Workable Models of Standard Performance in English & Spanish

2 June 2005

EALTA

Voss, Norway

J. Bernstein, J. Balogh, M. Lennig, E. Rosenfeld

Ordinate Corporation

Presentation

What does it mean to “speak language X?”

Practical Problem: measure listening & speaking in a particular language (English or Spanish).

Describe development & evaluation of ***workable models*** of language performance

Application dictates Technology

- Requirement for large volumes (>100/day) and for fairness suggests fully automatic methods

- Fully automatic testing dictates explicit, simple models of language (to implement & train)

- New models and methods require evaluation

Types of Spoken Language Test

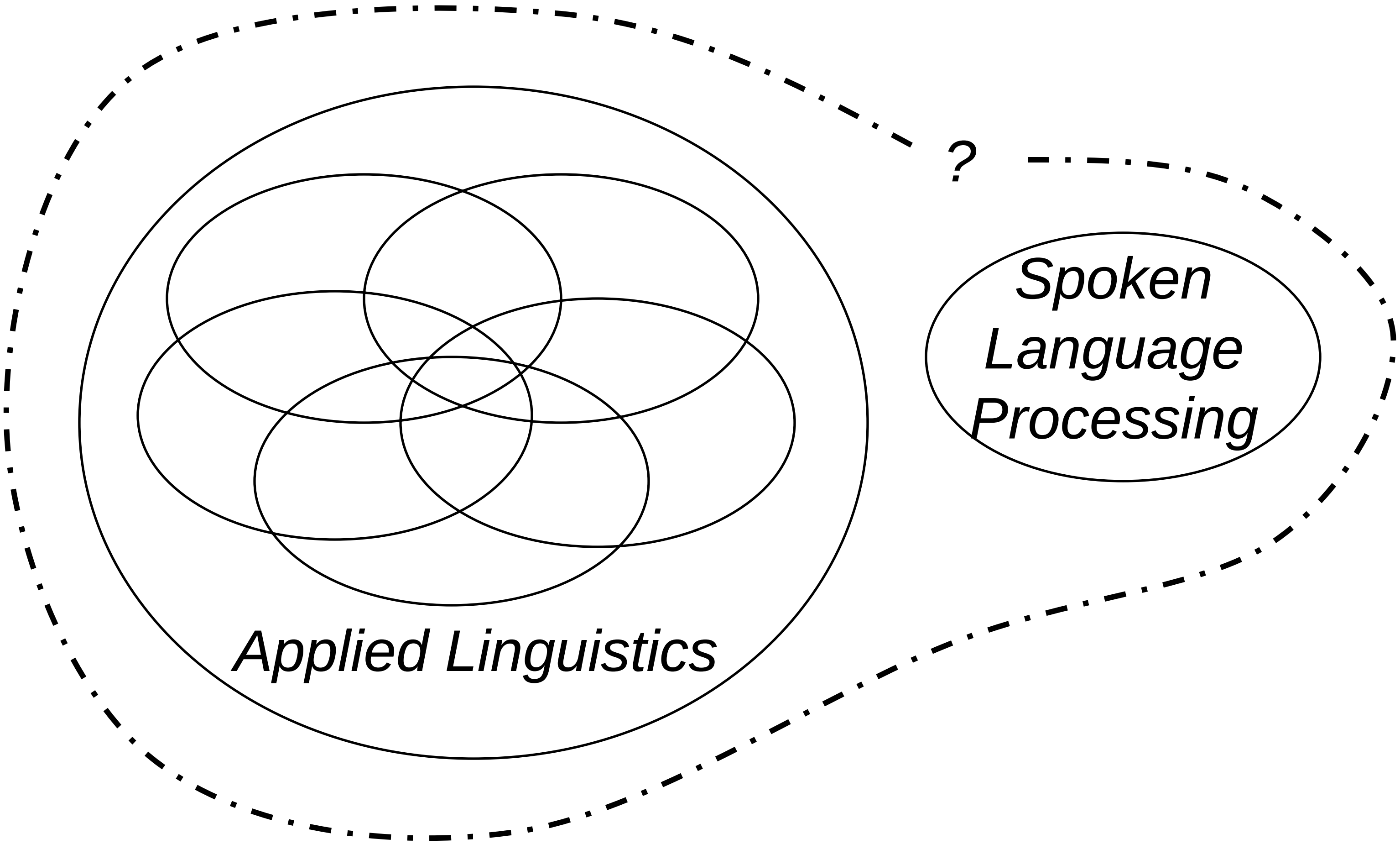
- Language Proficiency Interview (LPI)
  - Fully Human, operational construct definition
  - ILR OPI, ACTFL OPI, … TSE

- Automatic spoken language test
  - Fully automatic tests with *facility* construct
  - PhonePass SET-10, SST

Types of Spoken Language **Model**

- Oral Proficiency in Communication
  - Structure: applied linguistics research literature
  - Content: iterative expert judgment
- Performance with Language
  - Structure: General-purpose statistical estimation
  - Content: iterative training on performance data

Applied Linguistics ~ SLP



SLP History

Spoken Language Processing

"Simplicity (+ data) swamps insight."

Practical Goal: Human Machine Dialog

Original mainstream method was to implement expert meta-cognitive strategies.

Jelinek and others redefined the critical task as decoding speech to text on statistical basis.

First Overly Simple Model (1980s)

Utterance → words → phonemes → acoustics

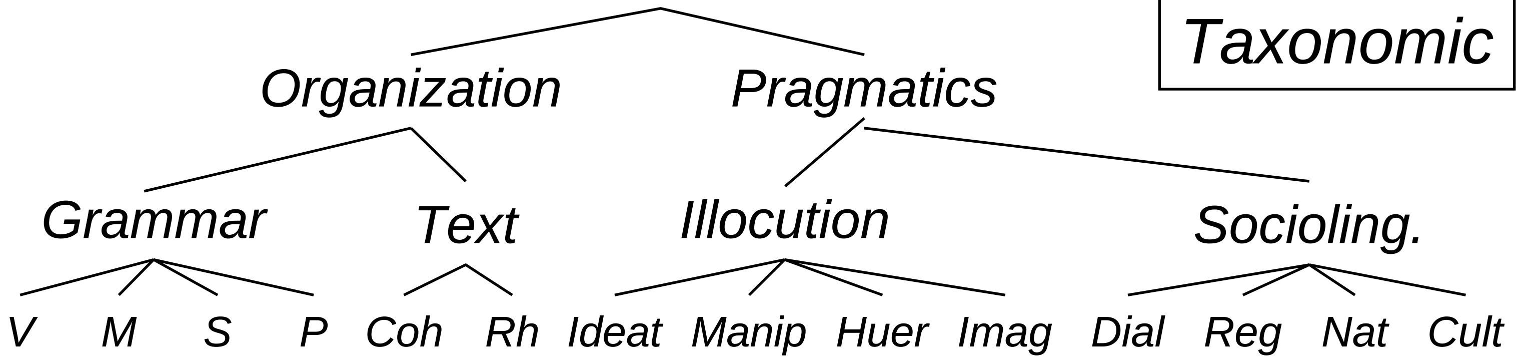
Grammar: p(w3) given (w1, w2)

Trained on large data sets, out-performed expert models based in insight.

The whole field changed and the statistical methods took over natural language processing.

Construct Comparison

COMMUNICATIVE COMPETENCE*



Organization Pragmatics

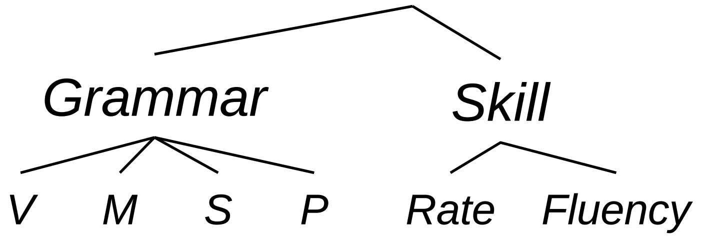
Grammar Text Illocution Socioling.

