Studies in Second Language Acquisition (2024), 1–23
doi:10.1017/S0272263124000305

RESEARCH ARTICLE

Coherence and Comprehensibility in Second Language Speakers' Academic Speaking Performance

Aki Tsunemoto[1,2] and Pavel Trofimovich[1]

[1]Department of Education, Concordia University, Montreal, Quebec, Canada and [2]Faculty of Foreign Language Studies, Kansai University, Suita, Osaka, Japan
Corresponding author: Aki Tsunemoto; Email: a.tsune@kansai-u.ac.jp

(Received 20 November 2023; Revised 20 February 2024; Accepted 14 March 2024)

Abstract

This study examined the role of discourse organization in second language (L2) comprehensibility ratings. Twelve English for Academic Purposes teachers listened to 60 L2 speech samples elicited through a TOEFL–type integrated speaking task, evaluating each sample for comprehensibility and coherence (perceived interconnectedness of ideas). The samples were analyzed for the occurrence of discourse features at micro and macro levels. Results revealed a strong association between coherence and comprehensibility (r = .70). Whereas L2 speakers' use of additive connectives (e.g., *and*) uniquely predicted comprehensibility, ordering of ideas and source–speech similarity in speakers' performances predicted coherence. Lexical overlaps predicted both constructs. Findings underscore the importance of coherence to comprehensible academic L2 speech demonstrating that the two constructs include partially overlapping yet distinct characteristics.

Introduction

Second language (L2) speakers are often encouraged or explicitly taught to produce coherent, logically connected speech, because disjointed ideas can be difficult for listeners to follow. Coherence, which refers to "the representational relationships of a text in the mind of a reader or listener" (Crossley, Salsbury, McCarthy, & McNamara, 2008, p. 1906), is a key component of academic oral performance, such as giving a presentation or providing an argument. According to Canale and Swain (1980), coherence is part of sociolinguistic competence encompassing language use. A speaker needs to understand how to connect ideas logically, making the discourse cohesive in production, and also to combine communicative functions of utterances, creating the intended understanding of the discourse in comprehension. Given its role in definitions of language competence, coherence has been extensively studied in L2 writing and reading research. Coherent texts are better comprehended and are evaluated more favorably (Crossley & McNamara, 2011; McNamara & Kintsch, 1996; Richards, 1990),

with coherence predicted through various linguistic features such as connectives and discourse markers in L2 written discourse (Crossley, Greenfield, & McNamara, 2008). Thus far, however, little is known about how coherence in L2 oral discourse is related to listeners' perception of speech, which is unfortunate, because coherence—along with other dimensions of speech (e.g., comprehensibility, fluency)—is considered a component of L2 oral proficiency (Hulstijn, Schoonen, & de Jong, 2012) and targeted in speaking tests (e.g., International English Language Testing System [IELTS], Test of English as a Foreign Language [TOEFL]). Comprehensibility (i.e., difficulty with which listeners understand L2 speech), in particular, has been discussed as a useful and intuitive perceptual measure of L2 speech (Kennedy & Trofimovich, 2019; Saito, 2021). However, it is unclear whether coherence and comprehensibility are related and whether listener perceptions of coherence and comprehensibility are predicted by a distinct set of discourse features (e.g., connectives and signposting). Our goal was therefore to address these issues. By investigating which discourse features feed into rater judgements of comprehensibility versus coherence, we wished to contribute to research on L2 speech development and language assessment in academic contexts.

Background literature

Coherence in L2 written and oral discourse

According to research in L2 writing and reading, textual cohesion, which captures a writer's use of cohesive devices (e.g., connectives such as *and* or *because*), enhances the readability of texts and reduces the processing effort for readers (Crossley et al., 2008; Kuiken & Vedder, 2017; Richards, 1990). For instance, McNamara and Kintsch (1996) compared texts that differed in degree of cohesion, where a low-cohesion text lacked explicit connections between sentences whereas a high-cohesion text included elements that made logical relationships explicit by identifying anaphoric reference, such as a pronoun, word, or phrase that refers back to a noun mentioned earlier in the text (McNamara, Crossley, & Roscoe, 2013), or by repeating the same term to describe a concept. Participants who read the low-cohesion text scored significantly worse in comprehension quizzes and took longer to read the text compared to those who read the high-cohesion text. In addition, lexical overlaps between sentences facilitate text processing, because lexical repetition helps readers track conceptual relationships within and across texts (Halliday & Hasan, 1976). For instance, using a natural language processing tool called Coh-Metrix (Graesser, McNamara, & Louwerse, 2004), Crossley et al. (2008) explored the occurrence of cohesive devices in L2 texts. They found that readers' difficulty with academic texts was predicted by the frequency with which content words overlapped between two adjacent sentences, which explained 63% in text difficulty scores. Also, a writer's use of connectives (e.g., *because*, *therefore*) can help readers make logical links between sentences (Halliday & Hasan, 1976). For example, readers spend significantly less time reading a text where connectives are provided, suggesting that connectives help them process it more easily (van Silfhout, Evers-Vermeuol, & Sanders, 2015).

Coherence is relevant to not only written texts but also spoken discourse (Halliday & Hasan, 1976). Coherent academic L2 speech is said to have a clear structure and to include discourse markers, both of which contribute to discourse organization (Tyler, 1992; Tyler & Bro, 1992, 1993). Williams (1992) found that explicit use of discourse markers (e.g., *for example*, *I give you the definition of…*) helped listeners understand lectures given by L2 English-speaking international teaching assistants (ITAs), where

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

clear signals of discourse structure may have compensated for speakers' pronunciation difficulties in helping listeners understand the content. In a test-validation project, Brown, Iwashita, and McNamara (2005) asked experienced teachers of English for Academic Purposes (EAP) to evaluate English oral proficiency in 40 speaking performances elicited through a prototype of the eventual TOEFL integrated speaking task. The teachers frequently commented about the organization of ideas, such that logically ordered and structured performances were easier to follow and understand. They also favored the performances that included a clear introduction and a conclusion, commenting that it was important for speakers to use discourse markers (e.g., *first*, *second*, *finally*), which can signal a change in topics.

L2 pronunciation also contributes to listener perception of L2 discourse. For example, Hahn (2004) focused on lectures given by ITAs to investigate how violations in expected prosodic patterns (i.e., nuclear stress to indicate a given–new information contrast) impact listeners. Unlike those who were exposed to the expected prosodic patterns, listeners who experienced inaccurate or missing nuclear stress took longer to understand the speech and evaluated it more negatively, implying that violations of prosody make discourse less cohesive and therefore harder to process. Similarly, Tyler, Jeffries, and Davies (1988) found that ITAs who were perceived to be disorganized and unfocused by listeners tended to have problems with pronunciation and fluency (e.g., pausing, nuclear stress placement), which again suggests that listeners make use of pronunciation when evaluating coherence in L2 speech.

Coherence and comprehensibility in L2 speech

In addition to coherence, comprehensibility is often included in assessment rubrics of language exams, on the assumption that it is a key component of L2 oral proficiency contributing to ease of communication (Saito & Plonsky, 2019) and that it captures listeners' understanding of L2 speech (Kennedy & Trofimovich, 2019). When evaluating comprehensibility, listeners rely on various linguistic dimensions in L2 speech, including phonology, lexis, grammar, fluency, and discourse (e.g., Isaacs & Trofimovich, 2012). Although the most recent meta-analysis suggests that measures of phonology and fluency are the most relevant to comprehensibility, together accounting for 30% to 50% in listener-rated comprehensibility (Saito, 2021), the remaining variance might be explained through other linguistic dimensions, including discourse.

