

The Future of Shared Resources for the Automated Assessment of Spoken and Written Language

Keelan Evanini

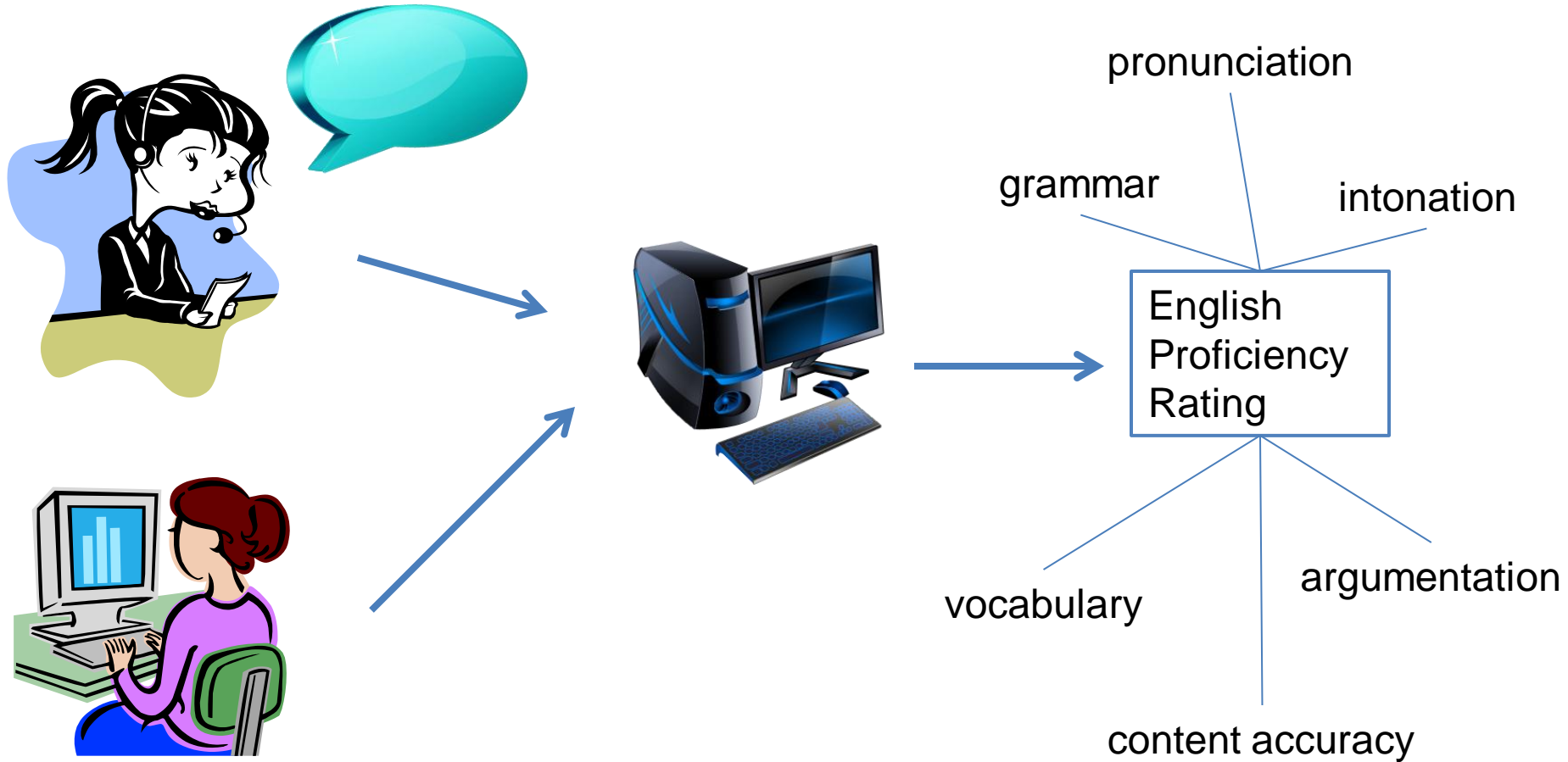
Educational Testing Service

**LDC's 20th Anniversary Workshop,
Sept. 6 - 7, 2012**

Outline

- Automated Language Assessment
 - What?
 - Why?
 - How?
- Samples of Available Resources
 - public
 - privately-held
- Shared Tasks
 - benefits / disadvantages
 - suggestions for the future

