BRITISH
COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

FEATURES OF DISCOURSE AND LEXICAL RICHNESS AT DIFFERENT PERFORMANCE LEVELS IN THE APTIS SPEAKING TEST

AR-G/2017/2

Noriko Iwashita, University of Queensland
Lyn May, Queensland University of Technology
Paul Moore, University of Queensland

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITOR: VIVIEN BERRY
ISSN 2057-5203

ABSTRACT

This study investigated features of discourse competence and vocabulary use across levels and tasks in the Aptis Speaking Test. These features are considered to be strongly linked with general performance and are in both the Common European Framework of Reference for Languages (CEFR) language level descriptors and the Aptis rating scale. Discourse competence was operationalised in terms of the textual features of cohesion and coherence. Selected aspects of cohesion and lexical richness were examined through quantitative and qualitative methods, while coherence was analysed employing qualitative methods.

The findings of the quantitative analyses show little variation in the use of cohesive devices across levels and tasks; however, some distinctive differences were observed in vocabulary use and features of coherence, including topic development. Qualitative analysis revealed that candidates' performances varied according to the task prompts and their approaches to the tasks.

These findings provide test developers with a more nuanced understanding of relevant aspects of the construct of L2 speaking that are elicited through the tasks used at different levels in the Aptis Speaking Test, and have the potential to inform future development/refinement of the speaking tasks and the rating scales currently used.

Authors

Noriko Iwashita is a senior lecturer in Applied Linguistics at The University of Queensland. Her research interests include task-based assessment, and cross-linguistic investigation of four major language traits and the interfaces of language assessment and SLA. Her work has appeared in *Language Testing, Language Assessment Quarterly, Applied Linguistics, Language Learning* and *Studies in Second Language Acquisition*.

Lyn May is a senior lecturer in TESOL at the Queensland University of Technology. Her research interests include L2 speaking assessment, test preparation, and the linguistic demands of tertiary study and professional contexts. Her work has appeared in *Language Testing, Language Assessment Quarterly* and *Assessment in Education: Principles, Policy and Practice*.

Paul Moore is a lecturer in Applied Linguistics at The University of Queensland. His research interests include classroom discourse (including the learner's perspective) and language testing discourse. He has published in *The Modern Language Journal, Assessment and Evaluation in Higher Education*, and in edited volumes published by Springer and John Benjamins.

Acknowledgements

The project team acknowledges the British Council's generous support for funding and guidance. We are also grateful for Wanhua (Cindy) Ma, Khoi Mai Ngoc, Annita Stell and Cheryl Jecht for their assistance with transcription, data sorting and coding.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. BACKGROUND AND RATIONALE

The study investigates key criteria for language levels assessed in the Aptis Speaking Test. In particular, the study compares features of discourse competence and vocabulary use observed in the test performances across levels and tasks. The first section of this report provides a brief overview of relevant aspects of the Common European Framework of Reference for Languages (CEFR) and its impact on the development of Aptis.

1.1 The Common European Framework of Reference for Languages (CEFR) and the testing of language proficiency

The Common European Framework of Reference for Languages: Learning, Teaching and Assessment (CEFR) was developed by the Council of Europe (Council of Europe, 2001) in order to provide “a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe” (p. 1). The CEFR’s influence in these areas has been extensive – it has been translated into more than 30 languages (Hulstijn, 2014) and has been adopted or adapted as part of language policy in countries far from Europe.

The six Common Reference Levels in the CEFR (Council of Europe, 2001) were developed as one part of the CEFR to describe levels of language proficiency as a point of reference across the languages of Europe. According to North (2014), “the prime function of the CEFR is not to get all tests reporting to the same scale, but to encourage reflection on current practice, and thus to stimulate improvement in language teaching and learning (and testing)” (p. 229). For historical reasons related to increased mobility in the region, the influence of the CEFR has been strongest in activities aiming to link tests to the CEFR.

While there has been much praise of this far-reaching exercise, several critiques of its application to language testing have been offered, both by observers (see Hulstijn’s [2014] argument on the lack of empirical support for the levels, related to his work on the construct of language proficiency), and by those involved in its development (e.g., Little, 2007; Trim, 2014; North, 2014). For example, Little (2007) notes that descriptors for writing were based on those for speaking, and that descriptors for speaking (especially fluency and intonation) were based on somewhat out-dated assumptions. The next section describes the approach taken in the linking of the Aptis Test to the CEFR, with a focus on the speaking tasks.

1.2 The Aptis Speaking Test

The computer-based Aptis Speaking Test consists of four tasks taking a total of 12 minutes to complete (British Council, 2013).

- Task 1 (90 seconds, target level CEFR A1/A2) involves short responses to three questions on personal topics.
- Tasks 2 and 3 each last 135 seconds, and target level CEFR B1. Task 2 involves short responses involving description and comparison based on a picture prompt. Task 3 involves description, comparison and speculation based on two picture prompts.
- Task 4 (180 seconds, target level CEFR B2) involves a longer, integrated response to three questions on an abstract topic.

1.3 Linking the Aptis Test to the CEFR

The Aptis Test was developed with reference to the CEFR and to the British Council/ European Association for Quality Language Services (EAQUALS) Inventory (North, Ortega & Sheehan, 2010), which itself was “an attempt to add detail to the CEFR descriptors” (O’Sullivan, 2015c, p. 14; cf. also North, 2014). In other words, this attempt involved the work necessary to apply the CEFR descriptors to a specific language, pointed out by Trim (2014).

The exercises aimed at linking the Aptis Test to the CEFR Levels (O’Sullivan, 2015c) involved the validation of the Aptis cut-scores which reflect different CEFR levels. The bases of this exercise included the following:

- theoretically, it drew on the socio-cognitive framework for test validation (O’Sullivan, 2011; O’Sullivan & Weir, 2011; Weir, 2005)
- methodologically, it drew on:
  - the recommendations of the CEFR Manual (Council of Europe, 2003; 2009)
  - previous CEFR linking experience (e.g., O’Sullivan, 2010)
  - the development process of Aptis itself (O’Sullivan, 2012; 2015a)
  - a recent linking exercise conducted by O’Sullivan (2010).

With regard to the Aptis Speaking test there was “complete agreement between the [CEFR] levels suggested by the expert panel and those suggested by the Aptis raters” (O’Sullivan, 2015a, p. 39). Nonetheless, O’Sullivan (2015c) points to the need for ongoing research aimed at gaining further insights into characteristics of performance at the different levels of the Aptis Test.

The current study focuses on features of discourse and vocabulary use, as these features are considered to be strongly linked with general performance (e.g., Bachman & Palmer, 1996; Harley, Cummins, Swain & Allen, 1990). By investigating examples of language production, with regard to discourse competence and vocabulary use, at different levels on the Aptis Speaking scale, this research will provide explicit insights into what constitutes spoken language performance at the various levels, as well as insight into the influence of tasks on test performance. As such, it is expected that the research will contribute to the iterative development of task specifications, tasks and marking criteria.

2. THEORETICAL FRAMEWORK AND LITERATURE REVIEW

2.1 Socio-cognitive framework for test validation

The research draws on Weir's (2005a) conceptualisation of context validity, in his socio-cognitive framework for test validation (see also O'Sullivan, 2011; O'Sullivan & Weir, 2011). This model is based on Bachman's (1990) model, but extended to take into account criticisms that it was "essentially psychological, with no reference to the social context of language use" (O'Sullivan, 2011, p. 261). Weir (2005a) notes that context validity represents a more social perspective on what has traditionally been referred to as "content validity" (i.e., whether the test tasks are representative of those in the content domain) defining it as follows.

> Context validity is concerned with the extent to which the choice of tasks in a test is representative of the larger universe of tasks of which the test is assumed to be a sample. This coverage relates to linguistic and interlocutor demands made by the task(s) as well as the conditions under which the task is performed arising from both the task itself and its administrative setting. (p. 19)

In particular, our investigation will provide insights into the underlying "linguistic demands" of the four speaking tasks as influenced by the "performance parameters" of each task. Linguistic demands of test tasks involve "the language of the input and the expected language of the output and can also include reference to the audience or interlocutor where appropriate," while performance parameters refers to "parameters such as timing, preparation, score weighting, and knowledge of how performance will be scored" (O'Sullivan, 2015c, p. 7).

Research into task-based language learning and teaching (e.g., Skehan, 1998, 2009; Robinson, 2011) and performance assessment (e.g., Norris, Brown, Hudson & Yoshioka, 1998) has investigated the influence of task types/task internal factors (e.g., monologic/dialogic, open/closed; personal/narrative/decision-making; topic familiarity) and task implementation conditions (e.g., contextualisation; supported/unsupported planning time) on task performance. More specifically, research (e.g., Elder & Iwashita, 2005) has drawn on Skehan's (1998, 2009) trade-off hypothesis and/or Robinson's (e.g., 2001, 2011) cognition hypothesis to investigate the influence of task complexity (e.g., code complexity, cognitive complexity and communicative demand) on the complexity, accuracy and fluency (CAF) of learners' language performance. While a rich volume of the literature in both second language acquisition and language testing has shown a complex relationship between tasks and characteristics of language observed in task performance mostly focusing on CAF, little research has been undertaken concerning features of discourse observed in task performance.

2.2 Analysis of candidate performance

When considering the construct of L2 speaking in test validation, it is widely acknowledged that different aspects of the language contribute differently to overall language proficiency (e.g., Higgs & Clifford, 1982; Iwashita, Brown, McNamara & O'Hagan, 2008). In particular, discourse competence, which involves the ability to interpret and produce appropriate language beyond the sentence level, is considered to be an essential aspect of communicative language ability (Bachman & Palmer, 1996). However, a detailed study of discourse competence in speaking assessment appears to have been somewhat neglected (cf. Purpura, 2008; van Lier, 1989). The notion of discourse competence can be articulated in terms of the candidate's ability to produce and comprehend unified pieces of oral or written text. Kang (2005) argues one way to explore discourse competence is to examine the degree of cohesion and coherence exhibited in performance.

2.2.1 Cohesion

Cohesion is a textual property that "defines a good text at the discourse level, as opposed to a sequence of sentences that would not be considered a text" (Kang, 2005, p. 264). It is generally agreed that skilled use of cohesive devices facilitates comprehension of the text. However, there have been conflicting results in the studies of cohesion in writing studies in both first and second languages. For example, McNamara, Crossley, and McCarthy (2011) evaluated features of writing which can predict high- and low-scoring essays using the automated analysis tool Coh-Metrix (Graesser, McNamara & Louwerse, 2004) and found that none of the cohesion devices analysed with the automated tool was correlated with essay ratings, but syntactic complexity, lexical diversity and word frequency are the features which discriminate high and low scored essays. Similarly, Todd, Khongput and Darawanga (2007) showed a lack of cohesive devices did not affect the writing quality. In their study the features of cohesion were analysed only for lexical cohesion based on Hoey's lexical analysis (Hoey, 1991). Liu and Braine (2005) investigated the quality of writing of 50 students in a writing course at a university in China and their analysis of use of cohesive devices shows a moderate relationship between the composition score and use of referential cohesion. It should be noted that these three studies investigated different cohesive devices, although some features overlap.

In spoken language, earlier studies examined the use of discourse markers used by international teaching assistants (ITA) in universities in the US to identify the source of difficulty in comprehending the speech of non-native speakers (e.g., Tyler, 1992; Williams, 1992). Some studies compared the quality of non-native speaker oral production with that of native speakers and other studies investigated factors influencing quality of speech, such as proficiency, tasks, and the amount of preparation time. The findings showed that infrequent or inappropriate use of discourse markers caused difficulties in comprehension (e.g., Fung & Carter, 2007; Tyler, 1992) and that the frequency and type of discourse markers differed according to the learner's proficiency, task types (Geva, 1992), and planning time (Williams, 1992).

2.2.2 Coherence

An earlier definition of coherence in written texts by Cameron, Lee, Webster, Munro, Hunt and Linton (1995) foregrounds the semantic relations between sentences within a text, which then provide the text with "a degree of unity". Focusing on coherence in the context of spoken discourse, Seedhouse and Harris (2011) state that the key indicators of a coherent text in the context of IELTS rating are "logical sequencing of sentences, clear marking of stages in a discussion, narration or argument, and the use of cohesive devices (e.g., connectors, pronouns and conjunctions) within and between sentences" (p. 4).

An additional aspect of coherence was articulated by Celce-Murcia et al, (1995, p. 15), who state that it is characterised by "the degree to which sentences or utterances in a discourse sequence are felt to be interrelated rather than unrelated". The inclusion of "felt to be" reflects the somewhat enigmatic nature of coherence, in that to some extent, particularly in spoken discourse, it rests in the understanding and perception of the listener. In the context of the semi-direct speaking tests, such as Aptis, the listener is the rater, and thus it is important to have access to rater comments, in addition to the transcribed responses of candidates to each task. This definition reflects the point made earlier by De Beaugrande and Dressler (1981) that coherence requires both knowledge a text presents and knowledge of the world that listeners possess. For a text to make sense, interaction between the content of the text and the knowledge of the listener is essential: thus, to some extent, coherence can be argued to be co-constructed. From these definitions, the key aspects of coherence are discourse, which is interrelated, unified and meaningful to the listener. Thus we can say that, while cohesion refers to the internal properties of a text, coherence refers to its "contextual properties; that is the way in which it relates to and makes sense in the situation it occurs" (Paltridge, 2000, p. 139).

Coherence appears to be something of a chimera in second language assessment and research. Given its complexity, it is unsurprising that the construct has proven difficult to operationalise in its entirety. In the CEFR levels: Qualitative aspects of spoken language use (http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf, p. 37–38), coherence is a separate criterion, although at times, the relationship between coherence and cohesion is not clearly articulated. For example, the descriptor for Coherence at Level C2 is: "can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices".

The operationalisation of coherence in standardised speaking tests reflects differing perspectives. Coherence is explicitly mentioned in the public version of the IELTS Speaking band descriptors, but is included in a joint criterion with Fluency. In the descriptor for "Fluency and Coherence" at Band 9, the points relevant to coherence are "speaks coherently with fully appropriate cohesive devices" and "develops topics fully and appropriately" (http://takeielts.britishcouncil.org/sites/default/files/IELTS_Speaking_band_descriptors.pdf). Thus the key aspects of coherence in the context of IELTS are the ability to use cohesive devices and develop topics in a way that is relevant to the task. It is interesting to compare this approach to operationalising coherence with that which is taken in the TOEFL iBT, where in the public version of the speaking rubrics, coherence is mentioned in the overarching "General Description" and also addressed in the separate criterion of "Topic Development". The descriptor for a score of 4 in the rubric for independent speaking tasks states that the "response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear" (https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf). Coherence is not explicitly mentioned in the Aptis rating scales, but seems to be implicit in the overarching descriptor for each level that refers to the extent to which responses are "on topic", along with a descriptor focusing on cohesion for Tasks 2, 3 and 4.

From the documentation of coherence in a range of speaking test rating scales, we can see that it is operationalised in different ways, and that the relationship between coherence and cohesion remains somewhat ambiguous. In order to explore the differing ways in which coherence has been analysed in recent studies, the methodology of one writing study and three speaking studies will be reported.

In a recent study on academic writing in the context of the TOEFL iBT, Knoch, McQueen and O'Hagan (2014) used Topic Structure Analysis (TSA) to evaluate coherence. The patterns they coded were:

- parallel progression: topics of successive sentences are the same (or synonymous)
- direct sequential progression: the comment of the previous sentence becomes the topic of the following sentence
- indirect progression: the topic or comment of the previous sentence becomes the topic of the following sentence; the topic or comment are only indirectly related (by inference)
- superstructure: coherence is created by a linking device instead of topic progression
- extended progression: the topic or comment before the previous sentence becomes the topic of the new sentence
- coherence break: attempt at coherence fails because of an error
- unrelated progression: the topic of a sentence is not related to the topic or comment in the previous sentence (p. 10).

Identification of these patterns enabled the analysis of the extent to which successive sentences in a text were related. Findings indicated that test-takers produced more coherent T-units in the integrated writing task than in the independent task, and that in the independent writing task, low-level writers produced "significantly fewer coherent T-units than the low medium, medium and high medium level writers" (p. 47). While the researchers acknowledged that TSA is "a simplification of the construct of coherence", it was still felt to "capture some of the essence of this important aspect of academic writing" (p. 10). This approach has much to offer for the systematic analysis of textual unity, particularly in written texts, where sentences form easily distinguishable units for analysis.

Two recent studies in L2 speaking operationalised aspects of generic structure to explore coherence. In a study of rater orientations to TOEFL speaking tasks, Brown, Iwashita and McNamara (2005) evaluated the quality of content through "looking at the overall schematic structure of the discourse, using the clause as the unit of analysis" (p. 60). This analysis included obligatory and optional moves that were expected and possible for a particular task. Analyses revealed that the extended integrated speaking tasks elicited identifiable text organisational patterns, including problem–solution, and thus were particularly suited to this analysis. Higher proficiency test-takers were able to present all crucial points in the speech in a well-illustrated and logically connected manner.

Iwashita and Vasquez (2015) also used text generic structure to explore coherence in the context of an IELTS speaking test. They used two measures: a detailed analysis of the development of ideas through the identification of theme and rheme; and an estimate of the degree of discourse compliance through a scale based on the description of prototypical text schematic structure presented in Paltridge (2000). Analyses showed that the performances varied in the number of constant theme sets displayed in the theme reiteration patterns where higher-level test-takers, in particular, included more sets. Additionally, more sophisticated patterns in more proficient performances tended to display a more complex configuration of the pattern itself.

A different approach was taken by Seedhouse and Harris (2011), who utilised Conversation Analysis (CA) to analyse topic development and management in the IELTS speaking test. They positioned the examiner-led interview as a "topic-based Q–A adjacency pair" (p. 1), and concluded that in order to succeed, the candidate must "understand the question they have been asked; provide an answer to that question; identify the topic inherent in the question; and develop the topic inherent in the question" (p. 1). Their findings identified features of high- and low-scoring performances in relation to topic development. An interesting aspect that emerged as relevant to lower-scoring performances was "topic trouble", where a problem may arise "in relation to the question or the topic inherent in the question, or both" (p. 21). They also found that through the way they developed topics and their lexical choices, candidates constructed an identity: "candidates who achieved a very high score typically developed topics that constructed the identity of an intellectual and a (future) high-achiever on the international stage" (p. 25).

2.2.3 Vocabulary use

It is well known that vocabulary plays an important role in communication (Wilkins, 1976) and it is also considered to be strongly linked to general performance (e.g., Bachman & Palmer, 1996; Harley et al, 1990). Features of vocabulary use, such as measures of lexical variation, sophistication, density and number of errors, are considered to "potentially have great value in allowing us to describe key features of spoken vocabulary in a quantitative manner that may provide useful comparisons between test-takers at different proficiency levels (Read & Nation, 2008, p. 7). Furthermore, lexis is considered to be at the heart of meaning-making in understanding discourse (Halliday & Hasan, 1976).

This study, therefore, aims to focus on features of candidate discourse and lexis that are elicited through different tasks and are distinctive at different levels. We thus examine these issues through in-depth analysis of candidate performance. An increasing volume of studies has undertaken an analysis of actual test performance to address issues of validity. Van Lier (1989), for example, stressed the importance of analysis of speech, especially the need to look at oral tests using data from the test performance (i.e., what candidates actually said). Furthermore, recent research has shown that scales which are developed based on sophisticated statistical analysis do not always capture the subtle differences observed in candidate performance (Frost, Elder, & Wigglesworth, 2011; Iwashita et al, 2008). Both the Aptis Speaking scale and the CEFR descriptors refer to aspects of coherence and cohesion. This study, with its focus on discourse competence in speaking, is therefore timely.

3. RESEARCH QUESTIONS

The research questions, which focus on features of candidate discourse and vocabulary use, are as follows:

1. *What are the distinctive characteristics of discourse and vocabulary use at different levels in spoken test performances?*
2. *To what extent do candidates' discourse and vocabulary use vary according to different tasks?*

Given that the Aptis Test is a new test, which is intended to be modified according to different uses in different contexts, O'Sullivan (2015c) notes "it would be naïve to think that the validation process ends with this report" (p. 53). He adds:

> It is important to note that this report describes recommendations made by the standard-setting panel during the final construction stage of the developmental process, at the same time as field trials. These recommendations will need to be revisited in the light of field-trial and operational data analysis and ongoing research and validation. (p. 9)

The current study aims to examine discoursal and lexical performance across these score ranges as part of the validity argument for Aptis. As such, it is expected that the research will contribute to the iterative development of task specifications, tasks and marking criteria.

4. METHODOLOGY

4.1 Data

The present study analysed speech samples provided by the Aptis Test Development team. The data of 249 speech samples comprised a total of 84 candidate performances on four different tasks corresponding to five different levels. The samples were rated according to the Aptis Speaking Task scales, and were accompanied by prompts, scores, and expert rater comments. They had been collected from operational test performances for the purpose of rater training (J. Dunlea, personal communication, 2 November 2014). The digitised data were first transcribed with standard orthography to identify features of discourse competence and lexical richness observed in the test performances. The four tasks in the Aptis Test are targeted at different CEFR levels, and therefore the score for each task performance is assigned differently. In order to compare task performances across levels and tasks, the scores awarded to each performance were converted as shown in Table 1.

CEFR level	Level for analysis	Task 1 Target level A2	Task 2 Target level B1	Task 3 Target level B1	Task 4 Target level B2
C2	5				6
C1	5				5
B2.2	4				4
B2.1	4		5 = B2 (or above)	5 = B2 (or above)	3
B1.2	3		4	4	2
B1.1	3	5 = B1 (or above)	3	3	1
A2.2	2	4	2	2	0 = A1/A2; performance not sufficient for B1
A2.1	2	3	1	1	
A1.2	1	2	0 = Below A2	0 = Below A2	
A1.1	1	1			
A0	0				

Table 1: Performance levels for analyses in the current study

The following table outlines the tasks undertaken by each candidate.

CEFR level	Level for analysis	Task 1 Target level A2	Task 2 Target level B1	Task 3 Target level B1	Task 4 Target level B2
C2	5				6 = P69, 77
C1	5				5 = P68, 72, 73, 75, 82
B2.2	4				4 = P67, 71, 79, 81, 84
B2.1	4		5 = P26, 31, 32, 38, 40	5 = P47, 49, 56, 58, 61	3 = P66, 76
B1.2	3		4 = P25, 27, 34	4 = P46, 48, 53	2 = P65, 70, 80, 83
B1.1	3	5 = P5, 7, 9, 14, 17, 19	3 = P24, 30, 28, 29	3 = P45, 50, 54, 60	1 = P64, 74
A2.2	2	4 = P4, 10, 16	2 = P44, 51, 55, 57, 59, 62	2 = P44, 51, 55, 57, 59, 62	0 = P63, 78
A2.1	2	3 = P3, 6, 12, 15, 20	1 = P43	1 = P43,	
A1.2	1	2 = P2, 11, 13, 18	0 = P42, 52	0 = P42, 52	
A1.1	1	1 = P1, 8			
A0	0				
(total)		*N*=20	*N* =21	*N* =21	*N* =22

Note. Each cell contains information on individual Aptis scores, equivalent CEFR level, and candidates in the data who performed at each level.

Table 2: Distribution of performances across levels and tasks with candidate ID

4.2 Task description

4.2.1 Task 1

Task 1 requires candidates to "answer three questions drawn from a bank of similar personal information exchange questions" (Aptis Training for Examiners), with no planning time permitted. As the three questions in each version of Task 1 are unrelated, each question gives the candidate the opportunity to make a fresh start with a new topic.

4.2.2 Task 2

Task 2 requires candidates to "answer three questions based on some visual input, with questions ranging from purely descriptive to more abstract concepts. The focus is mainly on the experiences of the test-taker, so the cognitive load is not very high" (Aptis Training for Examiners). The questions require description of a picture ("Describe this picture"), a recount ("Tell me about a time you gave or received a gift") and provide what, in a certain version of the task, could be interpreted either as an explanation or an exposition ("Why is it important to give people gifts on special occasions?") or, in another version of the task, is clearly an exposition ("Do you think people should pay to visit museums, or should they be free?").

4.2.3 Task 3

Task 3 requires candidates to "Answer three questions based on two visual prompts on the same theme", with questions ranging from purely descriptive to more abstract concepts. Candidates are required to "Describe, compare and speculate on topics familiar to the experiences of the test-taker, so the cognitive load is not very high" (Aptis Training for Examiners). Visual prompts for Task 3 are of two different activities (for example, golf and basketball) representing an overarching topic (for example, sport). Questions require candidates to describe the two pictures, speculate and give further information about the participants and then compare an aspect of the experience depicted by the pictures, stating and supporting an opinion and thus constructing an argument (which they would prefer, or which is most difficult etc.). The genres required to respond effectively to Task 3 are description and exposition. Although this may appear to be replicating the genres elicited through Task 2, the ability to speculate and compare is explicitly required only in Task 3.

4.2.4 Task 4

Task 4 requires candidates to answer three questions based on a visual prompt, which is not integral to the task, with one minute of planning time. With the addition of planning time, responses would be assumed to be more clearly signposted and topics more fully developed than those in response to Tasks 1, 2 and 3. This is not possible to ascertain, however, as the researchers did not have access to performances by the same candidates across the four tasks.

4.3 Analysis

The method of data analysis employed discourse measures used in Iwashita and Vasquez (2015), who investigated the discourse competence observed in IELTS Speaking Task 2 performance. As in the previous study, we operationalised discourse competence in terms of the textual features of cohesion and coherence, and examined selected aspects of cohesion employing both qualitative and quantitative methods and coherence with qualitative methods. Lexical richness was analysed with quantitative methods only.

4.3.1 Cohesion

For quantitative analysis we used Coh-Metrix (McNamara, Graesser, McCarthy & Cai, 2014). Coh-Metrix is a computational tool designed for interdisciplinary research into text cohesion and coherence, with cohesion defined as observable and measurable "linguistic, semantic and discourse characteristics of the text" (McNamara et al, 2014, p. 19), and coherence identified as "the consequences of cohesion in the mind of the reader". While Coh-Metrix is designed to be used for analysis of both written and (transcribed) "naturalistic oral discourse" (p. 9), the vast majority of research to date appears to have focused on written texts, with some studies on oral discourse involving asynchronous online interaction (McNamara et al, 2014) or "tutorial dialogue" e.g., human–human vs. human–computer; (Graesser, Jeon, Yan & Cai, 2007). In the current study, we used Coh-Metrix only for the analysis of cohesion.

All 249 samples taken from 84 candidates' performances were entered into Coh-Metrix and the results were compared across levels and tasks (except for 16 performances which were below the 100-word lower limit required for analysis in Coh-Metrix). Indices selected from Coh-Metrix are summarised in Appendix 1. When the speech samples were entered into Coh-Metrix, data corresponding to test-takers' responses to questions 1–3 in Tasks 1, 2 and 3 were combined and treated as one performance to make them comparable with the longer Task 4 performances. In total, 84 files were entered into Coh-Metrix for quantitative analysis as shown in Table 1 above. It should be noted that repetition, false starts and repair incidences were all removed before entering the data, as these features are not considered in the Coh-Metrix analysis.

For the quantitative analysis, the mean scores of the frequency count of cohesive measures yielded from Coh-Metrix were compared across the levels and tasks with both descriptive and inferential statistics in order to answer the two research questions. To provide deeper insight into, and exemplify, the results of the quantitative analyses, close examination of selected performances was conducted. Table 3 outlines the 15 performances which were selected for all analyses in the study. Candidates for analysis were chosen from performances which were close to the average of Coh-Metrix measures for cohesion, as well as those which answered all questions for each task, where possible. These performances were drawn on for qualitative analysis of both cohesion and coherence.