Thus far, researchers have focused on a handful of discourse measures in relation to comprehensibility. For instance, using Coh-Metrix, Appel, Saito, Isaacs, Webb, and Trofimovich (2019) examined the use of causal (e.g., *because*), logical (e.g., *and*), and additive (e.g., *furthermore*) connectives in relation to comprehensibility ratings in a picture narrative task and a TOEFL integrated speaking task. There was no relationship between the occurrence of connectives and comprehensibility in the picture narrative task, but a weak negative association ($r = -.25$) emerged between the two variables in the TOEFL task, although the use of connectives did not predict comprehensibility in a multiple regression analysis. Isaacs and Trofimovich (2012) explored three discourse features, including the frequency of adverbials (e.g., *suddenly*) and the number and diversity of distinct propositions produced, reporting moderate to strong relationships between discourse features and listener-rated comprehensibility in a picture narrative task (r = .50–.71), although lexis and phonology showed even stronger ties with comprehensibility. However, considering the number of cohesive devices explored in

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

L2 writing and reading research, L2 speech researchers have to date examined only a limited set of discourse features in relation to comprehensibility.

Like text, speech may contain multiple cohesive devices of relevance to listener perception of both coherence and comprehensibility. Tyler and Bro (1992) manipulated transcripts of an argumentative speech by a Chinese speaker, creating texts that differ in the use of logical connectives (e.g., *and*), tense and aspect forms (e.g., past tense indicating event sequence), lexical specification (e.g., pronoun *it*), and information order (e.g., arguments first vs. reasons first). The texts were subsequently rated for comprehensibility (defined as perceived ease of following the text) by English-speaking undergraduate students. Whereas the use of cohesive devices influenced comprehensibility, such that the text containing miscues (e.g., unspecified anaphoric reference, ambiguous use of *and*) was more difficult to follow than the text in which miscues were corrected, the order in which information was presented was unrelated to comprehensibility. In a follow-up study, Tyler and Bro (1993) used the same texts but introduced reading time to measure processing effort, showing that the text with miscues required more time to read than the reconstructed version. Even though these studies imply a link between discourse cohesion and comprehensibility, this work focused on transcribed, not actual, speech, because the researchers wished to avoid pronunciation as a major influence on listener perceptions.

Although the relationship between comprehensibility and various discourse features appears to be weaker than the link between comprehensibility and other linguistic dimensions, such as phonology and fluency (Saito, 2021), qualitative insights into listeners' perception reveal a potentially nontrivial role of discourse in their judgements. For example, in post-rating interviews, a rater commented that discourse structure (e.g., use of cohesive devices) was an important factor to her comprehensibility ratings (Isaacs & Trofimovich, 2012). Similarly, raters evaluating L2 comprehensibility dynamically in 2–3 min speech samples also frequently brought up discourse to explain their ratings (Nagle et al., 2019), describing coherence as a key reason for upgrading comprehensibility, such that the more logically the ideas were connected, the easier the speaker was to understand. Therefore, a coherent mental representation of speech may contribute to the ease with which it is understood. What is presently unknown, however, is whether and to what extent discourse cohesion, which is determined by how speakers use various cohesive devices, is important to listener-rated coherence and comprehensibility.

The current study

Coherence has attracted considerable attention from researchers and practitioners in L2 reading and writing, because coherence can impact how readers perceive text difficulty, evaluate its quality, and generally comprehend written discourse (Crossley et al., 2008; Halliday & Hasan, 1976; McNamara & Kintsch, 1996; van Silfhout et al., 2015). A coherent text includes various cohesive devices, such as anaphoric reference, lexical overlaps, and connectives (Crossley & McNamara, 2011). However, there is little knowledge about how coherence is related to or distinguished from comprehensibility in L2 speech. Because clearly structured L2 speech, often explicitly signalled through discourse markers, is considered as more comprehensible than less structured speech (Brown et al., 2005; Tyler, 1992; Tyler & Bro, 1992, 1993; Williams, 1992), analyzing spoken discourse for both local features (e.g., word- and sentence-level cohesive devices) and global features (e.g., signposting, organization of ideas) could help clarify

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

how various dimensions of discourse contribute to L2 coherence and comprehensibility. To date, researchers have shown only weak relationships between discourse features and comprehensibility, but these findings are tentative, because they pertain to only a few cohesive devices (Appel et al., 2019; Isaacs & Trofimovich, 2012), and many other micro-level (e.g., different connectives) and macro-level (e.g., structure, discourse markers) measures of spoken discourse remain underexplored.

Methodologically speaking, previous studies focusing on coherence have used Coh-Metrix (Graesser et al., 2004) to explore the occurrence of cohesive devices. However, the recently developed Tool for the Automatic Analysis of Cohesion (TAACO) 2.0 has enabled researchers to examine many more cohesive devices through a larger and more representative toolkit than Coh-Metrix (Crossley et al., 2019). In addition, most prior work on coherence and comprehensibility has targeted L2 writing (Crossley et al., 2008; McNamara & Kintsch, 1996; Richards, 1990); when L2 speech was examined, transcripts were used to avoid the influence of pronunciation on listeners' comprehensibility assessment (Tyler & Bro, 1992, 1993). However, in real-life settings, those who interact with or evaluate L2 speakers experience speech aurally, most frequently in the absence of textual support in the form of transcripts or subtitling. Thus, it would be important to explore whether discourse cohesion is relevant to listener-rated coherence and comprehensibility when listeners evaluate L2 speech rather than transcripts, while also controlling for speakers' pronunciation.

Our goal in this study was therefore to examine the relationship between listener ratings of speech coherence and comprehensibility in an academic L2 speaking task and to determine if these ratings can be distinguished through multiple discourse features of L2 speech. To develop a rating scale for the assessment of coherence, a pilot study with three EAP teachers was conducted first (described in Appendix A; all appendices are available in Online Supplementary Material). In the main study, 12 other EAP professionals provided evaluations of coherence and comprehensibility for 60 samples of audio-recorded TOEFL integrated task performances by university-level L2 speakers, and the same samples were analyzed for speakers' use of macro-level discourse features (i.e., structure, discourse markers) and their use of micro-level cohesive devices (i.e., connectives), as coded by the researcher or derived through TAACO 2.0. The EAP teachers also assessed the audio samples for accentedness, which captures how closely speakers approximate the target language variety (Derwing & Munro, 2015). Because accentedness ratings can be largely explained through measures of L2 phonology and fluency (Saito, 2021), these ratings served as a control covariate to account for between-speaker variability in pronunciation and thus to sidestep the limitation of prior research that relied on transcripts to evaluate coherence and comprehensibility. This study was guided by two research questions:

1. What is the relationship between L2 speakers' coherence and comprehensibility, as assessed by EAP teachers evaluating speakers' performance in an academic task?
2. To what extent are macro- and micro-level discourse features associated with EAP teachers' judgements of coherence and comprehensibility of L2 speech?

Method

Speech samples

The target audios were sampled from the Montreal Speech Corpus (Isaacs & Trofimovich, 2011), which contains audio recordings of 149 L2 English international

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

students from Canadian English–medium universities completing five academic tasks. The selected task was a 1-min academic speaking activity in which the speakers completed a publicly available version of the TOEFL integrated speaking task (Educational Testing Service, 2006). The speakers first took 45 s to read a 100-word passage about psychology or sociology (depending on task version) and then listened to an audio-recorded lecture related to the passage (see Appendix B). After 30 s of planning time, they produced a 1-min narrative, responding to the prompt asking them to integrate the information from the reading passage (which contained a brief definition and description of a specific psychological or sociological phenomenon) and the audio lecture (which provided several examples of that phenomenon).

Sixty audio samples featuring this task (20 women, 40 men) were drawn from the corpus, with the constraint that they illustrated performances by L2 speakers from different first language (L1) backgrounds, with similar representation from the selected groups (Farsi = 15; Hindi, Urdu, and Punjabi = 16; Mandarin = 14; and Spanish, French, and Portuguese = 15), that they were comparable in duration (i.e., approximately 60 s) and all exceeded 50 words minimally required for natural language processing analysis (Crossley et al., 2017), and that they were balanced across the two task versions (psychology = 28, sociology = 32). The speakers (M_{age} = 23.68 years, SD = 3.32) had studied English for about 11.16 years (SD = 5.11) and were enrolled in undergraduate (15) or graduate (45) programs at the time when the recording took place. As degree-seeking students, they had met the minimum English requirement for university admission, which was a TOEFL internet-based test (iBT) score of 75 (or equivalent). Using a 0–100% scale, they also reported a relatively high frequency of daily English use in speaking (M = 59%, SD = 23) and listening (M = 68%, SD = 22). The recordings were about 58.50 s in length (SD = 3.90) and included 117.97 word tokens on average (SD = 26.24).