V M S P Coh Rh Ideat Manip Huer Imag Dial Reg Nat Cult

Taxonomic

*Bachman

LANGUAGE FACILITY*



Grammar Skill

V M S P Rate Fluency

FSMs, HMMs
Metric in time

*SET-10

Construct Comparison

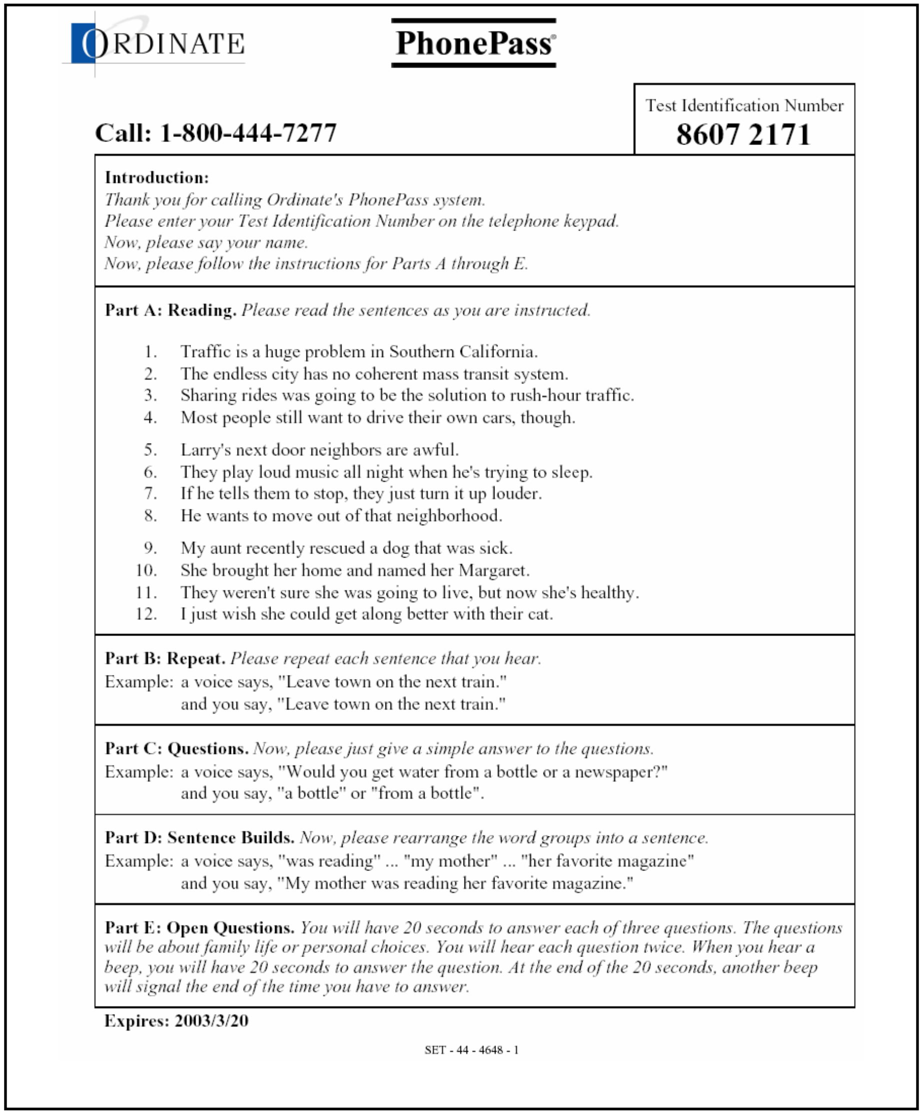
OPI Construct: Oral Proficiency as manifest in an Oral Proficiency Interview, but often with reference to *communicative competence* as reflected in the functional level and/or complexity of content accurately produced.

SET-10 Construct: *facility in spoken English* – the ability to understand spoken English and speak appropriately in response at a native-like pace on everyday topics.

SET-10 Format:

Test number (PIN)

Part A: reading	**8 items**
Part B: repeat Ss	**16 items**
Part C: short Qs	**24 items**
Part D: build Ss	**10 items**
Part E: open Qs	**3 items**



ORDINATE **PhonePass**®

Call: 1-800-444-7277

Test Identification Number
8607 2171

Introduction:
Thank you for calling Ordinate's PhonePass system.
Please enter your Test Identification Number on the telephone keypad.
Now, please say your name.
Now, please follow the instructions for Parts A through E.

Part A: Reading. *Please read the sentences as you are instructed.*

1. Traffic is a huge problem in Southern California.
2. The endless city has no coherent mass transit system.
3. Sharing rides was going to be the solution to rush-hour traffic.
4. Most people still want to drive their own cars, though.
5. Larry's next door neighbors are awful.
6. They play loud music all night when he's trying to sleep.
7. If he tells them to stop, they just turn it up louder.
8. He wants to move out of that neighborhood.
9. My aunt recently rescued a dog that was sick.
10. She brought her home and named her Margaret.
11. They weren't sure she was going to live, but now she's healthy.
12. I just wish she could get along better with their cat.

Part B: Repeat. *Please repeat each sentence that you hear.*
Example: a voice says, "Leave town on the next train."
and you say, "Leave town on the next train."

Part C: Questions. *Now, please just give a simple answer to the questions.*
Example: a voice says, "Would you get water from a bottle or a newspaper?"
and you say, "a bottle" or "from a bottle".

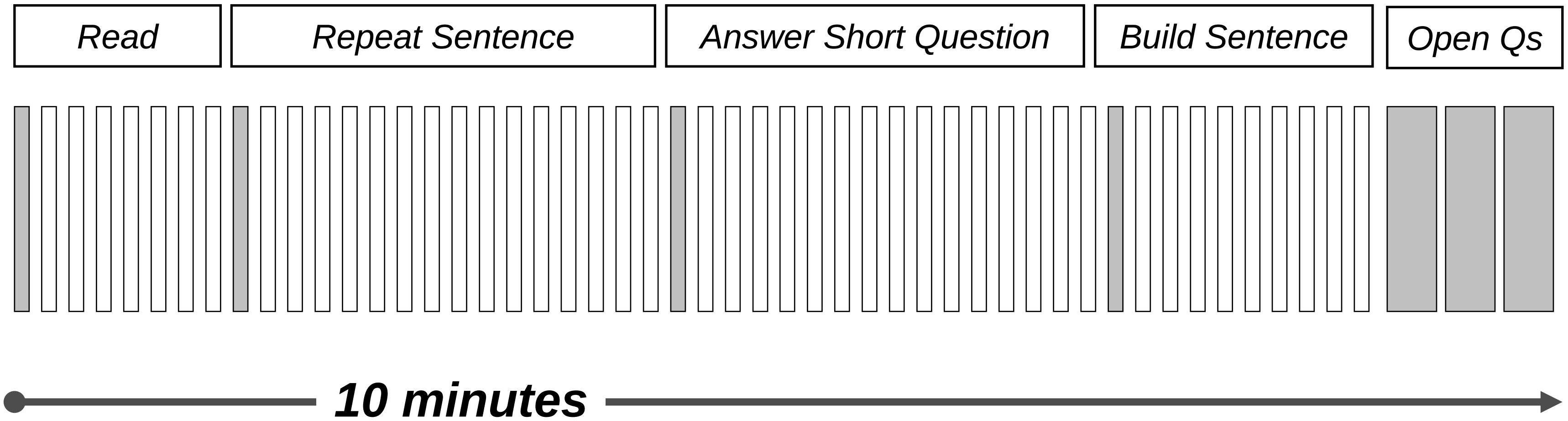
Part D: Sentence Builds. *Now, please rearrange the word groups into a sentence.*
Example: a voice says, "was reading" ... "my mother" ... "her favorite magazine"
and you say, "My mother was reading her favorite magazine."