To present this data systematically, they were segmented into C-units (communication unit) by two research assistants (inter-coder agreement was 96.5%). C-units are often used for oral discourse, as they are based on the clause (like T-units) but also include phrases which carry communicative meaning (Crookes, 1990), whether or not they are grammatical (Foster, Tonkyn & Wigglesworth, 2000). For the sake of exemplification of findings, we present examples of Task 4 performances across levels for RQ1 for cohesion, and examples across comparable tasks taken from Tasks 1, 2, 3 and 4 for coherence, while selected examples across tasks at a comparable level for cohesion are presented for RQ2. Applicable rater comments were drawn on for closer analysis of cohesion and coherence focusing on selected performances.

Levels for quantitative analysis					
	1 (A1)	2 (A2)	3 (B1)	4 (B2)	5 (C1 – C2)
Task 1 (target level A2)	P13 (125; 3)	P4 (93; 4)	P9 (118; 15)		
Task 2 (target level B1)	P21 (few cohesive devices; 71; 1)	P33 (simple, limited; 138; 11)	P24 (simple; 127; 13)	P31 (a range of cohesive devices; 130; 22)	
Task 3 (target level B1)	P52 (few cohesive devices are used; 126; 5)	P44 (simple cohesive devices only; 97; 8)	P46 (simple and more complex; 105; 17)	P56 (broad range; 91; 19)	
Task 4 (target level B2)		P63 (Simple; 71; 4)	P65 (Simple; 124; 17)	P81 (a range of cohesive devices; 97; 17)	P69 (a range of complex cohesive devices; 134; 20)

Note: information under the candidate IDs are rater comments (except for Task 1, where cohesion is not assessed); Coh-Metrix rating; raw count. P63 = Performance 63 – both are used interchangeably throughout the report; When referring to the individual whose performance is reported, the term "candidate" is used.

Table 3: Performances selected for qualitative analysis

Following Halliday and Matthiessen (2013), we identified three measures of cohesion: conjunction, reference and lexical cohesion.

4.3.1.1 Conjunction

This resource "creates cohesion by linking whole clauses or combinations of clauses" (Halliday & Matthiessen, 2013, p. 604). It represents logico-semantic relationships between components of a text at the clause level. In the quantitative analysis, seven indices in Coh-Metrix were included: causal "because", "so"; logical "and", "so"; contrastive "although", "whereas"; temporal and temporal expanded "first", "until"; additive "and", "moreover"; and the combined incidence of all conjunctions. It should be noted that the conjunctions analysed in Coh-Metrix are not strictly conjunctions in grammatical terms, which connect words, phrases and sentences. Adverbial terms such as "first," "actually" indicating transitions were also included in Coh-Metrix analyses.

4.3.1.2 Reference

Referential cohesion occurs when a noun or noun-phrase argument refers to another constituent in the text. In Coh-Metrix, this is identified by an overlap in noun or noun-phrase between local sentences. Four types of referential cohesion are included in Coh-Metrix. These are identified by computing overlap among nouns, arguments, stems (morpheme units), and content words as outlined below.

- Noun overlap: two sentences share one or more common nouns; the proportion of sentences in a text for which there are overlapping nouns, with no deviation in the morphological forms of the nouns (e.g., table/table).
- Argument overlap: overlap between the head nouns (e.g., table/tables) and pronouns.
- Stem overlap: overlap between a noun in one sentence and a content word (i.e., nouns, verbs, adjectives) in another sentence. The content word in the other sentence must share a common lemma (i.e., price/priced).
- Content word overlap: the proportion of explicit content words that overlap between pairs of sentences.

(McNamara et al, 2014 p. 64–65)

Coh-Metrix calculates referential cohesion (overlap) in two dimensions (i.e., local and global). Local cohesion is measured by assessing the overlap between adjacent sentences, and global cohesion by counting the overlap among all sentences in a text.

In assessing the types of overlap shown above, while the first three types of indices (i.e., noun, argument and stem overlaps) are binary (i.e., overlap is observed or not between a pair of adjacent sentences and all sentences), the last type, content word overlap was examined by calculating the proportion of content words (nouns, verbs, adverbs, adjectives and pronouns) that are shared between sentences.

4.3.1.3 Lexical cohesion

This resource operates at the lexical level and "it is achieved through the choice/selection of lexical items ... these cohesive relations [may] hold between single lexical units [or] wordings having more than one lexical item in them" (Halliday & Matthiessen, 2013, p. 642). In the current study, lexical cohesion was investigated in terms of hypernym–hyponym relationships. A *hypernym* is a word or phrase that can apply to the original word and others (e.g., a bird is a hypernym of pigeon, crow, eagle and seagull, and these are all hyponyms of bird). Coh-Metrix calculates instances of hypernymy for nouns, verbs and combinations of both nouns and verbs.

4.3.2 Vocabulary use

Vocabulary use was examined in terms of word frequency (all words, content words – nouns, verbs, adjectives, adverbs, and pronouns) and type–token ratios. Coh-Metrix provides two indices of type–token ratio (i.e., content words only and all words). For word frequency, because of the different required length of speech for each task, the data were converted to frequency score (per 60 seconds). For other word counts, Coh-Metrix calculates incidence per 1000 words. These measures were selected following successful use in previous studies (e.g., Iwashita & Vasquez, 2015; Iwashita et al, 2008).

Diversity of vocabulary used by test candidates was assessed according to the variety of words (types) used in a text in relation to the total number of word tokens. A smaller variety of words in a text indicate that more words are used multiple times across the text, and therefore the level of cohesion is considered to be higher than when a larger variety of words is observed in a text. It is well known that type–token ratio is correlated with text length (Malvern & Richards, 2000), and so we included the index (VOCD) in which text length is taken into account in analysis, in addition to type–token ratios.

4.3.3 Coherence

In the purely qualitative analysis of coherence, as in the qualitative analysis of cohesive devices explained above, we used candidate discourse, segmented into C-units, as our main data source. In addition, we also had access to visual and oral prompts for each task, scores awarded for the performance as a whole, and rater comments which were included in the Aptis rater training materials made accessible to us. Based on the literature on the analysis of candidate performance in L2 assessment studies introduced in Section 2.2.2, we identified the extent to which there was textual unity and a meaningful response to each question, in terms of three aspects:

1. relationship of successive C-units to each other, adapting TSA analysis from Knoch et al.'s 2014 study
2. topic development in relation to the question asked, adapting TSA analysis from Knoch et al.'s 2014 study
3. aspects of expected genre where relevant (e.g., Description and Recount in Task 1) and expected patterns of organisation (e.g., Comparison and Contrast in Task 3) applying genre analysis as used in Iwashita and Vasquez's 2015 study.

The analysis of coherence consisted of a “thick description” of each C-unit from performances of each task, selected as representing a range of performances, (including weak and strong), as evidenced through ratings and rater comments. Thus, in a total of 16 performances, textual unity was considered in relation to the relevance of the response to the questions asked, the extent to which the information in the text logically developed following expected generic structures (e.g., recount), patterns of organisation (e.g., comparison and contrast) and the ways in which each C-unit related to others in the response. C-units can be related through strands of experiential meaning, expressed in lexical chains, which are defined by Butt, Fahey, Feez and Spinks (2012, p. 249) as “chains of words related by repetition, similar or opposite meanings, association, composition (part/whole) or classification (types of) relationships”. Clauses can also be linked through strands of logical meaning by the use of conjunctions, while “strands of textual meaning keep track of people and things as the text unfolds, using reference chains of language elements such as pronouns, definite articles” (p. 249).

Inherent in the coherence analysis was the extent to which utterances related to the question, and thus, specific aspects of a topic and the way in which this topic was developed. Topic development is important in our study, as it is implicit in the construct of coherence, as the task, including the specific questions that candidates are required to address, forms the context for the response. An analysis of features of coherence must take this into account. It is important to note that in the Aptis Speaking Test the brevity of responses, particularly in Task 1, makes it difficult to comprehensively apply text generic structure. However, there are aspects of the description, exposition, explanation and recount genres implicit in tasks, and organisational patterns including comparison and contrast can be identified in response to specific questions.

4.3.4 Summary of the methods of analyses

4.3.4.1 Cohesion – Quantitative analyses

Frequencies for the cohesive devices identified and explained above (i.e., conjunction, references, lexical) were first obtained using Coh-Metrix. In addition to reporting the descriptive statistics for conjunctions and vocabulary use, we report the proportion of each of conjunction type and vocabulary type to see the distribution of different types of conjunction/vocabulary items. These results are reported in the end of the section. We used non-parametric correlation statistics (Spearman’s *Rho*) as the level data is ranked data. Because of the robustness of ANOVA statistics (Ito, 1980), we used parametric ANOVA (see the detailed results of normality tests in Appendix 2) to compare the mean score across levels and tasks, even though some data were not normally distributed.

The following is a summary of analyses in light of each research question.

- Descriptive statistics: frequency data of each feature of cohesion and vocabulary in each level (RQ1) and in each task (RQ2).
- Correlational statistics (Spearman’s *Rho*) (RQ1 only).
- Two-way ANOVA analyses
    - differences across the levels – RQ1
    - differences across the tasks – RQ2
    - interaction effect of levels and tasks – RQ2.
- Proportion of each of the conjunction type and vocabulary type to see the distribution of different types of conjunction or vocabulary items.

4.3.4.2 Cohesion – Qualitative analyses

Qualitative analyses of the selected candidate performances are presented after each set of related quantitative analyses of cohesion (i.e., conjunction, reference, and lexical cohesion) in the results for RQ1 and RQ2. These are presented as tables or figures including information on level of performance (Aptis and CEFR ratings), samples of candidate speech (separated into C-units, see Section 4.2.1 above), with cohesive devices of interest (i.e., those identified in Coh-Metrix automated analyses; see Sections 4.2.1.1 to 4.2.1.3 above) highlighted in bold text. Following is a summary of qualitative analyses (exemplification) of cohesion for both research questions.

- *Conjunction*: where they occur, specific instances of additive, causal, temporal, adversative, and other types of conjunction are identified and compared across levels (RQ1) for the same task, and tasks at a comparable level (RQ2).
- *Reference*: types of overlap, as per the definitions provided by the developers of Coh-Metrix (McNamara et al, 2014; outlined in Section 4.2.1.2 above) are identified and compared.
- *Lexical cohesion*: instances of hypernymy (included in Coh-Metrix), and meronymy (following Iwashita & Vasquez, 2015) are identified and compared.

4.3.4.3 Vocabulary use – Quantitative analysis

Analysis of vocabulary use was undertaken in the same way as quantitative analysis of cohesion devices explained above (Section 4.3.4.2).

4.3.4.4 Coherence – Qualitative analyses

Coherence, in terms of aspects characterising higher and lower performances and features associated with different tasks, was analysed in terms of three aspects:

- relationship of successive C-units to each other
- topic development in relation to the question asked
- aspects of expected genre where relevant (e.g., Description and Recount in Task 1) and expected patterns of organisation (e.g., Comparison and Contrast in Task 3).

5. RESULTS

In presenting the results of the examination of the two research questions addressed for the current study, we will report the results of the cohesion analysis (for both quantitative and qualitative analyses) and the analysis of vocabulary use in light of the research question in Sections 5.1 to 5.4.

When conducting a purely qualitative analysis of coherence, it was neither possible nor meaningful to separate the quality of the performance (RQ1) from the contextualised response to a particular task requiring a specific genre (RQ2). Task 1, for example, required candidates to engage with two genres: description and recount. In order to analyse coherence in performances which received high and low scores in Task 1, it was thus essential to view these performances in terms of the appropriateness of the structuring of information and development of the topic required for an effective description or recount. Section 5.1.2 contains analysis of coherence in a range of performances for Tasks 1, 2, 3 and 4.

5.1 Cohesion – Research question 1

The first research question examines whether discoursal features and lexical richness observed in candidate performances are varied according to assigned scores.

5.1.1 Conjunctions

Descriptive statistics of the frequency of conjunctions observed in task performances are summarised in Table 4. As shown in Table 4, higher-level candidates (i.e., Levels 4 and 5) used logical and adversative/contrastive conjunctions, more frequently than lower-level candidates (i.e., Levels 1 and 2). However, there is no pattern in the use of temporal, expanded temporal and additive conjunctions. For example, Level 1 candidates produced the largest token of expanded temporal conjunctions followed by Level 5 candidates. For additive conjunctions, the largest token was produced by Level 3 candidates. As shown in the large standard deviation scores, individual variation is relatively large. In order to investigate whether there is any difference in the frequency of these conjunctions across the levels, Spearman's correlation and ANOVA analyses were carried out and the results are reported in Tables 5 and 6 respectively.

Conjunction type	Level 1 (*N* = 16)		Level 2 (*N* = 16)		Level 3 (*N* = 15)		Level 4 (*N* = 14)		Level 5 (*N* = 23)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Causal	31.27	25.29	28.23	19.68	34.05	14.50	38.28	2.12	36.45	17.19
Logical	51.09	35.92	47.19	26.57	6.53	17.87	57.59	2.45	6.63	24.65
Adversative Contrastive	7.77	12.80	8.04	1.22	12.86	12.32	14.42	9.12	18.99	15.02
Temporal	13.38	18.09	13.14	1.92	11.83	7.51	12.62	6.03	15.19	9.88
Expanded temporal	19.69	19.45	18.77	2.15	1.58	11.29	12.18	11.64	18.42	7.36
Additive	63.13	49.04	65.33	33.81	75.85	24.87	56.76	12.82	56.77	22.23
All (Combined)	101.82	57.86	104.72	31.47	119.59	25.20	105.96	2.85	11.11	26.75

Table 4: Descriptive statistics of conjunction use in each level (per 1000 words)

The results of the correlational statistics are summarised in Table 5. The significant relationship between levels and frequency of conjunction was found only in adversative/causative and additive conjunctions. ANOVA analyses revealed no significant difference in the frequency of all seven types of conjunctions (Table 6).

	Causal	Logical	Adversative Contrastive	Temporal	Expanded temporal	Additive	All (Combined)
Level	-.128	-.006	.255*	-.053	.025	.331**	.209
Causal		.674**	.012	.079	.446**	-.219*	.302**
Logical			.483**	.288**	.256*	.051	.457**
Adversative Contrastive				.13	.105	.244*	.247*
Temporal					-.016	.041	.185
Expanded temporal						-.158	.098
Additive							.716**

Notes: * Correlation is significant at the .05 level (2-tailed); ** Correlation is significant at the .01 level (2-tailed).

Table 5: Correlation between level and frequency of conjunction (per 1000 words) (Spearman's rho)

Type	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial Eta Squared
Causal	614.57	4	153.64	.43	.79	.02
Logical	1709.71	4	427.43	.83	.51	.05
Adversative Contrastive	47.96	4	117.74	1.25	.30	.07
Temporal	328.75	4	82.19	.75	.56	.04
Expanded temporal	1226.28	4	306.57	1.46	.23	.08
Additive	5046.41	4	1261.60	1.55	.20	.08
All (Combined)	4576.12	4	1144.03	1.26	.30	.07

Table 6: Results of ANOVA analyses by level (conjunction use)

Table 7 and Figure 1 below display the proportion of each of the six conjunction types used by candidates of the five levels based on the assigned scores. While the proportion of additive conjunctions was the largest for Level 1, 2 and 3 candidates, Level 4 and 5 candidates used logical conjunctions most followed by additive conjunctions. Regardless of the levels, causal conjunctions were the third most frequently used conjunctions across the levels. While contrastive conjunctions were the least frequently used conjunctions among Level 1 and 2 candidates, Level 4 and 5 candidates used temporal conjunctions least frequently.

Type	**Level 1 (*N* = 16)**	**Level 2 (*N* = 16)**	**Level 3 (*N* = 15)**	**Level 4 (*N* = 14)**	**Level 5 (*N* = 23)**
Causal	16.78	15.62	16.55	19.95	17.66
Logical	27.42	26.12	29.43	3.02	29.37
Adversative Contrastive	4.17	4.45	6.25	7.52	9.20
Temporal	7.18	7.27	5.75	6.58	7.36
Expanded temporal	1.57	1.38	5.14	6.35	8.92
Additive	33.88	36.16	36.87	29.58	27.50

Table 7: Distribution of conjunction use (%) in each level

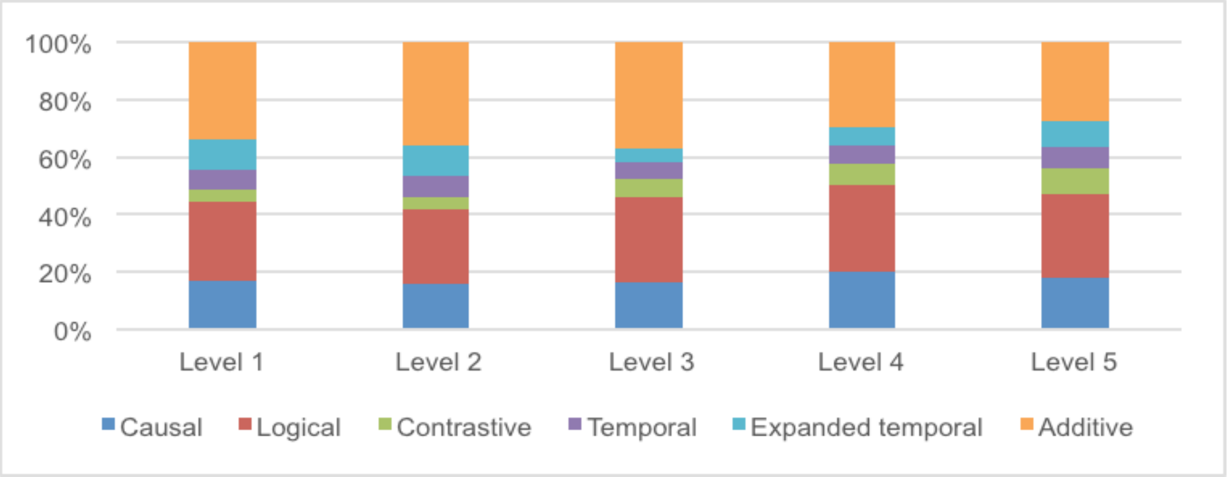


Figure 1: Distribution of conjunctions in each level

As noted above, in order to provide deeper insight into the findings of the quantitative analysis, close examination of selected samples was carried out. Raters' comments on the performance were also considered in the analysis. The examples below from Task 4 provide more specific exemplification of findings regarding the use of conjunctions. Each table title includes information regarding the level of the performance, including: the rating for our analysis (as outlined in Table 1 above); the Aptis rating of the script; and the equivalent CEFR level. Each table in the qualitative analysis below also includes extracts of the candidates' speech, with bold text highlighting features identified in Coh-Metrix automated analyses of conjunction, reference, and other features. In the examples below candidates across levels drew on prototypical conjunctions for each category of conjunction: additive – "and"; temporal – "when"; adversative/contrastive – "but"; and causal – "because," "so." The performances mainly drew on simple prototypical conjunctions. Employing both quantitative and qualitative analyses, conjunctions alone appear to provide little distinction of performances across levels.

	Candidate's speech	Conjunctions
1	Last year I visited Praha	
2	**and** visited a very tall building	Additive
3	it was very interesting to visit a very tall building	
4	I think there are many tall buildings	
5	**because** there are many people	Causal
6	**and**, er, ah, it's not much place to build	Additive
7	**so**, ah, it is necessary to build a tall building	Causal
8	ah, in this nice picture we see a lot of tall buildings	
9	it is very interesting	
10	those tall buildings, ah, are very nice	

Table 8: Performance 63, Task 4, conjunctions (Level 2; Aptis 0; CEFR A2)

Performance 63 (P63; Table 8) is far shorter than those scoring higher, and only provides four basic conjunctions – two additive (lines 2 and 6) and two causal (lines 5 and 7). As noted earlier, it was rated as involving "simple" use of cohesive devices.

	Candidate's speech	Conjunctions
1	Actually I think, em, while I'm alone at home I used to read books **and** used to hear music	Additive
2	**but** I think mostly in my childhood I used to live alone normally **because** my parents were not allowed to stay home alone **because** my parents, both my mum and father is having went … mm while they went to their job and my brother went to school and I be always free at home.	Adversative/ contrastive; Causal; Causal
3	I don't have much works to do.	
4	I be always I sit alone at home	
5	**and** do something else for (unint).	Additive
6	While watching, while seeing this picture I think this girl is concentrating or thinking something else **and** she's sad.	Additive
7	I think her facial expression seems that she's sad a bit.	Additive
8	She's, she's something thinking about something for her past will be, thinking about her parents or family or her friends and things.	Causal
9	**And** while I'm staying alone at home I used to read books.	Additive
10	I used to hear music	
11	**and** my main hobby is hearing music	Additive
12	**and** I'm interested in photography	Additive
13	**so** I just take some photos of interesting pics of ants doing something or birds flew, flew in the, in the trees or staying on the trees.	Causative
14	The main thing, one of the main things I remember while I was staying alone at home is that once I played some tricks over my grandmother while staying alone because I was staying alone that day	
15	**and when** my grandmother came this evening at home I just played some trick on him by throwing something, just throwing my pillows and (unint) on him or her.	Additive; temporal
16	**Actually** I think I just remember that times actually I was yes, always alone at home at those times.	Adversative/ contrastive
17	I was happy at that **but** eh, this woman,	Adversative/ contrastive
18	**actually** I like to be alone at home **because** being alone means you're always thinking about many of the opportunities which we will get …	Adversative/ contrastive; Causal

Table 9: Performance 65, Task 4, conjunctions (Level 3; Aptis 2; CEFR B1)

The longer performance in P65 (Table 9) was also commented as involving "simple" use of cohesive devices, though, as exemplified above, the range of conjunctions is more sophisticated than the preceding performance (e.g., the use of the adversative "actually", lines 16 and 18, compound conjunctions, line 15, and the use of a broader range throughout).

	Candidate's speech	Conjunctions
1	I will start with my favourite piece of clothes	
2	**and** I guess that is the shoes.	Additive
3	I love it.	
4	I love the shoes.	
5	I have a lot of shoes	
6	**and** I like it **because** you know **when** you're wearing a pair of shoes that they looks really good you look like, I don't know, like nice looking, like important person.	Additive; Causal; temporal
7	Eh, you can dress **but** if you're dressed in a nice pair of shoes like leather shoes with a lot of good quality,	Adversative/contrastive
8	sometimes they are expensive **but** eh, it's quite nice to spend some money on a good pair of shoes.	Adversative/contrastive
9	I love it.	
10	**And** I feel, as they say, like really secure, you know, secure of, of how I wear, you know.	Additive
11	I'm secure of, eh, everyone, um, everybody will pay attention of what I'm wearing	
12	**and** how it looks like	Additive
13	**and** this is quite important.	Additive
14	Me, people dress like a different ways it depends on their culture.	Additive
15	As I see in, in the picture here, it's a male that is wearing a leather (unint) can be from their Arabic **so** they always wear this kind of clothes.	Causal
16	**Also** they wear like a turban because it's important to keep your head, eh, from the sun **and** also from the temperature.	Additive Additive
17	Sometimes if you're wearing a lot of clothes it's **because** you are in a cold country	Causal
18	**and** you need to wear a special dress.	Additive
19	**Or** imagine that if you are living in a mountain or if you are living in a beach.	Additive
20	If you are living in a beach you don't need a lot of things to wear **so** sometimes it depends on the temperature, the weather, the country, even the culture.	Causal

Table 10: Performance 81, Task 4, conjunctions (Level 4; Aptis 4; CEFR B2)

P81 (Table 10) was identified by the rater as involving "a range of cohesive devices". This appears not to be more sophisticated than the previous performance, at least in the use of conjunctions, suggesting that other features of the performance may be involved in the rater's perspective. This will be further discussed later in the analysis of coherence.

	Candidate's speech	Conjunctions
1	Okay, I'm thinking about 45 years ago I went to China in a, you know it was a trip,	
2	it was an official trip,	
3	**and** I went to a couple of cities and one of them was Shanghai.	Additive
4	In Shanghai I I, I went to, I can't remember the name of there	
5	**but** it was Shanghai Tower or something.	Adversative/ contrastive
6	It was a tower with, you know, with the floor is in glass	
7	**so** you can step in there	Causal
8	**and** actually feel the emptiness that you can feel that, I don't know if I can use that word,	Additive
9	**but** anyway, to me it was a very tough experience	Adversative/ contrastive

10	**and** to be honest I wasn't able to step in there **because** I was kind of, I don't know what the right word for that **but** claustrophobia or something,	Additive;
		Causal;
		Adversative/ contrastive

11	**but** anyway I was scared, my legs were shaking,	Adversative/ contrastive
12	**so in order to** take a picture, **because** you want to take a picture of that moment, you want to record that moment, I just, I just had to actually crawl in my back using my hands and my legs and, you know, to get there	Causal; Causal
13	**and** I ask a friend to take me a picture.	Additive
14	I think tall buildings are necessary for, you know, **because** there is reduced space in, in big cities	Causal
15	**so** they need to find ways to, to give decent, you know, locations for working or just for living	Causal
16	**and** that's why they, they kind of stuck floor to floor **in order to** allow people to live in there **because** living in a big city it has a lot of um, um, I would say um, benefits.	Additive; Causal; Causal
17	**So, so** that's why, that's why they need to, to make room.	Causal
18	In fact, today they actually build a lot of ways in the, you know, kind of bridges	
19	**so** they can avoid a lot of traffic jams	Causal
20	**so, so** basically they're growing to, to go up.	Causal

Table 11: Performance 69, Task 4, conjunctions (Level 5; Aptis 6; CEFR C1-2)

P69 (Table 11) was identified as involving "a range of complex cohesive devices". At the level of conjunction, it appears slightly more complex than the previous performance, including the use of the complex conjunction "so in order to" (line 12). This suggests again that aspects related to text structure, such as staging transitions between text sections, were also involved in this perspective. This will be also discussed later in the analysis of coherence.

In summary, the quantitative analyses revealed a considerable number of the conjunctions used in performance are "additive" and "logical" conjunctions across the levels, and there is little difference across the levels in the frequency of the five types of conjunctions under investigation. The above qualitative findings relating to the use of conjunctions for Speaking Task 4 appear to distinguish the lowest rated performance from those at higher levels, but at the higher levels conjunctions alone appear to provide little distinction of performances across levels. More complex and effective spoken performances, as reflected in Aptis scoring and rater comments, appear to be achievable with only slightly more complex conjunctions than simpler, less effective performances.

5.1.2 Reference

As explained in the methodology section, referential cohesion was examined in two dimensions (i.e., local – between adjacent sentences, and global – among all sentences in a text). Descriptive statistics of the use of references observed in candidate performance at each level are summarised in Table 12 and presented visually in Figures 2 and 3. The use of reference between adjacent sentences was not very different across levels and individual variation is very large as shown in the large standard deviations. For reference observed among all sentences, slightly larger differences than reference between adjacent sentences were found. As shown in Figures 2 and 3, across the levels in both dimensions (i.e., local and global), argument (i.e., overlap between the head nouns and pronouns) was the most frequently observed referential use, and content word overlap was least frequently observed.

