

Using Big Data and AI to Look at Small Conversations – The Value of Data in Teaching Conversational English

Tony McEnery



@tonymcenery.bsky.social

Email: anthony.mcenery@polyu.edu.hk

Introduction

- What is AI?
- It is more than GenAI
- Advanced methods for pattern discovery
- How can that be helpful to language teaching?
- How can it be useful in language assessment?
- How can it be useful in materials development?
- How can we get a more authentic view of spoken language performance?
- Corpus data is a key tool for the production of language teaching materials and assessments. But that data needs to be processed.

Home

What is a Corpus?

Corpus-based publications

Cambridge International Corpus

Cambridge Learner Corpus

Resources

ELT website

Ordering

Help

Corpus-based publications

Cambridge University Press Corpus-based books (Chart 1)

Level	Adult	Secondary	Exams	Professional English	Cambridge Copy Collection
Beginner					
Elementary	 	 	 		
Pre-Intermediate	 				
Intermediate	 		  	 	
Upper Intermediate			      		
Advanced			  		

Problems

- A lot of the data and the associated research focuses on writing
- Speech is neglected because it is hard to collect
- Learner spoken language interactions are difficult to analyse
- Learner spoken interaction assessment is difficult to access

Yet interaction is important

- It is co-operative and co-constructed
- Apparent disfluencies and incidental features are, in fact, key elements of conversation

S1: I used to uh like to listen to classical music uh

S2: what?

S1: um I used to like listening to classical and world and jazz and weird music under it's like a is that a tre.. is that a bass clef or something?

S2: oh yeah it's in like a the number three's in like a alto <short pause> matt <short pause> come on now <short pause> no it's <short pause> it's an alto thank you very much

S1: but if the if the bottom bit uh goes then it that's like a

S2: that's an

S1: if you j... if you cut off the bottom bit it looks like a like a bass clef

S2: yeah <short pause> that's true <short pause> with the dots yeah <short pause> is it even playing anything? Or is just the quiet bit of the thing?

S1: no this is Radio Three man <short pause> radio Three often just doesn't do anything

S2: just silence?

S1: yeah <short pause> because si... in music silence silence is more important than noise

S2: listen hang on <short pause> some background pipes

Conversation is cooperative and it is co-constructed.

The Corpus used in this Talk

- AI can help to explore spoken language – but only if we have the right data in the first place
- Trinity Lancaster Corpus – 4.1 million words
- Based on the Trinity College London GESE (Graded Examination of Spoken English) Exam



L2 users: English proficiency

Three levels

- Advanced: Grades 10, 11, 12 (CEFR: C1 & C2)
- Intermediate: Grades 7 & 8 (CEFR: B2.1 & B2.2)
- Threshold: Grade 6 (CEFR: B1.2)

How measured?

- ✓ Proficiency by task (A-D)
- ✓ Proficiency overall (A-D)