Part E: Open Questions. *You will have 20 seconds to answer each of three questions. The questions will be about family life or personal choices. You will hear each question twice. When you hear a beep, you will have 20 seconds to answer the question. At the end of the 20 seconds, another beep will signal the end of the time you have to answer.*

Expires: 2003/3/20

SET - 44 - 4648 - 1

SET-10 Task Structure



Read | Repeat Sentence | Answer Short Question | Build Sentence | Open Qs

10 minutes

(Grey items not scored).

•Integrated ***"listen → speak"*** items

•Items require *real-time* processing

SLP Paradigm in SET & SST

Integrated model of linguistic performance

- embedded phoneme, word, and phrase networks
- quantitative models of criterion judgment and data-driven performance criteria

Corpus-based content and scoring

- Content is restricted by corpus occurrence
- Explicit model of target interlocutor
- Explicit, metric combination score elements

How SET, SST model a language

Hidden Markov Model framework (FSM, HMM)

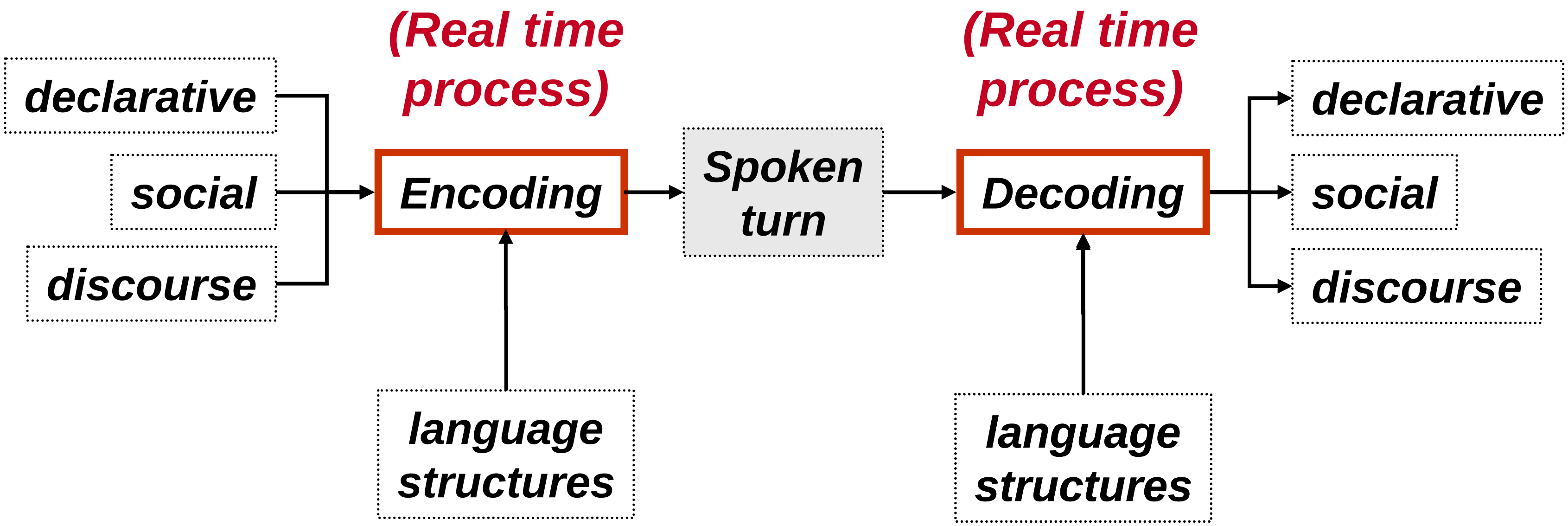
- Embedded stochastic networks
  - Lexicon; metric phrase & clause networks
  - Prosodic and segmental performance models
- Scoring is inherently disjunctive

Item Response Theory

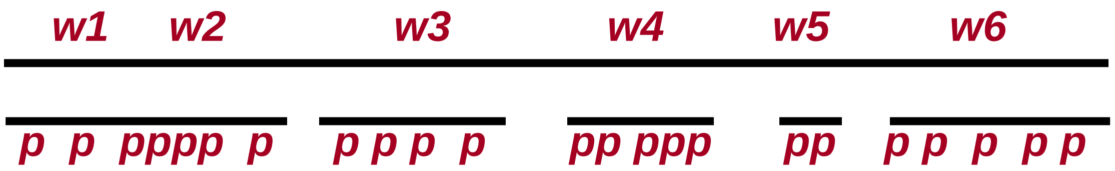
- Logistic regression (data-driven implicature)

Construct and Model

Facility in spoken English: ability to track what is said, extract meaning in real time, and formulate and produce relevant, intelligible responses, at a conversational pace

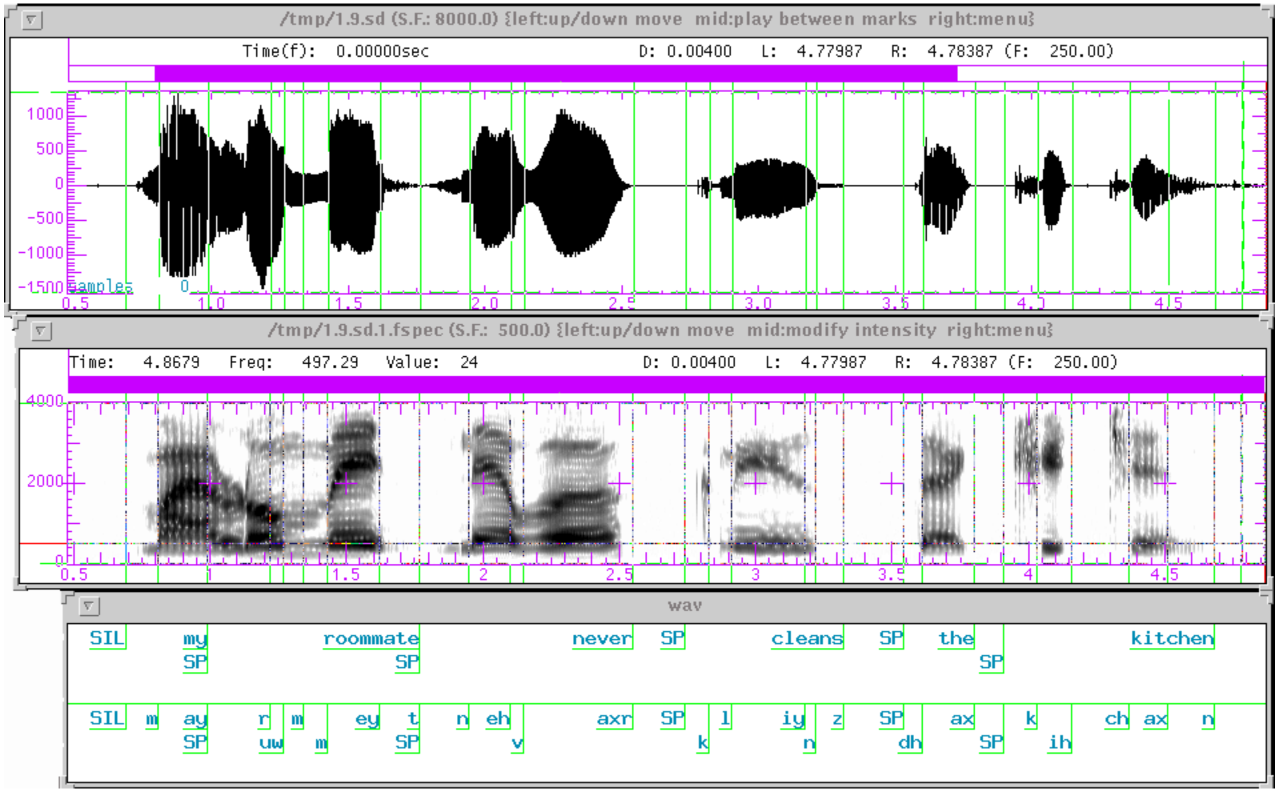


Phoneme & Word Alignment



75-90 Words/Min

5.8 Phones/Sec



/tmp/1.9.sd (S.F.: 8000.0) {left:up/down move mid:play between marks right:menu}

Time(f): 0.00000sec D: 0.00400 L: 4.77987 R: 4.78387 (F: 250.00)