Reference (overlap) type	**Level 1 (*N* = 16)**		**Level 2 (*N* = 16)**		**Level 3 (*N* = 15)**		**Level 4 (*N* = 14)**		**Level 5 (*N* = 23)**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Adjacent sentences (local)										
Noun	.26	.19	.34	.22	.29	.17	.35	.22	.32	.11
Argument	.53	.25	.66	.25	.70	.17	.78	.15	.78	.16
Stem	.27	.20	.37	.23	.40	.22	.43	.21	.42	.16
Content	.22	.10	.22	.08	.21	.06	.23	.05	.19	.04
All sentences (global)										
Noun	.18	.12	.26	.23	.23	.13	.29	.21	.23	.11
Argument	.43	.15	.55	.24	.63	.17	.69	.17	.67	.19
Stem	.19	.15	.30	.25	.32	.17	.39	.22	.38	.25
Content	.16	.08	.18	.07	.17	.05	.19	.05	.14	.03

Table 12: Descriptive statistics of references in each level (per 1000 words)

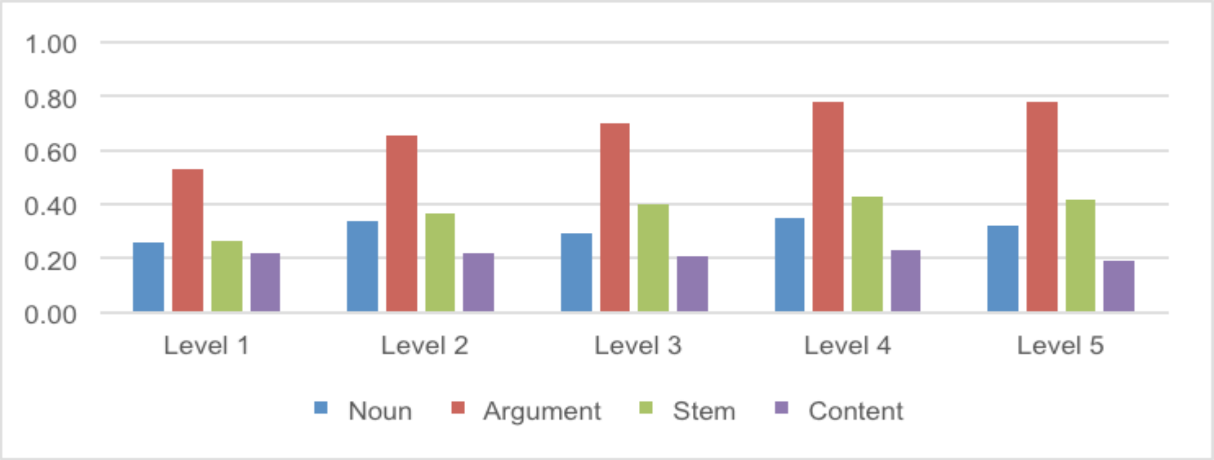


Figure 2: Distribution of references (adjacent sentences – local)

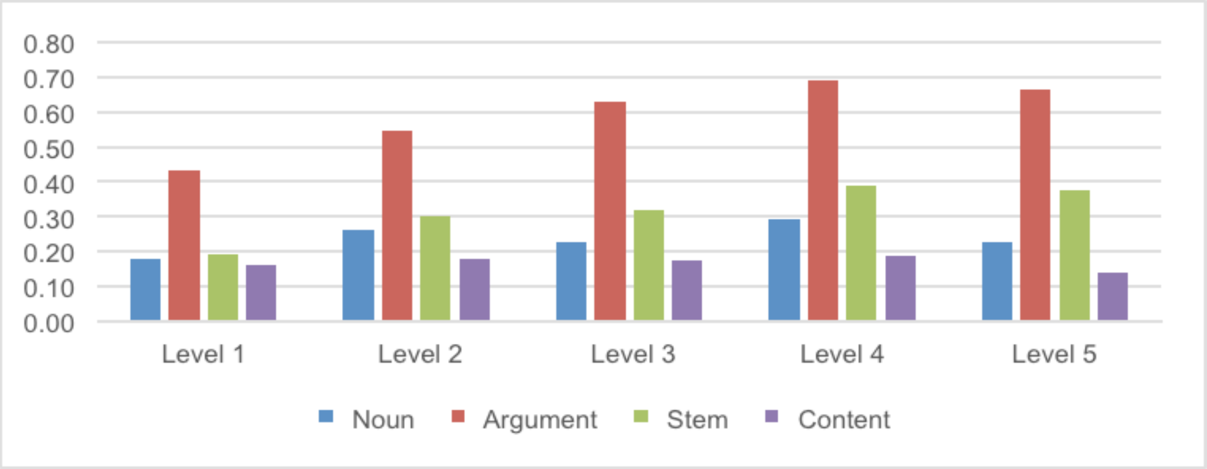


Figure 3: Distribution of references (global)

The results of correlation analyses are summarised in Tables 13 and 14. None of the referential expressions, both between adjacent sentences and among all sentences in a text, were significantly correlated with the assigned level. However, all referential expressions, both between adjacent sentences and among all sentences, are significantly correlated. That means candidates who used referential expression in one category also employed the device in other categories.

	Noun	Argument	Stem	Content
Level	.15	.11	.13	.10
Noun		.342**	.856**	.720**
Argument			.332**	.355**
Stem				.739**

Notes: * Correlation is significant at the .05 level (2-tailed)
** Correlation is significant at the .01 level (2-tailed).

Table 13: Correlation between level and frequency of reference (adjacent sentences – local) (Spearman's rho)

	Noun	Argument	Stem	Content
Level	-.02	.16	.08	-.08
Noun		.558**	.491**	.688**
Argument			.216*	.427**
Stem				.773**

Notes: * Correlation is significant at the .05 level (2-tailed)
** Correlation is significant at the .01 level (2-tailed).

Table 14: Correlation between level and frequency of reference (all sentences – global) (Spearman's rho)

The results of ANOVA analyses in Table 15 show no significant difference in referential cohesions between adjacent sentences, but for referential cohesion among all sentences in a text for argument (i.e., overlap between the head nouns and pronouns) and content word overlap, there is a large effect size. In other words, the frequency of referential cohesion measured with argument and content words among all sentences are significantly different across the levels. Post hoc analyses show the differences were found between Levels 1 and 2 (*p* = .009) 3, (*p* = .004) and 4 (*p* = .001) for argument and between Levels 2 and 3 (*p* = .034), and Levels 2 and 4 (*p* = .045) for content overlap.

Type	Type III Sum of Squares	*df*	Mean Square	*F*	*p*	Partial *Eta* Squared
Adjacent sentences (local)						
Noun	.06	4	.01	.33	.86	.02
Argument	.26	4	.07	1.85	.13	.10
Stem	.03	4	.01	.14	.97	.01
Content	.02	4	.01	.98	.42	.05
All sentences (global)						
Noun	.13	4	.03	1.17	.33	.06
Argument	.32	4	.08	2.63	.04	.13
Stem	.12	4	.03	.85	.50	.05
Content	.03	4	.01	2.77	.03	.14

Notes: effect size .01 > small, .06 > medium, .13 > large

Table 15: Results of ANOVA analyses (references)

As with the conjunctions data above, Task 4 examples are provided for exemplification of referential cohesion, as these were the longest texts, and this task involved more closely related questions than the other three tasks. Argument overlap revealed the highest incidence across levels. This is most likely because it is an inclusive measure, incorporating overlap between head nouns (e.g., table ⇔ tables), and pronouns (table ⇔ it) while noun overlap requires identical morphological forms (e.g., table ⇔ table) (McNamara et al, 2014).

What appears to separate the lower from the higher-rated performances is the (over)representation in the lower-rated transcripts of repetition, (especially repetition of terms which appear in the question prompts; further discussed with regard to coherence below); and either lack of anaphora or errors in the use of anaphora.

	Candidate's speech	Type of overlap
1	Last year I visited Praha	
2	and **visited** a very tall building	Content
3	**it** was very interesting to visit a very **tall building**	Argument; content; noun
4	I think there are **many tall buildings**	Content; content; stem
5	because there are **many** people	Content
6	and, er, ah, it's not much place to **build**	Content
7	so, ah, it is necessary to **build** a **tall building**	Content; content stem
8	ah, in **this** nice picture we see a lot of **tall buildings**	Argument; content; stem
9	**it** is very **interesting**	Argument; content
10	**those tall buildings**, ah, are very nice	Argument; Content; noun

Table 16: Performance 63, Task 4, reference (Level 2; Aptis 0; CEFR A2)

The short text (75 words) in P63 in Table 16 above is notable in its repetition of the phrase “tall building/s” six times, comprising almost 20% of the text’s tokens; while the use of argument overlap is generally accurate, there is very little use of pronouns to link the few ideas in the text. This will be further discussed in the coherence analysis.

	Candidate’s speech	Type of overlap
1	Actually I think, em, while I’m alone at home I used to read books and used to hear music	

2	but I think mostly in my childhood I used to live **alone** normally because my parents were not allowed to stay **home alone** because my **parents**, both my mum and father is having went … mm while they went to their job and my brother **went** to school and I be always free at **home**.	Content;
		Noun; Content;
		Noun;
		Content; Noun

3	I don’t have much works to do.	
4	I be always I sit **alone** at **home**	Content; noun
5	and do something else for (unint).	
6	While watching, while seeing this picture I think **this** girl is concentrating or thinking something else and **she’s** sad.	Argument; Argument
7	I think **her** facial expression seems that **she’s sad** a bit.	Argument; argument, content

8	**She’s, she’s** something thinking about something for **her** past will be, **thinking** about **her** parents or family or **her** friends and things.	Argument;
		argument;
		Content;
		Argument;
		Argument

9	And while I’m staying **alone** at **home** I used to **read books**.	Content; noun; content; noun
10	I used to **hear music**	Content; noun
11	and my main hobby is **hearing music**	Stem; noun
12	and I’m interested in photography	
13	so I just take some **photos** of **interesting** pics of ants doing something or birds flew, flew in the, in the trees or staying on the **trees**.	Stem; stem noun

14	The main thing, one of the main **things** I remember while I was **staying alone** at **home** is that once I played some tricks over my grandmother while **staying alone** because I was **staying alone that** day	Stem
		Stem; content;
		noun; content;
		content; content;
		content; argument

15	and when my **grandmother** came **this** evening at **home** I just **played** some **trick** on him by throwing something, just **throwing** my pillows and (unint) on **him** or her.	Noun; argument;
		noun; content;
		stem

16	Actually I think I just remember **that** times actually I was yes, always **alone** at **home** at **those** times.	Argument;
		content; noun;
		argument

17	I was happy at that but eh, **this** woman,	Argument
18	actually I like to be **alone** at **home** because **being alone** means you’re always thinking about many of the opportunities which we will get …	Content; noun; stem; content

Table 17: Performance 65, Task 4, reference (Level 3; Aptis 2; CEFR B1)

P65 (Table 17) repeats the terms "alone" 11 times, "home" nine times and "stay" six times. With regard to the use of anaphora, the candidate refers to his grandmother as "him" (line 15); and includes other errors in the use of pronouns, e.g., "that times" (line 16), as well as sometimes having a lack of clarity with regard to the referent. The transcript also includes repetition of proposition – e.g., "while staying alone because I was staying alone" (line 14) – interrupting the flow of the text.

	Candidate's speech	**Type of overlap**
1	I will start with **my** favourite piece of clothes	Argument
2	and I guess **that** is the shoes.	Argument
3	I love **it**.	Argument
4	I love the **shoes**.	Noun
5	I have a lot of **shoes**	Noun
6	and I like **it** because you know when you're wearing a pair of **shoes** that **they** looks really good you look like, I don't know, like nice looking, like important person.	Argument; noun; argument
7	Eh, you can dress but if you're **dressed** in a **nice** pair of **shoes** like leather **shoes** with a lot of good quality,	Stem; content; noun; noun
8	sometimes **they** are expensive but eh, **it's** quite **nice** to spend some money on a good pair of **shoes**.	Argument; argument; content; noun
9	I **love it**.	Content; argument
10	And I feel, as they say, like really secure, you know, **secure** of, of how I **wear**, you know.	Content; stem
11	I'm **secure** of, eh, everyone, um, everybody will pay attention of what I'm **wearing**	Content Stem
12	and how it **looks** like	Content
13	and this is quite **important**.	Content
14	Em, people **dress** like a different ways it depends on their culture.	Content;
15	As I see in, in the picture here, it's a male that is **wearing** a leather (unint) can be from their Arabic so **they** always **wear** this kind of **clothes**.	Stem; argument; stem; noun
16	Also **they** wear like a turban because **it's** important to keep your head, eh, from the sun and also from the temperature.	Argument; argument
17	Sometimes if you're **wearing** a lot of **clothes** it's because you are in a cold country	Stem; noun
18	and you need to **wear** a special **dress**.	Stem; content
19	Or imagine that if you are living in a mountain or if you are **living** in a beach.	Content
20	If you are **living** in a **beach** you don't need a lot of things to wear so sometimes it depends on the **temperature**, the weather, the **country**, even the **culture**.	Content; noun; noun; noun

Table 18: Performance 81, Task 4, reference (Level 4; Aptis 4; CEFR B2)

P81 (Table 18) begins like P63, repeating the term shoes seven times in the first eight C-units. This performance is also distinct from that of P65, in the broader and more accurate use of argument overlap.

	Candidate's speech	Type of overlap
1	Okay, I'm thinking about 45 years ago I went to China in a, you know **it** was a trip,	Argument
2	**it** was an official **trip**,	Argument; noun
3	and I went to a couple of cities and one of **them** was Shanghai.	Argument
4	In **Shanghai** I I, I went to, I can't remember the name of **there**	Noun; argument
5	but it was **Shanghai** Tower or something.	Noun
6	**It** was a **tower** with, you know, with the floor is in glass	Argument;
7	so you can step in **there**	Argument
8	and actually feel the emptiness that you can **feel that**, I don't know if I can use **that** word,	Content; argument; argument
9	but anyway, to me **it** was a very tough experience	Argument
10	and to be honest I wasn't able to **step** in **there** because I was kind of, I don't know what the right word for **that** but claustrophobia or something,	Content; argument; argument
11	but anyway I was scared, my legs were shaking,	
12	so in order to take a picture, because you want to **take** a **picture** of **that** moment, you **want** to record **that moment**, I just, I just had to actually crawl in my back using my hands and my legs and, you know, to get **there**	Content; noun; argument content; argument; noun; argument
13	and I ask a friend to take me a picture.	
14	I think tall buildings are necessary for, you know, because there is reduced space in, in big cities	
15	so **they** need to find ways to, to give decent, you know, locations for working or just for living	Argument
16	and **that's** why they, **they** kind of stuck floor to **floor** in order to allow people to **live** in there because **living** in a big **city** it has a lot of um, um, I would say um, benefits.	Argument; argument; noun; content; content
17	So, so **that's** why, that's why they need to, to make room.	Argument
18	In fact, today **they** actually **build** a lot of ways in the, you know, kind of bridges	Argument; content
19	so **they** can avoid a lot of traffic jams	Argument
20	so, so basically **they're** growing to, to go up.	Argument

Table 19: Performance 69, Task 4, reference (Level 5; Aptis 6; CEFR C1-2)

Unlike the preceding transcripts, P69 in Table 19 effectively draws on argument overlap in the form of pronouns to develop cohesion. Like P81, false starts are limited and relatively unobtrusive, and the use of pronouns in argument overlap is broad and accurate.

In summary, the quantitative analysis of reference found no significant difference across levels and that argument overlap was commonly used across levels. The qualitative analysis of Task 4 performance suggests that higher-rated performances involved the use of a broader range of pronouns, and that these were used more effectively. Across levels, it was also found that repetition of nouns, where the use of pronouns might have been more relevant, may be implicated in a markedly less effective use of reference.

5.1.3 Lexical cohesion

Lexical cohesion was examined in terms of the two types of hypernymy (for nouns and verbs) and a combination of the two. Descriptive statistics of lexical cohesion observed in candidate performance at each level are summarised in Table 20. As shown in the frequency of referential expressions between sentences (local), the use of lexical cohesion was not very different across the levels. As shown in Figure 4, among the three types of lexical cohesion observed in the study, hypernymy for nouns was the predominantly observed lexical cohesion device in the study.

Hypernymy type	**Level 1 (*N* = 16)**		**Level 2 (*N* = 16)**		**Level 3 (*N* = 15)**		**Level 4 (*N* = 14)**		**Level 5 (*N* = 23)**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
For nouns	6.78	.77	6.61	.76	6.62	.41	6.62	.36	6.48	.19
For verbs	1.53	.74	1.48	.28	1.45	.17	1.39	.13	1.50	.06
A combination of both nouns and verbs	1.63	.29	1.55	.39	1.38	.23	1.34	.16	1.45	.19

Table 20: Descriptive statistics of lexical cohesion (hypernymy) in each level

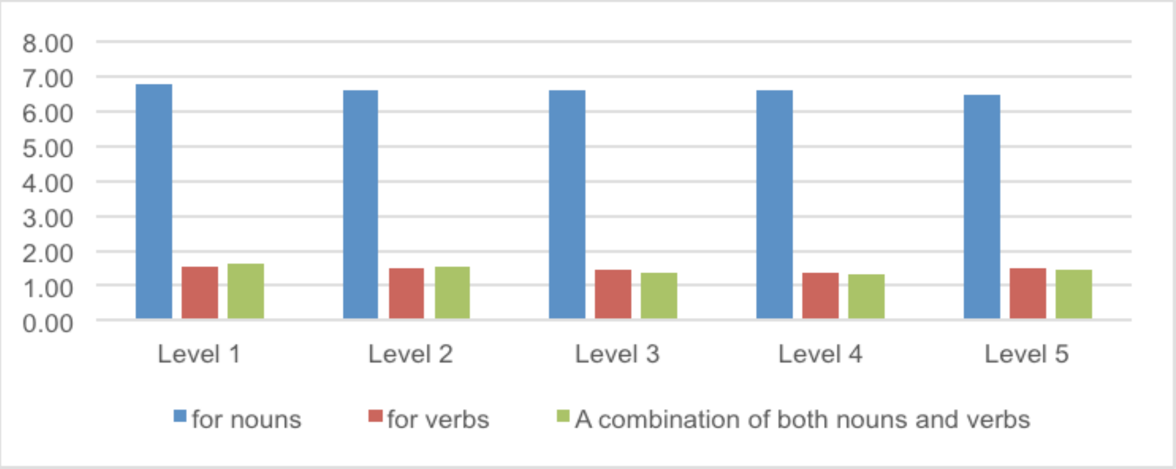


Figure 4: Distribution of lexical cohesion

Correlation analyses between level and frequency of lexical cohesion are summarised in Table 21. Little relationship between the assigned level and use of lexical cohesion was found. That means the assigned level does not discriminate the use of lexical cohesion. There was a significant correlation between nouns and a combination of both nouns and verbs, but no correlation between nouns and verbs was found.

	Nouns	Verbs	A combination of both nouns and verbs
Level	.13	.014	-.008
Nouns		.032	.364**
Verbs			.111

Notes: * Correlation is significant at the .05 level (2-tailed). ** Correlation is significant at the .01 level (2-tailed).

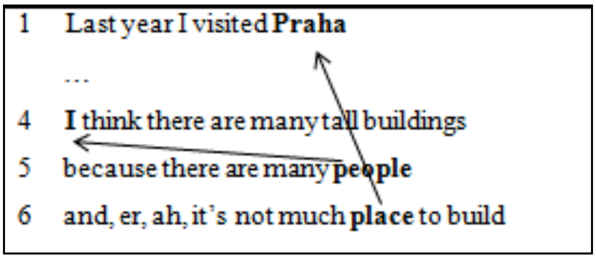
Table 21: Correlation between level and frequency of lexical cohesion (Spearman's rho)

As expected, due to the little difference in the descriptive statistics shown in Table 20 and Figure 4, ANOVA analyses revealed no statistical difference between groups as shown in Table 22.

Hypernymy type	*df*	Type III Sum of Squares	Mean Square	*F*	*p*	Partial *Eta* Squared
For nouns	.97	4	.24	.85	.50	.05
For verbs	.18	4	.05	.38	.83	.02
A combination of both nouns and verbs	.49	4	.12	1.54	.20	.08

Table 22: Results of ANOVA analyses (lexical cohesion)

We now present evidence of hypernymy and meronymy from the transcripts, followed by an extract of each transcript highlighting their use. Qualitative analysis revealed no clear evidence of hypernymy in the transcript of P63. Thus, it would appear that the quantitative results presented in Table 20 above may include false positive results, possibly related to the relationship between "place" and "Praha", "people", and "I" and "we."



1 Last year I visited **Praha**

...

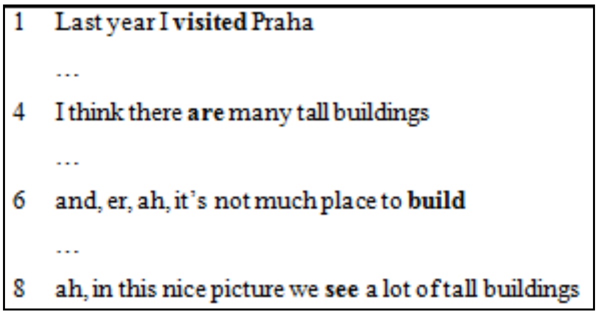
4 **I** think there are many tall buildings

5 because there are many **people**

6 and, er, ah, it's not much **place** to build

Figure 5: Transcript extract, Task 4, hypernymy – nouns (Performance 63)

It is difficult to see evidence of hypernymy in the verbs of the text (the copula, "visit", "build" and "see").



1 Last year I **visited** Praha

…

4 I think there **are** many tall buildings

…

6 and, er, ah, it's not much place to **build**

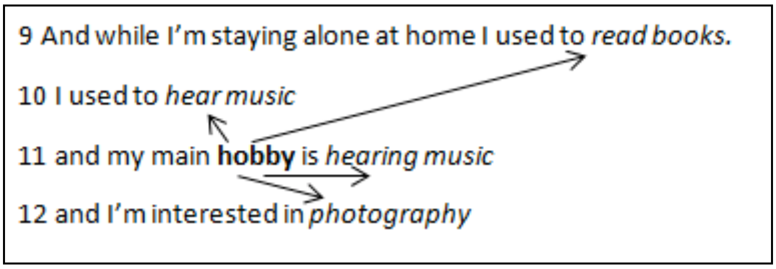
…

8 ah, in this nice picture we **see** a lot of tall buildings

Figure 6: Transcript extract, Task 4, hypernymy – verbs (Performance 63)

In contrast to P63, the transcript of P65 includes the following hyponyms, while the hypernyms may not be stated, but assumed, in the text:

- parents ⇨ mum, father
- mum, father, brother, grandmother;
- job, work
- hobby ⇨ photography, music, books
- photography ⇨ photos
- ants, birds
- watch, see
- hear, listen.



9 And while I'm staying alone at home I used to *read books.*

10 I used to *hear music*

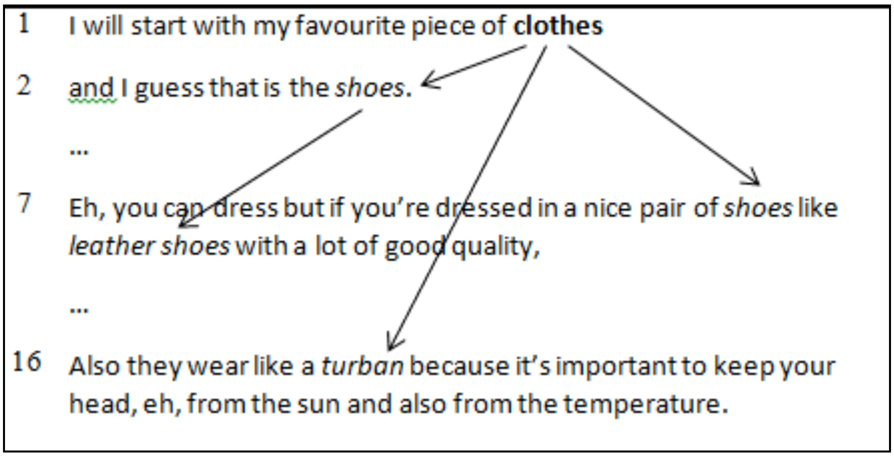
11 and my main **hobby** is *hearing music*

12 and I'm interested in *photography*

Figure 7: Transcript extract, Task 4, lexical cohesion (Performance 65)

Similarly, the transcript of P81 includes the following:

- clothes ⇨ dress, turban; shoes ⇨ leather shoes
- country ⇨ culture ⇨ Arabic
- weather ⇨ temperature
- mountain, beach
- wear, dress.



1 I will start with my favourite piece of **clothes**

2 and I guess that is the *shoes*.

...

7 Eh, you can dress but if you're dressed in a nice pair of *shoes* like *leather shoes* with a lot of good quality,

...

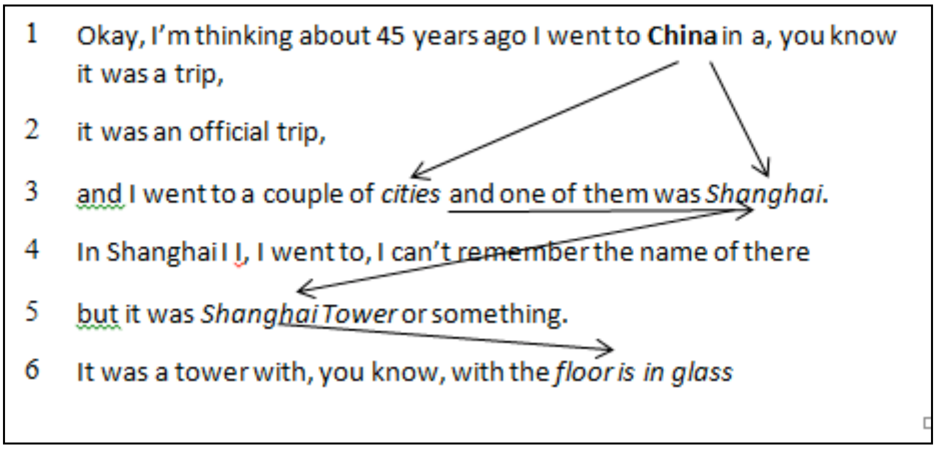
16 Also they wear like a *turban* because it's important to keep your head, eh, from the sun and also from the temperature.

Figure 8: Transcript extract, Task 4, lexical cohesion (Performance 81)

Finally, P69 includes the following:

- trip ⇨ China ⇨ city ⇨ Shanghai
- buildings ⇨ Shanghai Tower (including co-meronyms ⇨ floor ⇨ glass)
- buildings ⇨ Shanghai tower, bridges
- emptiness, claustrophobia
- (co-meronyms) hands, legs, back
- location, space
- step, crawl.

The difference here is that there are several hierarchical links which create a tight cohesive structure in the text as shown the diagram below.



1 Okay, I'm thinking about 45 years ago I went to **China** in a, you know it was a trip,

2 it was an official trip,

3 and I went to a couple of *cities* and one of them was *Shanghai*.

4 In Shanghai I I, I went to, I can't remember the name of there

5 but it was *Shanghai Tower* or something.

6 It was a tower with, you know, with the *floor is in glass*

Figure 9: Transcript extract, Task 4, lexical cohesion (Performance 69)

In summary, as with the findings regarding reference, while no significant quantitative difference were found across levels, higher-rated performances in the examples above appeared to involve greater depth and range of relationships of hypernymy and meronymy.

5.2 Vocabulary use – Research question 1

A variety of measures were employed to investigate vocabulary use in candidates' task performance as the descriptive statistics of each measure is presented in Table 23. Unlike the cohesion measures reported above, the frequency of a few measures (i.e., word count per 60 sec, verb and noun) calculated in each level was observed in linear fashion. In other words, while the number of words and verbs increases as the level goes up, the reverse trend was observed in the use of pronouns. The frequency of pronouns in Level 1 candidate performance was largest, and smallest in Level 5 candidate performance. Higher-level candidates used more adjectives and adverbs than lower-level candidates.

Type	**Level 1 (*N* = 16)**		**Level 2 (*N* = 16)**		**Level 3 (*N* = 15)**		**Level 4 (*N* = 14)**		**Level 5 (*N* = 23)**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Word count (per 60 sec)	5.08	25.61	115.32	38.70	201.69	64.78	261.65	76.06	277.43	31.52
Noun	244.61	6.22	216.77	71.57	174.89	42.57	168.11	21.96	195.86	32.04
Verb	109.18	49.24	116.47	28.36	122.01	28.70	126.65	23.41	127.81	13.92
Adjective	69.75	3.83	77.00	43.51	69.58	26.90	66.16	21.31	86.66	26.26
Adverb	57.65	41.54	71.05	43.28	67.43	24.58	69.95	24.13	78.84	19.05
Pronoun	165.41	5.88	132.22	51.10	145.74	36.57	143.61	23.62	111.85	24.89
Type–token ratio (content words)	.75	.13	.66	.12	.60	.11	.59	.08	.65	.05
Type–token ratio (all words)	.66	.14	.51	.10	.44	.08	.41	.06	.45	.04
VOCD	0	0	2.49	19.29	44.69	11.38	49.17	1.74	6.39	5.27

Table 23: Descriptive statistics (means and SDs) of measures of vocabulary use by levels

The results of correlation analyses are summarised in Table 24. There is a strong significant relationship between level and word count, noun, type–token ratio for content words and for all words and VOCD. That means, as explained above in the results of descriptive statistics, that the higher the level, the more words and nouns were produced. Also more content words and a greater variety of words were produced by candidates, as shown in the significant correlation for type–token ratio (content words) and (all words). Furthermore, when the length of speech is adjusted, as shown in the results of VOCD measure, higher-scored candidates produced a significantly greater variety of words. A significant correlation was also found between word count and noun, type–token ratio for both content and all words, and VOCD. The three lexical variety measures (i.e., type–token ratio for content words and all words and VOCD) are also significantly correlated.

