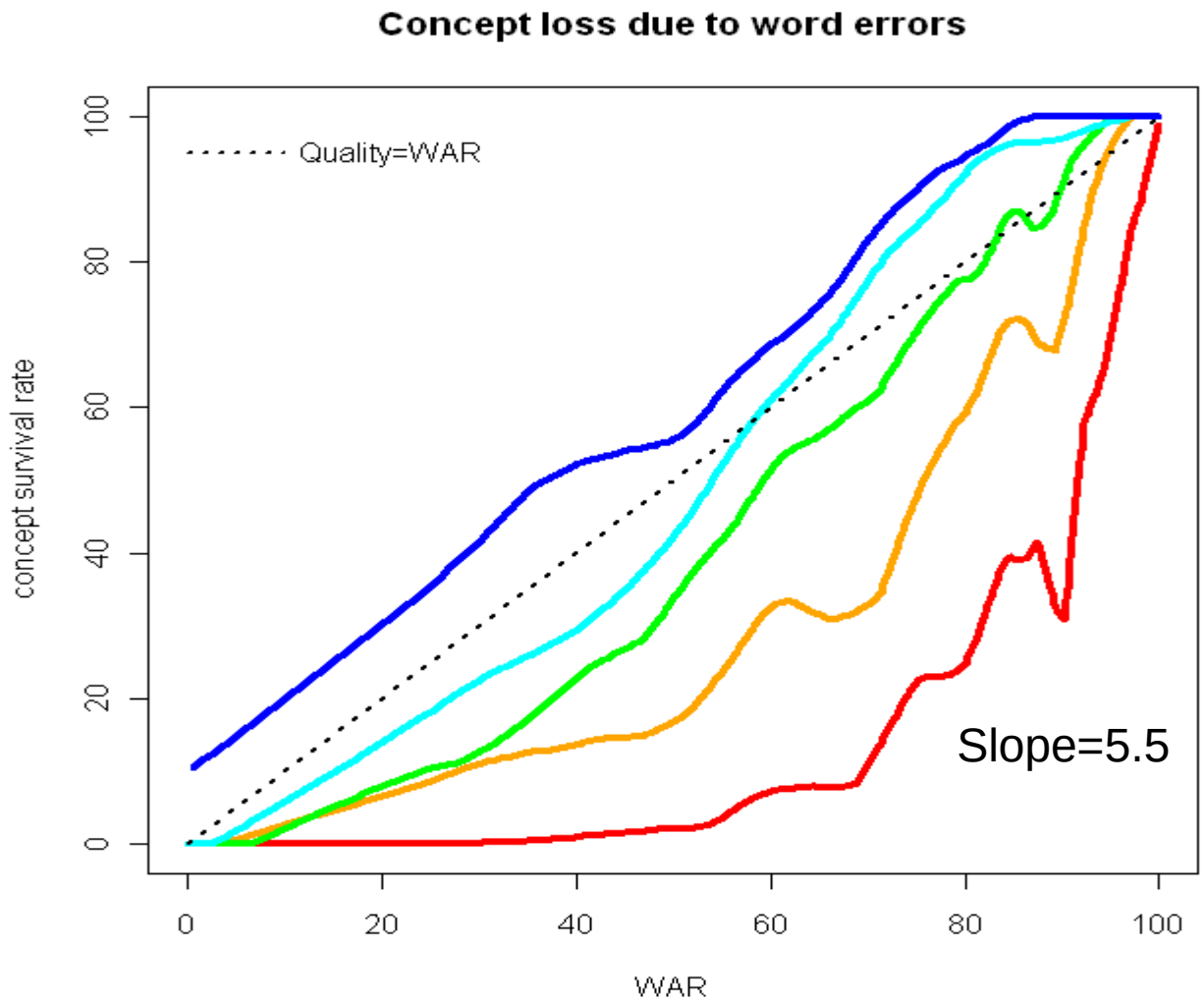
Most messages must achieve 60% WAR

???