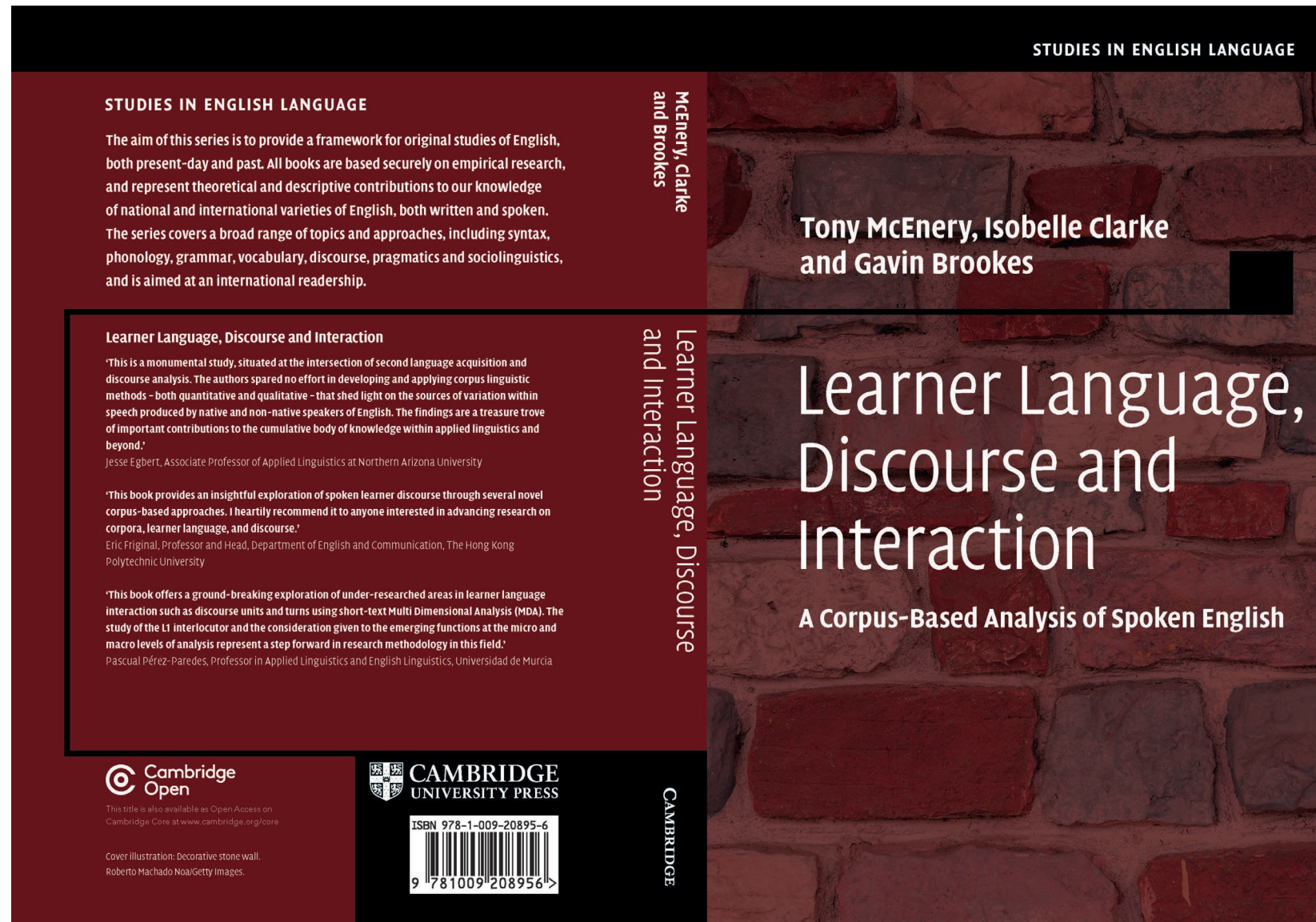
Four tasks - overview

	Topic familiarity	Interlocutor roles	Interactiveness	Formality
PRES	pre-selected topic	candidate-led	monologic	(semi-)formal
DISC	pre-selected topic	jointly-led	dialogic	semi-formal
INT	general topic	candidate-led	dialogic	semi-formal
CONV	general topic	jointly-led	dialogic	semi-formal

- Interactions dialogic
- Emphasis on continuing the conversation
- So we have the data – how can AI techniques help us process it?



- Our focus
- Our problems
- Our solution
- Our findings



STUDIES IN ENGLISH LANGUAGE

STUDIES IN ENGLISH LANGUAGE

The aim of this series is to provide a framework for original studies of English, both present-day and past. All books are based securely on empirical research, and represent theoretical and descriptive contributions to our knowledge of national and international varieties of English, both written and spoken. The series covers a broad range of topics and approaches, including syntax, phonology, grammar, vocabulary, discourse, pragmatics and sociolinguistics, and is aimed at an international readership.

McEnery, Clarke
and Brookes

Tony McEnery, Isobelle Clarke
and Gavin Brookes

Learner Language, Discourse and Interaction

'This is a monumental study, situated at the intersection of second language acquisition and discourse analysis. The authors spared no effort in developing and applying corpus linguistic methods - both quantitative and qualitative - that shed light on the sources of variation within speech produced by native and non-native speakers of English. The findings are a treasure trove of important contributions to the cumulative body of knowledge within applied linguistics and beyond.'