	Word count (per 60 sec)	Noun	Verb	Adjectives	Adverb	Pronoun	Type–token ratio (content words)	Type–token ratio (all words)	VOCD
Level	.815**	-.407**	.206	.027	.126	-.152	-.323**	-.522**	.785**
Word count (per 60 sec)		-.506**	.174	-.064	.222*	-.161	-.585**	-.802**	.737**
Noun			-.263*	.053	-.320**	-.220*	.466**	.578**	-.250*
Verb				-.246*	.199	.188	-.039	-.063	.229*
Adjective					.027	-.123	-.122	.065	-.072
Adverb						.121	-.332**	-.185	.165
Pronoun							.120	.075	-.215*
Type–token ratio (content words)								.857**	-.121
Type–token ratio (all words)									-.314**

Notes: * Correlation is significant at the .05 level (2-tailed); ** Correlation is significant at the .01 level (2-tailed).

Table 24: Correlation between level and vocabulary use (Spearman's rho)

As shown in Table 25, ANOVA analyses show significant differences were found in word count (per 60 sec), noun, pronoun, type–token ratio for both content words and all words, and VOCD with the large effect size. The results of post hoc analysis (Bonferroni correction) are summarised in Table 26. Significant differences were largely found in most measures between Level 1 and other levels.

Type	*df*	Type III Sum of Squares	Mean Square	*F*	*p*	Partial *Eta* Squared
Word count (per 60 sec)	352242.07	4	8806.52	38.97	.00	.69
Noun	27305.21	4	6826.30	2.79	.03	.14
Verb	463.94	4	1157.74	1.35	.26	.07
Adjective	3156.50	4	789.13	.87	.49	.05
Adverb	961.26	4	24.31	.21	.93	.01
Pronoun	17372.29	4	4343.07	3.16	.02	.16
Type–token ratio (content words)	.14	4	.03	3.93	.01	.19
Type–token ratio (all words)	.33	4	.08	12.47	.00	.42
VOCD	20634.48	4	5158.62	33.15	.00	.66

Notes: effect size .01 > small, .06 > medium, .13 > large

Table 25: Results of ANOVA analyses (vocabulary use)

Lexical richness measure	Differences between levels	*p*
Word count per 60 sec.	1 and 2, 1 and 3, 1 and 4, 1 and 5, 3 and 4,	.001
Nouns	1 and 3, 1 and 4	.001
	2 and 4	.033
Pronouns	1 and 5	.035
Type–token ratio for content words	1 and 3, 1 and 4	.001
Type–token ratio for content words	1 and 2, 1 and 3, 1 and 4, 1 and 5	.001
	2 and 3	.003
	2 and 4	.004
VOCD	1 and 2, 1 and 3, 1 and 4, 1 and 5, 2 and 3, 2 and 4, 2 and 5	.001

Table 26: Results of post-hoc analyses (lexical richness)

Proportions of each word type are summarised in Table 27 and graphically presented in Figure 10. While the five word types in lower-level candidate performance were more unequally distributed, the distribution of the five word types in higher-level candidate performances is more balanced. The figures of type–token ratios for both content words and all words are larger in lower-level candidate performances than higher-level candidates. However, the result is reversed in the VOCD measure where text length is taken into account in the analysis.

Type	Level 1 (*N* = 16)	Level 2 (*N* = 16)	Level 3 (*N* = 15)	Level 4 (*N* = 14)	Level 5 (*N* = 23)
Noun	37.83	35.33	3.17	29.26	32.59
Verb	16.88	18.98	21.05	22.05	21.26
Adjective	1.79	12.55	12.00	11.52	14.42
Adverb	8.92	11.58	11.63	12.18	13.12
Pronoun	25.58	21.55	25.14	25.00	18.61

Table 27: Distribution of measures of vocabulary use (%)

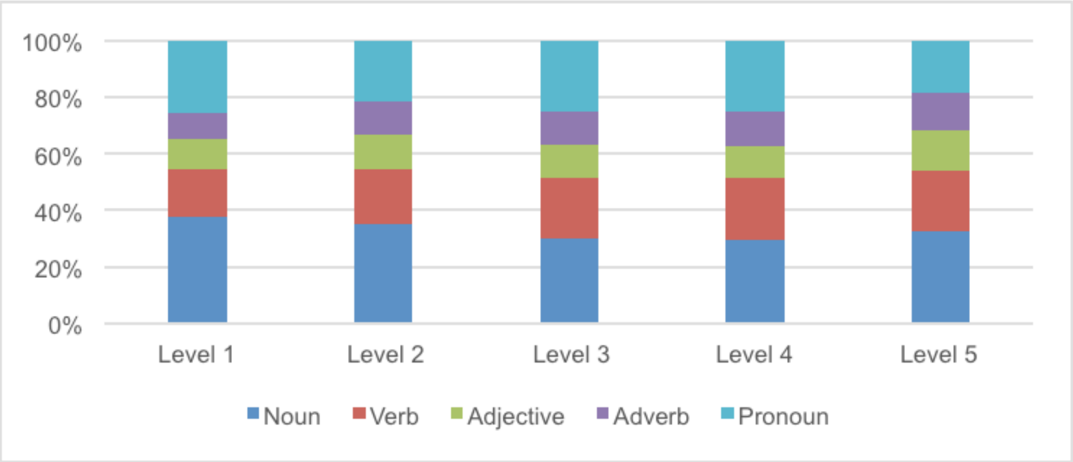


Figure 10: Distribution of measures of vocabulary use

5.3 Cohesion – Research question 2

The second research question addressed possible variations in candidates' vocabulary use and discoursal features according to the task types. For the quantitative analysis, performances across the four tasks were compared. As with the analysis of RQ1, qualitative analyses exemplifying the candidates' performances across tasks are provided at comparable levels, as represented in Table 3.

5.3.1 Conjunctions

Descriptive statistics of conjunction use in candidate performances in each task are summarised in Table 28. It is interesting to see considerable differences in the use of causal, logical, contrastive and temporal conjunctions between Task 1 and other task performances. Although the figures presented in Table 28 are all frequency data (per 1000 words), the frequency of these four conjunctions in Task 1 performance was substantially smaller than that observed in Task 2, 3, and 4 performances. The frequency of additive conjunctions observed in Task 1 and 4 performances was relatively smaller than Task 2 and 3 performances.

Conjunction type	**Task 1 (*N* = 20)**		**Task 2 (*N* = 21)**		**Task 3 (*N* = 21)**		**Task 4 (*N* = 22)**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Causal	28.13	24.11	32.14	17.54	31.60	13.94	4.30	18.27
Logical	41.15	27.52	52.54	21.07	63.87	21.73	61.87	22.49
Contrastive	4.23	7.27	13.35	11.45	15.45	12.02	15.40	12.23
Temporal	8.99	11.69	11.35	1.48	15.86	9.65	14.86	8.37
Expanded temporal	22.96	21.90	14.16	13.89	7.86	9.35	15.39	11.11
Additive	56.19	37.57	81.87	32.01	81.87	32.01	57.18	19.15
All (Combined)	28.13	24.11	32.14	17.54	31.60	13.94	4.30	18.27

Table 28: Descriptive statistics of conjunction use in each task (per 1000 words)

ANOVA analyses summarised in Table 29 revealed significant differences in performances in the frequency of logical, contrastive, temporal, additive conjunctions and all combined conjunctions with medium or large effect sizes. *Post hoc* analyses show differences in Tasks 1 and 3 (*p* = .001) and Tasks 1 and 4 (*p* = .003) for logical conjunctions, Tasks 1 and 2 (*p* = .002), Tasks 1 and 3 (*p* = .000), and Tasks 1 and 4 (*p* = .002) for contrastive conjunctions, and Tasks 1 and 2 (*p* = .003), and Tasks 2 and 4 (*p* = .004) for additive conjunctions. Furthermore, interaction effects between task and level show a significant difference, with large effect size in the frequency of contrastive conjunctions.

Type	*df*	Type III Sum of Squares	Mean Square	*F*	*p*	Partial Eta Squared
Task						
Causal	966.88	3	322.29	.90	.45	.04
Logical	7566.28	3	2522.09	4.88	.00	.18
Contrastive	1363.12	3	454.37	4.83	.00	.17
Temporal	942.34	3	314.11	2.88	.04	.11
Expanded temporal	1189.64	3	396.55	1.88	.14	.08
Additive	10374.13	3	3458.04	4.24	.01	.16
All (Combined)	1243.11	3	4143.37	4.54	.01	.17
Level * Task						
Causal	2887.62	7	412.52	1.15	.34	.11
Logical	6046.54	7	863.79	1.67	.13	.15
Contrastive	2631.39	7	375.91	4.00	.00	.29
Temporal	423.81	7	6.54	.56	.79	.05
Expanded temporal	1934.60	7	276.37	1.31	.26	.12
Additive	6409.67	7	915.67	1.12	.36	.10
All (Combined)	10878.83	7	1554.12	1.71	.12	.15

Notes: effect size .01 > small, .06 > medium, .13 > large

Table 29: Results of ANOVA analyses (conjunction use)

The proportion of each conjunction was also examined across the four tasks. As shown in Table 30 and Figure 11, the largest proportion of the conjunctions under investigation was additive in all tasks except Task 4, where the frequency of logical conjunctions was largest. While distribution patterns are similar in Task 2, 3 and 4 performances, Task 1 produced very few contrastive conjunctions, but a relatively large number of expanded temporal conjunctions were observed.

Type	Task 1 (*N* = 20)	Task 2 (*N* = 21)	Task 3 (*N* = 21)	Task 4 (*N* = 22)
Causal	17.40	15.65	14.60	19.66
Logical	25.46	25.58	29.50	3.18
Contrastive	2.62	6.50	7.13	7.51
Temporal	5.56	5.52	7.33	7.25
Expanded temporal	14.20	6.89	3.63	7.51
Additive	34.76	39.86	37.81	27.89

Table 30: Distribution of conjunction use by task (%)

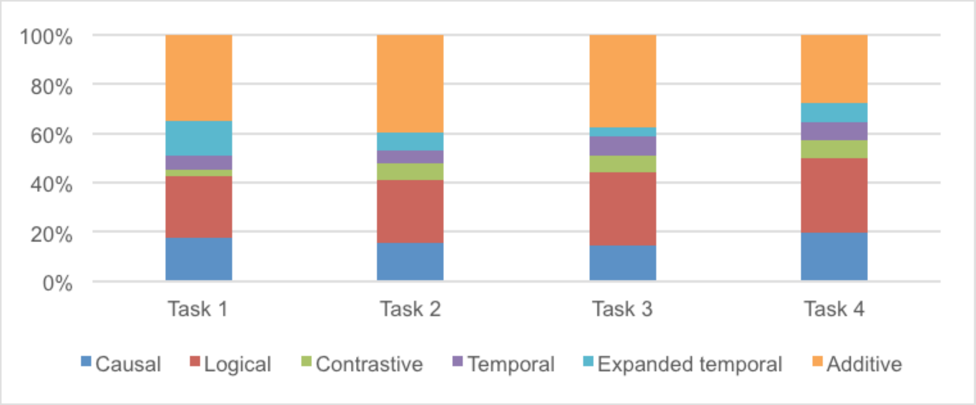


Figure 11: Distribution of conjunctions in each task performance

Possible explanations for the differences in overall use of conjunction reported in the quantitative analysis may be found in differences in task design. For example, Task 1 involves three short unrelated personal questions (though one of these relates to prior experience). In response to the short questions, candidates may not feel required to provide elaborate answers in a long sentence using conjunctions, as shown in the significant differences in the post-hoc analyses in logical, contrastive and additive conjunctions between Task 1 and other tasks. On the other hand, in Tasks 2 and 3 which involve separate questions including comparisons (explicitly in Task 3), time-related questions, and reasoning behind responses, candidates are more inclined to use contrastive conjunctions to make comparisons and temporal conjunctions to answer time-related questions, as shown in the significant differences in the post-hoc analyses of contrastive conjunctions. Also, logical conjunctions may be used to provide reasoning behind responses, as shown in the significant post-hoc test results between Task 1 and other tasks. In Task 4, where questions are integrated into one longer planned response, the result is a higher frequency of the all-combined category. The qualitative analysis below also reveals variation in individuals' responses.

	Candidate's speech	Conjunctions
1	This room is grey walls	
2	ah, there are a lot of tables	
3	**and** there are a lot of notebooks.	Additive
4	Eh, the floor is, is made in wooden.	
5	Last time **when** I went to the cinema I watched the film interesting film	Temporal
6	**and** I'm very …	Additive
7	I'm wearing blue jeans and black sweater with black boot	
8	**and** I'm wearing black glasses with with, with silver lens inserts.	Additive
9	I wearing …	

Note. Horizontal lines in the tables in this section indicate responses to separate prompts in the Aptis tasks.

Table 31: Performance 4, Task 1, conjunctions (Level 2; Aptis 4; CEFR A2)

As can be seen above, the candidate's responses in P4 responses are short, and independent, with the majority of the few conjunctions being additive (lines 3, 6 and 8), and one temporal conjunction (line 5), as might be expected from the task prompts (see Section 5.1.3 coherence analysis for details of task prompts).

	Candidate's speech	Conjunctions
1	It's a family, a happy family.	
2	On the left a, a child is, is the son of the family	
3	**and** she's a girl	Additive;
4	**and** in the middle of the picture are the father of the family.	Additive;
5	He is, he is happy **because** have a present in your (unint) your hand and …	Causal Additive
6	I would like to talk about this picture	
7	**and** normally, normally, in my family, mmm, (unint) a present **because** my father are poor	Additive
8	**and** never bought a present for me, only on Christmas and a very …	Additive
9	I think in, for difference peoples is an especial (unint) some days **because** she have a present of your family	Causal
10	**but** for me no present having **because** the present are only for rich	Adversative/ contrastive; causal
11	**but** is, I am very happy **when** I …	Adversative/ contrastive; temporal

Table 32: Performance 33, Task 2, conjunctions (Level 2; Aptis 1; CEFR A2)

P33 involves a mix of additive (lines 3, 4, 7, 8), adversative/contrastive (lines 10, 11), temporal (line 11) and adversative conjunctions, in line with the requisite of the task for candidates to talk about past experience and to provide reasons for responses. P46 to Task 3 might be expected to be similar, given the fact that both tasks have the same structure, as explained earlier.

	Candidate's speech	Conjunctions
1	(unint) a lot of people working with, with computers	
2	**and** this one is like ah, they are wearing mask and white clothes like doctors or (unint).	Additive
3	In the second one eh, they are speaking.	
4	There's one there alone	
5	**And** they are wearing matchings okay.	Additive
6	In the second one they're using telephone, keyboard, computer, mmm.	
7	Would like to working the second one, **because** I, I really love the computers.	Causal
8	I'm a student now, er, in, I'm a student in telecommunications, telecommunications, in, in um, high school	
9	**and** I think it's better than the first one **because** I saved, er, I saved the one where in wearing clothes different, I don't know.	Additive; Causal
10	I prefer the second one **because** I work with my, with other people very much, okay.	Causal
11	You can give help to other people	
12	**and** you can also receive this help, okay.	Additive
13	It's more enjoyable, um,	
14	**and** the … first one is more individual.	Additive

Table 33: Performance 44, Task 3, conjunctions (Level 2; Aptis 1; CEFR A2)

The candidate's response in P44, however, is notable in that comparisons are marked by means other than the use of conjunctions (e.g., repetitive use of "the second one" – lines 3, 6, 7, 10). The higher rated P46 (Level 3) included similar phrases, but also included contrastive conjunctions to make links between propositions.

Similarly, Task 4 responses at Levels 2 and 3 involved a range of conjunctions, though individual differences were represented in the data. For example, P65 (Level 3, reproduced in part below – see full transcript in section 5.1.1) included a range of conjunctions.

	Candidate's speech	Conjunctions
1	Actually I think, me, **while** I'm alone at home I used to read books **and** used to hear music	Temporal, Additive

2	**but** I think mostly in my childhood I used to live alone normally **because** my parents were not allowed to stay home alone **because** my parents, both my mum and father is having went … mm **while** they went to their job …	Adversative/ contrastive;
		Causal; Causal;
		Temporal

Notes: *P65 was rated slightly higher than the other three performances; This was chosen for exemplification here rather than P63, which was rated at a similar level to the other three performances, as it provides more data for comparison in terms of the measures chosen. Data for P63 (Rated "Below B1") is also available in the previous section. The ratings "Below B1" for Task 4, and "B2 or above" for Tasks 2 and 3 mean it is difficult to provide exact comparisons.

Table 34: Performance 65, Task 4, conjunctions (Level 3, Aptis 2; CEFR B1)*

In summary, the above findings related to the use of conjunctions across tasks appear to distinguish a lower use of conjunctions in Task 1 from the other tasks, with a similar distribution of conjunction types across the other three tasks. This is most likely because Task 1 was not designed to elicit a cohesive text, and indeed cohesion is not a criterion for this task. Comparisons across tasks revealed Tasks 2, 3 and 4 to have a similar range of conjunctions, as well as similar issues related to the accurate and effective use of those conjunctions.

5.3.2 Reference

Table 35 presents descriptive statistics of the frequency of referential cohesion observed in each task performance. Figures 12 and 13 present the descriptive statistics shown in Table 30 graphically. As for overlaps observed between adjacent sentences (local), the incidences of noun and argument overlap observed in Task 3 and 4 performances were much higher than those in Task 1 and 2 performances. However, stem and content overlaps were not different across the tasks.

For the overlaps observed in all sentences in a text, all four aspects of overlaps increased as task difficulty increased. In other words, the incidence of overlaps observed in Task 4 performances was much higher than that of Task 1. Across the four tasks, argument overlap was found to be the largest among the four types of overlaps in all sentences. A relatively smaller number of content overlaps was observed in all four task performances, and the difference across the tasks was not as large as other types of overlap.

Reference (overlap) type	**Task 1 (*N* = 20)**		**Task 2 (*N* = 21)**		**Task 3 (*N* = 21)**		**Task 4 (*N* = 22)**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Adjacent sentences (local)								
Noun	.25	.17	.34	.15	.31	.23	.36	.22
Argument	.63	.17	.61	.25	.71	.24	.80	.14
Stem	.26	.17	.40	.15	.38	.24	.38	.24
Content	.20	.09	.20	.06	.24	.07	.23	.06
All sentences (global)								
Noun	.12	.09	.24	.14	.30	.19	.30	.21
Argument	.49	.18	.52	.19	.63	.21	.73	.15
Stem	.14	.10	.33	.16	.35	.19	.43	.26
Content	.16	.07	.15	.05	.20	.06	.19	.05

Table 35: Descriptive statistics (means and SDs) of reference in each task

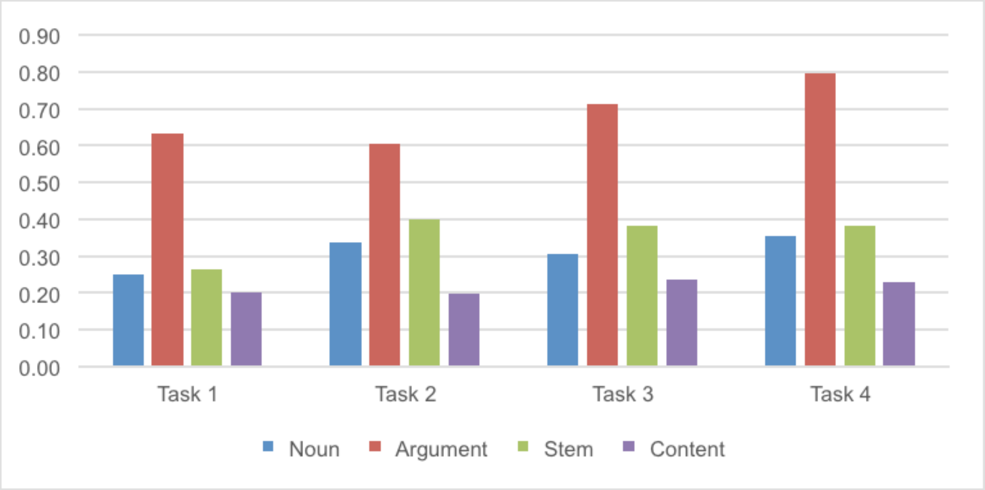


Figure 12: Referential cohesion between adjacent sentences (means)

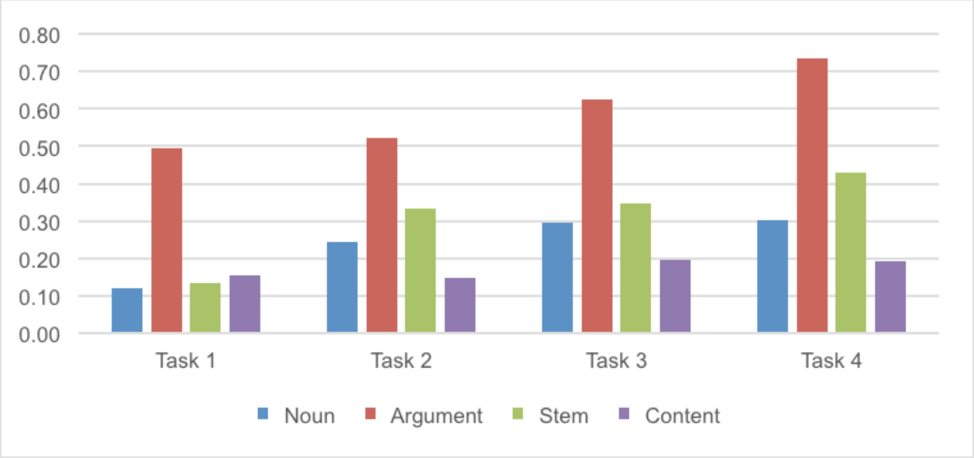


Figure 13: Referential cohesion between all sentences (means)

ANOVA analyses shown in Table 36 revealed statistical differences in all four overlaps in all sentences with large effect sizes. Post-hoc analyses found differences in Tasks 1 and 4 (*p* = .01) and Tasks 1 and 3 (*p* = .01) for noun overlaps, Tasks 1 and 4 (*p* = .001) and Tasks 2 and 4 (*p* = .001) for argument overlaps. Differences were observed in Tasks 1 and 2 (*p* = .001), Tasks 1 and 3 (*p* = .001) and Tasks 1 and 4 (*p* = .001) for stem overlaps, and Tasks 2 and 3 (*p* = .03), and Tasks 2 and 4 (*p* = .05) for content words overlaps. No interaction effect was observed between levels and tasks on all measures.

Type	*df*	Type III Sum of Squares	Mean Square	*F*	*p*	Partial *Eta* Squared
Task **Adjacent sentences (local)**						
Noun	.12	3.00	.04	1.07	.37	.04
Argument	.17	3.00	.06	1.43	.24	.06
Stem	.28	3.00	.09	1.96	.13	.08
Content	.02	3.00	.01	1.44	.24	.06
All sentences (global)						
Noun	.32	3.00	.11	3.73	.02	.14
Argument	.34	3.00	.11	3.72	.02	.14
Stem	.62	3.00	.21	5.70	.00	.20
Content	.04	3.00	.01	4.63	.01	.17
Level * Task						
Adjacent sentences (local)	.25	7.00	.04	.93	.49	.09
Noun	.34	7.00	.05	1.22	.30	.11
Argument	.17	7.00	.02	.52	.82	.05
Stem	.04	7.00	.01	1.03	.42	.10
Content	.25	7.00	.04	.93	.49	.09
All sentences (global)	.17	7.00	.02	.84	.56	.08
Noun	.32	7.00	.05	1.50	.18	.13
Argument	.19	7.00	.03	.74	.64	.07
Stem	.02	7.00	.00	1.07	.39	.10
Content	.17	7.00	.02	.84	.56	.08

Notes: effect size .01 > small, .06 > medium, .13 > large

Table 36: Results of ANOVA analyses (conjunction use)

Qualitative analyses of the selected samples provide further insights into the conjunctions use. For example, in P4 below limited overlap was observed given that Task 1 involves three separate short texts.

	Candidate's speech	Type of overlap
1	This room is grey walls	
2	ah, there are a lot of tables	
3	and there are a lot of notebooks.	
4	Eh, the floor is, is made in wooden.	
5	Last time when I went to the cinema I watched the film interesting **film**	Noun
6	and I'm very …	
7	I'm wearing blue jeans and black sweater with **black** boot	Content
8	and I'm **wearing black** glasses with with, with silver lens inserts.	Content; content
9	I **wearing** …	Content

Table 37: Performance 4, Task 1, reference (Level 2; Aptis 4; CEFR A2)

Although Task 2 involves three separate, but related, questions, the fact that they all relate to the same picture appears to be the reason for the more extensive overlap in P33. In P33, referential cohesion is mainly achieved through repetition: for example, "family" (lines, 2, 4, 7, 9), "present" (line 10). Argument overlap indicates the candidate's difficulty with accuracy in the use of pronouns (e.g., incorrect use of "she" in line 3, and "your" in line 5).

	Candidate's speech	Type of overlap
1	It's a family, a happy **family**.	Noun
2	On the left a, a child is, is the son of the **family**	Noun
3	and **she's** a girl	Argument
4	and in the middle of the picture are the father of the **family**.	Noun
5	**He** is, **he** is **happy** because have a present in your (unint) **your** hand and …	Argument, argument; argument; content
6	I would like to talk about this **picture**	Noun
7	and normally, **normally**, in my **family**, mmm, (unint) a **present** because my **father** are poor	Content; noun; noun; noun
8	and never bought a **present** for me, only on Christmas and a very …	Noun
9	I think in, for difference peoples is an especial (unint) some days because she have a **present** of your **family**	Noun; noun
10	but for me no **present** having because the **present** are only for rich	Noun; noun;
11	but is, I am very **happy** when I …	Content

Table 38: Performance 33, Task 2, reference (Level 2; Aptis 1; CEFR A2)

P44, involving comparisons across pictures (e.g., "this one", line 2, "the second one", line 3) uses both argument and noun overlap to achieve referential cohesion, but is somewhat repetitive in the use of both.

	Candidate's speech	Type of overlap
1	(unint) a lot of people working with, with computers	
2	and this **one** is like ah, **they** are wearing mask and white clothes like doctors or (unint).	Argument; argument
3	In the second **one** eh, **they** are speaking.	Argument; argument
4	There's **one** there alone	Argument
5	And **they** are wearing matchings okay.	Argument
6	In the second **one they're** using telephone, keyboard, computer, mmm.	Argument; argument
7	Would like to working the second **one**, because I, I really love the **computers**.	Argument; noun
8	I'm a student now, er, in, I'm a **student** in telecommunications, **telecommunications**, in, in um, high school	Noun noun
9	and I think it's better than the first **one** because I saved, er, I **saved** the **one** where in **wearing clothes** different, I don't know.	Noun; content; argument; content; noun
10	I prefer the second **one** because I work with my, with other **people** very much, okay.	Argument, noun; Noun
11	You can give help to other **people**	Noun
12	and you can also receive **this** help, okay.	Argument
13	**It's** more enjoyable, um,	Argument
14	and the … first **one** is more individual.	Argument

Table 39: Performance 44, Task 3, reference (Level 2; Aptis 1; CEFR A2)

As noted in the previous section, P65 (reproduced below) repeats the terms "alone", "home" and "stay," 11, nine, and six times, respectively. With regard to the use of anaphora, the candidate refers to his grandmother as "him"; and includes other errors in the use of pronouns, e.g., "that times", as well as sometimes having a lack of clarity with regards to the referent.