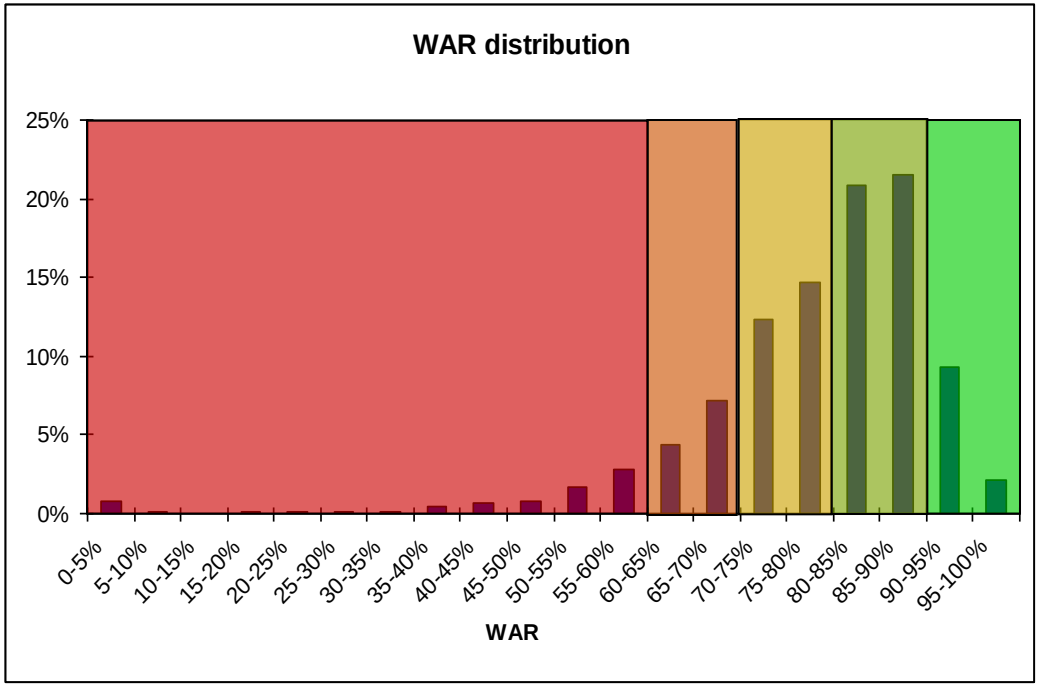
Average WAR is a correlated but faulty indicator of quality of service

Average WAR is a reasonable indicator of quality of service

Individual concept survival



Quality of voicemail service



Average WAR at 80%

Out of 10 messages,

1 excellent

4 good

3 OK

1 Poor

Average WAR at 85%

Out of 10 messages,

2 excellent

5 good

2 OK

1 Poor

Quality of voicemail conversion

Very Good (Accuracy 90-100%) = expectation that nearly all concepts intact

Good (80-90%) = most concepts intact; flow of message clear

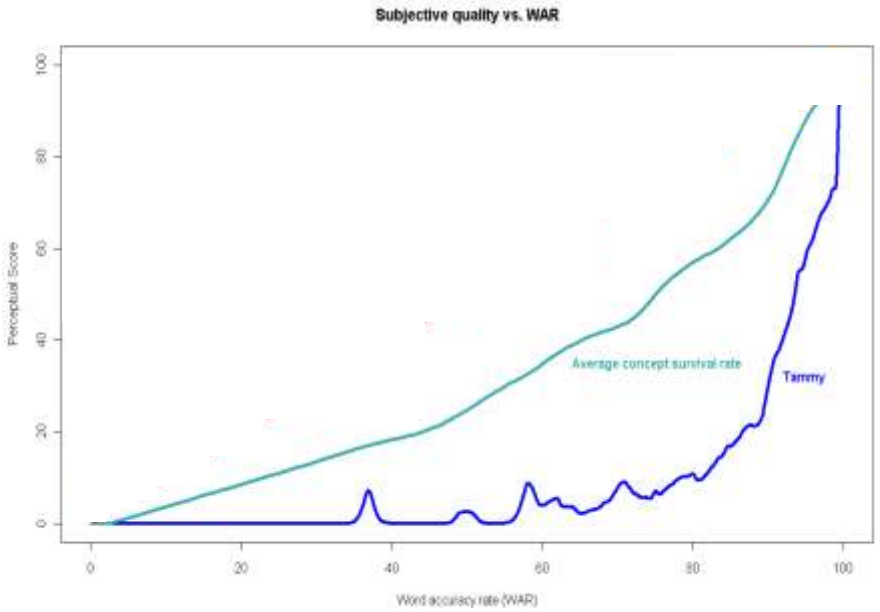
Somewhat Useful (70-80%) = ~half the concepts survive; flow of message may be disrupted