Jesse Egbert, Associate Professor of Applied Linguistics at Northern Arizona University

'This book provides an insightful exploration of spoken learner discourse through several novel corpus-based approaches. I heartily recommend it to anyone interested in advancing research on corpora, learner language, and discourse.'

Eric Prigal, Professor and Head, Department of English and Communication, The Hong Kong Polytechnic University

'This book offers a ground-breaking exploration of under-researched areas in learner language interaction such as discourse units and turns using short-text Multi Dimensional Analysis (MDA). The study of the L1 interlocutor and the consideration given to the emerging functions at the micro and macro levels of analysis represent a step forward in research methodology in this field.'

Pascual Pérez-Paredes, Professor in Applied Linguistics and English Linguistics, Universidad de Murcia

Learner Language, Discourse
and Interaction



This title is also available as Open Access on
Cambridge Core at www.cambridge.org/core

Cover illustration: Decorative stone wall.
Roberto Machado Noa/Getty Images.



ISBN 978-1-009-20895-6



CAMBRIDGE

Our Problems

- How do we approach the uppermost level of discourse micro-structure – the turn?
- But how can we develop a functional view of the turn which is undertaken automatically across millions of words of data?
- How could this view then show how co-construction of discourse works?

Why not just rely on ChatGPT?

but actually only occurs 35 times, but also only occurs 35 times, but and occurs 27 times - can these really account for the scale of difference seen and how do such relatively low figures count as high frequency?

No, they can't.

And your observation exposes a weakness in my earlier interpretation.

A possible solution – the multi-dimensional approach

- We can use AI to grammatically analyse the data using NLP tools.
- In principle, this explicitly meshes lexis and higher order structural units
- A pattern discovery approach which takes an automated grammatical analysis of words and then looks for patterns of co-occurrence to discover function in groupings of grammatical features
- Very successful at teasing apart differences in writing – e.g. showing what the features of successful essay writing may be, or of speech versus writing
- Could we not take the same approach to turns?

The problem of frequency

- Not as the technique stands – the problem of frequency in short texts.

E: <unclear text="you"/>

S: I'm fine ma'am

E: good

S: how are you ma'am?

E: I'm good thank you very much okay so erm good morning my name is <anon type="name"/> what's your name?

S: my name is <anon type="name"/>

E: <anon type="name"/>

S: yes ma'am

E: and you're a grade seven

S: yes

E: and can you just turn your oh you don't need to take it off

S: mm

E: okay thanks thanks for showing me that that's your ID card thanks <anon type="name"/>

- 68 words long
- BE as a main verb occurs 6 times
- Normalised frequency per 1,000 words of 88.2
- Frequency per 1,000 words of BE as a main verb is 37.64

Our solution

- Multiple correspondence analysis (Benzécri, 1979; Clarke and Grieve, 2017).
- A powerful pattern discovery technique that can work with relatively sparse data

Benzécri, J.P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données*, 4, 377–378.

Clarke, I. and Grieve, J. (2017) Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online* 1–10. Vancouver: ACL. <https://doi.org/10.18653/v1/w17-3001>

Dominance →

+N

-N

Function A	Turn 1	Dimension 1	Function B
Function C	Turn 1	Dimension 2	Function D
Function E		Dimension 3	Function F
Function G	Turn 1	Dimension 4	Function H
Function I		Dimension 5	Function J

Our Findings

- I will narrow my focus to grade 6, 7 and 8 (CEFR B1/B2), focusing on the conversation (unplanned) and discussion (planned) tasks
- Three sets of results are presented – examiner turn, learner turn and discourse unit functions
- We will see how function interacts with a number of supplementary variables
- Let's begin by looking at the examiner turn functions and then focus in on dimension two