/tmp/1.9.sd.1.fspec (S.F.: 500.0) {left:up/down move mid:modify intensity right:menu}

Time: 4.8679 Freq: 497.29 Value: 24 D: 0.00400 L: 4.77987 R: 4.78387 (F: 250.00)

wav

1000
500
0
-500
-1000
-1500
samples
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
4000
2000
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
SIL my roommate never SP cleans SP the kitchen
SP SP SP
SIL m ay r m ey t n eh axr SP l iy z SP ax k ch ax n
SP uw m SP v k n dh SP ih

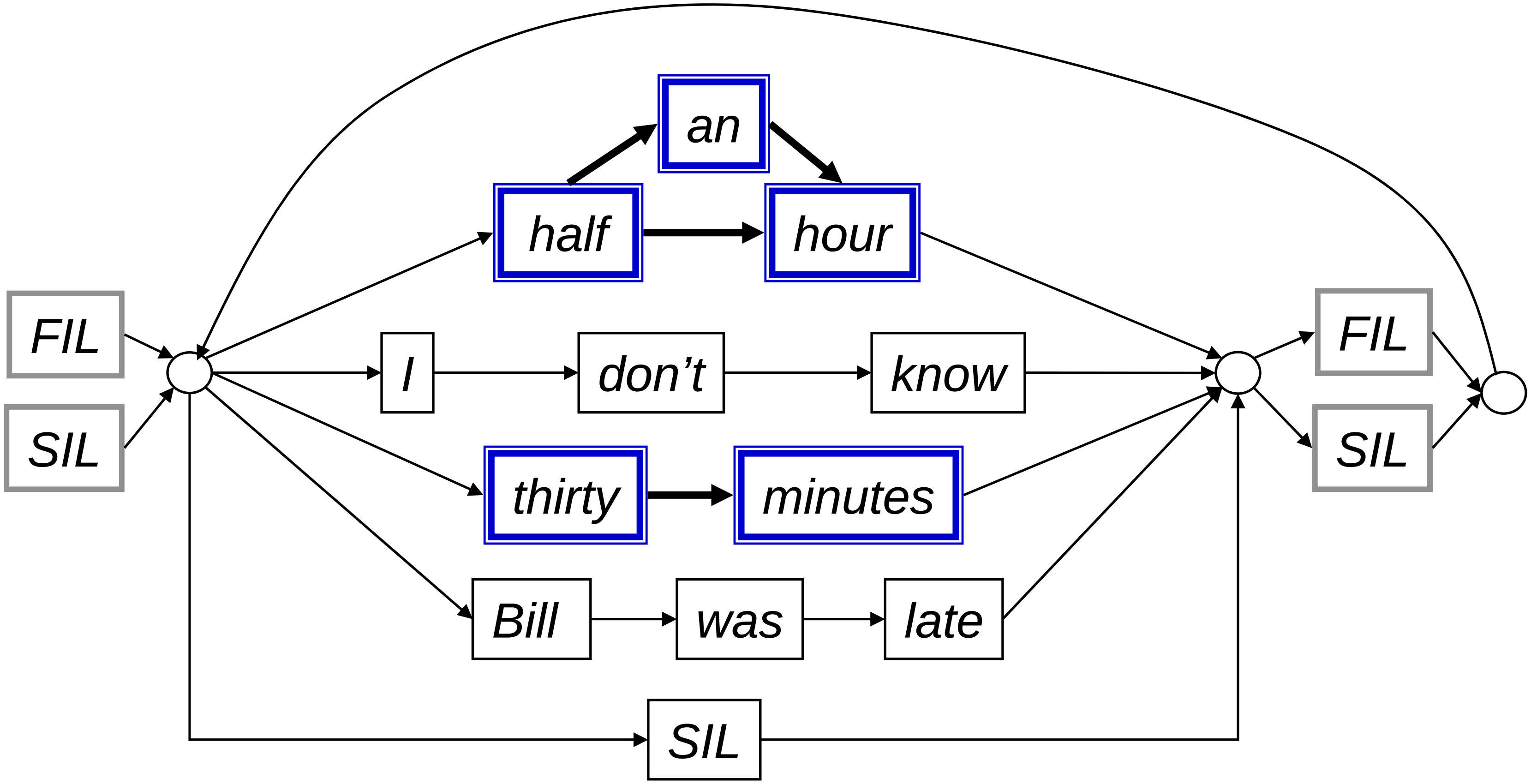
waveform

spectrum

words

segmentation

Simplified Response Network



Item Development Process

1. Bound lexicon to 1st 7000 lemmas in Switchboard
2. Sample sentences from N.American text or spoken transcripts; edit to fit in lexical bounds
3. Review text form in US, UK, Australia
4. Recitation recordings from diverse N.Americans
5. Pilot items on sample >= 50 natives/item (US, UK)

   If less than 90% correct, exclude the item

Spanish Item Process

1. Bound by lexicon to LDC counts (Sp, Ar, Mx)
2. Sample sentences from Argentine developer
3. Review text form; intersect Puerto Rico, Mexico, Venezuela, Spain, Argentina, and Ecuador

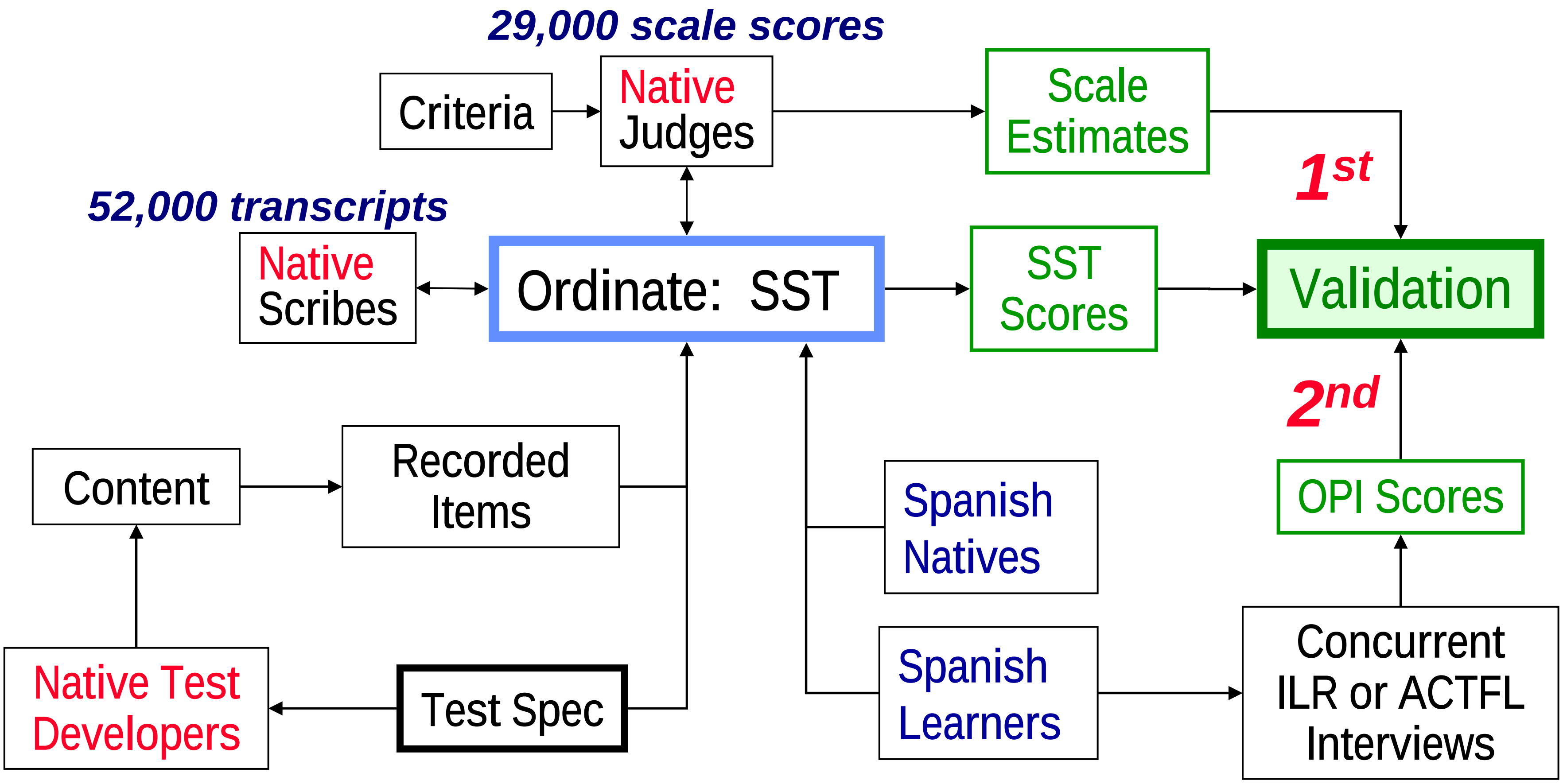
e.g. “*Aquellos eran otros tiempos*.”
“*Algunas veces se quedaba dormido*.”