Raters

In the main study, 12 EAP instructors participated as raters. EAP instructors were chosen as raters, because they represent domain experts who are in contact with university-level L2 speakers, who regularly evaluate students' academic performance, and who (unlike naïve, untrained raters) possess linguistic knowledge, teacher-training background, and pedagogical expertise to evaluate a complex, holistic construct such as coherence. All raters were recruited through pre-existing social media and professional email groups. The 12 instructors (8 female and 4 male) were all academic faculty (1) or graduate students (11) at English-medium universities in Montreal, Canada. They included nine L1 English speakers, two English–French bilinguals, and one English–Tagalog bilingual. Considering a growing number of L2 English-speaking instructors in higher education (Copland, Mann, & Garton, 2019), a mixed monolingual/bilingual group was representative of the general population of EAP instructors. The instructors (M_{age} = 36.08 years, SD = 7.03) had lived in Canada for about 31.12 years (SD = 13.80), and all had graduate degrees in applied linguistics or education and experience teaching EAP speaking skills in higher education (M = 7.41 years, SD = 4.30). Using 100-point scales (0 = *not at all familiar* and 100 = *very familiar*), they estimated their familiarity with accented L2 English by speakers from the target L1 backgrounds (Bergeron & Trofimovich, 2017), providing high ratings for Mandarin-accented English (M = 88.83, SD = 14.30), Farsi-accented English (M = 72.42, SD = 25.28), and Indian accents in English (M = 79.08, SD = 18.68). Because Montreal is a French–English bilingual city,

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

they provided separate familiarity ratings for French and Spanish (i.e., languages most frequently represented in the Romance speaker sample), again yielding high values for French-accented English (M = 91.33, SD = 25.61) and Spanish-accented English (M = 83.75, SD = 25.61).

Coding of speech samples

The 60 audios were analyzed for 12 measures to capture macro- and micro-level features of discourse potentially relevant to raters' evaluations of L2 coherence and comprehensibility. Analyses were carried out using transcripts of speakers' performance, with each first verified for accuracy to reflect the selected sample. Some measures were derived through manual coding, whereas others were computed through TAACO.

Macro-level measures

In terms of macro-level measures, there were two categories (all coded manually by the first author) focusing on discourse structure and discourse markers used for signposting (see Appendix C for all coded categories). These measures were targeted, because listeners might evaluate positively academic discourse in which the expected structural components of a task (e.g., introduction, example, summary) are present (Brown et al., 2005) and in which relationships among ideas are signalled through the order of these components or various discourse markers (Tyler & Bro, 1992; Williams, 1992).

1. Structure–ordering. This measure captured the occurrence and sequencing of the five components expected in TOEFL iBT integrated speaking task performances (Brown & Ducasse, 2019; Iwashita & Vasquez, 2015): (a) introduction, (b) concept explanation (from reading passage), (c) 1st example (from audio lecture), (d) 2nd example (from audio lecture), and (e) summary. A speaker's performance was assigned 1 point if the response contained all expected components and conformed to the expected schematic order (introduction, concept explanation, 1st example, 2nd example, summary) and 0 points if no obvious structure was present, with 0.5 deducted for unexpected order (e.g., example presented before introduction). This combined measure of the occurrence and ordering of task-essential components evolved from the pilot study, where the EAP instructors suggested that coherent L2 speech involves both the presence of essential components and their ordering. Initial inspection of the audios also revealed two broad patterns, where speakers either provided all five components in the expected order (i.e., score of 1), sometimes with a change in their sequencing (i.e., score of 0.5), or in fact omitted several components producing the remaining ones without any obvious sequential structure (score of 0). Therefore, a simple coding scheme appeared sufficient to capture broad variations in speakers' performance.
2. Signposting. This measure targeted the incidence of discourse markers relevant to the organization of L2 academic performances, including those elicited through the TOEFL iBT integrated speaking task (Jung, 2003): exemplifiers (e.g., *for example*), sequential markers (e.g., *the first time*), contrast expressions (e.g., *compared to*), summarizers (e.g., *in conclusion*), and source attributions (e.g., *in the reading passage*).

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

Two measures of text overlap were additionally derived through TAACO 2.0, because rater-assessed coherence and comprehensibility in an integrated speaking task may be predicted by the overall similarity between each of the two source texts (reading, lecture) and a speaker's performance (Crossley, Kyle, & Dascalu, 2019).

3. Similarity–reading. This measure captured the lexical overlap between a speaker's performance and the reading prompt. Computed through the Word2vec algorithm integrated in TAACO (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), this measure yields a similarity index ranging between 0 (completely different) and 1 (identical) that expresses the degree to which words co-occur in the prompt reading passage and a speaker's performance.
4. Similarity–listening. This was a similar measure of the lexical overlap between a speaker's performance and the prompt audio lecture, derived through the same algorithm.

Micro-level measures

Micro-level categories of discourse cohesion (all computed in TAACO 2.0) included measures of connectives, lexical overlaps, and givenness as indicators of anaphoric reference associated with greater comprehensibility and coherence in prior work (Crossley & McNamara, 2011; Tyler & Bro, 1992, 1993). To allow for automated analyses (Crossley, Clevinger, & Kim, 2014), transcripts were adjusted following the coding scheme adapted from Inoue and Lam (2021), such that metalinguistic notations (pauses, fillers) were removed but other disfluencies (false starts, repetitions, repairs) were retained.

The category of connectives included four measures illustrating relationships between words, clauses, and sentences; connectives were tallied separately across four unique sets to avoid counting the same token across multiple categories.

5. Disjunction. This measure targeted the frequency of the disjunctive connective *or* signaling a relationship between two distinct alternatives (e.g., "when people are being observed *or* they have the knowledge that they are being observed by others").
6. Addition. This measure captured the number of additive connectives (e.g., *and*, *also*) used to add information or to connect ideas (e.g., "the professor said that he was in a line *and* someone crosses the line").
7. Causal connectives. This measure focused on the number of causal connectives (e.g., *because*, *so*) linking a cause with an effect (e.g., "a group of people knew they were watched *so* they were a little bit of stress").
8. Opposition. This measure targeted the number of contrastive connectives (e.g., *but*) that signal opposing or contrasting ideas (e.g., "we don't know the situation *but* perhaps they have something").

Lexical overlaps refer to the repetition of words or phrases across sentences or texts (McNamara et al., 2013). This category included two measures (separately for nouns and verbs), because overlaps of different parts of speech might contribute differently to making discourse more or less coherent for a listener (Crossley et al., 2016). TAACO 2.0 calculates "overlap between words and set of word synonyms between sentences" (Crossley, Kyle, & McNamara, 2016, p. 1231) based on the WordNet database (Fellbaum, 1998). Therefore, TAACO-derived overlap indexes encompass not only

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

identical word repetitions but also overlaps between semantically related words; for instance, synonyms for *idea* include related nouns *thought* and *estimate*, and synonyms for *watch* comprise related verbs *observe* and *view*.

9. Noun overlaps. This measure captured the frequency with which nouns (including their synonyms, as determined through the WordNet database) were repeated within and across sentences (e.g., "according to the passage we read that *people* have different reactions toward their *behavior* and other *people behavior*").
10. Verb overlaps. This measure concerned the frequency with which verbs (including their synonyms) were repeated within and across sentences (e.g., "the social behavior of people when they are *watched* is clearly portrayed in the two examples… as people are being *observed*").