Dimension	Label	Features present
Dim 1 +ve	Long turns	Prediction modal, Infinitive, Private verb, Nominalisation, Quantifier, Auxiliary DO, Predicative Adjective, Subordinator, Definite article, General Determiner, Past tense, Analytic negation, Indefinite article, Contracted forms, Pronoun it, Third person singular verb, Coordinating Conjunction, General adverb, Preposition, Stative forms, First person pronoun, WH-word, Question mark, Subject pronoun, General verb, Second person pronoun, Attributive adjective, General noun.
Dim 1 -ve	Short turns	Only absent features (Subject pronoun, General noun, General verb)
Dim 2 +ve	Descriptive	Predicative adjective, Pronoun it, Contracted forms, Analytic negation, Stative forms, Demonstratives, Amplifier, Third person singular verb forms
Dim 2 -ve	Information seeking	WH-word, Auxiliary DO, Question marks, Second person pronoun
Dim 3 +ve	Guide future action	Nominalisation, Modal of prediction, Public verb, Quantifier, First person pronoun, Infinitive, General adverb, Definite article, Preposition
Dim 3 -ve	Discovering stance	WH-word, Auxiliary DO, Third person singular verb, Question, Predicative adjective, Past tense verb, Pronoun it, stative form

Dimension	Label	Features present
Dim 4 +ve	Stating stance	Analytic negation, Auxiliary DO, Negative interjection, Private verb, Third person pronoun, Infinitive, Subject pronoun, General verb
Dim 4 –ve	Discussing the here and now	Third person singular verb, Demonstrative, General determiner, Proper noun, Stative, Definite article, Preposition, Attributive adjective, General noun
Dim 5 +ve	Interjection (positive)	Positive interjection
Dim 5 -ve	Interjection (other)	Negative interjection, General interjection
Dim 6 +ve	Past orientation	Negative interjection, Indefinite article, Third person singular verb, Analytic negation, Third person pronoun, Past tense verb, Quantifier, Definite article, Attributive adjective, Amplifier, Preposition, General Noun
Dim 6 -ve	Future orientation	Modal of prediction, Contracted form, Infinitive, Predicative adjective, Demonstrative, General determiner, Second person pronoun, Subordinator, Positive interjection
Dim 7 +ve	Narrative	Third person pronoun, Negative interjection, Past tense verb, Public verb, Third person singular verb, General determiner, Subordinator
Dim 7 -ve	Stance seeking	Amplifier, Attributive adjective, Quantifier, Auxiliary DO, Private verb, Indefinite article, Second person pronoun