- Recitation recordings from diverse Latinos
- Pilot items on sample >= 50 natives

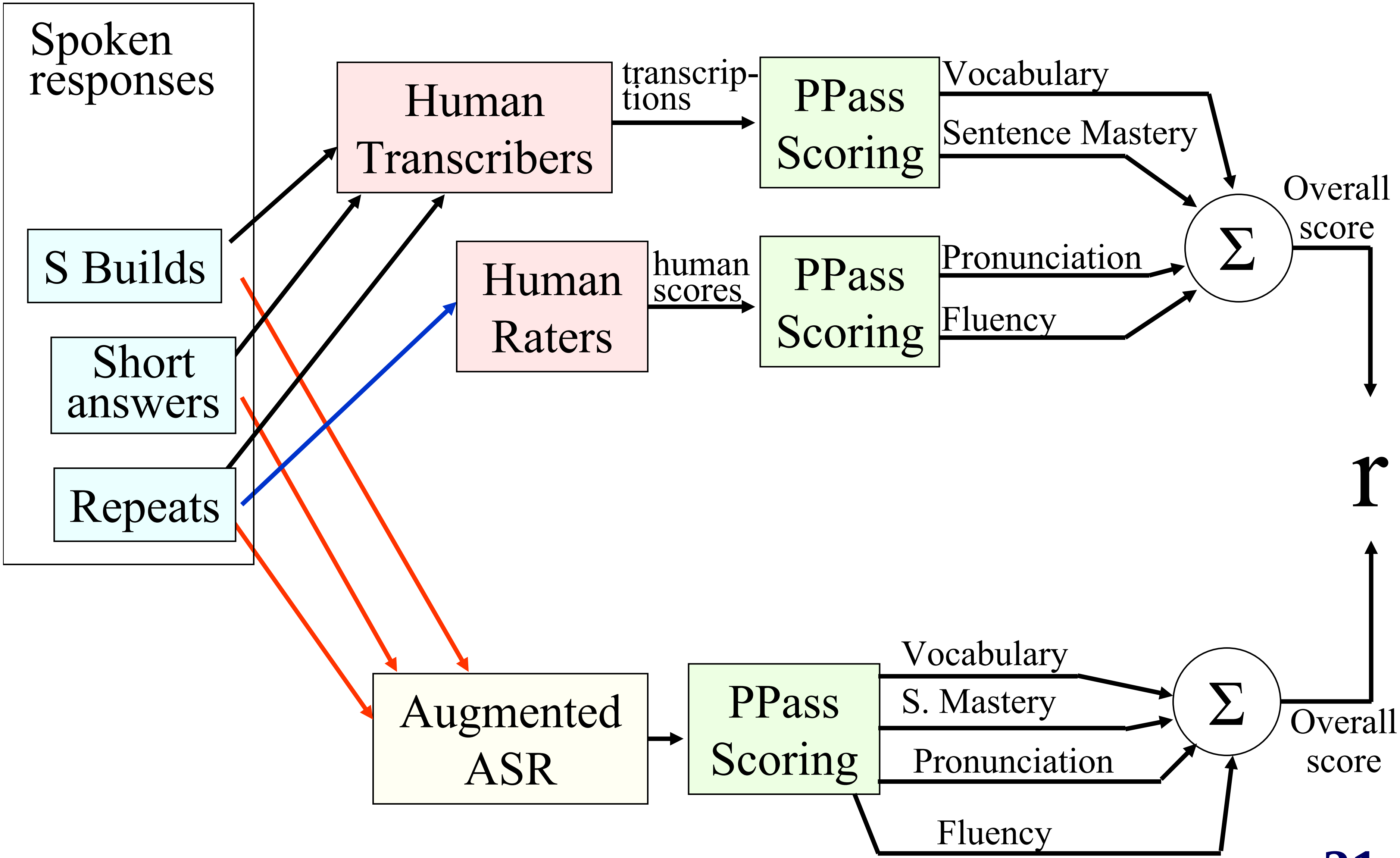
(Argentina, Mexico, Puerto Rico, Columbia,…)

