

---

# Experiments in Human Annotation of Usage Errors in Learner Text

---

Joel Tetreault  
Martin Chodorow  
Yoko Futagi

[Educational Testing Service]  
[Hunter College of CUNY]  
[Educational Testing Service]

---

# Questions

(1) Is it bad to use only one human rater?

- ⊖ Grammatical errors (e.g. Subject-Verb agreement) have high reliability
- ⊖ Usage errors (prepositions, collocations), we show to have low reliability

(2) Are there efficient methods for evaluating a system without having to annotate thousands of examples?

- ⊖ Low frequency usage errors
-

---

# Experiments

## (1) Is it bad to use only one human rater?

- ⊖ For prepositions usage errors: we double annotate TOEFL essays
- ⊖ Compute Agreement/Kappa measures
- ⊖ Evaluate system performance vs. two raters

## (2) Efficient methods?

- ⊖ Show how a sampling approach can reduce workload with results comparable to exhaustive annotation
-

---

# Outline

- ✓ Experiment Background
  - ✓ Experiment 1:
    - θ Annotation Scheme
    - θ Double Annotation
    - θ System Comparison
  - ✓ Experiment 2:
    - θ Sampling Approach
  - ✓ Conclusions
-

---

# Error Targeting

- ✓ Schemes target many different types of errors (Izumi, 2003), (Granger, 2003)
  - ✓ Problematic:
    - ⊖ High cognitive load on rater to keep track of dozens of error types
    - ⊖ Some contexts have several different errors (many different ways of correcting)
    - ⊖ Can degrade reliability
  - ✓ Targeting one error type reduces the effects of these issues
-

---

# Usage Error Detection Tasks

- ✓ Preposition Error Detection

- ⊖ Selection Error (“They arrived *to* the town.”)

- ⊖ Extraneous Use (“They came *to* outside.”)

- ⊖ Omitted (“He is fond this book.”)

- ✓ Collocation Error Detection

- ⊖ \*“a strong computer” vs. “a powerful computer”

- ⊖ \*“make an election” vs. “hold an election”

- ✓ Evaluation Domain: TOEFL essays

- ⊖ 20,000 preposition contexts

---

---

# Standard Annotation/Evaluation

- ✓ Rater presented with learner's sentence
  - ✓ For usage errors, rater judges acceptability of target word(s)
  - ✓ Acceptability can be binary or continuous
  - ✓ Annotation used as a gold standard to compare to system's judgments
  - ✓ Others:
    - ⊖ Verification (Gamon et al., 2008)
    - ⊖ Cloze/Choice Test Comparisons
-

---

# Experiment 1: Is One the Loneliest Number?

- ✓ Regardless of evaluation procedure, using one rater is problematic
    - ⊖ linguistic drift, age, location, fatigue, task difficulty
  - ✓ Experiment:
    - ⊖ Created annotation scheme (prep's, colloc's)
    - ⊖ Trained two native-speakers
    - ⊖ Every 2000 preposition contexts, both annotated the same overlap set of 100 contexts
-

---

# Annotation Scheme

- ✓ Annotators were presented sentences from TOEFL essays with each preposition flagged
  - ✓ Pre-Preposition annotation:
    - ⊖ Mark presence of spelling errors in context
      - ✓ +/- 3 words and commanding verb
    - ⊖ Determiner/Plural errors...
    - ⊖ Grammatical Errors
-

---

# Annotation Scheme

- ✓ Preposition Annotation:
    - ⊖ 0 – if preposition is extraneous
    - ⊖ 1 – if incorrect preposition is used (and mark substitution(s))
    - ⊖ 2 – preposition is perfect for that context
    - ⊖ e – preposition is perfect, but there are others that are acceptable as well (mark others)
    - ⊖ Then mark confidence in judgment
-