Dimension 2

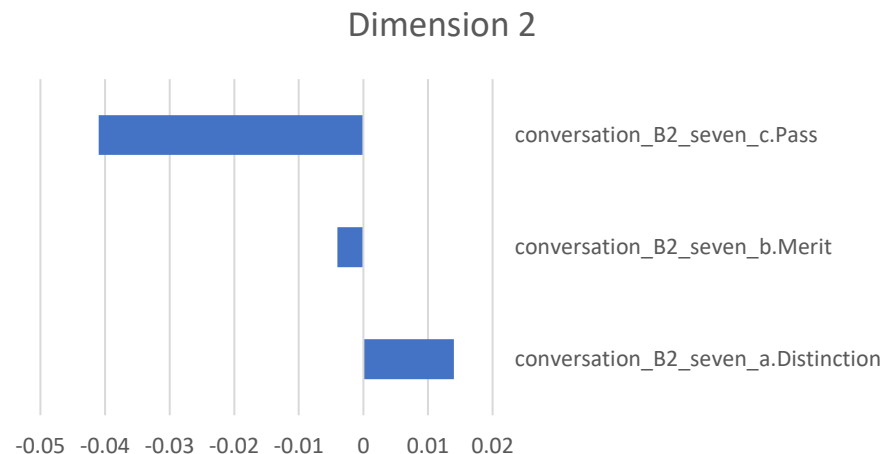
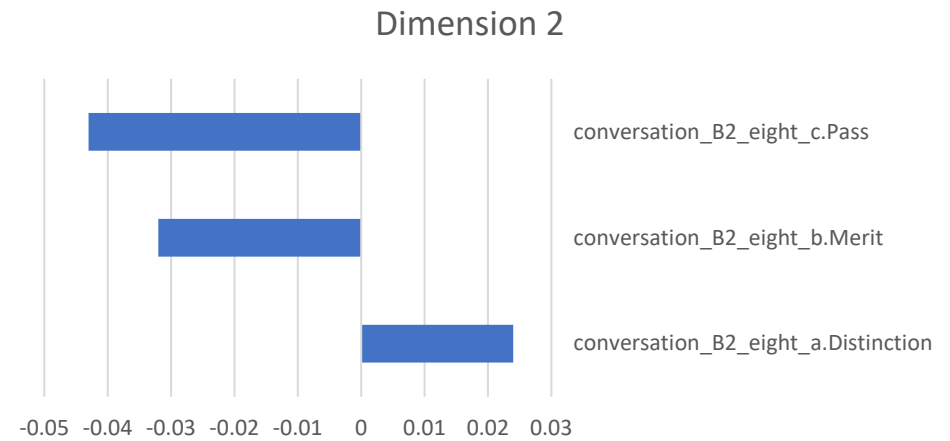
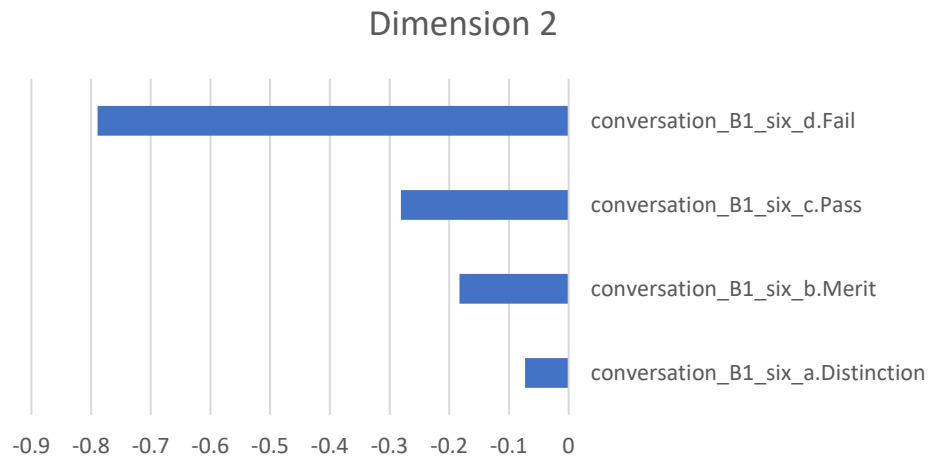
Positive: Descriptive

mm it sounds good erm yeah I've
never really watched anything
similar to that I've haven't watched
a lot of cartoons but some things
yeah I like to see yeah