Poor (60-70%) = some concepts survive; flow of message significantly disrupted

Bad = a fraction of concepts discernible among the noise of bad recognition

Study #2: severity of error

- Large scale study utilizing quality assurance agents
- Goal to correlate severity of error with score that represents expectations of near perfect transcription



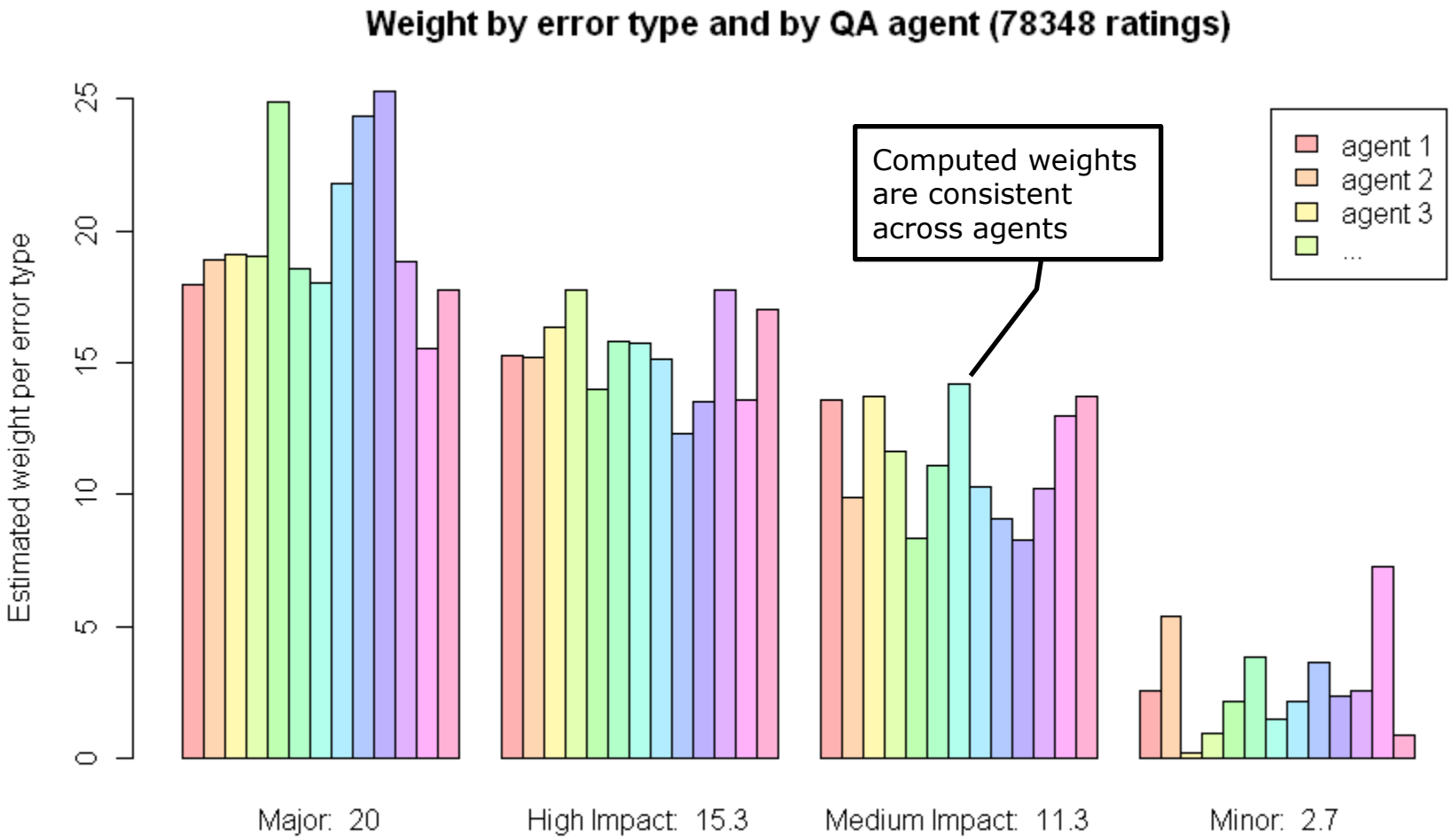
- *109,351 Spanish voicemails*
- *Ask agents to assess on a 1-4 scale (excellent, good, OK, poor – "Tammy scale")*
- *Piggy-back on existing workflow that requires agents annotate severity of errors (critical major, high, medium, low) as means of assessing production quality*