	Candidate's speech	Type of overlap
1	Actually I think, me, while I'm alone at home I used to read books and used to hear music	

2	but I think mostly in my childhood I used to live **alone** normally because my parents were not allowed to stay **home alone** because my **parents**, both my mum and father is having went … mm while they went to their job and my brother **went** to school and I be always free at **home**.	Content;
		Noun; Content;
		Noun; Content;
		Noun

3	I don't have much works to do.	
4	I be always I sit **alone** at **home**	Content; noun
5	and do something else for (unint).	
6	While watching, while seeing this picture I think **this** girl is concentrating or thinking something else and **she's** sad.	Argument; Argument
7	I think **her** facial expression seems that **she's sad** a bit.	Argument; argument, content

8	**She's, she's** something thinking about something for **her** past will be, **thinking** about **her** parents or family or **her** friends and things.	Argument;
		argument;
		Content;
		Argument;
		Argument

9	And while I'm staying **alone** at **home** I used to **read books**.	Content; noun; content; noun
10	I used to **hear music**	Content; noun
11	and my main hobby is **hearing music**	Stem; noun
12	and I'm interested in photography	
13	so I just take some **photos** of **interesting** pics of ants doing something or birds flew, flew in the, in the trees or staying on the **trees**.	Stem; stem Noun

14	The main thing, one of the main **things** I remember while I was **staying alone** at **home** is that once I played some tricks over my grandmother while **staying alone** because I was **staying alone that** day	Stem
		Stem; content;
		noun; content;
		content; content;
		content; argument

15	and when my **grandmother** came **this** evening at **home** I just **played** some **trick** on him by throwing something, just **throwing** my pillows and (unint) on **him** or her.	Noun; argument;
		noun; content;
		stem

16	Actually I think I just remember **that** times actually I was yes, always **alone** at **home** at **those** times.	Argument;
		content; noun;
		argument

17	I was happy at that but eh, **this** woman,	Argument
18	actually I like to be **alone** at **home** because **being alone** means you're always thinking about many of the opportunities which we will get …	Content; noun; stem; content

Table 40: Performance 65, Task 4, reference (Level 3; Aptis 2; CEFR B1)

In summary, with the exception of Task 1 where there is no expected relationship of propositions throughout the text, at the levels where comparisons can be made across all tasks, referential cohesion appears to have been achieved through the use of repetition and argument overlap. At least in the examples represented above, there appears to have been an overuse of repetition of nouns (noun overlap) rather than introducing a range of other referential features (e.g., accurate and effective argument overlap), and candidates at the levels of comparison, in support of RQ1 findings, were challenged in the accurate use of pronouns.

5.3.3 Lexical cohesion

The incidence of lexical cohesion was assessed in terms of hypernymy. As shown in the descriptive statistics in Table 41and Figure 14, the frequencies of hypernymy were similar across the four tasks for nouns, verbs and combinations of nouns and verbs. It should be noted that individual variation of hypernymy was not as great as other measures reported above.

Hypernymy type	Task 1 (*N* = 20)		Task 2 (*N* = 21)		Task 3 (*N* = 21)		Task 4 (*N* = 22)	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
For nouns	6.47	.73	6.80	.38	6.81	.53	6.43	.44
For verbs	1.43	.58	1.49	.21	1.48	.23	1.44	.15
A combination of both nouns and verbs	1.60	.31	1.51	.32	1.38	.29	1.36	.19

Table 41: Descriptive statistics of lexical cohesion (hypernymy) in each task

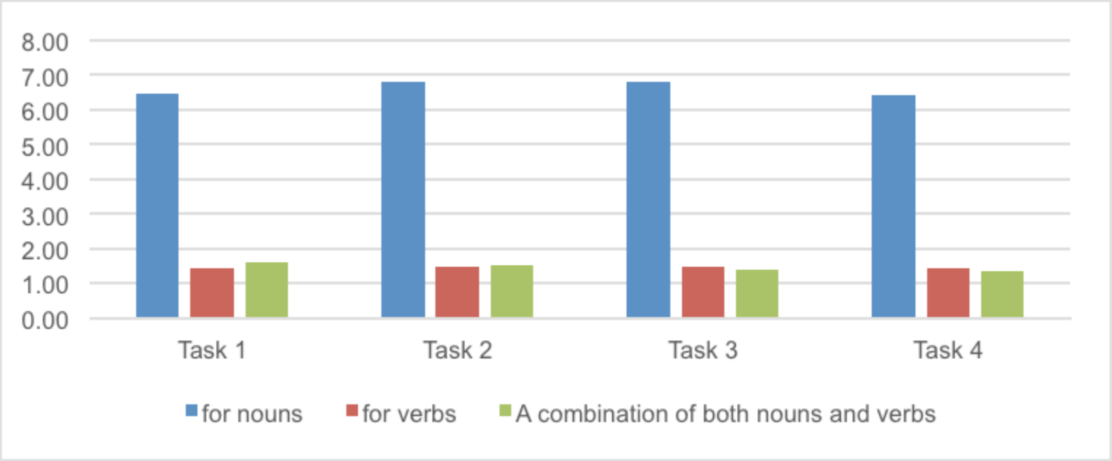


Figure 14: Distribution of lexical cohesion

The results of ANOVA analyses are summarised in Table 42. A significant difference with marginal effect size across tasks was found in hypernymy for nouns with medium effect size. This means the frequency of hypernymy for nouns observed in the four tasks varies significantly.

Hypernymy type	*df*	Type III Sum of Squares	Mean Square	*F*	*p*	Partial *Eta* Squared
Task						
For nouns	2.67	3	.89	3.14	.03	.12
For verbs	.05	3	.02	.15	.93	.01
A combination of both nouns and verbs	.46	3	.15	1.92	.14	.08
Level *Task						
For nouns	2.82	7	.40	1.42	.21	.13
For verbs	.31	7	.05	.37	.92	.04
A combination of both nouns and verbs	.28	7	.04	.50	.83	.05

Notes: effect size .01 > small, .06 > medium, .13 > large

Table 42: Results of ANOVA analyses (lexical cohesion)

As noted earlier, Halliday and Matthiessen (2013) also include the lexical relations of repetition, synonymy and meronymy under the broad heading of lexical cohesion, and the Coh-Metrix analysis for the current study measured hypernymy/hyponym relationships. As with the analysis related to RQ1, hypernyms may not be stated, but assumed, in the text. The qualitative analysis for the selected samples supports the quantitative analysis in that hypernym/hyponym relationships were evident across all tasks.

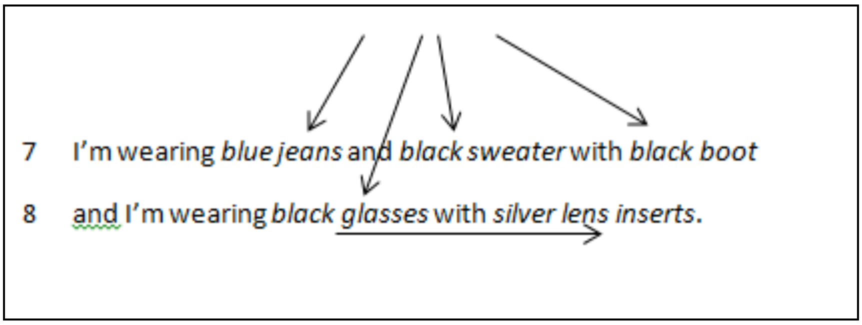
As with the analysis for RQ1, the following are lists of evidence of hypernymy and meronymy from the selected performances, followed by an extract of each transcript (Figures 15–18) highlighting their use.

The candidate's response to Task 1 in P4 included the following hypernym/hyponym relationships:

- floor ⇨ wooden
- cinema ⇨ film
- jeans, boots, glasses.

And the following examples of meronymy:

- room ⇨ walls, tables, floor
- glasses ⇨ lens inserts.



7 I'm wearing *blue jeans* and *black sweater* with *black boot*

8 and I'm wearing *black glasses* with *silver lens inserts.*

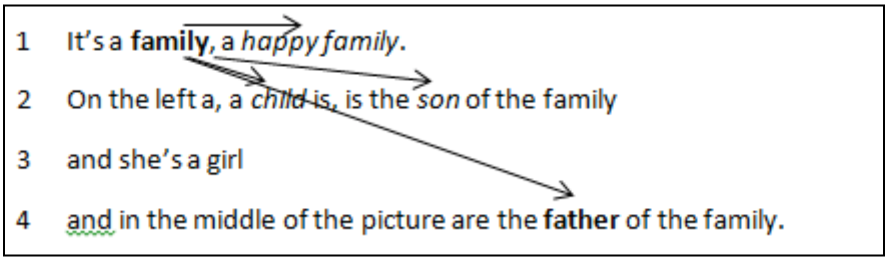
Figure 15: Transcript extract, lexical cohesion (Performance 4)

P33 (Task 2) included the following hypernym/hyponym relationships:

- rich, poor
- Christmas, "especial some days".

And the following examples of meronymy:

- family ⇨ father, child ⇨ son, girl
- father(person) ⇨ hand
- picture ⇨ middle of the picture.



1 It's a **family**, a *happy family*.

2 On the left a, a *child* is, is the *son* of the family

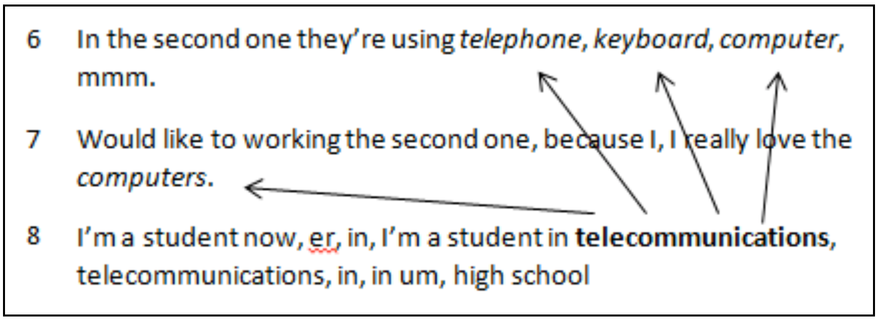
3 and she's a girl

4 and in the middle of the picture are the **father** of the family.

Figure 16: Transcript extract, Task 4, lexical cohesion (Performance 33)

P44 (Task 3) included the following hypernym/hyponym relationships:

- mask; white clothes; "matchings"
- doctors, student
- people, individual
- telecommunications ⇨ telephone, computer, keyboard.



6 In the second one they're using *telephone, keyboard, computer*, mmm.

7 Would like to working the second one, because I, I really love the *computers*.

8 I'm a student now, er, in, I'm a student in **telecommunications**, telecommunications, in, in um, high school

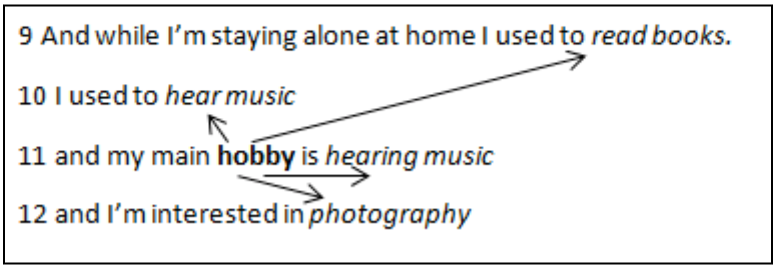
Figure 17: Transcript extract, Task 4, lexical cohesion (Performance 44)

P65 (Task 4; reproduced from previous section) included the following hypernym/hyponym relationships:

- job, work
- hobby ⇨ photography, music, books
- photography ⇨ photos
- ants, birds
- watch, see
- hear, listen.

And the following example of meronymy:

- mother, father, brother, grandmother.



9 And while I'm staying alone at home I used to *read books.*

10 I used to *hear music*

11 and my main **hobby** is *hearing music*

12 and I'm interested in *photography*

Figure 18: Transcript extract, Task 4, lexical cohesion (Performance 65)

As noted earlier, such hierarchical relationships can be seen as a bridge between the concepts of cohesion and coherence. In summary, as with the findings regarding reference, there was little difference found in the use of hypernymy and meronymy across all four tasks.

5.4 Vocabulary use – Research question 2

Descriptive statistics of measures of vocabulary are summarised in Table 43. As for word count (60 sec), Tasks 2, 3, and 4 produced a much larger number of word tokens than Task 1. However, this was found to be reversed in noun and pronoun tokens. Although the difference between Task 1 and Tasks 2–4 was not as large as word count, Tasks 2, 3, and 4 produced more adjectives than Task 1. There was no observable clear pattern in adverb tokens, type–token for content words and all words. The VOCD figure in Tasks 2, 3 and 4 was larger than that of Task 1.

ANOVA analyses reported in Table 44 show highly significant differences across tasks except verb and adverb tokens and VOCD figures, all with large effect sizes.

Post hoc analyses show differences in Tasks 1 and 2 (*p* = .001), Tasks 1 and 3 (*p* = .001), and Tasks 1 and 4 (*p* = .001) for word count; Tasks 1 and 3 (*p* = .001) and Tasks 1 and 4 (*p* = .004) for noun token; Tasks 2 and 3 (*p* = .003) for adjective token; Tasks 1 and 2 (*p* = .0014) and Tasks 1 and 3 (*p* = .0019) for pronoun tokens. For type–token ratio for content words, the difference was found between Tasks 1 and 2 (*p* = .003), Tasks 1 and 3 (*p* = .001), Tasks 1 and 4 (*p* = .002), Tasks 2 and 3 (*p* = .004) and Tasks 3 and 4 (*p* = .005). For type–token ratio for all words, the difference was found between Tasks 1 and 4 (*p* = .001) and Tasks 2 and 3 (*p* = .005). No interaction effect was observed between any measures.

Hypernymy type	**Task 1 (*N* = 20)**		**Task 2 (*N* = 21)**		**Task 3 (*N* = 21)**		**Task 4 (*N* = 22)**	
	M	***SD***	***M***	***SD***	***M***	***SD***	***M***	***SD***
Word count (per 60 sec)	102.35	64.41	186.86	95.41	204.62	98.56	204.73	77.39
Noun	234.02	59.79	202.05	69.44	171.29	41.73	179.99	32.34
Verb	13.00	32.14	115.57	34.77	113.10	27.55	122.27	26.30
Adjective	66.49	3.05	58.72	2.44	77.55	26.27	85.46	4.40
Adverb	6.72	3.95	73.47	35.69	65.19	37.00	73.78	25.52
Pronoun	167.28	43.53	13.54	41.80	131.87	34.73	138.70	38.93
Type–token ratio (content words)	.74	.11	.64	.11	.56	.10	.63	.07
Type–token ratio (all words)	.59	.13	.48	.11	.41	.10	.46	.05
VOCD	22.37	26.82	36.83	23.80	3.27	17.53	46.12	15.68

Table 43: Descriptive statistics (Means and SDs) of vocabulary use in each task

Type	***df***	**Type III Sum of Squares**	**Mean Square**	***F***	***p***	**Partial *Eta* Squared**
Task						
Word count (per 60 sec)	49993.74	3	16664.58	7.37	.00	.24
Noun	27981.85	3	9327.28	3.81	.01	.14
Verb	6274.76	3	2091.59	2.43	.07	.10
Adjective	9862.33	3	3287.44	3.61	.02	.14
Adverb	1843.83	3	614.61	.55	.65	.02
Pronoun	15742.83	3	5247.61	3.82	.01	.14
Type–token ratio (content words)	.20	3	.07	7.53	.00	.25
Type–token ratio (all words)	.14	3	.05	7.17	.00	.24
VOCD	924.95	3	308.32	1.98	.13	.08
Level * Task						
Word count (per 60 sec)	14229.78	7	2032.83	.90	.51	.08
Noun	14344.22	7	2049.17	.84	.56	.08
Verb	8371.43	7	1195.92	1.39	.22	.12
Adjective	7422.06	7	106.30	1.17	.33	.11
Adverb	5371.51	7	767.36	.68	.69	.07
Pronoun	16799.48	7	2399.93	1.75	.11	.15
Type–token ratio (content words)	.07	7	.01	1.08	.39	.10
Type–token ratio (all words)	.02	7	.00	.43	.88	.04
VOCD	1182.21	7	168.89	1.09	.38	.10

Notes: effect size .01 > small, .06 > medium, .13 > large

Table 44: Results of ANOVA analyses (vocabulary use)

Distributions of all five word types are slightly different across the four tasks as shown in Table 45 and Figure 19. Across the tasks, the three largest frequencies were noun followed by pronoun and verb. While Tasks 1, 3 and 4 produced more adjectives than adverbs, the order is reversed in Task 2. Tasks 3 and 4 and Tasks 1 and 2 show similar distributions respectively.

Type	**Task 1 (*N* = 20)**	**Task 2 (*N* = 21)**	**Task 3 (*N* = 21)**	**Task 4 (*N* = 22)**
Noun	35.54	34.81	3.64	29.99
Verb	19.74	19.91	2.23	2.37
Adjective	1.10	1.12	13.87	14.24
Adverb	9.22	12.66	11.66	12.29
Pronoun	25.40	22.49	23.59	23.11

Table 45: Distribution of measures of vocabulary use in each task (%)

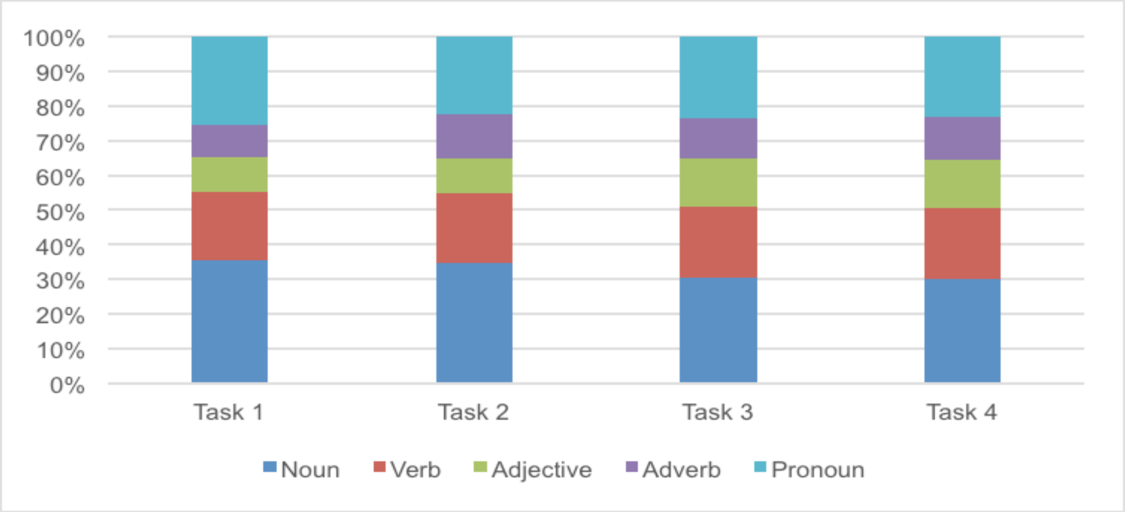


Figure 19: Distribution of lexical richness measures

5.5 Coherence

In order to explore the extent to which there were distinctive characteristics of coherence at different levels (RQ1), and in different tasks (RQ2), this analysis will focus on 16 performances representing a range of levels from each task: four from Task 1, four from Task 2, four from Task 3 and four from Task 4. As explained in the methodology section, coherence can only meaningfully be qualitatively analysed within the context of a response to a particular task; thus, the features of coherence which distinguish higher- and lower-scoring performances are presented in relation to the genres elicited through prompts in each task. Excerpts from these performances will be given to illustrate and support the points made regarding coherence. Relevant features of cohesion, including lexical chains and the use of conjunctions and pronouns which have been documented in the preceding cohesion analysis, will also be reported in explaining the results of the coherence analysis. Examples of selected performances are provided, with candidate speech segmented into C-units on the left and annotated aspects relevant to coherence on the right in each table. In order to substantively engage with the specific characteristics of each task, the features of high- and low-scoring tasks will first be considered within a discussion of findings for each task, followed by a summary of findings from different levels across all tasks.

5.5.1 Task 1

By examining four performances representing a range of levels in Task 1, it is possible to identify features of coherence, particularly topic development and adherence to genre, which were elicited. Sets of three questions in the different versions of Task 1 all required descriptions ("please describe this room") and a recount ("please tell me about the last time you visited friends"). Gerot and Wignell (1995) identify key moves in a description as the identification of a phenomenon to be described, followed by the description of parts, qualities and characteristics of this phenomenon, and key moves in a recount as an orientation, providing the setting and introducing participants, followed by the events, where the speaker describes what happened in sequence, and finally a re-orientation, which is optional and provides closure of events. In the four performances selected for analysis, more successful candidates were able to craft responses which clearly conformed to these genres.

In P11 in Table 46 below, which was scored as 2/5, the candidate is asked to "tell me about your first school" and is able to identify the phenomenon to be described "my school" (Segment 1), but is unable to further describe this phenomenon and develop the topic, as the following excerpt demonstrates:

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	My school is uh big school	Responds directly to Q1, identifies phenomenon
2	Ah my school, ah my school is ah, ah more more ah student	Expands on response with additional detail regarding size (enrolment) but appears hampered by lack of vocabulary
3	Ah my ah first day ah, (unint) remember	Attempts to further extend response but seems to lack vocabulary to do this. May have misunderstood "first school"
4	Ah my school also	Fragment that indicates intention to add to previous response but incomplete

Table 46: Performance 11, Task 1, coherence (Level 1; Aptis 2; CEFR A1)

In P4 (Table 47), which is mid–high scoring, the candidate is asked to "describe this room" and the following excerpt demonstrates the way in which the candidate is able to develop the topic and description through the use of lexical chains from Segments 1–4: "room" – "grey walls" – "tables" – "notebooks" – "floor – "wooden."

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	This room is grey walls	Directly responds to Q1, identifies the phenomenon, and begins description
2	Ah, there are a lot of tables	Provides additional details, builds description
3	And there are a lot of notebooks	Provides additional description, flagged explicitly by "and"
4	Eh, the floor is is made in wooden	Attempts to provide additional description, but seems hampered by limited syntax

Table 47: Performance 4, Task 1, coherence (Level 2; Aptis 4; CEFR A2)

P9, exemplifying a high scoring performance (5/5), includes a detailed description in response to "tell me about your first school". The candidate is able to extend the topic through sophisticated lexical chains in Segments 1–5 linking school and learning: "school" – "bi-lingual school" – "learned English" – "study translation" – "enjoy studying English", and clear signposting, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	My first school was located in my neighbourhood in Madrid	Responds directly to Q1, identifies phenomenon "school" and begins description by adding detail on location
2	and it was a bi-lingual school	Develops response to Q1 by providing additional details on the school, links this new information with "and"
3	so I learned English very well	Develops previous point by causal connection "so"
4	That's why I decided to study translation, because I, I quite enjoy studying English in School.	Develops previous point by causal connection "that's why" and develops topic of learning English
5	It was a very big school with very big rooms um, and with a very good computer room	Returns to initial description of the phenomenon, links with pronoun "it", elaborates with more details, connects additional information with "and"

Table 48: Performance 9, Task 1 excerpt 1, coherence (Level 3; Aptis 5; CEFR B1)

Task 1 also requires a recount, the social function of which is "to retell events for the purpose of informing" (Gerot & Wignell, 1995, p. 194). When asked the question "Please tell me about the last time you visited friends", the candidate in P13, who scored 2/5, is unable to produce a coherent response, instead attempting to describe a friend, relying heavily on the repetition of the word "friend" from the prompt, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
5	My **friend** name is, um (unint)	Attempts to respond to Q2, but does not address the recount aspect (last time you visited)
6	My mmm, **friend** so kind, kindly	Does not respond accurately to Q2 – describes friend, rather than the last visit, so is unable to develop the topic in a relevant way
7	My **friend** eh, so beautiful, eh, and, eh, beautiful and my **friend**, um, very, eh, innocent	Continues to describe friend – related to previous utterance but not the topic.
8	My, eh, **friend**, eh, is mmm, very talent	Continues to describe friend – related to previous utterance but not the topic

Table 49: Performance 13, Task 1, coherence (Level 1; Aptis 2; CEFR A1)

In response to the same prompt, P9, which scored 5/5, is a well-signposted recount addressing the topic by orienting the listener to the setting, introducing the participants, telling what happened in a sequence and finally reorienting the listener. The topic is developed through the lexical chain "visit" – "Italy" – "Verona" – "book a flight" – "see them". The effective use of pronouns to refer to participants: "my friends" – "them" – "them" – " them" and the focus on the temporal sequence: "a month ago" – "last year" – "it was a year that" are also characteristic of the recount genre. The following excerpt illustrates these points.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
6	I visited my friends in Italy a month ago	Responds directly to Q2, provides orientation to a recount, introducing participants and setting
7	I went to visit them because I went to last year to Verona	Expands on response providing reason, telling what happened and in what sequence
8	and I decided to, to book a flight and go to see them because it was a year that I didn't see them	Expands on previous utterance by providing details with "and", also gives justification for actions, linked by "because"
9	so I wanted to see them	Concludes by reiterating reason for visit, providing a re-orientation, ending the recount

Table 50: Performance 9, Task 1 excerpt 2, coherence (Level 3; Aptis 5; CEFR B1)

Thus the key differences between high- and low-scoring responses to Task 1, in terms of coherence, include the ability of the candidates to craft responses which conform to the required moves in the two genres elicited through the questions: description and recount. Moreover, lower-scoring candidates were unable to develop a topic effectively through extended lexical chains, often relying upon repetition of a limited range of vocabulary taken from the actual wording of the prompt. Pronoun use by lower-level candidates could sometimes lead to confusion, and the range of conjunctions used was limited. Higher-scoring candidates were able to develop topics through extended lexical chains, demonstrating a range and depth of relevant vocabulary, and use conjunctions and pronouns accurately and effectively to create a unified response to each of the three questions. Findings on cohesion from Sections 5.1 and 5.3 in the report provide a complementary analysis.

5.5.2 Task 2

Through the close examination of four performances representing a range of assigned marks in Task 2, it was possible to identify features of coherence, including the adherence of the response to the expected genre associated with the task and the ability to develop the topic, which were characteristic of higher- and lower-scoring candidates. The picture stimulus for P21, 33 and 24 is shown below.