Negative: Information seeking

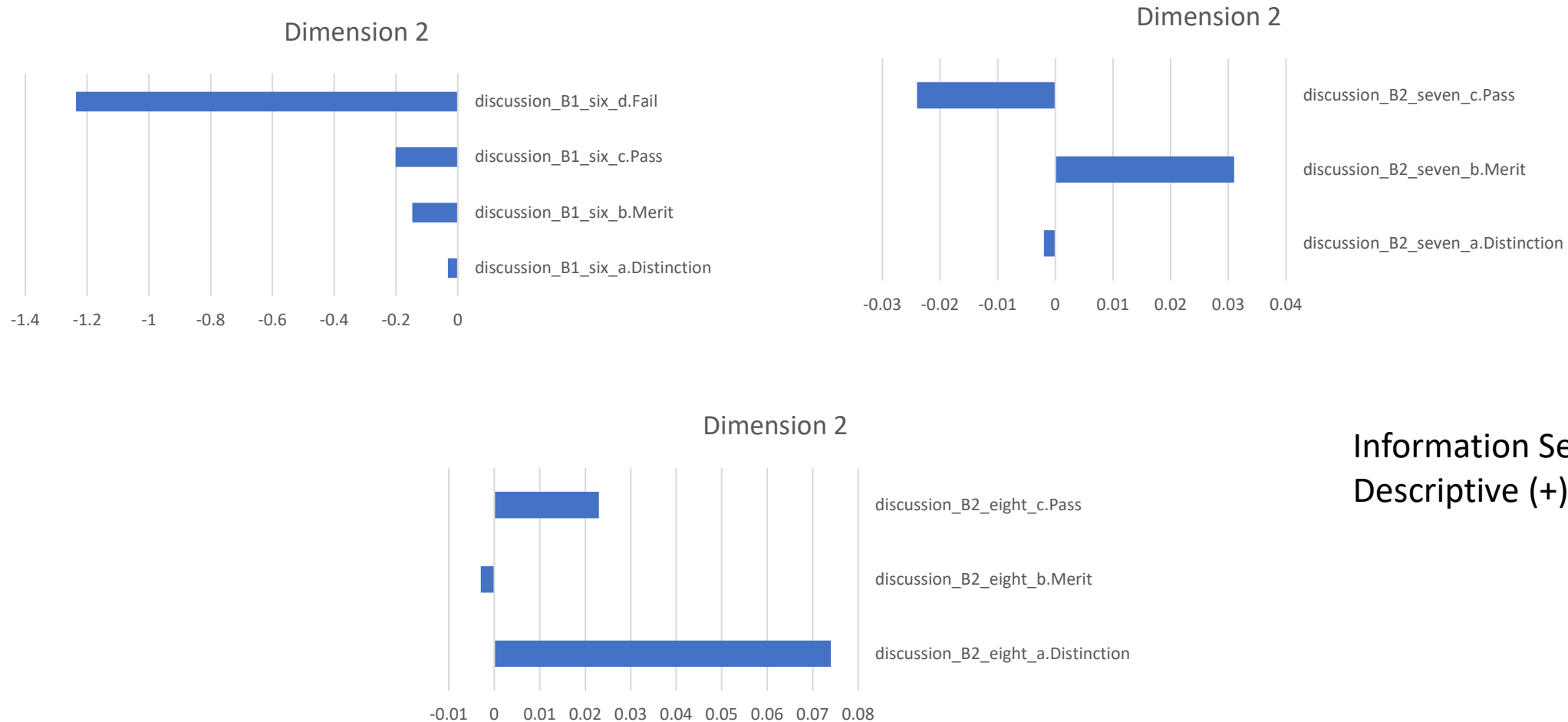
and if you travel to another country
what preparations do you have to
make?

Learners in the conversation task received a lower overall mark if the examiners produced more turns associated with information seeking as opposed to descriptive turns.



Information Seeking (-)
Descriptive (+)

Learners at grade 6 (B1) in the discussion task received a lower overall mark if the examiners produced more turns associated with information seeking as opposed to descriptive and stance turns. No other pattern.



E: yes okay alright then so i-i-if you go to a wedding what what kind of clothes do you wear?

S: if you go to the big s= er st= er shopping centre centre

E: yeah

S: Marineda City

SCAFFOLDING

E: yeah

S: are a lot of shops

E: <pause/> <unclear text="just"/> to to buy elegant clothes

S: yes

Learner turns

- Dimension 2: Involved vs. Informational
 - Dimension 3: Irrealis (unknown) versus Realis (known)
 - Dimension 4: Infer vs. Reveal
 - Dimension 5: Narrative vs. Non-narrative
 - Dimension 6: Opinionated Narrative vs. Situation Dependent Commentary
-
- Finding: a narrower repertoire of functions.
 - BUT – a different role in the interaction. Is there any evidence that proficiency may relate to the widening of the repertoire?
 - Look at task, grade and then L1 speakers performing the same tasks.