Givenness refers to "the amount of information that is recoverable from the preceding discourse" (Crossley et al., 2016, p. 1231). Considering that givenness is manifested through a speaker's use of several linguistic resources (McNamara et al., 2013), such as third-person pronouns that replace previously introduced information (*she* to refer to a store customer) or demonstrative pronouns that specify a noun (e.g., *this person* to refer to a particular individual), this category included two measures.

11. Pronoun–noun ratio. This measure was defined as the total number of third-person pronouns divided by total number of nouns (e.g., "the *professor* told us about the two *situations*, *one* in which *he* saw a *man* cutting the *line* in front of *him*"), where a smaller ratio indicates greater clarity, in the sense that content is expressed explicitly through nouns rather inferred from pronouns.
12. Attended demonstratives. This measure focused on the frequency of lexically specified demonstratives, which refer to any use of a demonstrative (*this*, *that*, *these*, *those*) directly preceding a noun phrase (e.g., "in *these* two situations…"). Here, more frequent use of attended demonstratives indicates greater clarity, meaning that relevant content is specified explicitly through a demonstrative determiner.

Procedure

The main study was carried out online in individual sessions using LimeSurvey (https://www.limesurvey.org), where raters evaluated the 60 target audios for speaker coherence, comprehensibility, and accentedness (control covariate). Coherence was defined for raters using the definition and examples developed through pilot testing (as described and illustrated in Appendix A):

> Coherence refers to how well a speaker makes links and expresses relationships between different ideas. If ideas are clearly related to each other, if they are logically connected, and there is no missing information, then a speaker is highly coherent. However, if ideas are not well connected, if they are presented out of sequence, and if you have to fill in gaps to piece together missing information, then a speaker is not coherent.

The remaining two dimensions were considered more intuitive and less complex than coherence, so they were introduced through previously established definitions. Comprehensibility was described as how effortful it is to understand an L2 speaker, while

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

accentedness was presented as how closely the speaker approximates the target language variety (Derwing & Munro, 2015). To evaluate these dimensions, raters used three 100-point scales with no numerical markings (for validation of similar scales, see Saito, Trofimovich, & Isaacs, 2017), apart from the anchor descriptors (to capture impressionistic judgements of speech), with the negative descriptor always on the left (corresponding to the rating of 0) and the positive descriptor on the right (corresponding to the rating of 100): coherence (*not coherent at all–very coherent*), comprehensibility (*hard to understand–easy to understand*), and accentedness (*heavily accented–not accented at all*).

Raters first read and signed the consent form and completed a background questionnaire (Appendix D), then proceeded to rate the audios. Because topic familiarity and background knowledge impact how information is processed (McNamara & Kintsch, 1996), raters were familiarized with the two task prompts (psychology and sociology). They read the text and listened to the lecture, then were asked to imagine answering each prompt based on the two sources. After reading the definitions of the terms, with examples of each construct, raters practiced assigning their ratings using two additional recordings. The 60 audios were presented in two blocks (organized by prompt), with six raters randomly assigned to one order (psychology first) and the remaining six raters assigned to the other order (sociology first). The audios, presented to raters in unique random order in each block, appeared as embedded audio files with the three rating scales placed under each file. The initial slider position was always in the middle. Raters could not stop or replay audios, and only one listening per audio was allowed. Between the two blocks, raters took a 5–10-min compulsory break to minimize fatigue. After evaluating all audios, they were asked to describe any concerns about their experience in a comment box.

Although online research tasks, compared to those administered in an in-person session, might arguably limit researchers' control over specific aspects of data collection (e.g., timing), online elicitation tools have shown high internal consistency, yielding datasets comparable to those obtained in a lab (Nagle, 2019; Nagle & Rehman, 2021). Nevertheless, several additional controls were implemented to increase data quality. For all questionnaire items, raters were not allowed to return to previous pages, change their answers, or skip questions, and their progress was time-tracked. Raters were also strongly encouraged to use headsets or earbuds, and they were advised to complete the survey in a quiet location.

Data analysis

All speech ratings were first checked for internal consistency using two-way, consistency, average-measures intraclass correlations, which yielded high values for coherence (.87), comprehensibility (.88), and accentedness (.89). For manually coded discourse measures, a trained research assistant independently coded 15 transcripts (25% of the dataset) to check inter-coder agreement. Because each variable occurred infrequently, mostly 0–2 times per audio, the counts were treated as categorical, and reliability was explored using Cohen's weighted kappa (κ). The κ values for individual structural components were .68–1.00, indicating substantial-to-perfect agreement (Landis & Koch, 1977). The κ values for signposting counts were .62–.84, which were again substantial-to-perfect in strength, except for contrast expressions (.47), where it was moderate. After resolving all disagreements through discussion, the coder then independently coded an additional six transcripts (10% of the dataset) for contrast

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

expressions, and the obtained κ value reached .80, which was high, with all transcripts re-coded for that measure to reflect the resolved disagreements. Because the incidence of the five signposting counts (exemplifier, sequential marker, contrast expression, summarizer, and source attribution) was limited per transcript (0–2, except for exemplifier, with a 0–4 range), an aggregate frequency count was computed per speaker for signposting. To enable meaningful comparisons across speakers, the manually coded aggregate signposting measure and all TAACO-derived measures were normalized by dividing each relevant value by the total number of words in each audio (Crossley et al., 2019). The structure–ordering measure (coded manually) and the source–speech similarity (computed through the Word2vec algorithm in TAACO) were not normalized, because they were presumably independent of sample length, given that all performances were about 1-min long (i.e., presence or absence of specific discourse components and their ordering was not considered to be a direct consequence of the speed and volume of content delivery).

To address the first research question, which examined the relationship between coherence and comprehensibility, linear mixed-effects models were computed in R (version 4.3.2, R Core Team, 2023) using the lme4 package (version 1.1-35.1; Bates, Maechler, Bolker, & Walker, 2015). Comprehensibility served as the outcome variable whereas coherence was as a fixed-effects predictor, and raters (12) and speakers (60) were entered as random-effects predictors. There were four missing data points due to a problem with audio playback, which yielded a total of 716 observations. As for control covariates, in addition to accentedness, which was used to capture speakers' pronunciation, speakers' L1 group (Farsi, Indian, Mandarin, or Romance), raters' familiarity with L2 accent (Farsi, Indian, Mandarin, or Romance), and prompt (psychology, sociology) were entered as fixed-effects covariates.

To address the second research question, which focused on discourse features associated with rater assessments of speakers' coherence and comprehensibility, another set of linear mixed-effects models was fitted, where comprehensibility and coherence served as separate outcome variables, and raters (12) and speakers (60) were entered as random-effects predictors. Because there was no expectation as to which discourse measures would be associated with coherence versus comprehensibility, all possible measures were considered as fixed-effects predictors in a single, exploratory model. As for control covariates, speakers' L1 group (Farsi, Indian, Mandarin, Romance) and prompt (psychology, sociology), which emerged as relevant in exploratory analyses for this question, were again entered as fixed-effects covariates. Because all predictors of coherence and comprehensibility were derived from written transcripts (rather than speech) and because the contribution of speakers' pronunciation to coherence and comprehensibility is explored under the first research question, speakers' accentedness was not considered as a covariate in this analysis. Correlations across predictors were checked to avoid multicollinearity (see Appendix E for a full matrix).

Whereas untransformed coherence and comprehensibility ratings served as the outcome variable in each model, all continuous predictors were *z*-transformed. Among categorical predictors, the Mandarin L1 group and the psychology prompt were designated as the reference (baseline) group. To perform multiple comparisons across the speakers' L1 (four levels), contrast coding was used, and post hoc comparisons were performed with Tukey-corrected *p* values using glht function of the multcomp package (version 1.4-25; Hothorn, Bretz, & Westfall, 2008). To determine statistical significance, *p* values were obtained using MuMIn package in *R* (version 1.47.5; Bartoń, 2023)

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

but 95% confidence intervals (CIs) were also examined to check the statistical significance of each parameter (interval does not cross zero).