Penalties computed using regression

Model	Correlation
WAR	0.3
type dependent regression (linear with frequency)	0.59
freqency and type dependent regression	0.68

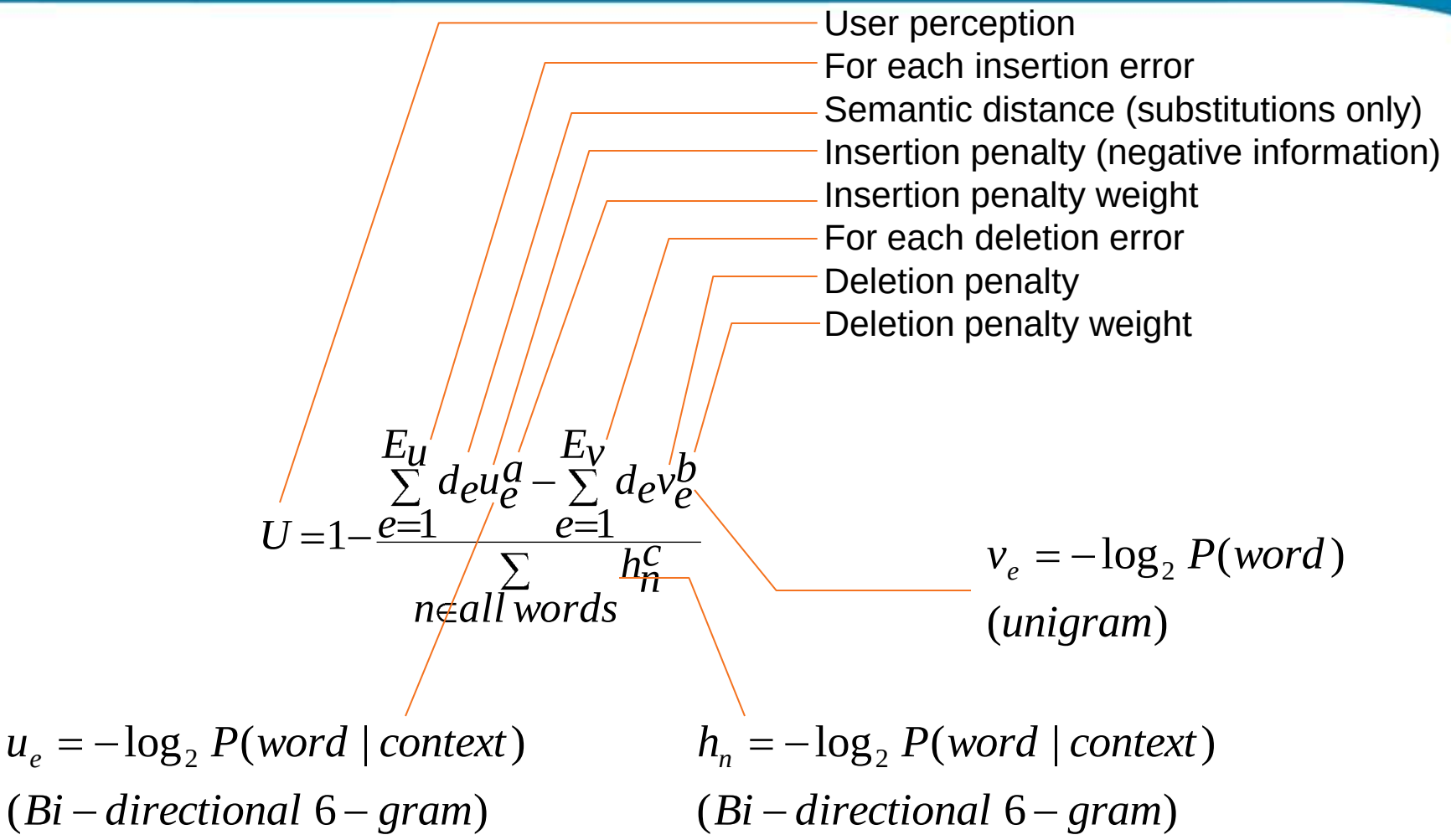
- Correlation between error classes requires penalties to flatten or drop for multiple events of the same type
- But high correlation coefficients achieved are suspect
  - Quantized scale and human annotation inconsistencies must account for significant portion of uncertainty.