---

# Procedure

- ✓ Raters given blocks of 500 preposition contexts
  - ✓ Took roughly 5 hours per block
  - ✓ After two blocks each, raters did an overlap set of 100 contexts (1800 contexts total)
  - ✓ Every overlap set was adjudicated by two other human raters:
    - ⊖ Sources of disagreement were discussed with original raters
    - ⊖ Agreement and Kappa computed
-

---

# How well do humans compare?

- ✓ For all overlap segments:
    - ⊖ “2” and “e” are collapsed to “ok”
    - ⊖ Contexts with grammar or spelling errors are excluded (resulting in 1336 contexts)
    - ⊖ Agreement = 0.952
    - ⊖ Kappa = 0.630
    - ⊖ Kappa ranged from 0.411 to 0.786
  - ✓ Including spelling/grammar:
    - ⊖ Kappa ranged from 0.474 to 0.773
-

# Confusion Matrix

Rater 1 \ Rater 2	Extraneous	“Error”	“OK”
Extraneous	17	0	6
“Error”	1	42	20
“OK”	4	33	1213

(Grammar and Spelling Excluded = 1336 contexts)

# Implications for System Evaluation

- ✓ Comparing a system (Chodorow et al., 2007) to one rater's judgments can skew evaluation results
- ✓ Test: 2 native speakers rated 2,000 prepositions from TOEFL essays:
  - ⊖ Diff. of 10% precision, 5% recall (rater as gold standard)

	<b>Precision</b>	<b>Recall</b>
System vs. Rater 1	0.778	0.259
System vs. Rater 2	0.677	0.205

---

# Multiple Prep's for Context

- ✓ “When the plant is horizontal, the force of gravity causes the sap to move \_\_\_\_\_ the underside of the stem.”
    - θ Writer: to
    - θ System: on
    - θ R1: toward
    - θ R2: onto
  - ✓ Shows another advantage of multiple raters ◇  
better able to list other acceptable prep's
-

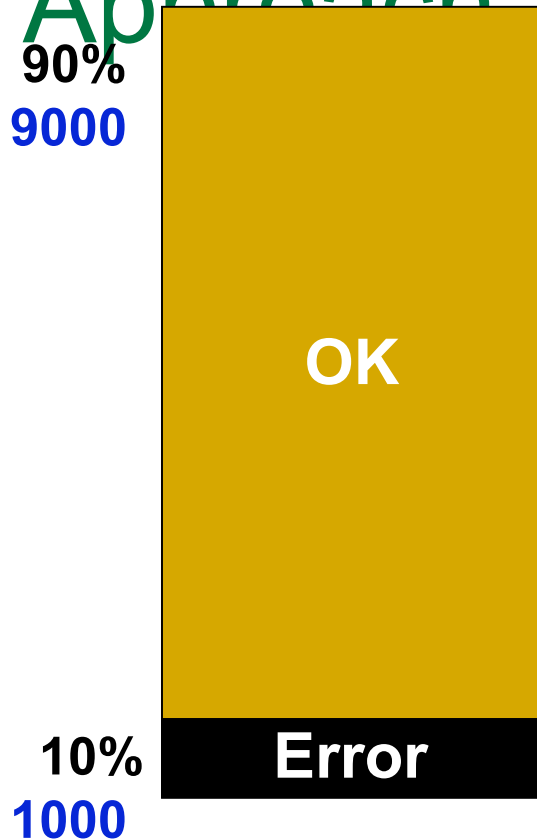
---

# Implications of Using Multiple Raters

- ✓ Multiple raters can indicate the variability of system performance
  - ✓ Standard annotation with multiple annotators is problematic:
    - ⊖ Expensive
    - ⊖ Time-Consuming (training, adjudication)
  - ✓ Is there an approach that can make annotation more efficient?
-