Results

Coherence–comprehensibility link

The first research question targeted the relationship between L2 speakers' comprehensibility and coherence in an academic speaking task. As shown in Table 1, comprehensibility was generally rated higher than coherence. Among the four L1 groups, the Mandarin group received the lowest ratings whereas the Romance group received the highest ratings for both comprehensibility and coherence.

As shown in Table 2, which summarizes the final mixed-effects model, coherence significantly predicted comprehensibility even after controlling for speakers' accentedness, their L1 background, and raters' familiarity with accented L2 speech (see Appendix F for a marginal effects plot illustrating the unique contribution of coherence to comprehensibility while taking control covariates into account). Prompt type (psychology vs. sociology) had no significant effect on comprehensibility. In terms of

Table 1. Descriptive Statistics for L2 Speech Ratings

	Comprehensibility			Coherence		
Speaker group	M	SD	95% CI	M	SD	95% CI
Overall (n = 60)	66.24	25.14	[64.40, 68.09]	59.60	22.82	[57.93, 61.28]
Mandarin (n = 14)	57.14	26.53	[53.05, 61.23]	51.99	23.12	[48.43, 55.56]
Farsi (n = 15)	69.27	23.72	[65.78, 72.76]	61.26	21.99	[58.03, 64.50]
Indian (n = 16)	63.18	25.38	[59.57, 66.80]	57.52	23.13	[54.23, 60.81]
Romance (n = 15)	74.78	21.54	[71.61, 77.95]	67.11	20.56	[64.08, 70.13]

Table 2. Full Mixed-Effects Model for Comprehensibility

Parameter	Estimate	SE	95% CI	t	p
(Intercept)	63.47	1.98	[59.37, 67.52]	31.99	<. 001
Coherence	9.25	0.59	[7.96, 10.51]	15.57	<. 001
Control covariates					
Accentedness	13.87	0.77	[12.34, 15.38]	18.04	<. 001
Prompt (psychology vs. sociology)	0.05	1.01	[–1.96, 2.07]	0.05	.958
Speakers' L1					
Mandarin vs. Farsi	2.68	1.44	[–0.16, 5.60]	1.86	.245
Mandarin vs. Indian	3.95	1.40	[1.19, 6.77]	2.83	.024
Mandarin vs. Romance	4.05	1.48	[1.14, 7.09]	2.74	.032
Listeners' L2 accent familiarity					
Mandarin familiarity	3.76	2.03	[–0.56, 8.09]	1.85	.089
Farsi familiarity	–11.19	5.78	[–23.46, 1.15]	–1.94	.077
Indian familiarity	1.33	2.25	[–3.46, 6.12]	0.59	.566
Romance familiarity	11.87	5.46	[0.22, 23.43]	2.18	.050

Random effects	Variance	SD	Criterion	Estimate
Rater (intercept)	30.43	5.52	Log–likelihood	–2794.0
Speaker (intercept)	2.72	1.65	AIC	5616.1
			BIC	5680.1

AIC = Akaike information criterion; BIC = Bayesian information criterion.

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

L1 group, speakers from Indian and Romance language backgrounds received higher comprehensibility ratings, compared to Mandarin speakers, with no significant differences between any other groups. Raters' accent familiarity only mattered for Romance speakers, where greater familiarity was associated with more comprehensible speech. Coherence, along with fixed-effects covariates, accounted for approximately 72% of the variance in comprehensibility (marginal R^2 = .72), and, together with random effects, it explained 78% of the variance in comprehensibility (conditional R^2 = .78).

As depicted in Figure 1 (see Appendix G for separate scatterplots by speakers' L1), the association between comprehensibility and coherence was positive and strong, r = .70, 95% CI [.66,. 74], p <. 001. Although the relationship was generally linear across the entire scale length, it appeared stronger (characterized by a steeper slope) at the lower scale end (for ratings below about 30). For speakers at a low performance level, where scores were more scattered and error estimates were wider, a small improvement in coherence seemed to be associated with greater benefit to their comprehensibility than for speakers at higher performance levels.

Discourse predictors of coherence and comprehensibility

The second research question explored the relationship between L2 speakers' use of discourse features and rater-assessed comprehensibility and coherence. As shown in Table 3, although the incidence of discourse features was generally low (M = 0.50–4.74), SD values (0.16–3.25) indicate that individual speakers varied in how they used those features.

Of the 12 discourse measures entered as fixed-effects predictors of coherence, only structure–ordering, verb overlaps, and the similarity between the reading source and a speaker's performance emerged as significant, after controlling for speakers' L1 and prompt type. As summarized in Table 4 (see Appendix H for full model), all relationships were positive, where speakers who provided the expected structural components

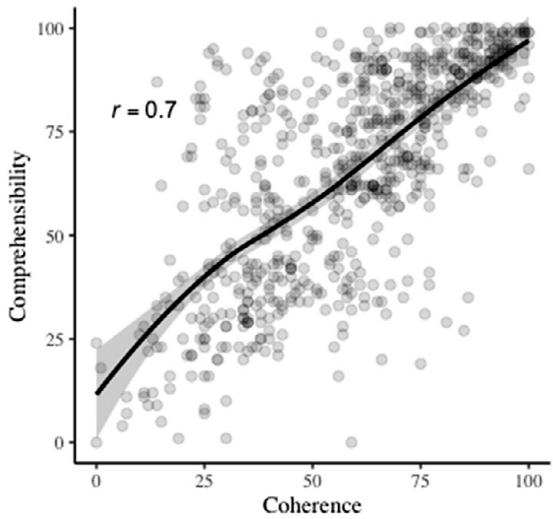

Figure 1. Scatterplot of the relationship between comprehensibility and coherence, with a Loess line and 95% CI estimates (shaded in gray) showing the best-fitting trendline.

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

Table 3. Descriptive Statistics for Discourse Measures

	Normalized values		Raw counts	
Discourse feature	M	SD	M	SD
Organization				
Structure–ordering	—	—	0.50	0.42
Signposting	0.02	0.02	2.68	1.83
Source–speech similarity				
Similarity–reading	—	—	0.68	0.16
Similarity–listening	—	—	0.68	0.13
Connectives				
Disjunction	0.01	0.01	0.71	1.01
Causal connectives	0.03	0.02	3.00	2.38
Opposition	0.01	0.01	1.60	1.30
Addition	0.04	0.03	4.74	3.25
Lexical overlaps				
Noun overlaps	0.25	0.24	—	—
Verb overlaps	0.45	0.44	—	—
Givenness				
Pronoun–noun ratio	0.68	0.35	—	—
Attended demonstratives	0.01	0.01	0.74	0.87

Note: Structure–ordering (scored as 0–1) and source–speech similarity (Word2vec similarity index) were not normalized. Raw counts do not apply to lexical overlaps and pronoun–noun ratios, which were computed as proportions of target items to total word counts.

Table 4. Summary of Mixed-Effects Model for Coherence and Discourse Features

Parameter	Estimate	SE	95% CI	t	p
(Intercept)	56.54	5.15	[46.17, 66.91]	10.99	<. 001
Organization					
Structure–ordering	4.22	1.42	[1.38, 7.06]	2.97	.004
Source–speech similarity					
Similarity–reading	4.55	1.91	[0.74, 8.36]	2.38	.021
Lexical overlaps					
Verb overlaps	3.26	1.54	[0.19, 6.33]	2.12	.039

Random effects	Variance	SD	Criterion	Estimate
Rater (intercept)	161.26	12.70	Log–likelihood	–3024.2
Speaker (intercept)	55.99	7.48	AIC	6088.3
			BIC	6179.8

in the anticipated sequence (e.g., introduction, concept explanation, 1st example), repeated verbs and their synonyms between sentences, and produced utterances that were similar to the source reading received higher coherence ratings. Fixed-effects predictors, along with covariates, accounted for approximately 14% of the variance in coherence (marginal $R^2 = .14$), and, together with random effects, they explained 56% of the variance in coherence ratings (conditional $R^2 = .56$).