If less than 80% correct, exclude the item

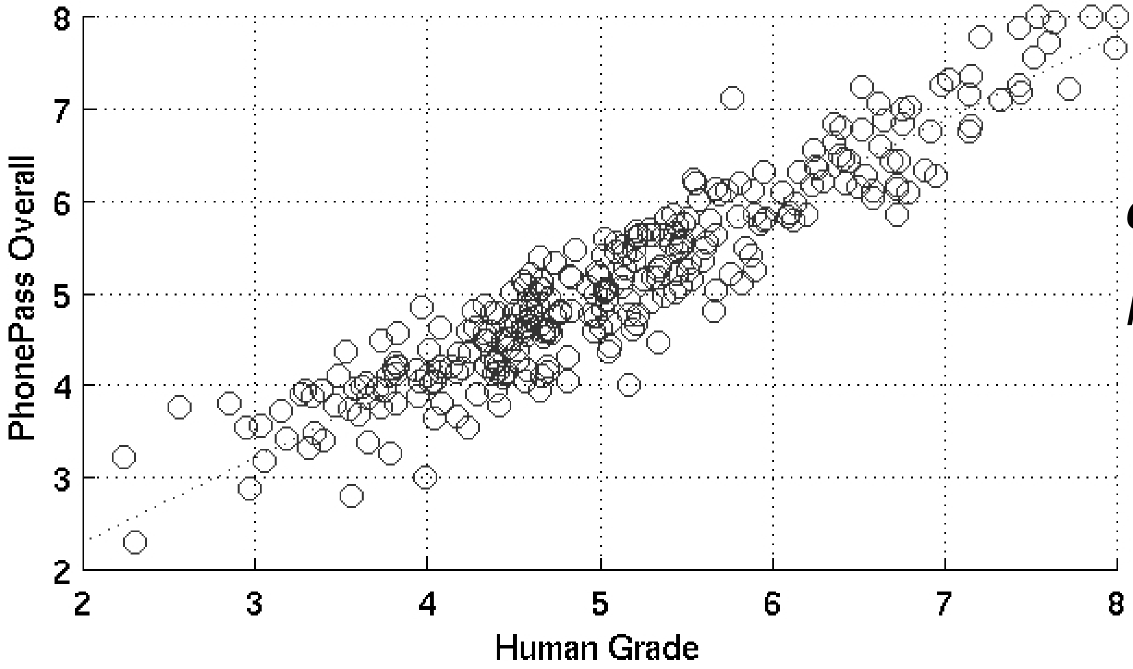
SST Development and Validation



1st Validation → Machine Estimates



1st Machine-Human Comparison

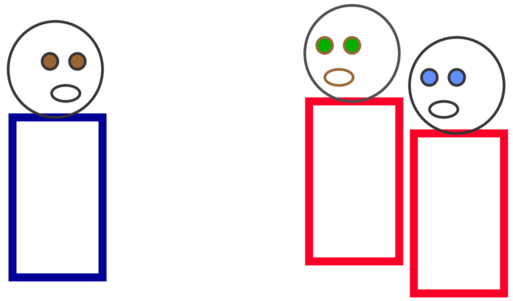


correlation = 0.94

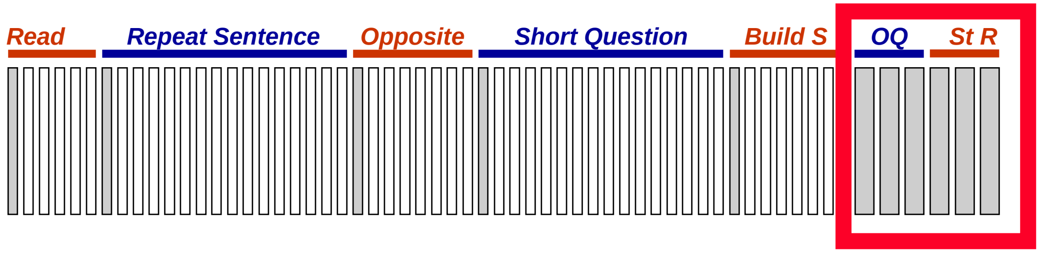
N = 288

Human scoring compared to machine-scoring

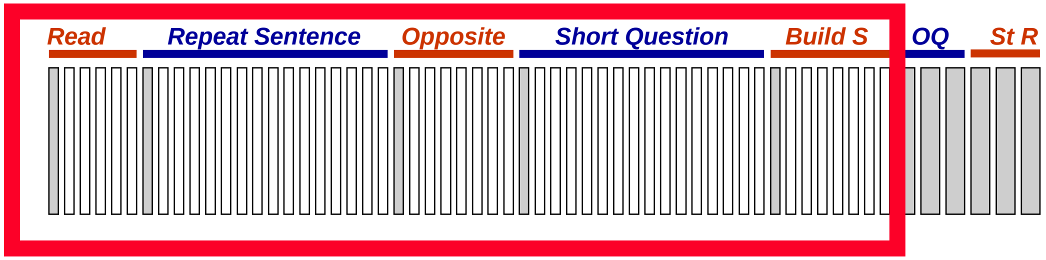
2nd Validation: Human ~ Machine Scores



ILR-SPT and ACTFL
Human Interview Scores



*ILR-SPT, **CEF***
Scale Estimates
(2 human raters per)

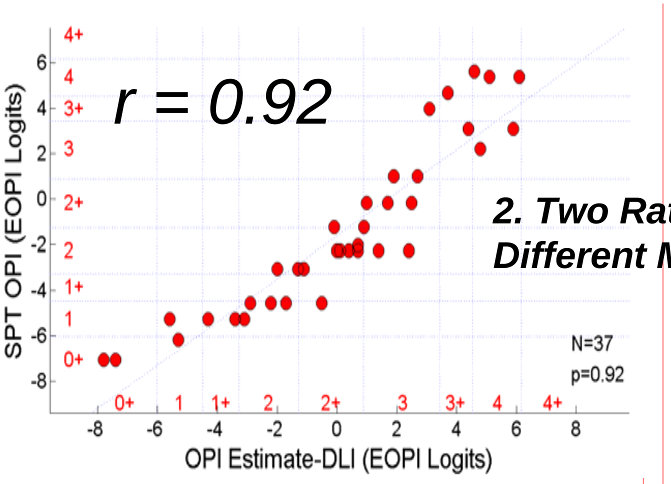


SST
Machine Scores

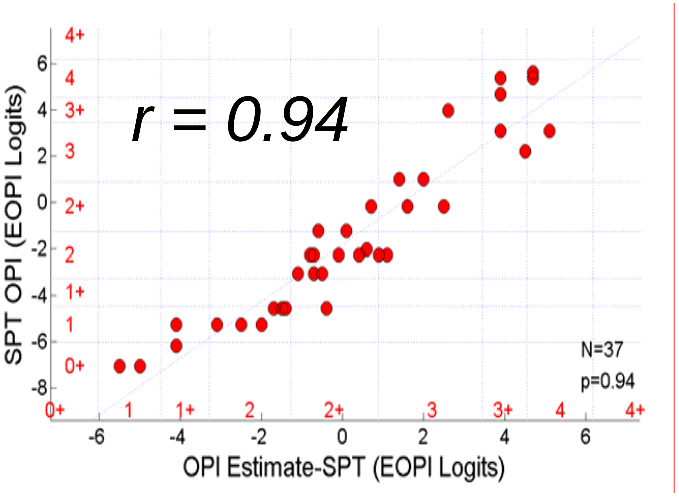
2nd Validation: Spanish Data (SST)

U.S. Government OPI Interviews

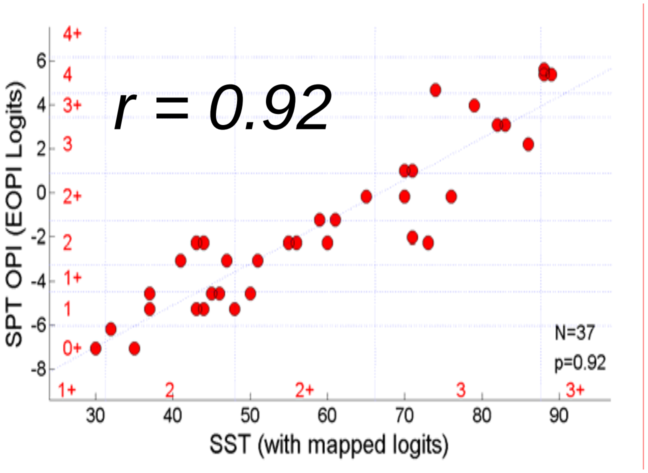
1. **OPI A-Raters ~ A-Raters Estimate**
2. **OPI A-Raters ~ B-Raters Estimate**
3. **OPI A-Raters ~ Machine score**



2. Two Rater Pairs Different Material

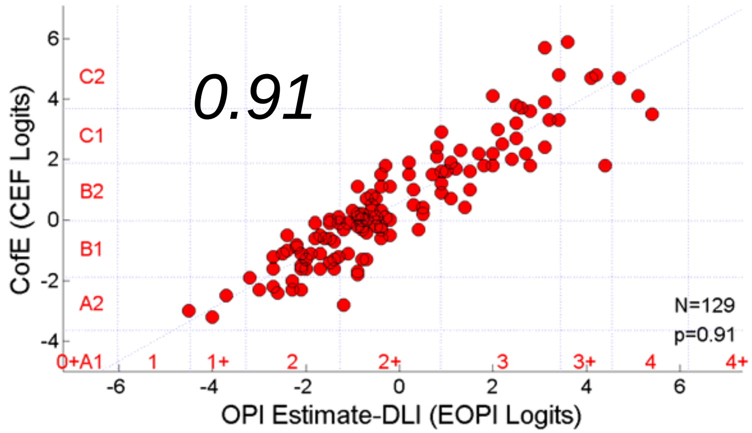


1. Same Raters Different Material



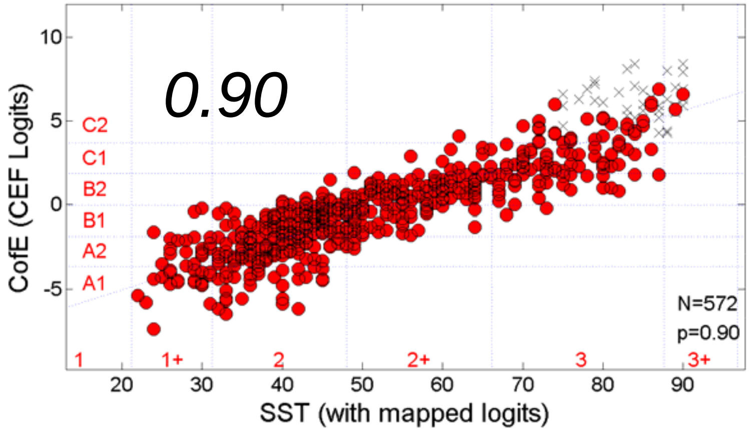
3. Machine ~ Two Raters Different Material