# Experiment 2: Sampling

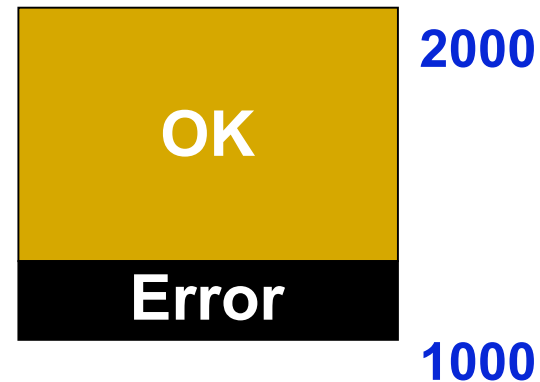
## Approach



### Sampling Approach:

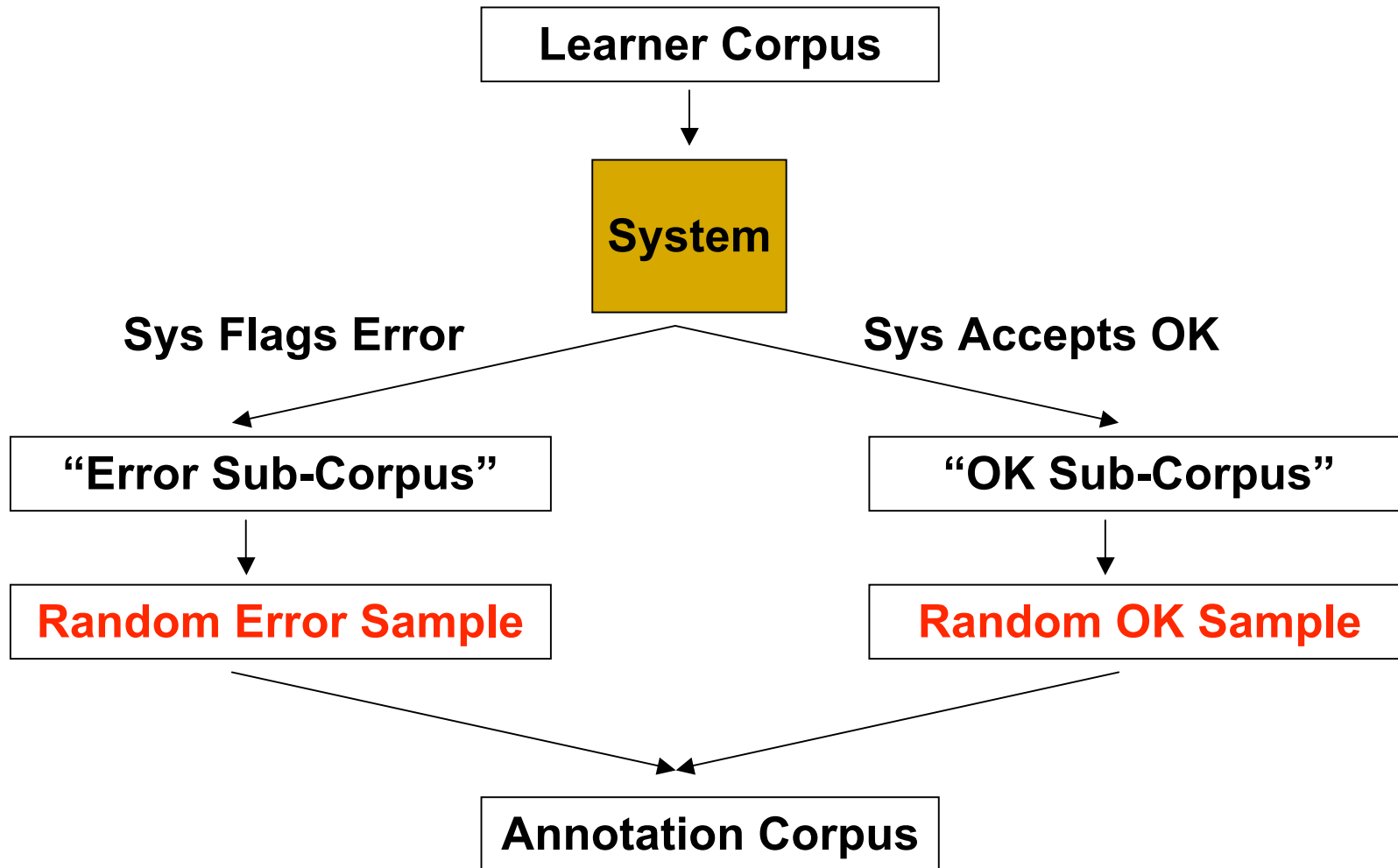
Sample system's output classifications  
Annotate smaller error-skewed corpus  
Estimate rates of hits, false positives, and misses