Of the 12 discourse measures entered as fixed-effects predictors of comprehensibility, only verb overlaps and additive connectives emerged as significant, after controlling for speakers' L1 and prompt type. As shown in Table 5 (see Appendix I for full model), speakers who repeated verbs and their synonyms within and across sentences were rated as more comprehensible than those whose speech contained fewer verb overlaps. However, speakers who used additive connectives more frequently (e.g., *and*) were

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

Table 5. Summary of Mixed-Effects Model for Comprehensibility and Discourse Features

Parameter	Estimate	SE	95% CI	t	p
(Intercept)	60.57	6.01	[48.30, 72.84]	10.09	<. 001
Connectives					
Addition	–2.72	1.34	[–5.40, –0.05]	–2.03	.047
Lexical overlaps					
Verb overlaps	3.75	1.48	[0.80, 6.70]	2.53	.014
Random effects	**Variance**	**SD**	**Criterion**		**Estimate**
Rater (intercept)	288.21	16.98	Log–likelihood		–3004.4
Speaker (intercept)	51.48	7.18	AIC		6048.8
			BIC		6140.3

perceived as harder to understand than those who used them to a lesser extent. For example, a 0.01 increase in the use of additive connectives (i.e., one additional connective used per 100 words) corresponded to a decrease of 2.72 points in the comprehensibility score on a 100-point scale. Fixed-effects predictors, along with covariates, accounted for approximately 12% of the variance in comprehensibility ratings (marginal R^2 = .12), and, together with random effects, they explained 66% of the variance in comprehensibility (conditional R^2 = .66).

Discussion

We explored the relationship between L2 coherence and comprehensibility in an academic speaking task, examining which discourse features predict these constructs. For L2 speakers performing an academic speaking task, coherence was strongly related to comprehensibility, such that performances evaluated by raters as more coherent also received higher comprehensibility ratings. In terms of discourse features associated with each rated dimension, coherence and comprehensibility emerged as overlapping yet partially distinct constructs. Whereas measures of discourse organization (structure–ordering), lexical repetition (verb overlaps, including synonyms), and similarity between the task prompt and a speaker's oral performance (reading similarity) contributed to coherence, lexical repetition (verb overlaps) and a speaker's use of connectives (additives) were linked to comprehensibility.

Relationship between coherence and comprehensibility

In this dataset, coherence and comprehensibility were strongly related, with approximately 50% of shared variance. In previous L2 writing research (e.g., Kuiken & Vedder, 2017), coherence and comprehensibility (defined as readers' ease of understanding texts) showed a greater overlap (76–87%), presumably because comprehensibility in L2 speech can be influenced not only by discourse features (in addition to such dimensions as lexis and grammar) but also by pronunciation, which is unique to speech. Our findings appear to provide the first direct evidence that coherence significantly contributes to comprehensibility—measured in spoken, not written, discourse—after controlling for several pronunciation-relevant speaker (accentedness, L1 background) and listener (accent familiarity) variables. Unlike text, which allows readers to revisit information or take additional time if they encounter a problem, speech affords listeners little opportunity to return to a specific segment or to easily slow down the

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

speech stream. Therefore, it appears reasonable that listeners find more structured and more repetitive speech easier to understand (Nagle et al., 2019).

Our findings clarify the relevance of various linguistic dimensions of L2 speech to comprehensibility. Comprehensibility is a multifaceted construct, associated with many linguistic properties of L2 speech (Isaacs & Trofimovich, 2012; Saito, 2021), where about 30–50% of variance in rater-assessed comprehensibility is explained through measures of phonology (segment substitutions, word stress accuracy) and fluency (frequency and location of pauses, speaking rate). As shown here, some remaining variance in comprehensibility may be partly explained through a measure of coherence, at least in an integrated academic speaking task. Because this task requires speakers to integrate several sources of information to produce clear, intelligible discourse to succeed, it seems reasonable that coherence was related to comprehensibility. However, this relationship might be different—in both strength and quality—in other tasks (e.g., picture narratives, argumentative speech) that vary in their demands and complexity (Bergeron & Trofimovich, 2017; Crowther, Trofimovich, Isaacs, & Saito, 2015), which needs to be investigated in future work.

Although the coherence–comprehensibility relationship was generally linear (see Figure 1), inasmuch as coherence had a proportionately similar contribution to comprehensibility throughout the entire scale length, the value of coherent speech to comprehensibility seemed to be magnified for speakers at the lower scale end (for ratings below about 30 on a 100-point scale). Thus, when a speaker's comprehensibility is low, likely because of various local issues such as phonemic substitutions, disfluencies, and lexical and morphosyntactic errors (Isaacs & Trofimovich, 2012; Saito, 2021), clearly organized and logically structured discourse might offer an especially valuable benefit to the speaker's comprehensibility, insofar as listeners' processing effort is concerned (see Nagle et al., 2019, for similar arguments). In fact, other linguistic dimensions similarly vary in their contribution to comprehensibility for L2 speakers of different ability levels, where grammar is a higher-order skill which might offset lower-level issues, such as fluency and vocabulary (Isaacs & Trofimovich, 2012). Put simply, clearly organized discourse might compensate for various local language difficulties of speakers, making their message easier for listeners to understand.

Discourse features underlying coherence and comprehensibility

Our second goal was to explore discourse-level predictors of L2 coherence and comprehensibility. Among the unique predictors of coherence were two discourse features, namely, the organization of task-essential structural elements and the similarity between the reading prompt and a speaker's performance, both positively associated with coherence. Just as interconnectedness of ideas was previously shown to predict L2 written discourse (Crossley et al., 2016), a speaker's use of structural components that are essential to task completion (e.g., providing an example to illustrate a concept) and that occur in the expected order (e.g., general information provided first, followed by examples) seems relevant to L2 spoken discourse. In fact, a TOEFL-type integrated speaking task requires test-takers to integrate a conceptual explanation (from the reading) with two examples (from the lecture), typically by first introducing abstract ideas followed by specific examples (Brown & Ducasse, 2019). In this sense, providing the essential components of an argument in the expected order corresponds to a view of coherence as being

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

concerned with conceptual congruence in both oral and written discourse (van Dijk & Kintsch, 1983).

Another unique predictor of coherence was the semantic similarity between the reading source and a speaker's performance. The semantic similarity index (Word2vec) captures the extent to which a speaker re-uses lexical content from a source text (Crossley et al., 2019). In integrated speaking tasks, source texts provide information necessary for speakers to explain the target phenomenon, so the similarity between the source text, with which all raters were familiar, and a speaker's performance seems to be a reliable index of coherence and in fact its strongest predictor, judging by the estimate values (*Estimate* = 4.55, see Table 4). For instance, for the psychology prompt, speakers were expected to explain a phenomenon (i.e., people's awareness of external observers) described in the reading by referring to two real-life examples provided in the lecture. Raters thus appreciated a speaker's effort to include the conceptual explanation from the reading to contextualize the two examples. And because raters appeared to factor only the reading content into their coherence ratings, they might have prioritized clarity in delivering task-relevant conceptual information from the reading over task-relevant details from the lecture.

When it comes to comprehensibility, it was uniquely predicted by speakers' use of additives (e.g., *also*, *and*). This measure had an inverse relationship with comprehensibility, which aligns with negative links between connectives (including additive connectives) and comprehensibility established previously (Appel et al., 2019). One possible reason for this relationship pertains to various ambiguous uses of *and*, which appear to interfere with listeners' information processing (Tyler, 1992). As illustrated in the example below (where *and* is italicized when used at idea unit boundaries), some speakers, like P139, repeatedly used *and* even when two adjacent idea units were not necessarily in additive relationships.