Agents naturally use similar weights



- **RISK** – possible contamination by internal quality guidelines & process

The Goal


User perception
For each insertion error
Semantic distance (substitutions only)
Insertion penalty (negative information)
Insertion penalty weight
For each deletion error
Deletion penalty
Deletion penalty weight
$U = 1 - \frac{\sum_{e=1}^{E_u} d_e u_e^a - \sum_{e=1}^{E_v} d_e v_e^b}{\sum_{n \in all\ words} h_n^c}$
$v_e = -\log_2 P(word)$
$(unigram)$
$u_e = -\log_2 P(word \mid context)$
$(Bi-directional\ 6-gram)$
$h_n = -\log_2 P(word \mid context)$
$(Bi-directional\ 6-gram)$

The Goal (layman's version)

$$WAR = 1 - \frac{S + D + I}{N}$$

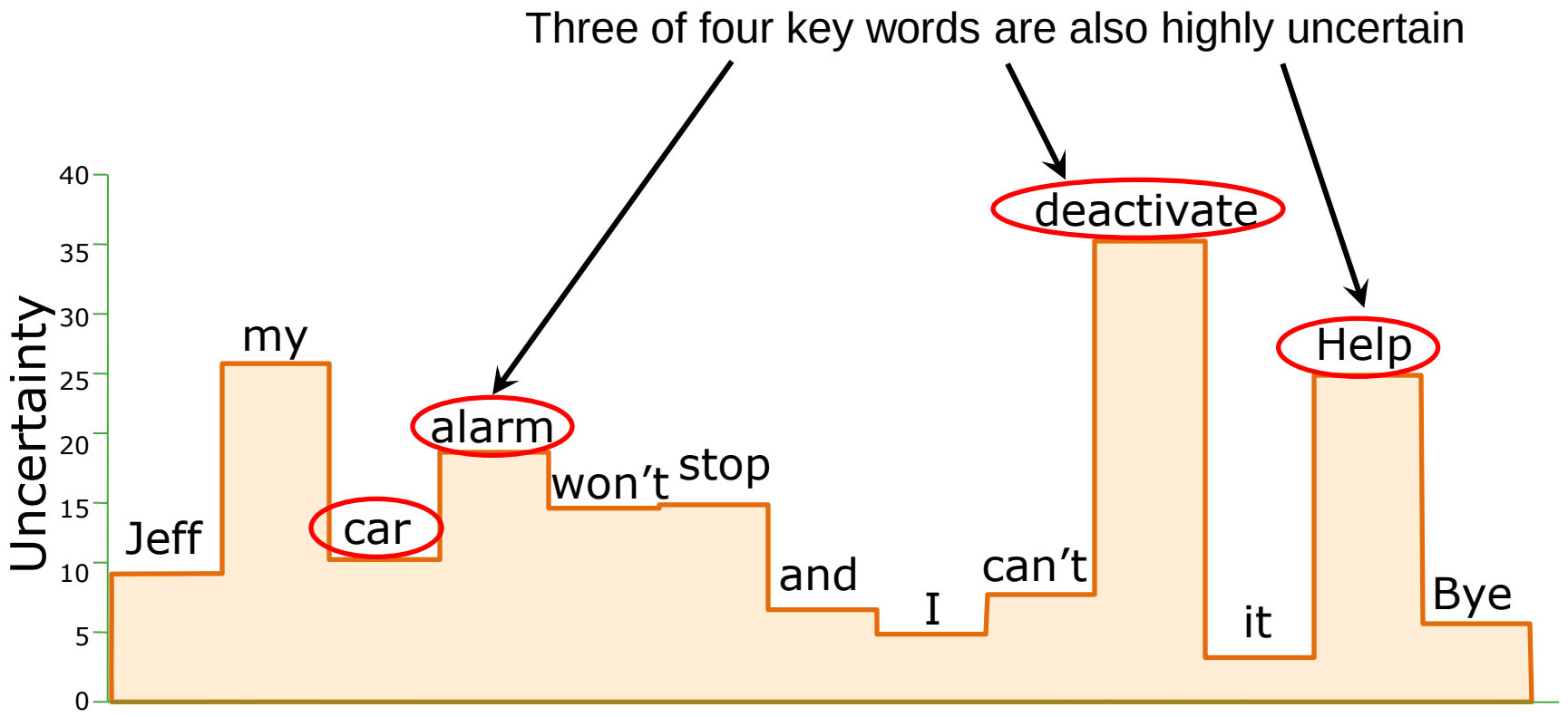
$$Weighted..WAR = 1 - \frac{wS + wD + wI}{wN}$$