◇ Can calculate precision and recall

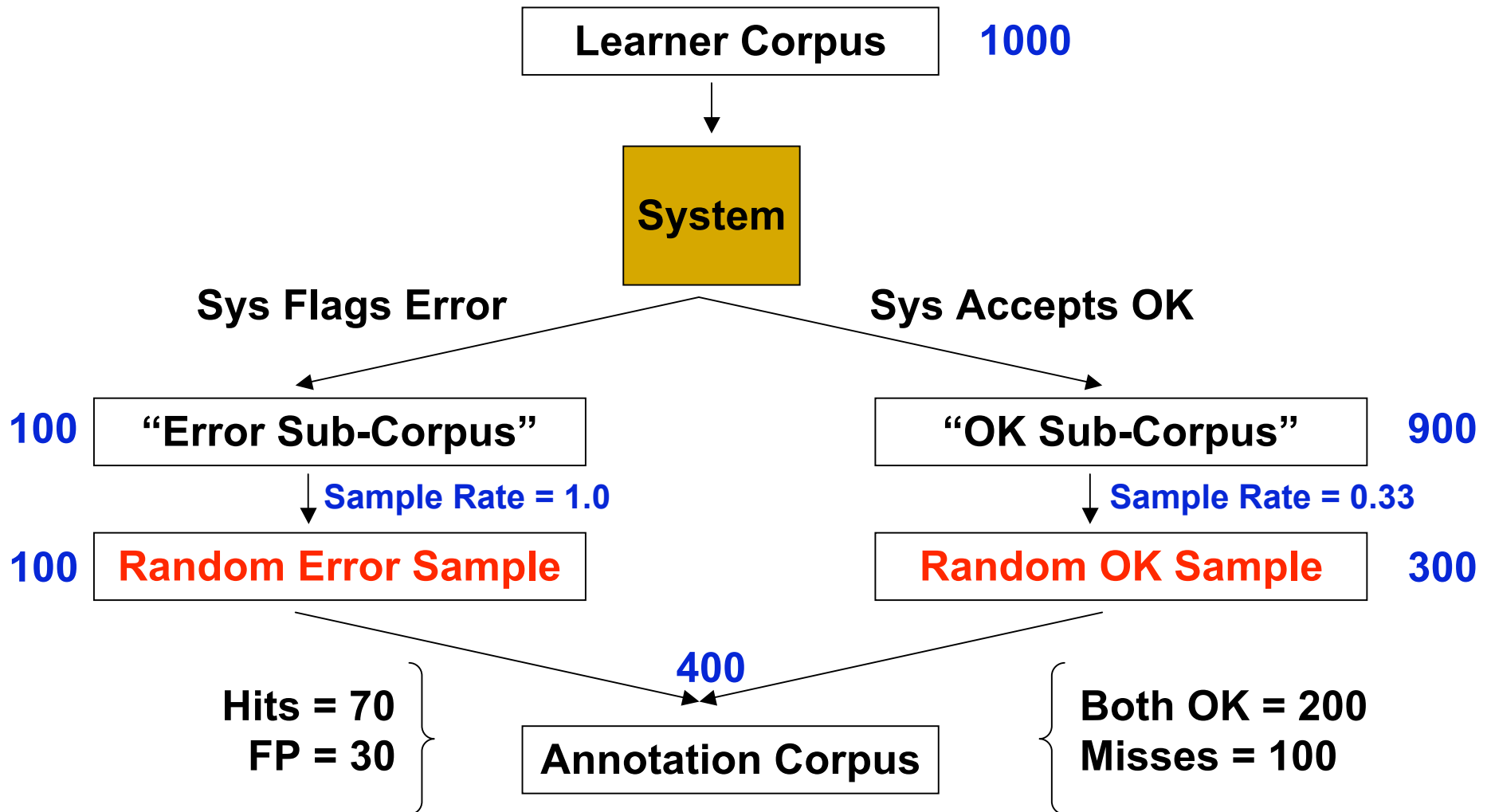


**Problem: to make an eval corpus of 1000 errors can take 100hrs!**

# Sampling Procedure



# Sampling Methodology



---

# Sampling Results

- ✓ Two raters working in tandem on sampled corpus
  - ✓ Compare against standard annotation
  - ✓ Results:
    - θ Standard:  $P = 0.79$ ,  $R = 0.18$
    - θ Sampling:  $P = 0.79$ ,  $R = 0.16$
-

---

# Conclusions

- ✓ Are two or more annotators better than one?
    - ⊖ Annotators vary in their judgments of usage errors
    - ⊖ Evaluation based on a single annotator under- or over-estimates system performance
  - ✓ Value of multiple annotators:
    - ⊖ Gives information about the range of performance
      - ✓ Dependent on number of annotators
    - ⊖ Multiple prep's per context handled better
  - ✓ Issues not unique to preposition task:
    - ⊖ Collocation kappa scores: 0.504 to 0.554
-

---

# Conclusions

- ✓ Sampling approach: shown to be a good alternative strategy to exhaustive annotation approach
    - ⊖ Requires less time through enriching the annotation set
    - ⊖ Results are similar to exhaustive annotation
    - ⊖ Avoid fatigue problem
-

---

# Future Work

- ✓ Do another sampling comparison to validate results
  - ✓ Leverage confidence annotations
  - ✓ Look at intra-annotator agreement
    - ⊖ How much does annotator agree with him/herself?
  - ✓ Cloze/Choice tests with learner text
    - ⊖ Easy way of assessing system performance