P139: The example given by the professor is { the } the professors waiting in line. *and* suddenly a people just get in front of him. *and* he think he is rude and kind of selfish. *and* this thinking is based on his own ground. *and* he also take another sort of the people which is businessman.

The use of *and* illustrated in this example is ambiguous (Halliday & Hasan, 1976), considering that it is not obvious whether *and* is meant to indicate a structural relation (functioning as a coordinate connective), to elaborate on the preceding utterance, or to introduce a new idea (functioning as an additive connective). To complicate matters further, *and* is also used as a filler in oral discourse (Iwashita & Vasquez, 2015), which listeners may perceive as obtrusive if not bothersome. Just as specific L2 grammatical errors are considered bothersome by listeners (Derwing, Rossiter, & Ehrensberger-Dow, 2002), frequent and sometimes confusing use of additives might be detrimental to rater judgments. Whatever reasons raters had for downgrading speaker comprehensibility, excessive, ambiguous, and inconsistent occurrences of additive connectives likely required raters to expend additional processing effort when listening to L2 speech.

How distinct are coherence and comprehensibility?

Considering similarities and differences between coherence and comprehensibility, the two constructs seem to be distinguishable at the level of macro- versus micro-structure of discourse, such that coherence was underpinned by measures of task-relevant

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

structural elements and their ordering and source–speech similarity, whereas comprehensibility was predicted through the use of a local cohesive device. This data-driven distinction was also supported through rater comments, where, for instance, one EAP instructor from the pilot study noted:

> [C]oherence for me is a little bit more about ideas… and the flow of the idea and the content. Can I understand what their overall meaning is? Whereas comprehensibility applies more to sort of chunks. Am I able to actually understand what they're saying, even though the big idea might not be as clear to me? (Rater 3)

This instructor's thoughts are supported through previous work on comprehensibility as a dynamic, time-sensitive measure of listening effort. For instance, raters reliably evaluate comprehensibility using brief speech excerpts (e.g., 5–30 s), often presented outside a larger discourse context, and raters can assess comprehensibility even for a single word (Uchihara, 2022). In fact, comprehensibility has recently been proposed to capture moment-to-moment fluctuations in processing effort as listeners (re)construct meaning in real time (Nagle et al., 2019). Thus, unlike coherence, which is concerned with listeners' global understanding of discourse, for example, in terms of how information is structured and how its various elements are interlinked, comprehensibility captures listeners' evolving close, phrase-level understanding.

Despite some differences, coherence and comprehensibility have a clear overlap. For instance, both sets of ratings were associated positively with the same discourse feature, namely, the frequency of a speaker's repeated use of verbs. Lexical repetition (including the use of synonyms) is known to enhance text cohesion, highlighting semantic relations within text (Halliday & Hasan, 1976). To illustrate, some speakers, such as P88, who received high coherence and comprehensibility ratings, tended to re-use the same verb forms (e.g., *were told*, *were being watched*). In contrast, for other speakers who elicited low coherence and comprehensibility ratings, individual verbs (or their derivatives) occurred only once per utterance.

P88: a group of students *were told* to, like, tie their shoes. and { one } one group *was told* that they *were being watched*…meaning that the ones that *were being watched* and were aware that *were being watched*, they did it like { their } their speed was increased which they performed what they *were told* to do

Because the task required speakers to explain how the two examples (from the lecture) illustrate the phenomenon (from the reading), verb overlaps might have helped listeners to easily identify contrasted examples and to achieve semantic unity across various narrative elements (Crossley et al., 2008; Halliday & Hasan, 1976), which also lessened their processing burden.

Lastly, even though 12 different discourse measures were examined in relation to coherence and comprehensibility, only a handful emerged as predictors. In fact, noun overlaps, signposting, connectives (disjunction, causal connectives, opposition), givenness (pronoun–noun ratio, attended demonstratives), and lexical similarity between the listening source and a speaker's performance showed no links to coherence or comprehensibility. These findings stand in contrast to results of prior work, which revealed a positive relationship between the frequency of noun overlaps and writing quality in an integrated writing task (Guo, Crossley, & McNamara, 2013) and between

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

measures of signposting and anaphoric reference and text comprehension difficulty (Tyler, 1992; van Silfhout et al., 2015). For one, the lack of strong associations is likely due to the infrequent occurrence of many discourse features, whose incidence ranged between 0.71 and 3.00 on average (see Table 3), which made it difficult for meaningful relationships to emerge.

Another reason for the lack of strong associations may stem from task effects, considering that the specific demands of a given task call for speakers' use of particular discourse features (Appel et al., 2019) or create different expectations of coherence (Richards, 1990). Although lexical overlaps are arguably helpful for processing L2 speech, in this study, only overlaps in verbs, not nouns, were relevant to coherence and comprehensibility, presumably because verbs carry a heavier information burden for the listener than nouns (Miller & Fellbaum, 1991). For example, in the integrated speaking task, whereas nouns appearing as agents (e.g., a professor) and objects (e.g., shoes) were most likely similar across all speakers, verbs may have been used more uniquely by individual speakers as they combined different ideas to structure their discourse. In addition, speakers frequently use signposting in oral presentations and lectures (Jung, 2006; Williams, 1992), so listeners may associate coherence in these genres with a measure of signposting (Brown et al., 2005; Jung, 2003). Unlike longer speech samples, such as a 14-min lecture (Jung, 2006), a 1-min integrated speaking task might provide few opportunities for speakers to produce clear signposting cues or to emphasize new versus old information, so these features would predictably be of less relevance to listener-rated coherence in a brief task. Similarly, in a speaking task which targets a conceptual phenomenon illustrated through two examples, speakers' use of specific connectives (e.g., disjunctives to contrast alternatives) might be less important or relevant, relative to a compare-and-contrast oral narrative.

Limitations and future work

Several limitations of this exploratory work should be acknowledged. First, just as the linguistic dimensions of comprehensibility can be task specific (Crowther et al., 2015), so can individual discourse features vary in their relevance to coherence as a function of a speech elicitation instrument. In future work, researchers should therefore explore different speaking tasks, including those which are used in academic and non-academic contexts for learning and assessment purposes, to develop a descriptive toolkit of discourse grammars and their features relevant to speaker coherence in those tasks. Second, L2 speakers' accentedness was included as a covariate to explore the relationship between coherence and comprehensibility. Although this approach sidestepped the methodological limitation of studying coherence in speech through transcripts, it was unclear which aspects of speakers' pronunciation contributed to listener perception of coherence and in which way, such as by focusing listeners' attention to spoken discourse or in fact by distracting them from it. An acoustic analysis of speakers' performance, combined with an investigation of discourse features, would clarify the role of pronunciation in listener-assessed coherence. Finally, because only several discourse features predicted coherence and comprehensibility, it would be important to investigate these rater-assessed dimensions in relation of other aspects of spoken discourse such as the accuracy and appropriateness of speech content (i.e., distinct ideas produced) and to examine previously targeted measures, including connectives and signposting, through qualitative rather than quantitative analyses.

https://doi.org/10.1017/S0272263124000305 Published online by Cambridge University Press

Conclusion

For EAP teachers evaluating L2 speakers' performances in an academic speaking task, coherence and comprehensibility were strongly related, even after controlling for pronunciation-relevant speaker (accentedness, L1 background) and listener (accent familiarity) variables. In terms of specific discourse features, coherence was associated with speakers' creating a logically structured and ordered argument and their use of repeated verb forms. Comprehensibility was also positively linked to repeated verb forms but was predicted negatively by speakers' use of additive connectives (e.g., *and*). These findings underscore the importance of coherence to comprehensible L2 speech, suggesting that these constructs are partially distinct. Underpinned by macro-level measures of discourse organization, coherence captures listeners' global level of understanding. In contrast, comprehensibility is supported through micro-level discourse features and targets local, phrase-level understanding.