Automated Language Assessment



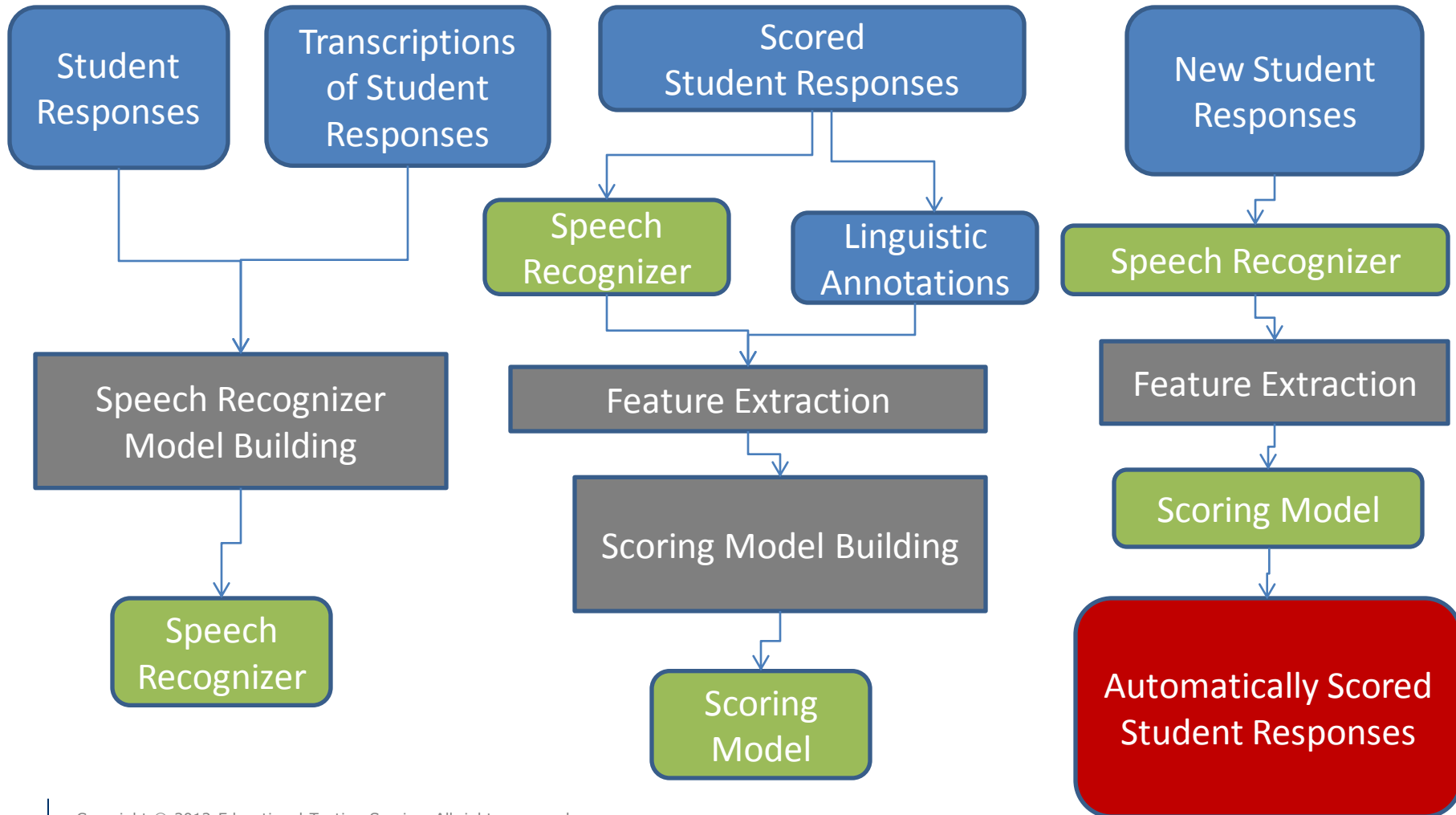
Automated Scoring at ETS

- **e-rater**: automated essay scoring
 - used in TOEFL and GRE (with human score)
 - features: grammar, usage, mechanics, style, discourse (little content)
- **Criterion**: writing feedback
 - e-rater engine, but no score -> highlights errors
- **SpeechRater**: automated speech scoring
 - used in TOEFL Practice Online
 - features: pronunciation, fluency (no content)
- **c-rater**: automated short answer scoring
 - assesses content accuracy

Pros and Cons

- Pros
 - faster score turnaround times
 - reduced scoring expenses
 - higher score reliability
- Cons
 - large initial R&D effort
 - cannot evaluate all aspects of language that a human can evaluate
 - potential for gaming the system

Automated Spoken Response Scoring



Children's Speech Corpora

- necessary for developing ASR component of reading assessment systems

Corpus	# Students	L1	Type of Speech	Citation
CMU Kids	76	NS	RA	LDC97S63
CSLU Kids' Speech	1100	NS	RA+SP	LDC2007S18
IBM Kid-speak	800	NS+NNS	RA	Kantor et al. (2012)
SRI-internal	> 400	NS+NNS	RA	Franco, <i>p.c.</i>
ETS-internal	3385	NNS	RA+SP	N/A

Non-Native Speech

- necessary for developing ASR component of non-native speech assessment systems

Corpus	# Speakers	Size	Citation