    - ⊖ Have done evaluation on well-formed text
-

---

# Plugs

- ✓ ACL Workshop: Innovative Uses of NLP for Building Educational Applications (ACL-BEA)
  - ⊖ Deadline: this Friday, March 21
  - ⊖ Still time!



---

# Sampling Procedure

- (1) Use system to classify targets as either errors or correct usage (“OK”)
  - (2) Create two subcorpora
    - ⊖ Error Sub-Corpus
    - ⊖ OK Sub-Corpus
  - (3) Differentially sample from two sub-corpora:
    - ⊖ Take a higher percentage of Error Sub-Corpus in order to enrich the annotation set with a larger proportion of errors than would be found in the original corpus
    - ⊖ Take a lower percentage of OK Sub-Corpus
  - (4) Have annotator rate each sentence in the standard way (applies to all three eval approaches)
-

---

# Verification

- ✓ System-only Verification

- ⊖ Rater checks system's output (Gamon et al, 2008)

- ✓ Blind Verification

- ⊖ Rater presented with system's and writer's choices (in random order)

- ⊖ Rater indicates which choice is preferred, or if they are equal

“My early experiments with shark behavior [at / on] Cape Haze during the 1950s....”



---

# NLP Evaluation

- ✓ Number of non-native speakers in US schools rising in the past decade
  - ✓ Highlights need for NLP tools to assist in language learning
  - ✓ Evaluation of NLP learner tools
    - ⊖ Important for development
    - ⊖ Error-annotation: time consuming and costly
    - ⊖ Usually one rater (Izumi, 2003; Eeg-Olofsson, 2002)
-