"Deconstructing" the linguistic factors that underlie listeners' L2 comprehensibility ratings

Talia Isaacs, University of Bristol, UK

Pavel Trofimovich, Concordia University, Canada

EALTA, Siena, Italy, May 6, 2011

Comprehensibility

Listeners' perceptions of how easily they understand L2 speech

- A major construct in L2 pronunciation research
- Central to the goal of helping learners be more understandable to their interlocutors

Comprehensibility

Listeners' perceptions of how easily they understand L2 speech

▸ Little is known about the linguistic dimensions that

- most influence listeners' perceptions
- discriminate between different levels

3

Comprehensibility

Listeners' perceptions of how easily they understand L2 speech

1 2 3 ④ 5 6 7 8 9

Difficult to
understand

Easy to
understand

Definitional distinctions

<u>Narrow sense</u> (Munro & Derwing, 1999)

▸ <u>Comprehensibility</u> - Listeners' *perceptions* of understanding → Rating scale

▸ <u>Intelligibility</u> – More objective measure of listener understanding → e.g., Dictation

<u>Broad sense</u> - synonymous (Levis, 2006)

Definitional distinctions

✓ **<u>Narrow sense</u>** (Munro & Derwing, 1999)

▸ <u>Comprehensibility</u> - Listeners' *perceptions* of understanding → Rating scale

▸ <u>Intelligibility</u> – More objective measure of listener understanding → Dictation

Oral proficiency scales - high-stakes tests

Shortcomings: L2 proficiency scales

▶ Pronunciation is omitted from scale descriptors altogether
- Pronunciation is too difficult to model

CEFR descriptors of benchmark levels

Shortcomings: L2 proficiency scales

▸ Vague descriptors

- "Mispronunciations… cause some difficulty for the listener" <mark>IELTS (public version)</mark>

▸ Comprehensibility & accentedness conflated

- "Pronunciation is easily understood; Many features… are 'nativelike' "

<mark>Certificate of Proficiency in English</mark>

Comprehensibility & L2 instruction

▸ L2 teachers could benefit from

- guidance on linguistic factors that most influence listeners' comprehensibility judgments

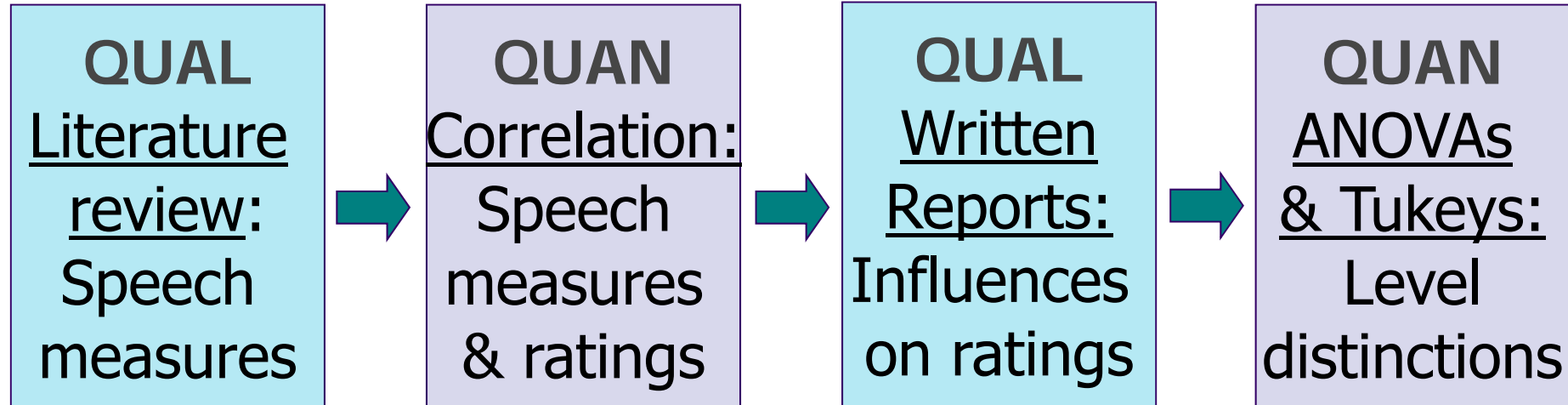- L2 comprehensibility scale for pedagogical purposes

The present study

Research questions

▸ Which linguistic dimensions are most strongly related to listeners' L2 comprehensibility ratings?

▸ Which measures best discriminate between different levels of L2 comprehensibility?

Research design

▶ Sequential mixed methods design (Creswell & Plano-Clark, 2007)

QUAL Literature review: Speech measures	→	**QUAN** Correlation: Speech measures & ratings	→	**QUAL** Written Reports: Influences on ratings	→	**QUAN** ANOVAs & Tukeys: Level distinctions

Speakers

▸ 40 adult L1 French speakers of English from Quebec, Canada

- 1st exposure to English: 8.7 years (0-17)
- English use: 15% (0-100%)
- Speaking proficiency: 6.9 (1-9)

L2 speaking task

▶ 60 native English speaking undergraduate students studying at a Canadian university
  - no TESL experience

English use: 92% of the time ($SD = 8$)

French proficiency: 3.4 on 9 pt scale ($SD = 2$)

(Derwing et al., 2008)

Novice raters

▸ 60 native English speaking undergraduate students studying at a Canadian university
- no TESL experience

English use: 92% of the time ($SD = 8$)

French proficiency: 3.4 on 9 pt scale ($SD = 2$)

L2 speaking task

▸ Picture narrative (Derwing et al., 2008)