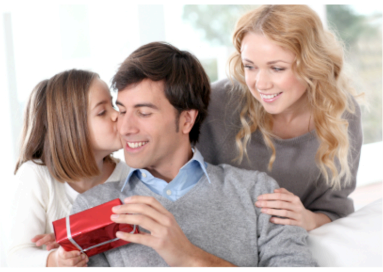


Figure 20: Task 2 Stimulus for Performance 21, 33 and 44

P21, which received 0/5, exhibits brief responses, with the candidate seeming to lack the vocabulary to develop topics beyond a basic level. The candidate is most successful in addressing the first question "Tell me about this picture", being able to develop the topic and provide a description through the lexical chain of "friendly family" – "mother", "daughter" – "father" – "his daughter" as the excerpt below demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	On this picture I see a friendly family	Responds directly to Q1, identifies phenomenon to be described and begins description
2	There are the mother and eh, eh, eh, daughter	Develops description by identifying family members
3	Eh, the father mmm, eh, eh, takes eh, present, eh, for eh, his daughter, eh	Continues to develop description by describing the action of the father

Table 51: Performance 21, Task 2 excerpt 1, coherence (Level 1; Aptis 0; CEFR A1)

However, the candidate is unable to develop the topic in response to the request for a recount ("Tell me about a time you gave or received a present"), as the following excerpts demonstrate.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
4	Ah, I eh, received a gift, ah, on my birthday	Responds directly to Q2, uses past tense and orients to a recount
5	It was ah in winter	Provides additional information, but of questionable relevance: focus on "time", interpreting this as season, not "present"

Table 52: Performance 21, Task 2 excerpt 2, coherence (Level 1; Aptis 0; CEFR A1)

P33, which received a slightly higher overall mark (1/5), demonstrates the ability to further develop the description, but encounters "topic trouble" which may be a result of the disjuncture between the image in the picture and the candidate's life experience. In response to the request for a recount, which assumes that the giving and receiving of presents is a normal part of peoples' lives, the candidate states that this is not the case for him, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
7	I would like to talk about this picture	Appears to continue to address Q1 – does not respond to Q2
8	and normally, normally, in my family, mmm, (unint) a present because my father are poor and never bought a present for me, only on Christmas and a very …	Responds to Q2, providing reason why this question is not relevant to candidate's life/lived experience

Table 53: Performance 33, Task 2 excerpt 1, coherence (Level 2; Aptis 1; CEFR A2)

This candidate experiences further "topic trouble" when asked to explain why gift-giving on special occasions is important, as the following excerpt demonstrates:

	Candidate speech, segmented in C-units	Aspects relevant to coherence
9	I think in, for difference peoples is an especial (unint) some days because she have a present of your family	Attempts to answer Q3 but seems to misunderstand question
10	but for me no present having because the present are only for rich	Attempts to addresses topic through personal response, comparing self to the "reality" that is presented in the picture
11	but is, I am very happy when I …	Attempts to develop topic by describing feelings, but response incomplete

Table 54: Performance 33, Task 2 excerpt 2, coherence (Level 2; Aptis 1; CEFR A2)

The candidate appears to misunderstand the question (Segment 9) and there is also the possibility that this visual prompt and topic have resulted in further "topic trouble" for this candidate. There appears to be a dissonance between the candidate's life experience and the assumptions underlying the prompt/topic, as the candidate comes from a poor family "because my father are poor and never bought a present for me" (Segment 8) and this is reiterated in Segment 10: "but for me no present having because presents are only for the rich". It would be interesting to have the candidate's framing of this task and topic, and to understand the extent to which negative affect may have impacted on this candidate's performance, in terms of constructing a coherent response to the visual stimulus and the topic.

P24, which received a higher mark of 3/5, demonstrates the ability to more fully develop the description of the picture, through lexical chains: "family" – "father" – "mother" – "a little daughter" – "father" – "his daughter" – "little girl" – "her father" and extensive use of pronouns when building the description relating to the family: "her" – "his" – "they" – "they" – "her" – "she". The description is developed beyond identifying members of the family, extending to their emotions, actions, and surroundings, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	On the picture we can see a family.	Responds directly to Q1, identifies phenomenon to be described
2	There's father, mother and a little daughter	Develops description by naming family members
3	and the father, ah, is giving her, his daughter a present	Builds on previous utterance to describe action, linked with "and", linked topically through family members
4	and they look happy	Builds on previous response to describe emotion, linked with "and", uses "they" for family
5	Ah, also I can tell that they are at home and sitting on the sofa.	Builds on previous response and the overall description by introducing the surroundings, linked with "also" & "and", uses "they" for family
6	Ah, and the little girl is, eh, kissing her father because, eh, she is happy because of the present she, she gives to her. Also …	Builds on previous response and connects to actions in Segment 3 and emotion in Segment 4, linked with "and" for the action and "because" to explain the emotion. The "daughter" from previous utterances is now "her" and "she"

Table 55: Performance 24, Task 2 excerpt 1, coherence (Level 3; Aptis 3; CEFR B1)

While this candidate was able to provide a fully developed description of the picture, she was not able to effectively respond to the question requiring a recount "tell me about a time when you gave or received a gift". Instead of producing a recount of a particular event, the candidate generalises about giving and receiving gifts, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
7	When I give a gift to somebody I tell, eh, some congratulations to a person or to whom I give this gift.	Attempts to respond to Q2, but in terms of generalisations, not the recount of a specific event, as the question requires
8	And also when somebody gives me a gift during receiving the gift I'm smiling and happy	Builds on previous point, now providing a generalisation in terms of receiving a gift, and the emotions this invokes, connects points with "and". Does not develop a recount
9	and, ah, also I like to receive some handmade gifts and …	Builds on previous point to the kind of gifts s/he likes to receive, linked by "and". Does not develop recount

Table 56: Performance 24, Task 2 excerpt 2, coherence (Level 3; Aptis 3; CEFR B1)

In contrast, P31, which received a score of 5/5, demonstrates the ability to effectively craft a recount. The picture stimulus for the version of Task 2 which this candidate responded to is shown below.

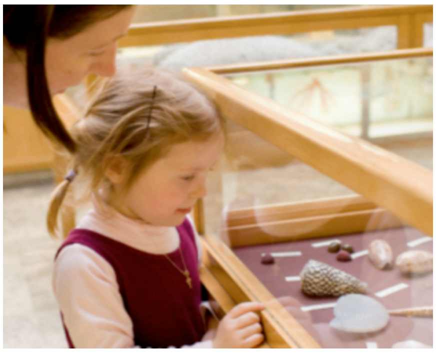


Figure 21: Task 2 Stimulus for Performance 31

The candidate was asked for a recount through the prompt "Tell me about a time when you visited a museum". The candidate is able to produce an extended response to this prompt, effectively crafting a recount, with the key moves of orientation and sequencing of events, followed by the optional move of re-orientation to provide a closure of events. The topic is developed through lexical and reference chains connecting to the museum: "museum" – "museum" – "it" – "it" – "ancient museum"; connecting to knowledge of ancient civilisations: "how a historical man lived" – "how man came about" – "all the facts" – "historical man"; and also connecting to family: "family members" – "family friends" – "we" – "we". The relevant annotated excerpt is below.

	Candidate speech, segmented in C-units	**Aspects relevant to coherence**
9	I visited a museum when I was in (unint)	Begins to respond directly to Q2, conforms to recount genre by beginning the orientation
10	and I went to the museum with my family members and some of my family friends	Builds on response through details of a past visit, conforms to recount genre by introducing participants, connects explicitly to previous C-unit with "and"
11	It was really beautiful as we didn't know how a historical man lived	Develops recount by telling what happened, providing more details about visit to the museum, which now becomes "it" and the family are "we"
12	and it was a real experience because we knew how actually man came and become (unint)	Develops recount by explaining why the experience was memorable, signalled through "because". The family are again referred to as "we" and visiting the museum by "it"
13	all the facts and all the things and I thought that the historical man in olden times	Develops previous point with more detail on the significance of the museum visit
14	and it was a really ancient museum which I went	Builds on previous C-unit by giving additional information about the museum, connects explicitly to previous point with "and"
15	It was in Bahrain.	Signals the end of the recount with a re-orientation. The museum is again referred to as "it"

Table 57: Performance 31, Task 2 excerpt 1, coherence (Level 4; Aptis 5; CEFR B2)

The third question in all versions of Task 2 required a response of either an explanation or exposition. The generic structure for an exposition is "the statement of a position, followed by a series of arguments to support that position. The point of each argument is introduced and then elaborated with supporting evidence" (Butt et al, 2012, p. 273). While lower-scoring candidates experienced difficulty with this aspect of Task 2, higher-scoring candidates were able to develop a topic in a manner consistent with the expectations of an exposition to some extent. In the performances selected for analysis, the highest-scoring performance was in response to what was clearly a question designed to elicit an expository spoken text. In order to illustrate the features of a high-scoring performance in response to the request for an exposition, P31 will continue to be analysed.

The third question, which elicited the exposition, was "Do you think people should pay to visit museums, or should they be free?" P31 demonstrates the ability to initially clearly state a position "it should be free" and this response is supported through the importance of people learning about history through museums. The lexical chain to develop the topic is specialised: "historical artefacts" – "museum," as the following except demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
16	I think it should be a free affair because if it is a free (unint) value it much	Statement of position, responds directly to Q3, followed by first argument signposted by "because"
17	and people won't know much about actually the historical artefact if they don't visit the museum	Builds on argument by providing additional point regarding importance of visiting museums, explicitly connected by "and"
18	and I think the museum is a great place to go to.	Builds on previous C-unit and argument by giving personal opinion on value of museums, signposted by "and I think"

Table 58: Performance 31, Task 2 excerpt 2, coherence (Level 4; Aptis 5; CEFR B2)

However, P31 then develops an argument for the other side of the case (that museums should not be free), with the candidate perhaps misunderstanding that the question requires an argument, rather than a discussion of two options, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
19	It should be paid because, a paid affair because the artefacts are come by dearly more, are costlier than things cost if some free person goes there.	Introduces an opposing view – that museums should not be free. However, this has not been clearly signposted
20	If a person who (unint) goes there and they might destroy the place if it is (unint)	Attempts to build on previous utterance with a conditional, but logic – that if museums are free the people who visit might destroy them – is questionable
21	and people who started the museum must also have a revenue from the museum.	Builds on previous point by providing additional reason – revenue – why museums should not be free. Explicitly connects this to previous point through "and"

Table 59: Performance 31, Task 2 excerpt 3, coherence (Level 4; Aptis 5; CEFR B2)

In response to question three, the initial argument that museums should be "a free affair" is contrasted with "should be paid", "a paid affair", but as the change to presenting this opposite view was not explicitly signposted, there is potential confusion around this. Interestingly, the rater does not seem to realise that the candidate has not fully developed an argument either *for* or *against* museums being free, but instead has discussed both options. The rater comments "all three responses are on topic".

Thus we can see that lower-scoring candidates in Task 2 could provide a basic description of a picture, but were unable to develop the description. They were also unable to respond effectively with the recount and explanation/exposition genres required in this task, whereas higher-scoring candidates were able to develop the description of a picture, craft a recount in connection with the topic introduced in the picture and either explain or argue a position in relation to this topic.

5.5.3 Task 3

By examining four performances – one high-scoring, two mid-scoring and one low-scoring – in Task 3, it is possible to identify features of coherence, particularly the extent to which candidates were able to respond to the requirements of the expected genres associated with the task and develop topics effectively. Candidates who scored poorly on Task 3 were unable to develop the description of the two pictures. P52 received a score of 0/5. The visual prompts for the version of the task undertaken by this candidate were of people playing golf and people playing basketball.

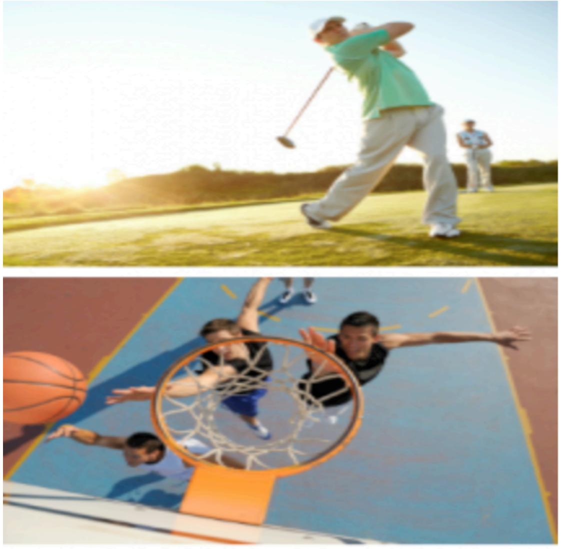


Figure 22: Task 3 stimulus for Performance 52

The candidate was unable to go beyond the identification of the phenomenon to be described (golf and basketball), and the response is characterised by repetition of vocabulary and ideas, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	So about um … So one picture is golfing, yeah playing golfing	Responds directly to Q1 by identification of the phenomenon to be described
2	And other picture is basketball … is playing basketball, play.	Identifies the second phenomenon to be described, connects to first utterance with "and"
3	One picture, one picture is play golfing … some picture	Repeats the gist of the first C-unit, unable to develop description
4	And another picture um … play basketball	Repeats the gist of the second C-unit, connecting with "and", unable to develop description

Table 60: Performance 52, Task 3, coherence (Level 1; Aptis 0; CEFR A1)

Whereas pictures of people playing basketball and golf could be assumed to require little interpretation (although the assumption that golf is a sport which is “familiar to the experience of the test-taker” is questionable), candidates of both lower- and higher-scoring performances experienced “topic trouble” in response to the following visual prompts.

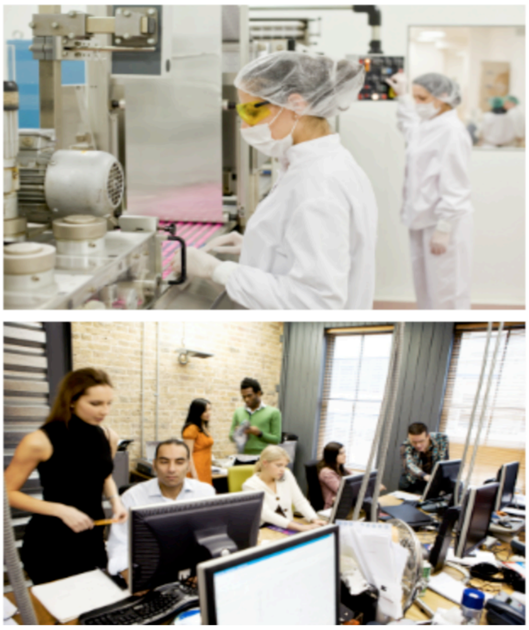


Figure 23: Task 3 stimulus for Performance 44 and 46

In P44, which received 2/5, the candidate was unsure about what was actually depicted in the first picture, interpreting it in Segment 2 as people “wearing masks and white clothes like doctors”, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	(unint) a lot of people working with, with computers	Responds directly to Q1, describing the second picture
2	and this one is like ah, they are wearing mask and white clothes like doctors or (unint).	Continues description in response to Q1, describing the first picture, signalling with “and this one”, connecting to the previous response with “they”. Appears to misinterpret the photo
3	In the second one eh, they are speaking.	Signals move to describing the second picture, provides more details
4	There’s one there alone	Moves back to describing the first picture, but does not explicitly signal this
5	and they are wearing matchings okay.	Builds on description of the first picture, connects to previous point with “and”, but appears hampered by limited vocabulary, using “matchings” instead of “uniforms”
6	In the second one they’re using telephone, keyboard, computer, mmm.	Signals move to describing second picture, with topic developed through listing items in the picture

Table 61: Performance 44, Task 3 excerpt 1, coherence (Level 2; Aptis 2; CEFR A2)

The candidate in P46, who responds to the same visual prompts and receives an overall score of 4/5, is also unsure about what is depicted in the first photo, stating in Segments 2–3 “I’m not sure what they are doing, but it seems like a really hard job”. The relevant excerpt is shown below.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	Okay in the first one it seems to be a laboratory or maybe a factory,	Responds to Q1 by attempting to describe phenomenon – the workplace in the first picture, indicates uncertainty with “maybe”
2	I’m not sure what they are doing.	Builds on previous C-unit by acknowledging that he is unsure of exactly what job the picture is depicting, going from place to people “they”
3	But it seems to be a really hard job.	Builds on previous point by giving details on the job “they” are doing, connecting with “but” to contrast what the candidate is sure about

Table 62: Performance 46, Task 3, coherence (Level 3; Aptis 4; CEFR B1)

Thus candidates were unsure about whether the first picture represented the production line in a factory or work taking place in a laboratory/medical setting. Having to interpret and respond quickly to visual prompts in the context of a timed speaking test is potentially stressful. The impact of viewing a picture which may have quite different interpretations makes this task even more stressful. This highlights the importance, in task design, of selecting visual stimulus that is easy and unambiguous to interpret, allowing the candidate to spend time and energy on describing a particular phenomenon, rather than having to second guess what that phenomenon might be. The subsequent questions in Task 3 build upon the topics from the visual prompts, thus potentially compounding the impact of confusion regarding exactly what is being depicted in the initial visual prompts.

The second feature required for an effective response to Task 3 is the ability to speculate, with questions following from the initial description of the visual prompts by asking “What would it be like to work in these two places?” (following from the pictures of the factory and office) and “What kind of people play these two sports?” (following from the pictures of golf and basketball). Lower-scoring candidates were unable to develop the topic and speculate, as the following excerpt from P44 demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
5	Yeah, so one mm, one sport is golf	Attempts to respond to Q2, but does not address the question directly
6	Ah the one couple … couple is plays golf	Attempts to respond to the question about the “kind of people” who play golf, but instead erroneously describes the two people in the picture as a “couple”
7	The other people is mm they’re playing baseball	Attempts to respond to the question about the “kind of people” who play basketball, but is unable to go beyond a repetition of a description of the action in the picture, erroneously describing it a “baseball”. Links to the previous utterance by “the other people”
8	So maybe they (unint) um play baseball (unint)	Repeats previous statement about baseball, trying to communicate that people who play basketball are also likely to play baseball “maybe they”

Table 63: Performance 44, Task 3 excerpt 2, coherence (Level 2; Aptis 2; CEFR A2)

The wording of Q2 “What kind of people” may have caused topic trouble for this candidate, as he appears unable to speculate about the kind of person who plays golf, instead commenting that there is a “one couple … couple is plays golf” (Segment 6). It is interesting that Task 3 claims to be testing topics familiar to the experience of the test-taker, when it is possible that golf may not connect to the experience of this test-taker. Only the perspective of the candidate himself would enable us to understand whether this response is caused by a lack of vocabulary, a misunderstanding of the question, or golf being too unfamiliar to his life experience to enable him to speculate about the “kind of people” who might play it. The candidate is able to attempt to speculate about the “kind of people” who play basketball, suggesting in Segment 8 that “maybe they play baseball”. However, the response to Q2 is confusing, as the candidate incorrectly states at one point that the people in the second picture are playing baseball (Segment 7).

The third prompt in Task 3 requires the candidate to produce an exposition based on comparison, in response to a question such as “Which of these places is it better to watch films? Why?” The candidate in P56, who scored the highest possible mark for this task (5/5), responded to this question with regard to the following visual prompts.

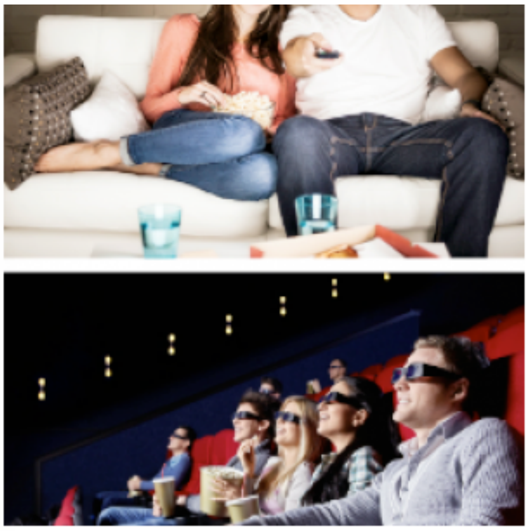


Figure 24: Task 3 Stimulus for Performance 56

This candidate was able to structure a well-organised and signposted response, clearly stating a position and supporting this position with arguments and examples, thus effectively responding to the expository genre, and explicitly comparing the two possibilities (watching a movie at home and watching a movie in the cinema), as the excerpt below demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
19	Well, I would usually prefer to watch movies at home because it's in the comfort of my own home.	Responds directly to Q3, signalling new topic through "well", stating position and providing a reason through explicit signalling "because"
20	I can move freely, I can go around, I can be in my PJs.	Builds on previous utterance by providing examples to support the comfort in being at home
21	I don't have to be in my jeans like, like going out on the street.	Builds on previous utterance by emphasising comfortable clothes that can be worn at home, compares this to having to dress to go out
22	The kitchen is right there, the bathroom is right there, it's all private.	Introduces new aspect of "comfort" argument through point that facilities, including kitchen and bathroom, are close at hand and private
23	It's an evening I can eat whatever I want.	Builds on "comfort" argument by describing food possibilities at home
24	The kitchen is, is next door	Reiterates the benefit of having the kitchen so close
25	so I can just fix myself any meal and go watch my movie and enjoy my own sofa	Builds on previous utterance by explaining benefit, explicitly signalled through "so"
26	Also I can do any phone calls I want, whereas in the movies when I can't speak to say anything out loud or use my cell phone whenever I feel like it.	Builds argument by additional point regarding phone usage, explicitly signalled through "also". Provides a direct comparison to other situation through "whereas"

Table 64: Performance 56, Task 3, coherence (Level 4; Aptis 5; CEFR B2)

The topic of the exposition (benefits of watching a movie at home) was developed through lexical chains relating to home: "home" – "comfort of my own home" – "kitchen" – "bathroom" "kitchen is next door" – "my own sofa", and comfortable clothes – "PJs," which were contrasted with "jeans" that have to be worn if going outside the house. Lexical chains also emphasise the freedom that is linked to the comfort of home: "move freely" – "be in my PJs" – "eat whatever I want" – fix myself any meal" "do any phone calls I want" – "say anything out loud" – "use my cell phone whenever I feel like it". This resulted in a well-developed and extended response to the question, engaging substantively with the topic and exhibiting key features of the required expository genre.

Thus, we can see that Task 3 required description, comparison, speculation and exposition, making it potentially quite a complex task, particularly as it involves two pictures and the continued engagement with one overarching topic, with no planning time given. Candidates who received lower scores experienced difficulty developing topics in relation to the pictures and were unable to compare, speculate or develop an argument effectively.

5.5.4 Task 4

Performances which received low scores overall in Task 4 were characterised by the brevity of response, a repetition of vocabulary from the prompt, and an inability to develop topics in a way that was consistent with the requirements of the task and genre. A detailed analysis of P63 in response to the task, which required a recount of an event ("Tell me about a time when you visited a very tall building"), a description of emotions associated with the event to build on the recount ("How did you feel about it?") and an explanation of causal links or an exposition, depending on how the candidate interpreted the question ("Why do you think so many cities have tall buildings?"*)* illustrates this.

P63, which received a score of 0/6, is characterised by relatively brief responses and repetition of words/phrases from the prompt ("visited a very tall building", "visit a very tall building", "tall building", "build a tall building", "a lot of tall buildings", "those tall buildings", "visit a city which has a tall building"), with limited pronoun use to link concepts as the text unfolds. The frequent repetition was also commented on in the cohesion analysis (see Table 13, for example). The candidate responds with basic information to the topics raised in Questions 1, 2 and 3. In response to Question 3, the candidate demonstrates the ability to speculate and attempts to explain a causal link, as is shown in the following excerpt.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
4	I think there are many tall buildings because there are many people	Responds to Q3, speculates, tries to explain causal link, using "because"
5	and, er, ah, it's not much place to build	Attempts to build on previous point to develop explanation, links with "and". Potential confusion of the message by error in pronoun "its"
6	so, ah, it is necessary to build a tall building	Continues to develop explanation in response to Q3, concludes and builds logically on previous utterance with "so"

Table 65: Performance 63, Task 4 excerpt 1, coherence (Level 2; Aptis 0; CEFR A2)

However, the candidate then proceeds to describe the picture – "in this nice picture we see a lot of tall buildings", which is not relevant to this task. Utterances which follow this build upon each other and develop the description of the picture, but are unrelated to the questions.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
7	Ah, in this nice picture we see a lot of tall buildings	Describes picture – unrelated to task
8	It is very interesting	Gives opinion – builds on previous utterance but unrelated to task
9	Those tall buildings, ah, are very nice	Gives opinion – builds on previous utterances, links ideas "those tall buildings", but response is unrelated to task
10	Ah, next year I also, ah, want to visit a city which has a tall building	States future intentions – related to previous utterance but unrelated to task

Table 66: Performance 63, Task 4 excerpt 2, coherence (Level 2; Aptis 0; CEFR A2)

It would be interesting to know, from the candidate's perspective, whether this "topic trouble" was a misunderstanding of the requirements of the task, as previous tasks (Tasks 2 and 3) had required a picture description, or occurred because the candidate was unable to develop the topics further. The rater's comments reflect this behaviour as: "The candidate answers all three questions in a basic way then goes off topic by describing the picture and is unable to speak for the full two minutes".

P65, which also received a low score (2/6), contains a longer response to all questions in a different version of Task 4, but also includes a description of the stimulus picture, which is unrelated to the task.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
6	While watching, while seeing this picture I think this girl is concentrating or thinking something else and she's sad.	Describes stimulus picture, but this response is not relevant to the task
7	I think her facial expression seems that she's sad a bit	Builds on previous utterance, but not relevant to task: "this girl" and "she" from Segment 6 are connected through pronouns "her" and "she"
8	She's, she's something thinking about something for her past will be, thinking about her parents or family or her friends and things	Builds on previous topic, continuing to refer to the girl "she," but not relevant to task

Table 67: Performance 65, Task 4, coherence (Level 3; Aptis 2; CEFR B1)

Higher-scoring performances, including P81 and P69, are characterised by longer responses, a greater range of vocabulary which results in extended lexical chains, effective use of pronouns to enhance cohesion and the ability to develop a topic in a way that is relevant to questions and genre, and is well sign-posted and logical.

The candidate in P81, who received a relatively high score (4/6), clearly signposted responses to all three questions. In addition to effectively using a range of conjunctions, including “also”, “and”, “but”, “because”, and “or”, the candidate is able to structure his responses to a question requiring explanation (“Why do people dress in such different ways?”) by a series of conditionals, as he speculates: “if you’re wearing a lot of clothes”, “if you are living in a mountain”, “if you are living in a beach.”

	Candidate speech, segmented in C-units	Aspects relevant to coherence
17	Sometimes if you’re wearing a lot of clothes it’s because you are in a cold country	Directly responds to Q3, builds on previous point (weather and clothes) and provides reason, linked by “because”
18	and you need to wear a special dress.	Builds on previous utterance, adding to explanation with “and”
19	Or imagine that if you are living in a mountain or if you are living in a beach	Speculates on other situations/contexts, responds to Q3, signposts with “or”
20	If you are living in a beach you don’t need a lot of things to wear so sometimes it depends on the temperature, the weather, the country, even the culture	Provides possible explanations, directly responds to Q3, uses causal link “so” and summarises total response

Table 68: Performance 81, Task 4 excerpt 1, coherence (Level 4; Aptis 4; CEFR B2)

The candidate also constructs lexical chains which enable the topic to be discussed at length and thus more fully developed in response to each question. For example, in response to Questions 1 (“Tell me about your favourite piece of clothing”) and 2 (“How do you feel when you wear it?”), the candidate goes from “favourite piece of clothes”, to “the shoes”, “it”, “love the shoes”, “a lot of shoes”, “wearing a pair of shoes”, “dressed in a nice pair of shoes”, “leather shoes”, “good quality”, “expensive”, “a good pair of shoes”. This excerpt, along with commentary, is given below.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
1	I will start with my favourite piece of clothes	Begins to address Q1, clearly signals with “I will start”
2	and I guess that is the shoes	Identifies favourite clothes, signals with “and”
3	I love it.	Builds on response to Q1, linking to Q2 with feelings
4	I love the shoes	Repeats point, replacing “it” with “the shoes”
5	I have a lot of shoes	Expands on previous C-unit, related to Q1
6	and I like it because you know when you’re wearing a pair of shoes that they looks really good you look like, I don’t know, like nice looking, like important person	Builds on previous C-unit by explaining the positive feeling provides causal link with “because”, responds to Q2
7	Eh, you can dress but if you’re dressed in a nice pair of shoes like leather shoes with a lot of good quality	Builds on previous C-unit through conditional “if you are dressed in a nice pair of shoes”, providing extra detail of shoes “leather”, “good quality”
8	sometimes they are expensive but eh, it’s quite nice to spend some money on a good pair of shoes.	Builds on previous C-unit by explaining feelings: “expensive”, “a good pair of shoes”, related to Q2

Table 69: Performance 81, Task 4 excerpt 2, coherence (Level 4; Aptis 4; CEFR B2)

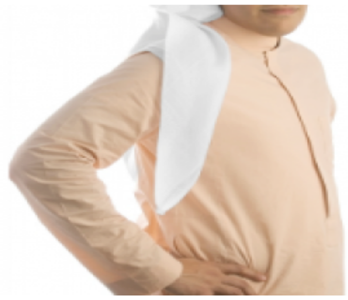


Figure 25: Task 4 Stimulus for Performance 81

The photo (above) of a man wearing what appears to be Arabic dress elicits stereotypes from the candidate in Segments 15 and 16 of the candidate's response "they're Arabic so they always wear this kind of clothes", "Also they wear like a turban", as the excerpt below demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
15	As I see in, in the picture here, it's a male that is wearing a leather (unint.) can be from they're Arabic so they always wear this kind of clothes	Describes picture – tangential relation to topic, and elicits a stereotype "Arabic … they always wear"
16	Also they wear like a turban because it's important to keep your head, eh, from the sun and also from the temperature	Adds to previous utterance with another stereotype: "also they wear". Builds on previous utterance, provides reason, directly responds to Q3

Table 70: Performance 81, Task 4 excerpt 3, coherence (Level 4; Aptis 4; CEFR B2)

It is interesting to note that the prompt for Q3 is "Why do people dress in such different ways?" implying that the dress of the gentleman in the picture is somehow "different" or an example of the dress of "others". It would be helpful to learn the candidate's perspective on the picture and the framing of this task.