Comparisons to CEF



ILR Estimate-DLI ~ CEF

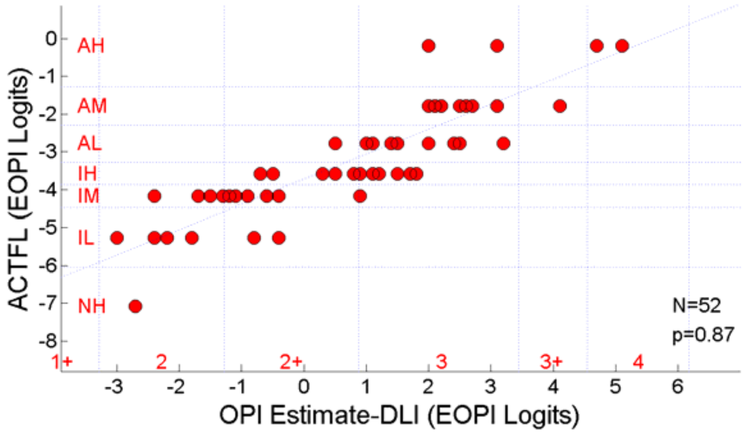
Two Rater Pairs
Same Material



SST ~ CEF

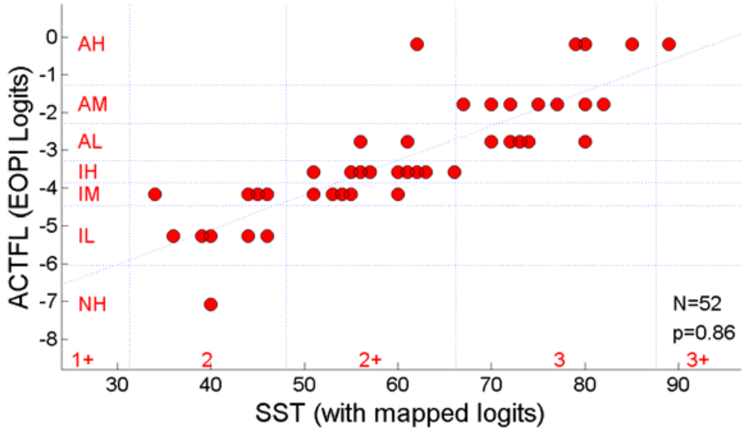
Machine ~ Two Raters
Different Material

ACTFL Interviews



ILR Estimate-DLI ~ ACTFL

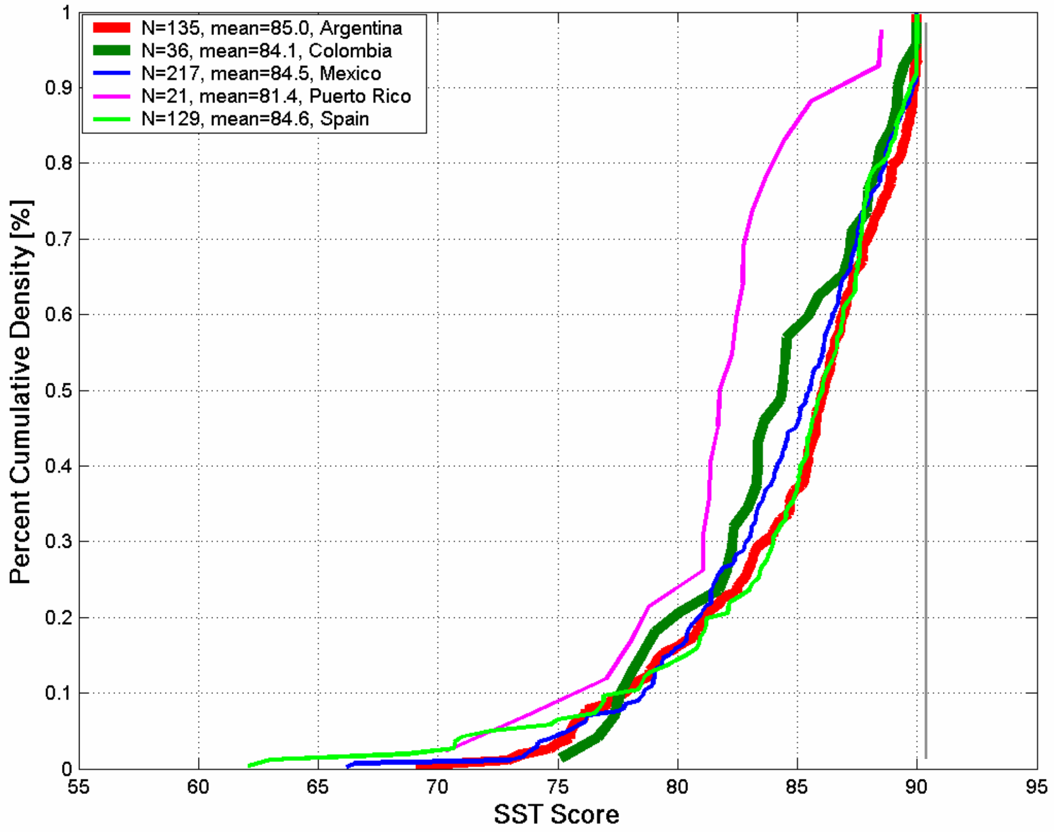
Two Rater Pairs
Different Material



SST ~ ACTFL

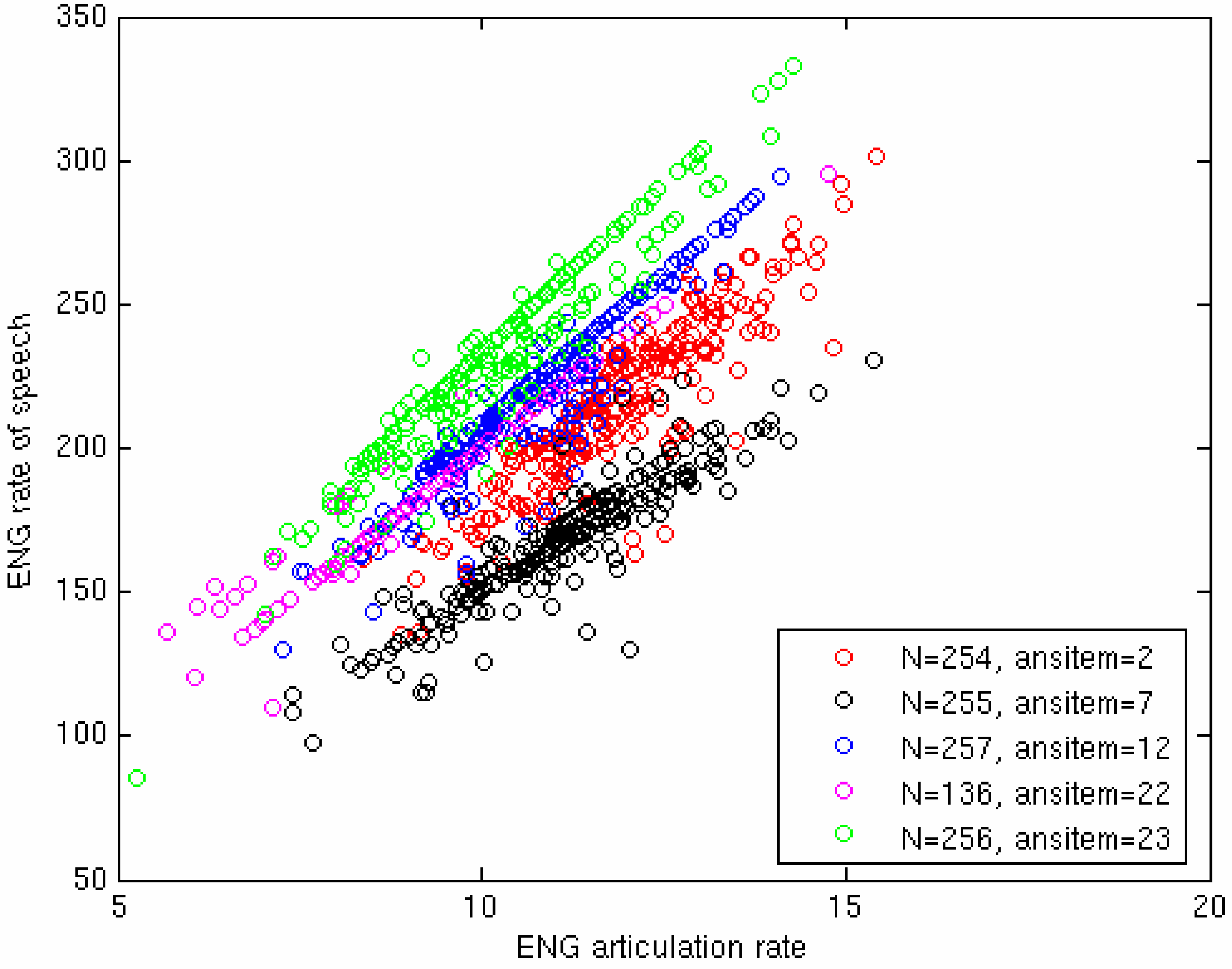
Machine ~ Two Raters
Different Material

Model Fits New Dialect Performances



CDF by Country

Item-specific models are sharper

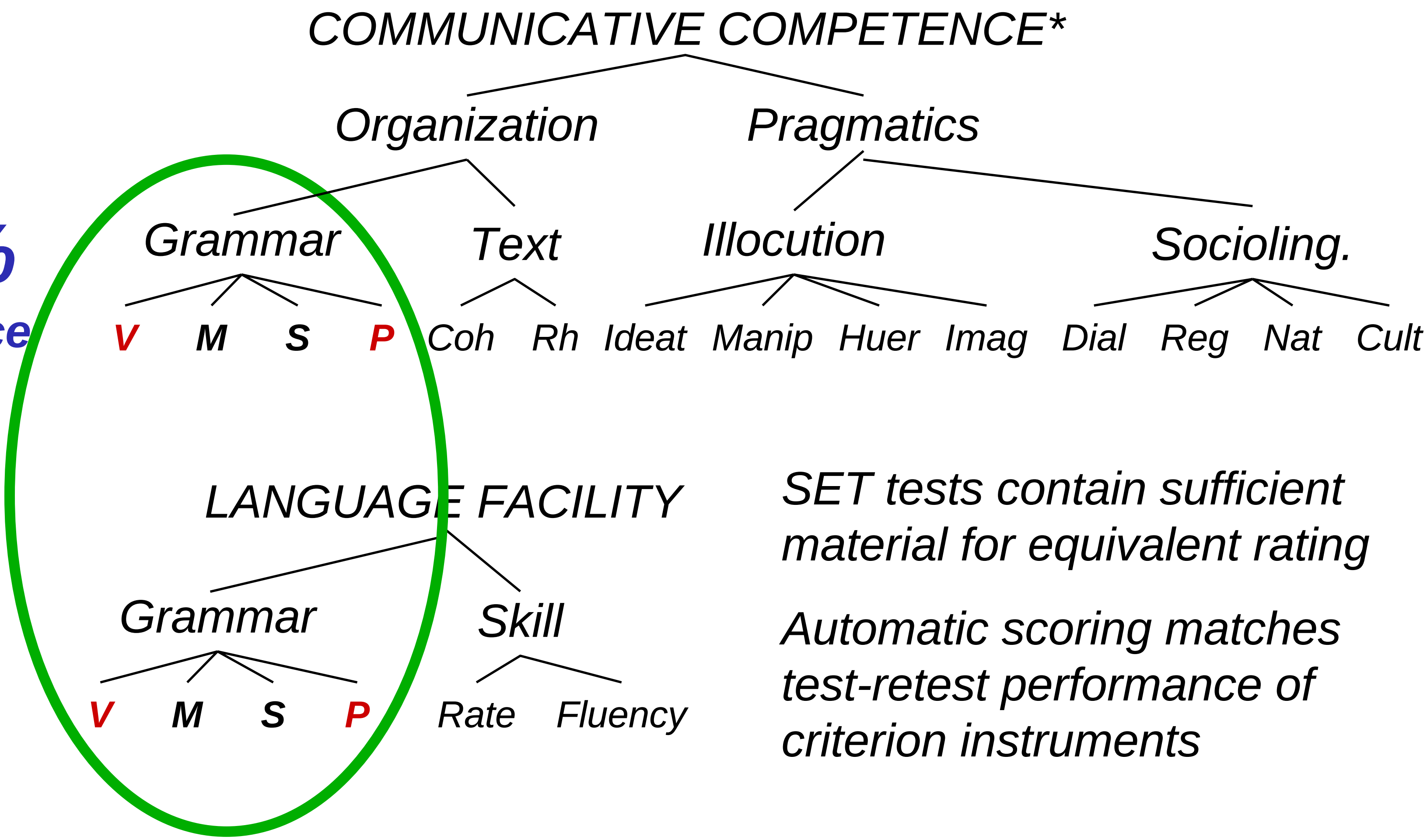


SST Summary & Conclusions

SST (Spoken Spanish Test) contains

- Material sufficient for ILR or ACTFL estimate
    - 49 constrained responses are adequate
    - Six 30-second responses also adequate
  - – Automatic scoring: strong predictor from 49 responses
  - – SST consistently assigns high scores to natives
  - – SST distributes learners of Spanish over a wide range
- Useful alignment with ILR, CEF, ACTFL levels
  - – SST scores can estimate >80% variance of CEF scores

2nd Validation → Performance Puzzle



COMMUNICATIVE COMPETENCE*

Organization

Pragmatics

Grammar

Text

Illocution

Socioling.

V M S P

Coh Rh

Ideat Manip Huer Imag

Dial Reg Nat Cult

LANGUAGE FACILITY

Grammar

Skill

V M S P

Rate Fluency

~80%

of variance

SET tests contain sufficient material for equivalent rating

Automatic scoring matches test-retest performance of criterion instruments

Cross-Construct Puzzle

- The communicative frameworks (e.g. ILR, CEF) generally look for the maximum complexity level of material or function that can be expressed (without time constraint)