ISLE	46	11484 utt., 18 hours	Menzel et al. (2000), ELRA-S0083
multiple (mostly academic)	mostly small (< 20 hours)		Raab et al. (2007)
ETS-internal	> 40,000	> 250,000 utt., > 3000 hours	N/A
Pearson-internal	several million utterances		Bernstein, <i>p.c.</i>

- other large, privately-held, non-native corpora: IBM, Rosetta Stone, Carnegie Speech, etc.

Types of Linguistic Annotations

- Syntactic parses
 - Sentiment / opinion
 - Discourse structure
- existing corpora are out-of-domain
- Grammatical errors
 - Segmental errors in pronunciation
 - Lexical stress errors
 - Ratings for different aspects of proficiency: vocabulary, intonation, etc.

Pronunciation Error Corpora

- necessary for training systems for pronunciation error detection and training

Corpus	# Speakers	Size	Citation
ISLE	46	8000 utt., 8 hours	Menzel et al. (2000), ELRA-S0083
SRI-internal (Spanish L2)	206	3,500 utt., 200,000 phones	Bratt et al. (1998)
IBM Kid-speak	163	14,000 utt., 21 hours	Kantor et al. (2012)

- Again, other large, privately-held pronunciation error corpora at Rosetta Stone, Carnegie Speech, etc.

Grammatical Error Corpora

- Leacock et al. (2010) list 10+ corpora of learner English with grammatical error tags
- different corpora used for different studies
- comparative evaluation of methodologies is difficult
- shared task / corpus necessary

Why Share Data?