Acknowledgements. We are grateful to Yoo Lae Kim for coding discourse features, to Sara Kennedy and Concordia Applied Linguistics lab members, including Kym Taylor Reid, Lauren Strachan, Pakize Uludag, Rachael Lindberg, Oguzhan Tekin, Chen Liu, Yoo Lae Kim, Anamaria Bodea, and Thao-Nguyen Nina Le who helped us develop materials. An earlier version of this work was presented at the annual meeting of the American Association for Applied Linguistics (AAAL) 2023 conference.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000305.

Data availability statement. The experiment in this article earned an Open Materials badge for transparent practices. The materials are available at https://osf.io/ugr75/.

Funding statement. This research was supported through a Social Sciences and Humanities Research Council of Canada (SSHRC) grant to Pavel Trofimovich (430-2020-1134).

Competing interest. The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Appel, R., Saito, K., Isaacs, T., Webb, S., & Trofimovich, P. (2019). Lexical aspects of comprehensibility and nativeness from the perspective of native-speaking English raters. *International Journal of Applied Linguistics*, *170*, 24–52. https://doi.org/10.1075/itl.17026.app

Bartoń, K. (2023). *MuMIn: Multi-Model Inference. R package version 1.47.5.* https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*, 547–566. http://doi.org/10.1111/flan.12285

Brown, A., & Ducasse, A. M. (2019). An equal challenge? Comparing TOEFL iBT™ speaking tasks with academic speaking tasks. *Language Assessment Quarterly*, *16*, 253–270. https://doi.org/10.1080/15434303.2019.1628240

Brown, A., Iwashita, N., & McNamara, T. F. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *TOEFL Monograph* (No. MS29), i-157, Educational Testing Service. http://doi.org/10.1002/j.2333-8504.2005.tb01982.x

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*, 1–47.

Copland, F., Mann, S., & Garton, S. (2019). Native-English-speaking teachers: Disconnections between theory, research, and practice. *TESOL Quarterly*, *54*, 348–374. http://doi.org/10.1002/tesq.548

Crossley, S., Barnes, T., Lynch, C., & McNamara, D. S. (2017). Linking language to math success in an on-line course. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), *Proceedings of the10th International Conference on Educational Data Mining* (pp. 180–185). Educational Data Mining Society.

Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*, 250–270. https://doi.org/10.1080/15434303.2014.926905

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*, 475–493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*, 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*, 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S., & McNamara, D. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 1236–1241). Cognitive Science Society.

Crossley, S. A., Salsbury, T., McCarthy, P. M., & McNamara, D. S. (2008). LSA as a measure of coherence in second language natural discourse. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1906–1911). Cognitive Science Society.

Crowther, D., Trofimovich, P., Isaacs, T., & Saito, K. (2015). Does a speaking task affect second language comprehensibility? *The Modern Language Journal*, *99*, 80–95. https://doi.org/10.1111/modl.12185

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.

Derwing, T. M., Rossiter, M. J., & Ehrensberger-Dow, M. (2002). "They speaked and wrote real good": Judgements of non-native and native grammar. *Language Awareness*, *11*, 84–99. http://doi.org/10.1080/09658410208667048

Educational Testing Service. (2006). *The official guide to the new TOEFL iBT*. McGraw Hill.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. MIT Press.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *36*, 193–202. https://doi.org/10.3758/BF03195564

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218–238. http://doi.org/10.1016/j.asw.2013.05.002

Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly*, *38*, 201–223. http://doi.org/10.2307/3588378

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*, 346–363. https://doi.org/10.1002/bimj.200810425

Hulstijn, J. H., Schoonen, R., & de Jong, N. H. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing*, *29*, 203–221. http://doi.org/10.1177/0265532211419826

Inoue, C., & Lam, D. M. K. (2021). *The effects of extended planning time on candidates' performance, processes, and strategy use in the lecture listening-into-speaking tasks of the TOEFL iBT® Test* (TOEFL Research Report No. RR-21-09). Educational Testing Service. https://doi.org/10.1002/ets2.12322

Isaacs, T., & Trofimovich, P. (2011). *Montreal speech corpus*. Concordia University.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*, 475–505. http://doi.org/10.1017/S0272263112000150

Iwashita, N., & Vasquez, C. (2015). An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2. *IELTS Research Reports Online*, *5*, 1–44.

Jung, E. H. (2003). The role of discourse signaling cues in second language listening comprehension. *The Modern Language Journal*, *87*, 562–577. http://doi.org/10.1111/1540-4781.00208

Jung, E. H. (2006). Misunderstanding of academic monologues by nonnative speakers of English. *Journal of Pragmatics*, *38*, 1928–1942. http://doi.org/10.1016/j.pragma.2005.05.001

Kennedy, S., & Trofimovich, P. (2019). Comprehensibility: A useful tool to explore listener understanding. *Canadian Modern Language Review*, *75*, 275–284. http://doi.org/10.3138/cmlr.2019-0280

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*, 321–336. https://doi.org/10.1177/0265532216663991

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

McNamara, D.S., Crossley, S.A. & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*, 499–515. https://doi.org/10.3758/s13428-012-0258-1

McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*, 247–288. https://doi.org/10.1080/01638539609544975

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119. https://doi.org/10.48550/arXiv.1310.4546

Miller, G. A., & Fellbaum, C. (1991). Semantic networks of English. *Cognition*, *41*, 197–229. http://doi.org/10.1016/0010-0277(91)90036-4

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, *5*, 294–323. https://doi.org/10.1075/jslp.18016.nag

Nagle, C. L., & Rehman, I. (2021). Doing L2 speech research online: Why and how to collect online ratings data. *Studies in Second Language Acquisition*, *43*, 916–939. http://doi.org/10.1017/S0272263121000292

Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, *41*, 647–672. https://doi.org/10.1017/S0272263119000044

R Core Team. (2023). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Richards, J. C. (1990). *The language teaching matrix*. Cambridge University Press.

Saito, K. (2021). What characterizes comprehensible and nativelike pronunciation among English-as-a-second-language speakers? Meta-analyses of phonological, rater, and instructional factors. *TESOL Quarterly*, *55*, 866–900. https://doi.org/10.1002/tesq.3027

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652–708. http://doi.org/10.1111/lang.12345

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*, 439–462. http://doi.org/10.1093/applin/amv047

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly*, *26*, 713–729. https://doi.org/10.2307/3586870

Tyler, A., & Bro, J. (1992). Discourse structure in nonnative English discourse: The effect of ordering and interpretive cues on perceptions of comprehensibility. *Studies in Second Language Acquisition*, *14*, 71–86. https://doi.org/10.1017/S0272263100010470

Tyler, A., & Bro, J. (1993). Discourse processing effort and perceptions of comprehensibility in nonnative discourse: The effect of ordering and interpretive cues revisited. *Studies in Second Language Acquisition*, *15*, 505–522. https://doi.org/10.1017/S0272263100012407

Tyler, A., Jefferies, A. A., & Davies, C. E. (1988). The effect of discourse structuring devices on listener perceptions of coherence in non-native university teacher's spoken discourse. *World Englishes*, *7*, 101–110. http://doi.org/10.1111/j.1467-971X.1988.tb00223.x

Uchihara, T. (2022). Is it possible to measure word-level comprehensibility and accentedness as independent constructs of pronunciation knowledge? *Research Methods in Applied Linguistics*, *1*, 100011. http://doi.org/10.1016/j.rmal.2022.100011

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. Academic Press.

van Silfhout, G., Evers-Vermeul, J., & Sanders, T. (2015). Connectives as processing signals: How students benefit in processing narrative and expository texts. *Discourse Processes*, *52*, 47–76. https://doi.org/10.1080/0163853X.2014.905237

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly*, *26*, 693–711. http://doi.org/10.2307/3586869

Cite this article: Tsunemoto, A., & Trofimovich, P. (2024). Coherence and Comprehensibility in Second Language Speakers' Academic Speaking Performance. *Studies in Second Language Acquisition*, 1–23. https://doi.org/10.1017/S0272263124000305