W usually an indicator of class (Name…)

$$Perception..Score = 1 - \frac{f()S + g()D + h()I}{w()N}$$

f(), g() and h() should be functions of perplexity, easily extractable semantic tags and "neighborhood error behavior"

w() should normalize



Measurement	Correlation to 1-4 ranking
WAR	0.3
WAR – weighted by log unigram	0.35

Summary

- Perception depends on expectations
- WAR an OK metric when service expectations are low to moderate
- WAR a correlated but flawed metric when quality expectations are very high
- Amounts of annotated data are approaching quantities that may allow complex metrics to objectively model user perception
  - Indications that simple statistical measurements (e.g. perplexity) can be incorporated as weighting mechanisms

Appendix

Validity of QA agents as judges

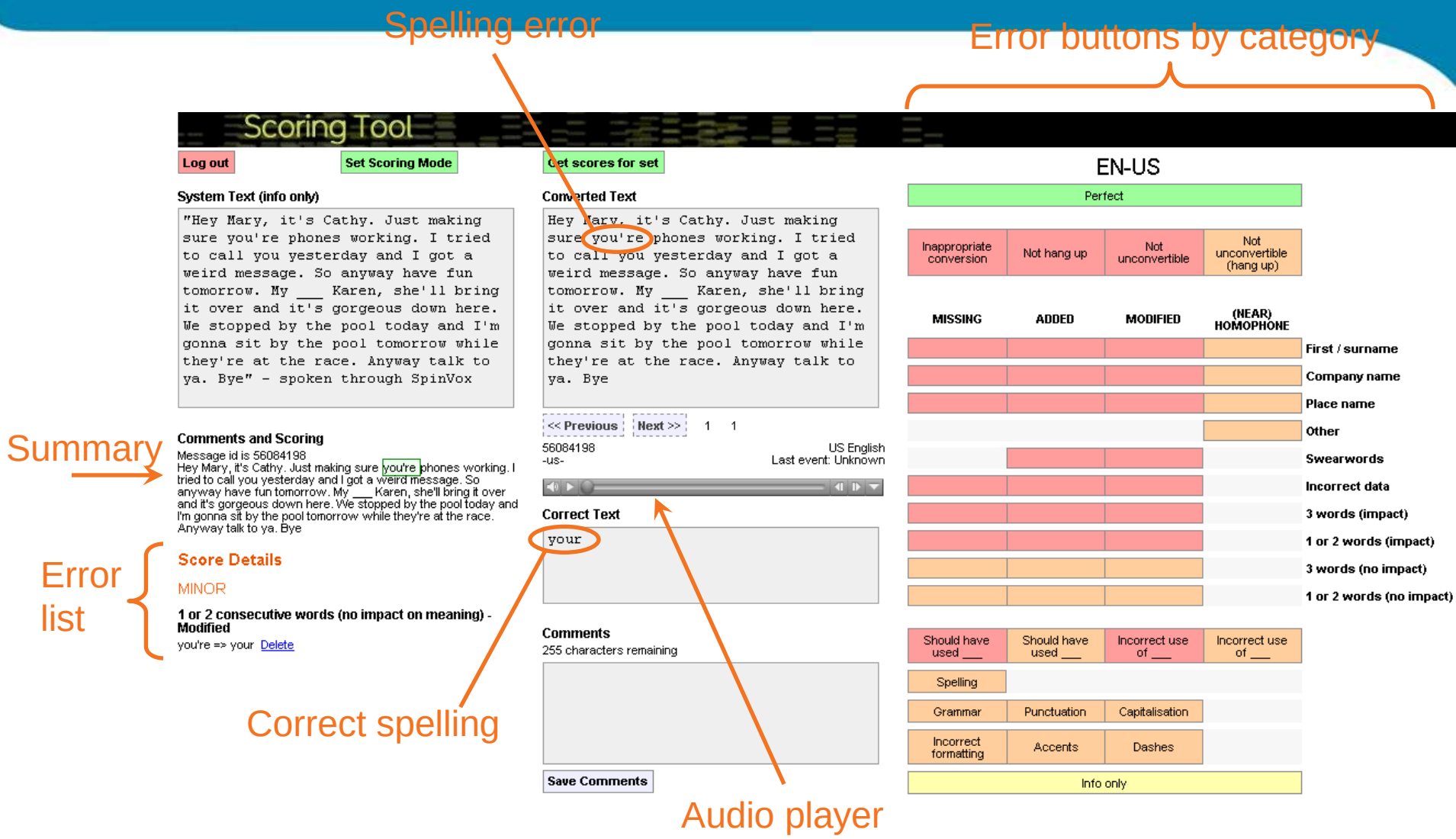
Pros:

- They listen to the audio (users have to guess)
- Trained - their job is to evaluate quality
- Experienced - very consistent results

Cons:

- Agents are not users
  - Not paying for the service
  - Content do not affect them personally
  - Must guess at *a priori* information
- Possible contamination from QA methodology
- Possible bias from editing message text

QA Tool


Spelling error
Error buttons by category
Scoring Tool
Log out
Set Scoring Mode
Get scores for set
System Text (info only)
"Hey Mary, it's Cathy. Just making sure you're phones working. I tried to call you yesterday and I got a weird message. So anyway have fun tomorrow. My ___ Karen, she'll bring it over and it's gorgeous down here. We stopped by the pool today and I'm gonna sit by the pool tomorrow while they're at the race. Anyway talk to ya. Bye" - spoken through SpinVox
Converted Text
Hey Mary, it's Cathy. Just making sure you're phones working. I tried to call you yesterday and I got a weird message. So anyway have fun tomorrow. My ___ Karen, she'll bring it over and it's gorgeous down here. We stopped by the pool today and I'm gonna sit by the pool tomorrow while they're at the race. Anyway talk to ya. Bye
<< Previous
Next >>
1 1
56084198
-us-
US English
Last event: Unknown
Summary
Comments and Scoring
Message id is 56084198
Hey Mary, it's Cathy. Just making sure you're phones working. I tried to call you yesterday and I got a weird message. So anyway have fun tomorrow. My ___ Karen, she'll bring it over and it's gorgeous down here. We stopped by the pool today and I'm gonna sit by the pool tomorrow while they're at the race. Anyway talk to ya. Bye
Error list
Score Details
MINOR
1 or 2 consecutive words (no impact on meaning) - Modified
you're => your Delete
Correct Text
your
Comments
255 characters remaining
Save Comments
Correct spelling
Audio player
EN-US
Perfect
Inappropriate conversion
Not hang up
Not unconvertible
Not unconvertible (hang up)
MISSING
ADDED
MODIFIED
(NEAR) HOMOPHONE
First / surname
Company name
Place name
Other
Swearwords
Incorrect data
3 words (impact)
1 or 2 words (impact)
3 words (no impact)
1 or 2 words (no impact)
Should have used ___
Should have used ___
Incorrect use of ___
Incorrect use of ___
Spelling
Grammar
Punctuation
Capitalisation
Incorrect formatting
Accents
Dashes
Info only