- increased transparency of methods
- face validity of automated scoring systems
- state-of-the-art advances faster
- push for open-source methodology in recent contracts (K-12 assessments)
- dissemination of brand
- ETS mission statement: *“Our products and services ... support education and professional development for all people worldwide.”*

TOEFL Public Use Dataset

- corpus of spoken and written language produced by TOEFL iBT examinees
- non-native English
- also includes
 - scores
 - demographic information
 - test materials
- available to researchers who submit a research proposal to ETS

TOEFL Public Use Dataset

- Speech
 - 2880 spoken TOEFL responses from 480 examinees
 - 6 responses per examinee
 - 44 hours of audio
 - each response has score provided by expert raters (1 – 4 scale)
- Writing
 - 960 written TOEFL essays from 480 examinees
 - 2 responses per examinee
 - each essay has score provided by expert raters (1 – 5 scale)

TOEFL Public Use Dataset

- not very large
- no transcriptions of spoken responses
- only annotations are proficiency scores
- not very easily attainable

→ efforts at ETS to release more data

TOEFL11

- new corpus of 11,000 essays
- 1000 essays from each of 11 L1s
- useful for Natural Language Identification
- also contains proficiency ratings

- will be distributed through LDC soon!

ETS Written Corpora

- > 1.5M essays
- > 500M words
- ca. 2/3 non-native English, 1/3 native

- potential to release more beyond TOEFL11 corpus

Corpus Based Reading and Writing Research group

- *"We" are a group of linguists, psychologists, computer scientists, and writing-program professionals; and we believe that that a large collection of student writing, as part of a larger collection of texts and annotations, would provide an essential basis for many important kinds of research.*
- *Our general idea is to create an open and evolving dataset of both student writing and expert writing, combined with an open and evolving collection of layers of annotation*
- <http://languagelog ldc.upenn.edu/nll/?p=3964>

Shared Tasks

- Instrumental in spurring innovation in many sub-fields of speech / NLP:
 - speech synthesis (Blizzard)
 - machine translation (WMT)
 - many semantic analysis tasks (SenseEval / SemEval)
 - etc.
- Until recently, no shared tasks for the field of automated language assessment
- HOO, ASAP

Helping Our Own (HOO)

- “we want to ‘help our own’ by developing tools which can help non-native speakers of English (NNSs) (and maybe some native ones) write academic English prose of the kind that helps a paper get accepted.” (Dale & Kilgarriff 2010)
- new on-going shared task aimed at *grammatical error detection* and *prose revision tools* more generally

HOO 2012

- HOO pilot in 2011 and full task in 2012
- Results presented at NAACL-BEA workshop
- 14 groups participated
- Messy gold-standard data
 - participants allowed to request revisions during eval.
 - in total, 205 revisions made to error annotations out of a total of only 473 instances (Dale et al. 2012)
- Still, better than evaluating on different corpora
- Enabled direct comparison of precision / recall across different methodologies

Automated Student Assessment Prize (ASAP)

- Sponsored by the Hewlett Foundation
- Phase 1: Automated Essay Scoring
- Phase 2: Short Answer Scoring
- Total prize money of \$100,000 for each phase
- 8 commercial vendors also took part in a separate competition in Phase 1

ASAP

Automated Student Assessment Prize
Phase One: Automated Essay Scoring

The Hewlett Foundation: Automated Essay Scoring

Finished

Friday, February 10, 2012

\$100,000 • 156 teams Monday, April 30, 2012

Dashboard

Home/Info

Get the Data

Make a submission

Leaderboard

Forum (60 topics)

What approach did you use?

intellectual property

Public Leaderboard Performance Over Time

Congratulations and DC Conference

Final model scoring

Award Ceremony Expectations?

Competition Details » [Get the Data](#) » [Make a submission](#)

Develop an automated scoring algorithm for student-written essays.

The William and Flora Hewlett Foundation (Hewlett) is sponsoring the Automated Student Assessment Prize (ASAP). Hewlett is appealing to data scientists and machine learning specialists to help solve an important social problem. We need fast, effective and affordable solutions for automated grading of student-written essays.