- SET-10 measures automaticity of perception and production for relatively simple material
- Yet SET-10 predicts communicative measures at or near their reliability limit

Message complexity depends, in part, on automaticity

- If one measures communicative competence by the functional level or relative complexity of the messages that are communicated, what are the bases of this complexity?

  1. Adequate language-independent cognition
  2. Adequate control of the language system
  3. In **listening** and **speaking**, adequate automaticity of encoding and decoding

Linguistics Reconceived

Read Chapter 1 in:

C. Manning & H. Schutze (1999) Foundations of Statistical Natural Language Processing. MIT Press.

The question is: “what might a person say?”

rather than “what is the structure of the language?”

Linguistics may be coming back to language use, but not thru the lens of Hymes’ communicative competence.

Model Characteristics

- Explicit and predictive
- Language focused; not IQ, not social skills

- Advantages of this kind of modeling
  - Equivalent scoring across time and location
  - Expandable capacity – up to 1000's of tests per day
  - Open to continuous audit – reliability & accuracy
  - Periodic re-estimation of parameters
    - e.g. item difficulty, subscore combination

Automatic Spoken Language Testing

SET-10 and SST build models of native and high-proficiency non-native behavior.

Tests work because models of proficiency-dependent aspects of performance spread the L2 speakers but don’t differentiate L1 speakers (even new dialect samples).

Sentence-level structural diffs >> social and supra-sentential diffs for common L pairs (hypothesis).

. . .