P69, which scored 6/6, is characterised by lengthy responses that develop the topic of each answer by providing additional detail and the explicit and accurate use of causal conjunctions to structure explanations. Topics are also developed through lexical chains, with the "tall building" in the prompt leading to "tower" and "floor is in glass". As mentioned in the cohesion analysis earlier, hierarchical relations "China" – "a couple of cities" – "Shanghai" – "Shanghai Tower" – "tower" help to contextualise and structure the response in a recognisable pattern from general to specific, as the excerpt below demonstrates.

	Candidate speech, segmented in C-units	*Aspects relevant to coherence*
1	Okay, I'm thinking about 45 years ago I went to China in a, you know it was a trip, it was an official trip	Contextualises past event, with temporal connection "about 45 years ago", conforms to recount structure, directly related to Q1
2	and I went to a couple of cities and one of them was Shanghai	Builds on previous utterance, giving events in the recount, links information with "and"
3	In Shanghai I I, I went to, I can't remember the name of there	Attempts to develop previous utterance, describing past action, linked through "Shanghai"
4	but it was Shanghai Tower or something	Builds on previous utterance, naming tall building, contextualising further, responds to Q1
5	It was a tower with, you know, with the floor is in glass	Builds on previous utterance using pronoun "it" to link, provides further detail on the tower, responds to Q1

Table 71: Performance 69, Task 4 excerpt 1, coherence (Level 5; Aptis 6; CEFR C1–2)

It is interesting to note that this high-scoring candidate positions himself as an experienced professional, possibly having worked for the government when he states "about 45 years ago I went to China in a, you know it was a trip, it was an official trip".

In response to the prompt "How do you feel about it?", the candidate develops the topic of feelings evoked by the visit through low frequency lexis including "emptiness" and "claustrophobia", and continues to build on this by describing the "tough experience," and being so "scared" that "my legs were shaking", thus enabling the candidate to substantively develop the recount. There is a metacognitive aspect to this description of feelings, with the candidate commenting on his own choice of vocabulary "I don't know if I can use that word, but anyway … " in a way that does not detract from the coherence of the response. The full excerpt is below.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
7	and actually feel the emptiness that you can feel that, I don't know if I can use that word,	Builds on previous utterance and responds to Q2, explicitly describing feeling associated with event, as part of the recount
8	but anyway, to me it was a very tough experience	Links to point in previous utterance with "but", and continues to focus on feelings. Responds directly to Q2, developing recount
9	and to be honest I wasn't able to step in there because I was kind of, I don't know what the right word for that but claustrophobia or something	Builds on previous utterance about feelings, links to previous point with "and", develops description of experience in accordance with recount, explains feelings with "because", responds directly to Q2
10	but anyway I was scared, my legs were shaking	Builds on previous utterance about feelings and in giving more details of experience, responds directly to Q2
11	so in order to take a picture, because you want to take a picture of that moment, you want to record that moment, I just, I just had to actually crawl in my back using my hands and my legs and, you know, to get there	Adds to recount by describing particular activity – taking a photo – linked to feelings evoked, and thus develops response to Q2
12	and I ask a friend to take me a picture	Builds on previous utterance, but not directly related to questions

Table 72: Performance 69, Task 4 excerpt 2, coherence (Level 5; Aptis 6; CEFR C1–2)

The third question in Task 4 requires candidates to produce what could be interpreted as either an explanation or exposition, in response to a question such as "Why do you think so many cities have tall buildings?" Again drawing on P69, it is evident that the task has elicited the expository genre from a successful candidate, as the following excerpt demonstrates.

	Candidate speech, segmented in C-units	Aspects relevant to coherence
13	I think tall buildings are necessary for, you know, because there is reduced space in, in big cities	States position, provides first supporting argument, signalled with "because", directly responds to Q3
14	so they need to find ways to, to give decent, you know, locations for working or just for living	Builds on previous point, provides causal explanation signalled with "so", directly responds to Q3
15	and that's why they, they kind of stuck floor to floor in order to allow people to live in there because living in a big city it has a lot of um, um, I would say um, benefits	Elaborates on previous utterance, provides causal explanation "and that's why", directly responds to Q3
16	So, so that's why, that's why they need to, to make room	Builds on previous utterance, provides causal explanation, "so that's why" directly responds to Q3
17	In fact, today they actually build a lot of ways in the, you know, kind of bridges	Extends topic from tall building to bridges, tangentially related to Q3
18	so they can avoid a lot of traffic jams	Builds on previous utterance, provides causal link, but lacks relevance to Q3
19	so, so basically they're growing to, to go up	Builds on previous utterance, links back to Q3 and topic of necessity of tall buildings

Table 73: Performance 69, Task 4 excerpt 3, coherence (Level 5; Aptis 6; CEFR C1–2)

It is interesting to note that, after developing an effective exposition, the candidate then goes on to a tangential topic in Segments 17–18, before orienting the response back to the actual task in Segment 19. The rater comments: "the response addresses all three questions fully and is very logically structured and organised". Thus we can see that Task 4, which incorporates planning time, enables more proficient candidates to substantively engage with topics, developing well-connected responses that conform to the recount, explanation and expository genres.

5.5.5 Genres elicited through each task

The analysis of coherence in performances at different levels in Tasks 1, 2, 3 and 4 demonstrates that there are distinctive features associated with low and high scoring performances. As discussed in the preceding analysis, key points of difference were: the length of the response; the extent to which a candidate was able to correctly interpret the task and thus develop the topic in a relevant way which conformed to the genre required; the use of cohesive devices, including pronouns and lexical chains to develop topics; and the use of conjunctions to signpost the organisation of information between and within C-units. These aspects of cohesion have also been extensively documented in the relevant sections of the report.

To further explore RQ2, we now focus on the wording of questions and the genres elicited in each task. From the analysis of Tasks 1, 2, 3 and 4, it is clear that while there are overlapping genres in tasks, for example, descriptions, recounts and expository genres, each task contains a different mix of these genres and thus requires candidates to engage with and develop topics in rather different ways. A summary of the genres required for each task as identified in the 15 performances that were selected for qualitative analysis is below. Questions from each task are provided alongside the genre.

Task 1

- Description *Please tell me about your first school; Please describe this room.*
- Recount *Please tell me about the last time you visited friends; Please tell me about the last time you went to the cinema.*
- Description: *Please tell me about your favourite singer*; *What are you wearing today?*

Task 2

- Description of a picture: *Describe this picture.*
- Recount: *Tell me about a time when you gave or received a gift; Tell me about a time when you visited a museum.*
- Explanation or Exposition (depending on the wording of the task) *Why is it important to give people gifts on special occasions? Do you think people should pay to visit museums, or should they be free?*

Task 3

- Extended description of two pictures (involving comparison and speculation): *Tell me what you see in the two pictures. What kind of people play these two sports? What would it be like to work in these two places? What would it be like to see a film in these two places?*
- Exposition involving comparison: *Which of these two sports is more difficult to play? Why? Which of these two places would you prefer to work in? Why? Which of these places is it better to watch films? Why?*

Task 4

- Extended recount or description: *Tell me about a time when you visited a very tall building. How did you feel about it? Tell me about a time when you were on your own. How did you feel about it? Tell me about your favourite piece of clothing. How do you feel when you wear it?*
- Explanation, exposition or description (depending on the wording of the question) *Why do you think so many cities have tall buildings? What are some of the ways of passing the time on your own? Why do people dress in such different ways?*

It is potentially problematic that parallel versions of Tasks 2 and 4 appear to require different genres. If we compare two questions asked as the third question in parallel versions of Task 2, "Why is it important to give people gifts on special occasions?" pre-supposes that it is important to do this, and could be interpreted by the candidate as requiring an explanation (explaining what is) or an exposition (stating a position and constructing an argument to support this stand), whereas the wording of the parallel task "Do you think people should pay to visit museums, or should they be free?" clearly flags that this requires the candidate to take a position and support that position, thus conforming to the expository genre.

In the wording for the third question in Task 4, candidates were asked for what could be interpreted as either an explanation or an exposition "Why do you think so many cities have tall buildings?" "Why do people dress in such different ways?" or to provide an extended description "What are some ways of passing time on your own?". Information would be expected to be structured in different ways for each of these genres, so this is an important consideration for test developers.

5.6 Summary of results

5.6.1 Quantitative analyses of cohesion

Quantitative analyses of cohesion devices under study are summarised in Table 74.

Category	Sub-category	RQ1				RQ2		
		Correlation with levels	Difference across the five levels	Effect size	Post-hoc analysis	Difference across the four tasks	Effect size	Post-hoc analysis
Cohesion								
Conjunction	Causal							

	Logical					✓	.018	T1 ≠ T2;
								T1 ≠ T3;
								T1 ≠ T4

	Adversative					✓	.017	T1 ≠ T3; T1 ≠ T4
	Contrastive							
	Temporal							
	Expanded temporal							
	Additive	✓				✓	.016	T1 ≠ T2; T2≠ T4
	All (Combined)					✓	.017	T1 ≠ T2; T2≠ T4
Reference Adjacent sentences (local)	Noun							
	Argument							
	Stem							
	Content							
All sentences (global)	Noun					✓	.14	T1 ≠ T3; T1 ≠ T4

	Argument		✓	.013	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4
						✓	.14	T1 ≠ T4;
								T2 ≠ T4

	Stem					✓	.20	T1 ≠ T2;
								T1 ≠ T3;
								T1 ≠ T4

	Content		✓	.014	L2 ≠ L3; L2 ≠ L4	✓	.17	T2 ≠ T3; T2 ≠ T4
Lexical cohesion (hypernymy)	for nouns					✓	.12	
	for verbs							
	a combination of both nouns and verbs							

Vocabulary use	Word count (per 60 sec)	✓	✓	.69	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4;
					L1 ≠ L5;
					L3 ≠ L4
						✓	.24	T1 ≠ T2;
								T1 ≠ T3;
								T1 ≠ T4

	Noun	✓	✓	.14	L1 ≠ L3; L1 ≠ L4	✓	.14	T1 ≠ T3; T1 ≠ T4
	Verb	✓						
	Adjective					✓	.14	T1 ≠ T4; T2 ≠ T3
	Adverb					✓	.14	T1 ≠ T2; T1 ≠ T3

	Pronoun					✓	.25	T1 ≠ T2;
								T1 ≠ T3;
								T1 ≠ T4;
								T2 ≠ T3;
								T3 ≠ T4

	Type–token ratio (content words)	✓	✓	.19	L1 ≠ L3; L1 ≠ L4			

Category	Sub-category	RQ1				RQ2		
		Correlation with levels	Difference across the five levels	Effect size	Post-hoc analysis	Difference across the four tasks	Effect size	Post-hoc analysis

	Type–token ratio (all words)	✓	✓	.42	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4;
					L1 ≠ L5;
					L2 ≠ L3
						✓	.24	T1 ≠ T4;
								T2 ≠ T3

	VOCD	✓	✓	.46	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4;
					L1 ≠ L5;
					L2 ≠ L3;
					L2 ≠ L4;
					L2 ≠ L5			

Notes: L in L1, L2 etc., and T in T1, T2 in the post hoc analyses results column (Column six and nine) denotes Level and Task respectively. Effect size .01 > small, .06 > medium, .13 > large

Table 74: Summary of the results for quantitative analyses

5.6.2 Qualitative analysis of cohesion

As noted in Section 4.3.1, the qualitative analysis of measures of cohesion (conjunction, reference and lexical cohesion) aimed to add depth to the insights of the quantitative analysis, through close examination of these phenomena in the transcribed candidate performances across levels (RQ1) and tasks (RQ2), supported by rater comments. Across levels, it was found that learners drew on simple conjunctions in their attempts to achieve cohesive responses. Reflecting the fact that little difference was found in the frequency of the different types of conjunction, qualitative analysis revealed that more effective responses, as reflected in the Aptis scoring and rater comments, appear to be achievable without relying on the use of more complex conjunctions. Similarly, while no significant difference was found in the use of reference across levels, qualitative analysis revealed that repetition of nouns, where the use of pronouns may have been more relevant, may be a feature of lower-rated performances. With regard to lexical cohesion, while Coh-Metrix only measures hypernymy, and again no significant differences were found across levels, qualitative analysis suggests that there may be greater depth and range in relationships of both hypernymy and meronymy across levels (cf. Iwashita & Vasquez, 2015).

Across tasks, Task 1 performances were found to include fewer conjunctions than the other three tasks; an unsurprising finding, given that Task 1 involves three unrelated questions, and does not include cohesion as a criterion for raters. Based on the qualitative analysis, the use of different types of conjunction appeared to be linked to the questions being asked; for example, a question beginning "Tell us about a time when …" might be expected to elicit more temporal conjunctions than a question asking candidates to "Compare and contrast …". With regard to referential cohesion, the finding that noun and argument overlap were used more extensively in Tasks 3 and 4 appears to be reflected in the qualitative data presented in Section 5.2.1.2. In addition, it appears that the effective use of these forms of overlap (e.g., use of pronouns) was a challenge for candidates across tasks. Finally, as with findings for RQ1, little difference was found in the use of hypernymy and meronymy across tasks.

5.6.3 Vocabulary use

Vocabulary use was analysed quantitatively only and the results are summarised in Table 75.

Category	Sub-category	RQ1				RQ2		
		Correlation with levels	Difference across the five levels	Effect size	Post-hoc analysis	Difference across the four tasks	Effect size	Post-hoc analysis

Vocabulary use	Word count (per 60 sec)	✓	✓	.69	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4;
					L1 ≠ L5;
					L3 ≠ L4
						✓	.24	T1 ≠ T2;
								T1 ≠ T3;
								T1 ≠ T4

	Noun	✓	✓	.14	L1 ≠ L3; L1 ≠ L4	✓	.14	T1 ≠ T3; T1 ≠ T4
	Verb	✓						
	Adjective					✓	.14	T1 ≠ T4; T2 ≠ T3
	Adverb							
	Pronoun					✓	.14	T1 ≠ T2; T1 ≠ T3

	Type–token ratio (content words)	✓	✓	.19	L1 ≠ L3;
					L1 ≠ L4
						✓	.25	T1 ≠ T2;
								T1 ≠ T3;
								T1 ≠ T4;
								T2 ≠ T3;
								T3 ≠ T4

	Type–token ratio (all words)	✓	✓	.42	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4;
					L1 ≠ L5;
					L2 ≠ L3
						✓	.24	T1 ≠ T4;
								T2 ≠ T3

	VOCD	✓	✓	.46	L1 ≠ L2;
					L1 ≠ L3;
					L1 ≠ L4;
					L1 ≠ L5;
					L2 ≠ L3;
					L2 ≠ L4;
					L2 ≠ L5			

Notes: L in L1, L2 etc. and T in T1, T2 in the postdoc analyses results columns (columns 6 and 9) denote level and task respectively. Effect size .01 > small, .06 > medium, .13 > large

Table 75: Summary of the results for quantitative analyses (vocabulary use)

5.6.4 Coherence

Qualitative analysis of coherence revealed that higher-scoring candidates in all tasks were able to craft responses conforming to the structuring of information characteristic of particular genres. In Task 1, these genres were description and recount. Higher-scoring Task 1 performances were characterised by longer responses in which topics were effectively developed through extended lexical chains and crafted into a unified response through the effective use of conjunctions and pronouns. Lower-scoring performances were characterised by candidates relying on a noticeably limited range of vocabulary, often relying on repetition from the prompt. There appeared to be difficulty connecting ideas, extending descriptions and describing past events.

Task 2 required a description, recount and either an explanation or exposition, depending on the version of the task encountered. Lower-scoring performances were brief, with candidates able to provide only basic descriptions and experiencing difficulty in crafting an extended recount, explanation and/or exposition. Higher-scoring candidates were able to develop topics effectively through lexical and reference chains in extended descriptions and recounts, and structure a logical and unified explanation or exposition through clearly stating and supporting a position in relation to the topic.

Task 3 required a description of two pictures and an exposition, with the explicit need for comparison and speculation, differentiating this task from Task 2. In lower-scoring performances candidates were unable to go beyond the identification of the phenomenon to be described, with responses characterised by repetition of vocabulary and ideas. Topic trouble emerged as candidates struggled to interpret pictures and engage with topics which did not seem to reflect their life experiences. Higher-scoring candidates were able to effectively compare and speculate as they engaged substantively with topics and extended their responses, structuring them in ways which were appropriate to the required genres.

Task 4, which included planning time, initially required an extended recount or description, followed by an explanation, exposition or further description, depending on the wording of the questions in each version of this task. Performances which received low scores were characterised by the brevity of response, a repetition of vocabulary from the prompt and an inability to develop topics and structure information in a way that was relevant to the topic inherent in the question and the required genres. Although a picture was provided, in contrast to previous tasks, candidates were not asked to describe the picture. Lower-scoring candidates appear to have misunderstood this, and proceeded to describe the picture, which was irrelevant to the task requirements. Higher-scoring performances were characterised by longer, well-signposted and logical responses, a greater range of vocabulary resulting in extended lexical chains, effective use of pronouns to enhance cohesion and thus unify a response, and the ability to develop the topic and structure information in a way that was relevant to the required genres.

6. DISCUSSION

This study investigated features of discourse competence and vocabulary use across levels and tasks in the Aptis Speaking Test, by comparing features of discourse competence and vocabulary use across levels and tasks, features which appear in both the CEFR language level descriptors and the Aptis rating scales. In the study, discourse competence was examined in terms of features of cohesion and coherence. For cohesion we investigated the role of conjunction, reference, lexical cohesion and vocabulary use with both quantitative and qualitative methods, while coherence was qualitatively examined in terms of relevance of candidates' responses to the questions asked, and textual unity. In the qualitative analyses, we conducted in-depth analysis of candidate performance and also took raters' comments into consideration.

RQ1 investigated the aforementioned features of discourse and vocabulary across different levels observed in the Aptis Speaking Test performances. Use of cohesive devices under study in all performances was not very frequent regardless of the type of measures. The quantitative analyses revealed little statistical difference in the use of various cohesive devices across levels. As shown in the qualitative analysis, on the whole, the raters commented more positively on higher-rated performances, and the close examination of the selected samples show some differences across the levels. These findings are consistent with the studies in L2 assessment introduced earlier (Brown et al, 2005; Iwashita and Vasquez, 2015), and also with writing studies in both first and second languages (e.g., McNamara et al, 2010; Todd, et al, 2007), but not with other studies (e.g., Connor, 1984; Geva, 1992; Liu & Braine, 2005) which found some relationship between use of cohesive devices and learner proficiency. For example, in Liu and Braine's (2005) study, the most notable difference found in the use of cohesive devices was repetition, which we did not investigate because Coh-Metrix does not include it in the analysis. It is generally acknowledged that greater cohesion facilitates comprehension (Gernsbacher, 1990), but considering the short length of speech produced in all four task performances and the nature of oral language, it seems that candidates did not employ various cohesive devices to make their speech comprehensible.

As explained in the methodology section, the level assignment was done based on the score awarded to each performance, and therefore, there may be a potential discrepancy between candidates' use of conjunctions, and raters' interpretation of the use of "cohesive devices", suggesting the need for further investigation regarding the interpretation and role of "cohesive devices" in the rating scales.

As for vocabulary use, unlike cohesion devices, the results showed that, while word-length increased, the use of pronouns decreased as the level of performance rose. The measure VOCD, which takes word length into account, revealed a linear increase in type–token ratio as the level of performance rose. These results are consistent with Iwashita and Vasquez (2015) and Iwashita et al. (2008), which indicate vocabulary use is a good predictor to discriminate candidate performance according to the proficiency level. These findings show that the features that characterise good performance are not aligned with cohesive features, but vocabulary use discriminates the candidate performance well.

With regard to coherence across levels, there were clear differences between high- and low-scoring performances in each task. These differences could be seen in terms of the length of responses, the way in which topics were developed through lexical chains, the ways that people, places, objects and ideas were linked through pronouns and the ways in which points within and between C-units were linked through conjunctions. These findings are consistent with those of Seedhouse and Harris (2011) in their analysis of topic development in the long turn section of the IELTS speaking test. It was noticeable that high-scoring candidates were able to conform to genres elicited through particular questions, a finding which accords with that of Iwashita and Vasquez (2015). These genres included description, recount, explanation and exposition. The Aptis rating scales do not specifically refer to coherence: it seems that this is implicit in the overarching descriptors related to the extent to which a candidate addresses the topic. Whether coherence could be operationalised more explicitly is an issue for the test developers.

"Topic trouble" (Seedhouse & Harris, 2011) was evident in the response of some lower-scoring candidates in Tasks 2 and 3, who encountered difficulty with assumptions of their life experiences. For one candidate, there appeared to be a dissonance between his life experience and the assumption that gift-giving is a normal part of life and that it is important to give presents on significant occasions, as he came from a "poor family" which could not afford to buy gifts. It would be interesting to have the candidate's framing of this task and topic, and to understand the extent to which negative affect may have impacted on this candidate's performance, in terms of constructing a coherent response to the visual stimulus and the topic. Another candidate experienced difficulties in Task 3 when attempting to describe a picture depicting golfers, and could not engage with the follow-up question which required him to speculate about "the kind of people who play this sport [golf]". The Aptis Training manual states that Task 3 requires candidates to "describe, compare and speculate on topics familiar to the experiences of the test-taker, so the cognitive load is not very high". The different life experiences and topics familiar to candidates from very diverse backgrounds is difficult to predict; however, the topic trouble encountered by these candidates points to the need for careful consideration of the assumptions that may be made regarding effective and appropriate visuals and topics.

A different type of topic trouble was encountered by candidates in Task 3, who were unable to identify what kind of workplace was depicted in a photo, which they felt could have been either a factory or a laboratory. As the subsequent questions in the task were all built upon a clear understanding of the workplace depicted, this may have impacted negatively on candidate performance throughout the task. Another interesting aspect related to visuals was the stereotypes which appeared to be elicited from a version of Task 4 featuring a picture of a man in a particular type of clothing, in response to a question about why people dress "differently". It would be useful to explore the ways in which candidates actually interpret and frame visuals provided in the test. It was also important to note that having been required to describe the pictures provided in Tasks 2 and 3, lower-scoring candidates also described the picture in Task 4, although this was only provided as a stimulus, and a description was not required. It would be helpful to understand how these candidates framed Task 4, in terms of the relation to the picture provided and their response. Specifications and task design related to the selection of visuals is thus an area that would reward further consideration.

Seedhouse and Harris (2011, p. 25) posit that, in terms of identity construction, candidates who achieve very high scores in the IELTS speaking test "typically developed topics that constructed the identity of an intellectual and a (future) high achiever on the international stage". In two of the analysed performances, it was notable that a low-scoring candidate constructed an identity of lower socio-economic status, being brought up in a family who were too poor to engage in gift-giving, whereas a high-scoring candidate constructed an identity of a well-travelled, influential government official. It would be helpful to have the rater response to this aspect of identity formation, and the extent to which the identity that a candidate constructs through topic development may relate to their score.

RQ2 investigated the same features of discourse and vocabulary, including cohesion and coherence across the four speaking tasks in the Aptis Speaking Test. Unlike the findings of RQ1 reported above, some differences across the four types of tasks were reported. The frequency across tasks of conjunctions revealed that Task 1, which does not include conjunction as a feature in its rating scale, has a much lower frequency than the other three tasks. Task design was discussed as a possible reason for difference in conjunction use found across tasks. As noted earlier, Task 1 involves three short unrelated personal questions, Tasks 2 and 3 involve separate questions involving comparisons, time-related questions, and reasoning behind responses, and Task 4 questions are integrated into one longer planned response. These findings are consistent with Williams (1992) who investigated the use of discourse markers used by international teaching assistants (ITA) at a university in the US in order to identify the source of difficulty in comprehending the speech of non-native speakers.

Individual differences were also found in the qualitative analysis, including the use of forms of reference (e.g., referring to one of two photos in the prompts for Task 3) where conjunctions might otherwise be used to develop argumentation. Task 4 might be expected to provide more opportunity for conjunction use, given that it is an extended monologue. With regard to reference, there was a significant but small difference found in the incidence of noun and argument overlap only between the first two tasks (1 and 2) and the last two tasks (3 and 4). The reliance on repetition and argument overlap was also found in the qualitative analyses, which also found accuracy of referential terms to be implicated in the performances at the level of comparison. As with the findings across levels, the use of hypernymy was similar across levels, as was the use of meronymy in the qualitative analyses.

Finally, measures of lexical richness revealed that word length was greater for Tasks, 2, 3 and 4 than Task 1, as were type–token ratios. As with findings across levels, VOCD revealed a linear increase in type–token ratio from Task 1 to Task 4. As noted, effect sizes were minimal or small for all results. These findings are consistent with Brown et al. (2005).

With regard to coherence across tasks, a different mix of genres was required in each task. Task 1 required descriptions and a recount, Task 2 required a description of a picture, a recount and either an explanation or an exposition, Task 3 required a description of two pictures, speculation and an exposition involving comparison, Task 4 required an extended recount and either an explanation or exposition – or in one version of the task, an extended description. The extent to which a candidate was able to successfully develop a response depended on their ability to recognise and conform to the requirements of the expected genres. The wording of parallel versions of the prompt in Task 2 and Task 4 could lead to difficulty in candidates interpreting whether they were required to respond with an explanation or an exposition. An explanation is defined by Gerot and Wignell (1995) as having the social function of explaining "the processes involved in the formation of workings of natural or sociocultural phenomena", with the generic structure of a general statement to position the reader, followed by a sequenced explanation of why or how something occurs (p. 212). An exposition has the social function of "persuading the reader or listener that some should or should not be the case", and is structured by a thesis, followed by arguments, and ending with a recommendation (p. 209). In parallel versions of Task 2, candidates are asked: "Why is it important to give people gifts on special occasions?" or "Do you think people should pay to visit museums, or should they be free?" While the first of these questions could be interpreted as requiring either an explanation or an exposition, the second question clearly indicates that an exposition is required.

In parallel versions of Task 4, following the extended recount, a candidate could be asked “Why do you think so many cities have tall buildings?” or “Why do people dress in such different ways?” or “What are some of the ways of passing the time on your own?” Whereas the first two of these prompts could be interpreted as requiring either an explanation or an exposition, the third prompt seems to require an extended description. This points to the need for parallel versions of a task to be truly parallel, in terms of linguistic and cognitive demands on a candidate, and consistently and clearly designed to elicit the same generic spoken texts. Thus there are implications for specifications and task design. Aspects of task design which may impact on task complexity, including the selection and use of visuals, the provision of planning time, the genre of spoken responses elicited and the topics chosen, thus lend themselves to further examination.

7. CONCLUSION

This study analysed features of discourse competence and vocabulary use in Aptis Speaking Test performances. The results were co-referenced with key criteria described in the Common European Framework of Reference for Languages (CEFR) and compared across levels and tasks. The findings show little variation in the use of cohesive devices across levels and tasks, but some distinctive differences were observed in the vocabulary use and coherence feature (i.e., topic development). Qualitative analyses show candidate performance varied according to the task prompts, and their approaches to tasks. These findings provide further insights into the complex interplay of linguistic and task demands imposed on candidates as well as their approaches to tasks.