Hewlett is sponsoring the following prizes:

- \$60,000: 1st place
- \$30,000: 2nd place
- \$10,000: 3rd place

- **Description**
- [Background](#)
- [Evaluation](#)
- [Rules](#)
- [Prizes](#)
- [Help](#)
- [Data Description](#)
- [Submission Instructions](#)
- [Team](#)
- [Timeline](#)

<http://www.kaggle.com/c/asap-aes>

ASAP

Automated Student Assessment Prize
Phase Two: Short Answer Scoring

The Hewlett Foundation: Short Answer Scoring

21 hours to go

Monday, June 25, 2012

\$100,000 • 152 teams Wednesday, September 05, 2012

Dashboard

Home/Info

Get the Data

Make a submission

Leaderboard

Prospect

Forum (59 topics)

Public leaderboard / Private
leaderboard

TY

Visualization Prospect

Test Data Released

Will You Please Provide Public Leader

Competition Details » [Get the Data](#) » [Make a submission](#)

Develop a scoring algorithm for student-written short-answer responses.

The William and Flora Hewlett Foundation (Hewlett Foundation) is sponsoring the Automated Student Assessment Prize (ASAP) in hopes of discovering new tools to support schools and teachers. The competition aspires to solve the problem of the high cost and the slow turnaround of hand scoring thousands of written responses in standardized tests. As a result many schools exclude written responses in favor of multiple-choice questions, which are less able to assess students' critical reasoning and writing skills. ASAP has been designed to help determine whether computerized systems are capable of grading written content accurately for schools and teachers to adopt those solutions. ASAP aspires to inform key decision makers, who are already considering adopting these systems, by delivering a fair, impartial and open series of trials to test current capabilities and to drive greater

- [Description](#)
- [Background](#)
- [Evaluation](#)
- [Rules](#)
- [Prizes](#)
- [Help](#)
- [Data Description](#)
- [Submission Instructions](#)
- [Team](#)
- [Timeline](#)

<http://www.kaggle.com/c/asap-sas>

ASAP

- Phase 1
 - ca. 22,000 student essays (grades 7,8, 10) from 8 prompts
 - completed April 30
 - best-performing systems exceed human-human agreement (Shermis & Hamner 2012)
- Phase 2
 - completed September 5
 - results available soon

Lessons from ASAP Phase 1

- important to figure out intricacies of data set
 - no carriage returns (from transcribed data)
 - errors in scores
 - global deletion of capitalized tokens (attempt at ensuring anonymity)
- potential for reduced focus on deeper scientific issues
- emphasis on single evaluation metric (weighted κ) limiting

Recommendations

- public release of more privately held learner corpora
- especially data with annotations
 - error markings
 - more general linguistic information
- more shared tasks using these corpora
- especially sub-components of overall assessment system (error detection)

References

- Bratt, H., L. Neumeyer, E. Shribert, and H. Franco. Collection and detailed transcription of a speech database for development of language learning technologies. *Proceedings of ICSLP*, Sydney.
- Dale, R., I. Anisimoff, and G. Narroway. 2012. HOO 2012: A Report on the Preposition and Determiner Error Correction Shared Task. *Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*.
- Dale, R. and A. Kilgarriff. 2010. Helping Our Own: Text Massaging for Computational Linguistics as a New Shared Task. *In Proceedings of the International Natural Language Generation Conference 2010, Dublin, Ireland*.
- Kantor, A., M. Cernak, J. Havelka, S. Huber, J. Kleindienst, D. Gonzalez. 2012. Reading Companion: The Technical and Social Design of an Automated Reading Tutor. *Proceedings of the InterSpeech Workshop on Child, Computer, and Interaction*.
- Leacock, C., M. Chodorow, M. Gamon, and J. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Synthesis Lectures on Human Language Technologies, Vol. 3, No. 1.
- Menzel, W., E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter. The ISLE corpus of non-native spoken English. *Proceedings of LREC*.
- Raab, M., R. Gruhn, and E. Noeth. 2007. Non-native speech databases. *Proceedings of ASRU*.
- Shermis, M. and B. Hamner. 2012. Contrasting state-of-the-art automated scoring of essays: Analysis. *Paper presented at NCME*.