Method

▸ Analyzed short excerpts of 40 L2 speech samples using 19 measures

   Extended Iwashita et al.'s (2008) measures

- *Phonology
- Fluency (temporal)
- Linguistic resources (lexico-grammatical)
- *Discourse-level

▸ Correlated with 60 raters' mean L2 comprehensibility judgments

Phonology: 6 measures

1. Segmental error ratio
   - e.g., _fa_n for _fu_n

2. Syllable structure error (deletion, epenthesis)
   - e.g., _pologize for _a_pologize

3. Word stress error ratio
   - e.g., _sky-scra-PER_ for _SKY-scra-per_

Phonology: 6 measures

4. Vowel reduction ratio
 - *in a CI-ty there were TWO PEO-ple*

5. Pitch contour (boundary tones)
 - *It was a sunny day in the city.* [falling tone]

6. Pitch range (boundary tones)
 - Difference – highest & lowest F0

Fluency: 6 measures

7. Total filled pauses

- *It's a nice sunny day in uh uh New York.*

8. Total unfilled pauses

- *They* [unfilled pause] *hit each other.*

9. Pause error ratio

- Filled & unfilled pauses within clauses

Fluency: 6 measures

10. Repetitions/self-corrections
- *I I see a a lot of buildings.*

11. Pruned syllables per second
- Total syllables produced excluding dysfluencies

12. Mean length of run
- Mean syllables produced between pauses

Linguistic resources: 4 measures

13. Grammatical accuracy
    - *They falled on the floor.*

14. Lexical errors ratio
    - *Circulation* instead of *traffic.*

15. Token Frequency (total words produced)
16. Type Frequency (total unique words)

Discourse: 3 measures

17. Story cohesion
- *Suddenly, But, Hopefully*

18. Story breadth – No. of propositions
- Stein & Glenn's (1979) scheme

19. Story depth – No. of proposition types
- *Setting, Attempt, Consequence, Reaction.*

Preliminary analyses

▸ 2nd coding of 40% of speech data
- Intraclass correlation: ≥.9 for all measures
  - Exception: lexical error ratio (.85)

▸ Interrater reliability (60 raters)
- Intraclass correlation: .97

Correlations between speech measures & comprehensibility ratings

Speech measures	r
Type frequency	.78
Word stress error ratio	−.76
Rhythm ratio	.74
Mean length of run	.71
Story breadth	.71
Grammatical accuracy	−.63

18 measures:
$p < .05$

Confirming quantitative measures

▸**Introspective reports** of 3 experienced native speaking ESL teachers

- 10–12 years TESL experience

1	2	3	4	5	6	7X	8	9

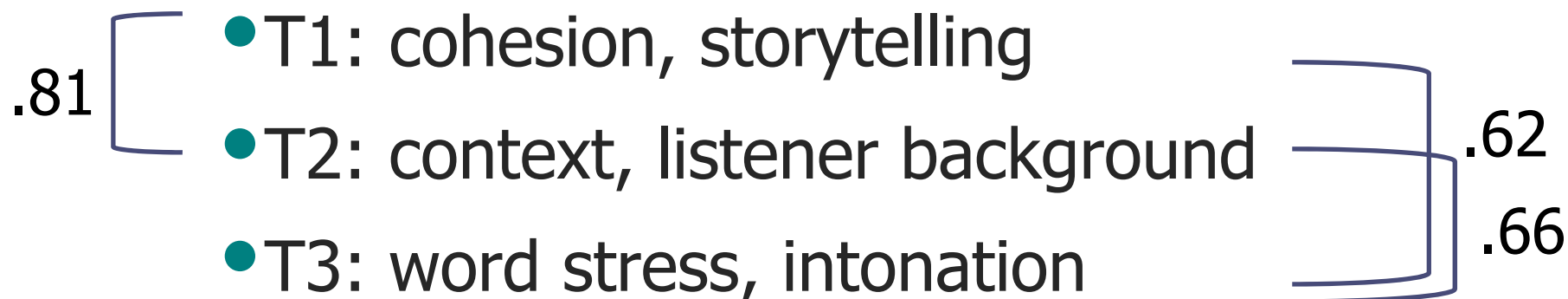1 = hard to understand   9 = easy to understand

Typed influences on ratings in textbox.

Analysis of introspective reports

▸ 10 coded categories
  - Exact intercoder agreement: 95%

▸ *Grammar, vocabulary, fluency
  - T1: cohesion, storytelling
  - T2: context, listener background
  - T3: word stress, intonation

.81

.62

.66

Intraclass correlations with 60 raters: .8 to .9

Data triangulation

▸ **Goal**: Identify linguistic features that most efficiently distinguish between 3 levels of comprehensibility

▸ **Retaining variables for possible inclusion in scale**

- Strong correlation with 60 listeners' mean comprehensibility rating ($r > .7$)
- Referred to in teachers' reports ⟹ ANOVAs

LEVEL	SPEECH MEASURES that DISTINGUISH 3 COMPREHENSIBILITY LEVELS		
High	Word stress	MLR Types	Grammar Propositions
Intermediate	Word stress		Grammar
Low	Word stress	MLR Types	Propositions

Major findings

▶ A wide range of measures feed into listeners'
L2 comprehensibility judgments
- Phonology
- Fluency (temporal)
- Linguistic resources (lexicogrammatical)
- Discourse-level

Major findings

▸ Raters would benefit from more guidance on the defining features of comprehensibility for construct validity reasons

▸ The linguistic factors that most influence listeners' comprehensibility judgments could help teachers set instructional targets

Future directions

▸ Validation studies – Determine generalizability of linguistic criteria across
  - Different L1 groups
  - Different task types

▸ Develop a formative assessment tool
  - Diagnose learner difficulties, monitor learning