7.1 Limitations

The quantitative and qualitative analyses of selected cohesive and coherence devices and lexical richness provide a comprehensive picture of candidates’ performances according to the level and four types of tasks. However, as noted above, there are a number of limitations. First, while it may be preferable that performances be measured independently of *a priori* ratings and comments, the data provided by the Aptis test development team included such ratings and comments to ensure a range of performances were used – as with Knoch et al’s (2014) study, the lack of external information on candidates’ ability meant that this was the only data available. Second, the dataset provided included no information on the background of candidates, or, indeed which candidates performed which tasks. As such, it was not possible to track individual candidate’s performances across tasks. Third, the fact that the different sections of the Aptis Speaking Test are targeted at different performance levels made it difficult to compare performances at precisely the same level.

The use of the computational tool Coh-Metrix allowed for the analysis of several measures related to cohesion in a large number of transcripts in the study. As with other studies of this kind (e.g., McNamara, et al, 2010), the underlying constructs are necessarily reduced for the purpose of measurement. As noted in Section 4.2, this is represented in Coh-Metrix in the following ways: the existence of conjunctions is measured, but not accuracy or appropriateness; reference is measured in terms of the existence of overlap (repetition) of related forms, without making links between pronouns and their actual referents; and hypernymy is the only measure of lexical cohesion measured.

7.2 Recommendations

Based on our findings we are able to make the following recommendations for the design and assessment of future speaking tasks and rating scales. First, there are several interpretations of the constructs, cohesion and coherence, in the literature. Our quantitative analysis was informed by the necessarily reduced operationalisation of these terms in Coh-Metrix, while somewhat broader interpretations informed our qualitative analyses. Our study has provided a clearer understanding of how cohesion and coherence are represented in Aptis performances across levels and tasks. We recommend a research-based interpretation of these constructs be developed to inform the design and implementation (e.g., rater training) of future iterations of the Aptis Speaking Test. The fact that the data used in our study are used for rater training means that we have provided a rich mix of data analysed with a variety of quantitative and qualitative methods which can be drawn on for such training. The finding that relatively simple conjunctions are prominent across all levels of spoken performance has implications for the design of rating scales and assessor training, if this finding is replicable in future research. This may reduce the need for raters to look for the use of "diverse and complex cohesive devices" to distinguish performance at different levels – at least as far as conjunctions are concerned.

Our analysis of coherence in terms of candidate responses indicates that some visual prompts (pictures of a workplace that could be interpreted as a factory or a laboratory, pictures of people playing golf) and topics (the "kind of people" who play golf, the giving and receiving of presents) may not be in the life experience of candidates and can thus cause topic trouble. It is suggested that continued emphasis be placed on selecting visual prompts and topics that minimise the risk of candidates encountering topic trouble.

The coherence analysis also revealed potential confusion between explanation and expository genres, in terms of the way in which questions in parallel versions of the particular tasks were worded. Thus, for example, in Task 4 candidates could be required to produce an extended description, exposition or explanation, depending on the version of the task they encountered. It is suggested that the wording of questions be closely examined so that, irrespective of which version of a task a candidate encounters, the same mix of genres is explicitly elicited. Moreover, if at some point the Aptis Test developers would like to assess discourse characteristics more prominently to distinguish between test-takers' ability, significant revisions will be needed to the task types in order to be able to elicit and capture such differences.

Our research could be extended in several ways in future research projects. Major research questions coming from our study include:

- What is the relative importance of cohesion and coherence in determining the level of performance?
- How do raters arrive at a score for a candidate?
- How is candidate performance influenced by the media, e.g., "interaction" with and through technology; the use of photos which invite different interpretations in prompts?
- How is "topic trouble" implicated in candidates' performance?

Such investigations may provide a clearer understanding of factors informing the socio-cognitive model underlying the ongoing development of Aptis (e.g., O'Sullivan, 2015c, see Section 2.1 above). For example, research into candidates' framing of visual prompts, proposed topics, expected genres, and identity construction, may provide a more contextualised understanding of candidate performance. This, in turn, may inform the construction of tasks, as well as their choice in the localisation of task choice for particular contexts.

Methodologically, we would suggest future studies include full data sets for each candidate, with performance of each task, and preferably the performance by the same candidate on different versions of the same task.

Drawing on a relatively new quantitative methodological tool, as well as qualitative approaches used in recent research, this study has provided insights into cohesion and coherence on the Aptis Speaking Test, offering some clarification of the variable roles of features of discourse and vocabulary which are implicated in spoken test performance and their ratings.

REFERENCES

Alderson, J. C. (2007). The CEFR and the need for more research. *The Modern Language Journal, 91*(4), 659–663. doi: 1.1111/j.1540–4781.2007.00627_4.x

Alderson, J. C. (Ed.). (2002). *Common European Framework of Reference for Languages: Case studies*. Strasbourg: Council of Europe Publishing.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly, 3*(1), 3–3. doi: 1.1207/s15434311laq0301_2

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports, Vol 7.* IELTS Australia, Canberra and British Council, London.

British Council. (July 2013). *Aptis candidate guide: Online version.* Retrieved 9 January 2014 from: http://www.britishcouncil.org/sites/britishcouncil.uk2/files/aptis-candidate-guide-web.pdf

Brown, A., Iwashita, N., & McNamara, T. (2005). *An Examination of Rater Orientations and Test-taker Performance on English for Academic Purposes Speaking Tasks*. [Monograph Series MS-29]. Princeton, NJ: Educational Testing Service.

Butt, D., Fahey, R., Feez, S., & Spinks, S. (2012). *Using Functional Grammar: An explorer's guide*. South Yarra: Palgrave Macmillan.

Byram, M. & Parmenter, L. (Eds.). *The Common European Framework of Reference: The globalisation of language education policy*. Bristol: Multilingual Matters.

Celce-Murcia, M., Dörnyei, Z. & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied Linguistics, 6*, 5–35.

Cizek, G. J. & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage.

Cobb, T. (2013). *VP Classic (Version 4)* [Computer program]. University of Québec, Montréal.

Connor, U. (1984). A study of cohesion and coherence in English as a second language students' writing. *Research on Language and Social Interaction, 17*(3), 301–316.

Council of Europe. (2001). *The Common European Framework of Reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2003). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)*. DGIV/EDU/LANG (2003) 5, Strasbourg: Council of Europe.

Council of Europe. (2009). *Relating examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A manual*. Retrieved from: http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf

Council of Europe. (2011). *Manual for language test development and examining*. Retrieved from: http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics*, *11*(2), 183–99.

De Beaugrande, R. & Dressler, W. U. (1981). *Introduction to text linguistics*. New York: Longman.

De Jong, J. H. A. L. & Van Ginkel, L. W. (1992). Dimensions in oral foreign language proficiency. In L. T. Verhoeven & J. H. A. L. de Jong (Eds.), *The construct of language proficiency: Applications of psychological models to language assessment* (pp. 187–205). Amsterdam: John Benjamins.

De Jong, N. H., Steinel, M. P, Florijn, A., Schoonen, R. & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*, 5–34.

Dix, B. P. & de Mejía, A.-M. (2012). Policy perspectives from Colombia. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalisation of language education policy* (pp. 140–148). Bristol: Multilingual Matters.

Elder, C. & Iwashita, N. (2005). Planning for test performance: does it make a difference? In Ellis, R. (Ed.), *Planning and Task Performance in a Second Language* (pp. 219–237). Amsterdam: John Benjamins.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477–485. doi: 1.1093/elt/ccs037

Figueras, N., North, B., Takala, S., Verhelst, N., & Avermaet, P. V. (2005). Relating examinations to the Common European Framework: a manual. *Language Testing, 22*(3), 261–279. doi: 1.1191/0265532205lt308oa

Foster, P. Tonkyn, A., & Wigglesworth, G. (2000). A unit for all reasons: The analysis of spoken interaction. *Applied Linguistics, 21*, 354–74,

Frost, K., Elder, C., and Wigglesworth, G. (2011). Investigating the validity of an integrated listening– speaking task: A discourse-based analysis of test-takers' oral performances. *Language Testing, 29*(3), 345–369.

Fulcher, G. (2004): Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly 1*(4), 253–266.

Fung, L. & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410–439.

Galaczi, E. & ffrench, A. (2011). Context validity. In Taylor, L. (Ed.) *Examining speaking: Research and practice in assessing second language speaking. Studies in Language Testing*, Vol 3. Cambridge: Cambridge University Press.

Gerot, L., & Wignell, P. (1994). *Making sense of functional grammar*. QLD: Gerd Stabler.

Gernsbacher, M. A. (1990). *Language comprehension as structure building.* Hillsdale, NJ: Erlbaum.

Geva, E. (1992). The role of conjunctions in L2 text comprehension, *TESOL Quarterly, 26*(4), 731–747.

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal, 15*, 199–213.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers 36*, 193–202.

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

Halliday, M. A. K. & Matthiessen, C. (2013). *Halliday's introduction to functional grammar* (4th ed.). Routledge, New York.

Harley, B., Cummins, J., Swain, M., & Allen, P. (1990). The nature of language proficiency. In B. Harley, P. Allen, J. Cummins & M. Swain (Eds.), *The development of second language proficiency* (pp. 7–25). Cambridge: Cambridge University Press.

Higgs, T. & Clifford, R. (1982). The push towards communication. In Theodore V. Higgs (Ed.), *Curriculum, competence, and the foreign language teacher* (pp. 57–79). Lincolnwood, IL: National Textbook Company.

Huhta, A., Kushik, G. B., Alderson, J. C., Nieminen, L. & Ullakonoja, R. (2015). *Exploring the Linguistic Basis of the Common European Framework of Reference Levels in EFL Writing*. Paper presented at American Association for Applied Linguistics Congress, Fair Mount Hotel, Toronto, 20–24 March.

Hoey, M. (1991). *Patterns of lexis in text*. Oxford: Oxford University Press.

Hulstijn, J. H. (2007). The shaky ground beneath the CEFR: Quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal, 91*, 663–667.

Hulstijn, J. H. (2011). Language proficiency in native and non-native speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229–249. doi: 1.1080/15434303.2011.565844

Hulstijn, J. H. (2014). The Common European Framework of Reference for Languages: A challenge for applied linguistics. *International Journal of Applied Linguistics, 165*(1), 3–18.

Hulstijn, J. H. (2015). *Language Proficiency in Native and Non-native Speakers: Theory and Research*. Amsterdam: John Benjamins.

Hulstijn, J. H., Schoonen, R., de Jong, N. H., Steinel, M. P. & Florijn, A. (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*(2), 203–221.

Ito, P. K. (1980). Robustness of ANOVA and MANOVA test procedures. *Handbook of Statistics 1*, 1999–236

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, 29(1), 24–49.

Iwashita, N. & Vasquez, C. (2015). An examination of discourse competence at different proficiency levels in IELTS Speaking Task 2. *IELTS Research Report Vol 7,* IELTS Australia, Canberra and British Council, London.

Kang, J. Y. (2005). Written narratives as an index of L2 competence in Korean EFL learners. *Journal of Second Language Writing*, 14, 259–279.

Knoch, U., Fairbairn, J, & Huisman, A. (2015). *An evaluation of the effectiveness of training Aptis raters online.* Retrieved 4 August 2015 from: http://www.britishcouncil.org/sites/britishcouncil.uk2/files/evaluation-of-effectiveness-aptis-online-training.pdf

Little, D. (2007). The Common European Framework of Reference for languages: Perspectives on the making of supranational language education policy. *The Modern Language Journal, 91*(4), 645–655. doi: 1.1111/j.1540–4781.2007.00627_2.x

Little, D. (2011). The Common European Framework of Reference for languages: A research agenda. *Language Teaching, 44*(3), 381–393.

Liu, M. & Braine, G. (2005). Cohesive features in argumentative writing produced by Chinese undergraduates. *System, 33*, 623–636.

Malvern, D. D. & Richards, B. J. (2000). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Martyniuk, W. (ed.) (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual.* Studies in Language Testing 33. Cambridge: Cambridge University Press.

McNamara, D. S., Crossley, S. A. & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*, 57–86.

McNamara, D. S., Graesser, A. C., McCarthy, P. & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

Negishi, M., Takada, T. & Tono, Y. (2011). A progress report on the development of the CEFR-J. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring language frameworks: Proceedings of the ALTE Krakow Conference, July 2011, Studies in language testing, 36* (pp. 135–163). Cambridge, UK: Cambridge University Press.

Ngo, X. M. (2014). *Diffusion of the CEFR among Vietnamese teachers: A mixed methods investigation.* Unpublished MA Dissertation, University of Queensland.

Norris, J. M., Brown, J. D., Hudson, T. D. & Yoshioka, J. K. (1998). *Designing second language performance assessment.* Honolulu: University of Hawai'i Press.

North, B. (2014). Putting the Common European Framework of Reference to good use. *Language Teaching 47*(2), 228–249. doi:1.1017/S0261444811000206

North, B. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal, 91*(4), 656–659. doi: 1.1111/j.1540–4781.2007.00627_3.x

North, B., Ortega, A. & Sheehan, S. (2010). *A core inventory of general English*. British Council/EAQUALS.

O'Sullivan, B. (2010). The City & Guilds Communicator examination linking project: A brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 33–49). Studies in Language Testing 33. Cambridge: Cambridge University Press.

O'Sullivan, B. (2011). Language Testing. In Simpson, J. (Ed.) *Routledge handbook of applied linguistics*. Oxford: Routledge.

O'Sullivan, B. (2015a). *Linking the Aptis reporting scales to the CEFR*. TR/2015/001: British Council.

O'Sullivan, B. & Weir C. J. (2011). Test development and validation. In O'Sullivan, B. (Ed.) *Language testing: Theories and practices* (pp. 13–32). Basingstoke: Palgrave MacMillan.

O'Sullivan, B. (2015b). *Aptis Formal Trials Feedback Report.* TR/2015/002: British Council.

O'Sullivan, B. (2015c). *Aptis Test Development Approach.* Aptis Technical Report TR/2015/003: British Council.

Paltridge, B. (2000). *Making sense of discourse analysis*. Gerd Stabler: Gold Coast, QLD.

Purpura, J. (2008). Assessing communicative language ability: models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Language testing and assessment* (2nd Ed., Vol. 7, pp. 53–68). New York, NY: Springer Science and Business Media.

Read, J. & Nation, P. (2008). *An investigation of lexical dimension of the IELTS speaking tests.* IELTS Research Reports, Vol 8, (Ed.) J. Osborne, IELTS Australia, Canberra.

Robinson, P. (2001). Task complexity, cognitive resources and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). New York: Cambridge University Press.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*(Supplement 1), 1–36.

Seedhouse, P. & Harris, A. (2011). Topic development in the IELTS Speaking Test. *IELTS Research Reports, Vol 12,* IDP:IELTS Australia and British Council.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532.

Todd, R. W., Khongput, S. & Darasawanga, P. (2007). Coherence, cohesion and comments on students' academic essays. *Assessing Writing, 12*, 10–25.

Trim, J. (2012). The Common European Framework of Reference for Languages and its background: A case study of cultural politics and educational influences. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalisation of language education policy* (pp. 14–33). Bristol, UK: Multilingual Matters.

Trim, J. (2014). Three decades of work for the Council of Europe. Talking with John Trim (1924–2013): Part II/ Interviewers D. Little & L. King. *Language Teaching, 47*(1), 118–32.

Tyler, A. (1992). Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly, 26*(4), 713–729.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly, 23*(3), 489–508.

Weir, C. J. (2005a). *Language testing and validation: An evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weir, C. J. (2005b). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22*(3), 281–30.

Wilkins, David. (1976). *Notional syllabus*. Oxford: Oxford University Press.

Williams, J. (1992). Planning, discourse marking, and the comprehensibility of international teaching assistants. *TESOL Quarterly*, *26*(4), 693–711.

Wisniewski, K. (2013). The empirical validity of the CEFR fluency scale: the A2 level description. In E. D. Galaczi & C. J. Weir (Eds.), *Exploring Language Frameworks: Proceedings of the ALTE Krakow Conference, July 2011, Studies in Language Testing, 36* (pp. 253–272). Cambridge: Cambridge University Press.

Wu, J. (2012). Policy perspectives from Taiwan. In M. Byram & L. Parmenter (Eds.), *The Common European Framework of Reference: The globalisation of language education policy* (pp. 213–223). Bristol: Multilingual Matters.

APPENDIX 1:
Measures used for quantitative analyses

Category	Ref	Label	Description
Cohesion			
Conjunction	50	CNCAll	All conjunctions incidence
	51	CNCCaus	Causal conjunctions incidence
	52	CNCLogic	Logical conjunctions incidence
	53	COCADC	Adversative and contrastive conjunctions incidence
	54	CONTemp	Temporal conjunctions incidence
	55	CONTEMPEXi	Expanded temporal conjunctions incidence
	56	CNCAdd	Additive conjunctions incidence
Reference	28	CRFNO1	Noun overlap adjacent sentences
	29	CRFAO1	Argument overlap adjacent sentences
	30	CRFBS1	Stem overlap adjacent sentences
	31	CRFNOa	Noun overlap all sentences
	32	CRFAOa	Argument overlap all sentences
	33	CRFBSa	Stem overlap all sentences
	34	CRFCWO1	Content word overlap adjacent sentences proportional mean
	36	CRFCWOa	Content word overlap all sentences proportional mean
Lexical cohesion	101	WRDHYPn	Hypernymy for nouns
	102	WRDHYPv	Hypernymy for verbs
	103	WRDHYPnv	Hypernymy for nouns and verbs
Lexical richness	3	DESWC	Word count number of words
	82	WRDNOUN	Noun incidence
	83	WRDVERB	Verb incidence
	84	WRDADJ	Adjective incidence
	85	WRDADV	Adverb incidence
	86	WRDPRO	Pronoun incidence
	46	LDTTRc	Lexical diversity type–token ratio content words
	47	LDTTRa	Lexical diversity type–token ratio all words
	49	LDVOCDa	Lexical diversity d (VOCD) all words

Note: adopted from McNamara et al 2014 pp. 249–251; Ref – reference number to Coh-Metrix

APPENDIX 2: Test of normality

	Level	Shapiro-Wilk	*df*	*sig*	Task	Shapiro-Wilk	*df*	*sig*
Conjunctions								
Causal	*1*	.914	12	.24	*1*	.89	20	.02
	2	.951	22	.324	*2*	.96	21	.50
	3	.967	26	.548	*3*	.95	21	.27
	4	.942	17	.345	*4*	.97	22	.80
	5	.882	7	.234				
Logical	*1*	.957	12	.743	*1*	.93	20	.14
	2	.982	22	.942	*2*	.96	21	.58
	3	.965	26	.505	*3*	.95	21	.28
	4	.942	17	.349	*4*	.97	22	.72
	5	.936	7	.603				
Adversative	*1*	.681	12	.001	*1*	.64	20	.00
Contrastive								
	2	.781	22	0	*2*	.89	21	.03
	3	.932	26	.087	*3*	.92	21	.10
	4	.963	17	.683	*4*	.92	22	.09
	5	.949	7	.721				
Temporal	*1*	.715	12	.001	*1*	.79	20	.00
	2	.916	22	.062	*2*	.90	21	.03
	3	.949	26	.218	*3*	.89	21	.02
	4	.972	17	.845	*4*	.95	22	.33
	5	.954	7	.767				
Expanded temporal	*1*	.9	12	.159	*1*	.89	20	.03
	2	.819	22	.001	*2*	.89	21	.02
	3	.857	26	.002	*3*	.79	21	.00
	4	.904	17	.08	*4*	.94	22	.16
	5	.932	7	.571				
Additive	*1*	.946	12	.585	*1*	.94	20	.26
	2	.989	22	.995	*2*	.96	21	.56
	3	.976	26	.784	*3*	.96	21	.46
	4	.913	17	.11	*4*	.865	22	.01
	5	.878	7	.217	*1*	.97	20	.80
All (Combined)	*1*	.962	12	.819	*2*	.87	21	.01
	2	.972	22	.762	*3*	.97	21	.74
	3	.935	26	.1	*4*	.93	22	.15
	4	.949	17	.437				
	5	.919	7	.461				
References								
Adjacent sentences (local)								
Noun	*1*	.901	12	.161	*1*	.94	20	.19
	2	.971	22	.736	*2*	.94	21	.26
	3	.962	26	.428	*3*	.92	21	.11
	4	.889	17	.044	*4*	.92	22	.06
	5	.898	7	.322		.94	20	.26

	Level	Shapiro-Wilk	df	sig	*Task*	Shapiro-Wilk	df	sig
Argument	*1*	.963	12	.82	*1*	.97	21	.66
	2	.93	22	.123	*2*	.92	21	.08
	3	.961	26	.401	*3*	.94	22	.21
	4	.947	17	.405	*4*	.93	20	.13
	5	.939	7	.625		.94	21	.26
Stem	*1*	.88	12	.086	*1*	.95	21	.27
	2	.97	22	.714	*2*	.96	22	.51
	3	.966	26	.528	*3*	.94	20	.25
	4	.914	17	.116	*4*	.92	21	.09
	5	.934	7	.585		.88	21	.02
Content	*1*	.963	12	.821	*1*	.87	22	.01
	2	.881	22	.012	*2*	.94	20	.19
	3	.917	26	.038	*3*	.94	21	.26
	4	.759	17	.001	*4*	.92	21	.11
	5	.917	7	.446				
All sentences (global)								
Noun	*1*	.907	12	.197	*1*	.96	20	.62
	2	.951	22	.336	*2*	.95	21	.35
	3	.966	26	.512	*3*	.98	21	.91
	4	.935	17	.263	*4*	.97	22	.70
	5	.967	7	.877		.95	20	.39
Argument	*1*	.951	12	.646	*1*	.91	21	.05
	2	.904	22	.035	*2*	.94	21	.25
	3	.921	26	.048	*3*	.93	22	.12
	4	.823	17	.004	*4*	.94	20	.20
	5	.827	7	.075		.99	21	1.00
Stem	*1*	.918	12	.268	*1*	.96	21	.58
	2	.966	22	.623	*2*	.96	22	.48
	3	.927	26	.065	*3*	.94	20	.24
	4	.942	17	.344	*4*	.94	21	.24
	5	.943	7	.669		.84	21	.00
Content	*1*	.911	12	.222	*1*	.98	22	.89
	2	.957	22	.423	*2*	.96	20	.62
	3	.969	26	.606	*3*	.95	21	.35
	4	.946	17	.399	*4*	.98	21	.91
	5	.924	7	.498				
Lexical cohesion								
for nouns	*1*	.982	12	.99	1	.88	20	.02
	2	.878	22	.011	2	.97	21	.83
	3	.897	26	.013	3	.94	21	.21
	4	.878	17	.029	4	.94	22	.16
	5	.987	7	.986				
for verbs	*1*	.648	12	0	1	.59	20	.00
	2	.971	22	.73	2	.94	21	.19
	3	.972	26	.687	3	.94	21	.25
	4	.984	17	.987	4	.95	22	.38
	5	.935	7	.593				

	Level	Shapiro-Wilk	df	sig	Task	Shapiro-Wilk	df	sig
A combination of both nouns and verbs	*1*	.869	12	.063	1	.96	20	.48
	2	.934	22	.151	2	.80	21	.00
	3	.966	26	.519	3	.96	21	.49
	4	.958	17	.589	4	.98	22	.84
	5	.875	7	.203				
Vocabulary use								
Word count (per 60 sec)	*1*	.911	12	.219	*1*	.93	20	.12
	2	.945	22	.251	*2*	.95	21	.29
	3	.905	26	.02	*3*	.94	21	.26
	4	.968	17	.775	*4*	.91	22	.05
	5	.875	7	.204				
Noun	*1*	.951	12	.649	*1*	.89	20	.03
	2	.824	22	.001	*2*	.79	21	.00
	3	.965	26	.495	*3*	.99	21	.98
	4	.966	17	.746	*4*	.92	22	.09
	5	.7	7	.004				
Verb	*1*	.94	12	.494	*1*	.92	20	.10
	2	.952	22	.349	*2*	.98	21	.90
	3	.967	26	.549	*3*	.96	21	.44
	4	.959	17	.606	*4*	.89	22	.02
	5	.897	7	.314				
Adjective	*1*	.948	12	.606	*1*	.95	20	.35
	2	.967	22	.634	*2*	.93	21	.12
	3	.908	26	.023	*3*	.92	21	.11
	4	.89	17	.046	*4*	.97	22	.78
	5	.915	7	.432				
Adverb	*1*	.927	12	.348	*1*	.93	20	.16
	2	.936	22	.166	*2*	.97	21	.65
	3	.934	26	.098	*3*	.89	21	.03
	4	.957	17	.581	*4*	.97	22	.80
	5	.955	7	.778				
Pronoun	*1*	.941	12	.512	*1*	.96	20	.55
	2	.956	22	.413	*2*	.88	21	.01
	3	.98	26	.873	*3*	.93	21	.12
	4	.916	17	.125	*4*	.94	22	.24
	5	.879	7	.222				
Type–token ratio (content words)	*1*	.983	12	.992	*1*	.96	20	.46
	2	.98	22	.909	*2*	.97	21	.64
	3	.932	26	.086	*3*	.83	21	.00
	4	.968	17	.775	*4*	.97	22	.78
	5	.96	7	.817				

	Level	Shapiro-Wilk	df	sig	*Task*	Shapiro-Wilk	df	sig
Type–token ratio (all words)	*1*	.959	12	.767	*1*	.95	20	.30
	2	.965	22	.59	*2*	.85	21	.00
	3	.939	26	.129	*3*	.80	21	.00
	4	.936	17	.279	*4*	.94	22	.24
	5	.847	7	.116				
VOCD	*1*	.851	22	.004	*1*	.76	20	.00
	2	.978	26	.819	*2*	.88	21	.02
	3	.938	17	.296	*3*	.90	21	.03
	4	.942	7	.66	*4*	.92	22	.08
	5	.911	12	.219				

APPENDIX 3:

Example of coherence analysis (Task 1, Performance 9, Mark: 5/5)

Question 1 – Please tell me about your first school.

Question 2 – Please tell me about the last time you visited friends.

Question 3 – Please tell me about your favourite singer.

	Candidate speech, segmented in C-units	**Aspects relevant to coherence**
1	My school is uh big school	Responds directly to Q1, identifies phenomenon
2	Ah my school, ah my school is ah, ah more more ah student	Expands on response with additional detail regarding size (enrolment) but appears hampered by lack of vocabulary
3	Ah my ah first day ah, (unint) remember	Attempts to further extend response but seems to lack vocabulary to do this. May have misunderstood “first school”
4	Ah my school also	Fragment that indicates intention to add to previous response but incomplete
6	I visited my friends in Italy a month ago	Responds directly to Q2, provides orientation to a recount, introducing participants and setting
7	I went to visit them because I went to last year to Verona	Expands on response providing reason, telling what happened and in what sequence
8	and I decided to, to book a flight and go to see them because it was a year that I didn’t see them	Expands on previous utterance by providing details with “and,” also gives justification for actions, linked by “because”
9	so I wanted to see them	Concludes by reiterating reason for visit, providing a re-orientation
10	My favourite singer is a man called (UNCLEAR)	Responds directly to Q3, identifies phenomenon to be described
11	I discovered his songs in a film	Develops description by adding details
12	and since then I, I have been following his career	Develops description, links with “and”
13	and I quite like his songs because he’s a very romantic singer	Links to previous point with “and”, develops topic of her favourite singer by causal explanation with “because”
14	and he sings in English so it helps me to learn	Extends previous point in C-unit with “and”. Develops topic further by explaining additional reasons for liking the singer with “so”

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

FEATURES OF DISCOURSE AND LEXICAL RICHNESS AT DIFFERENT PERFORMANCE LEVELS IN THE APTIS SPEAKING TEST

AR-G/2017/2

Noriko Iwashita, Lyn May and Paul Moore

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203