Dimension 2

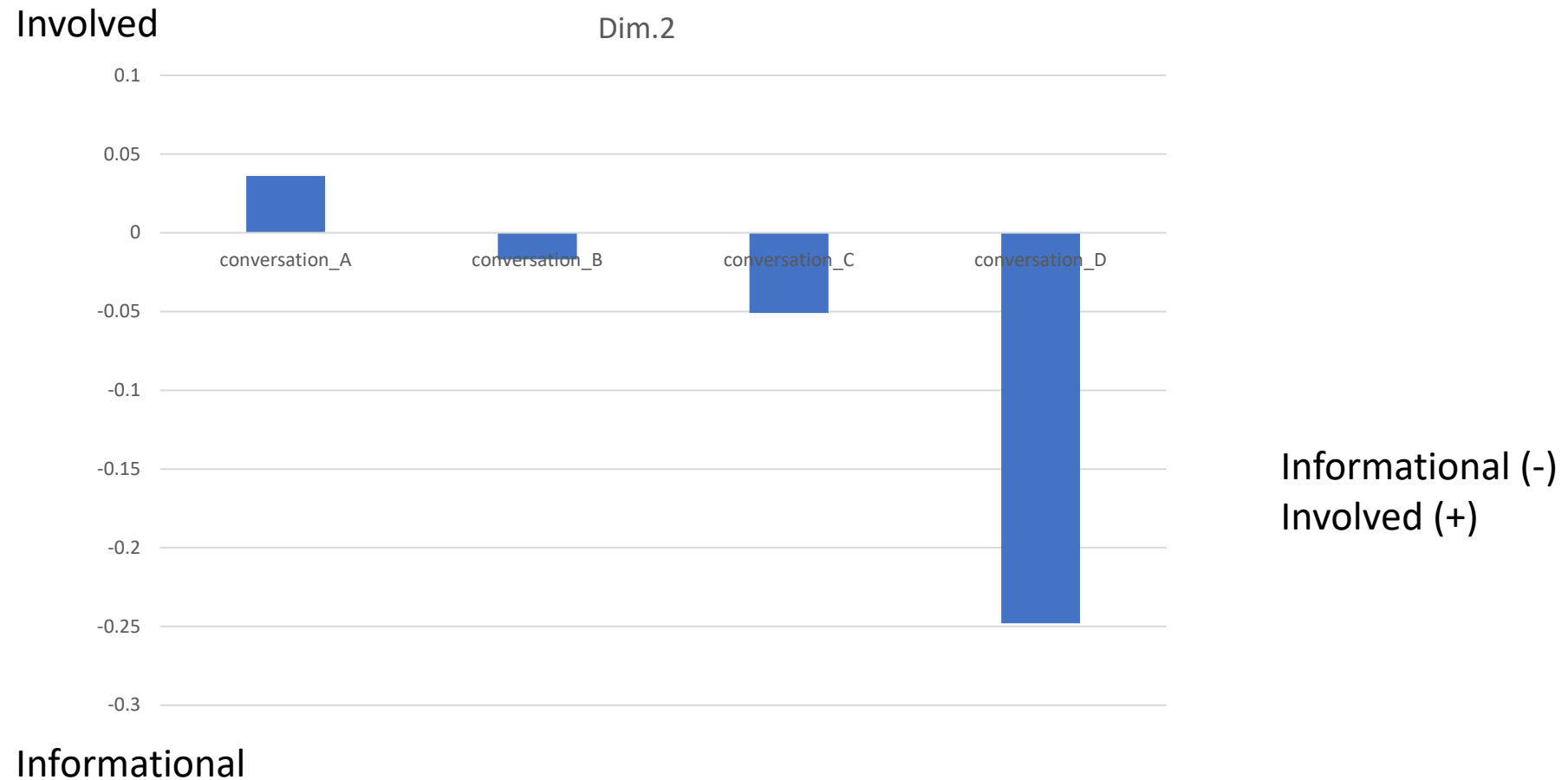
Positive: Involved

- Copular Verb (she *looks* beautiful)
- Existential *there* (*there* were very cold but sunny days)
- Comparative (she looks *more* beautiful)
- Predicative Adjective (she looks *beautiful*)
- Passive (*it is exported* to a lot of countries...)
- Cause Subordinator (...*because* it's very famous)
- Pronoun IT (*it*)
- Third person singular verb (*looks*)
- Contrastive Conjunction (*but*)
- Demonstrative Pronoun (*that*)

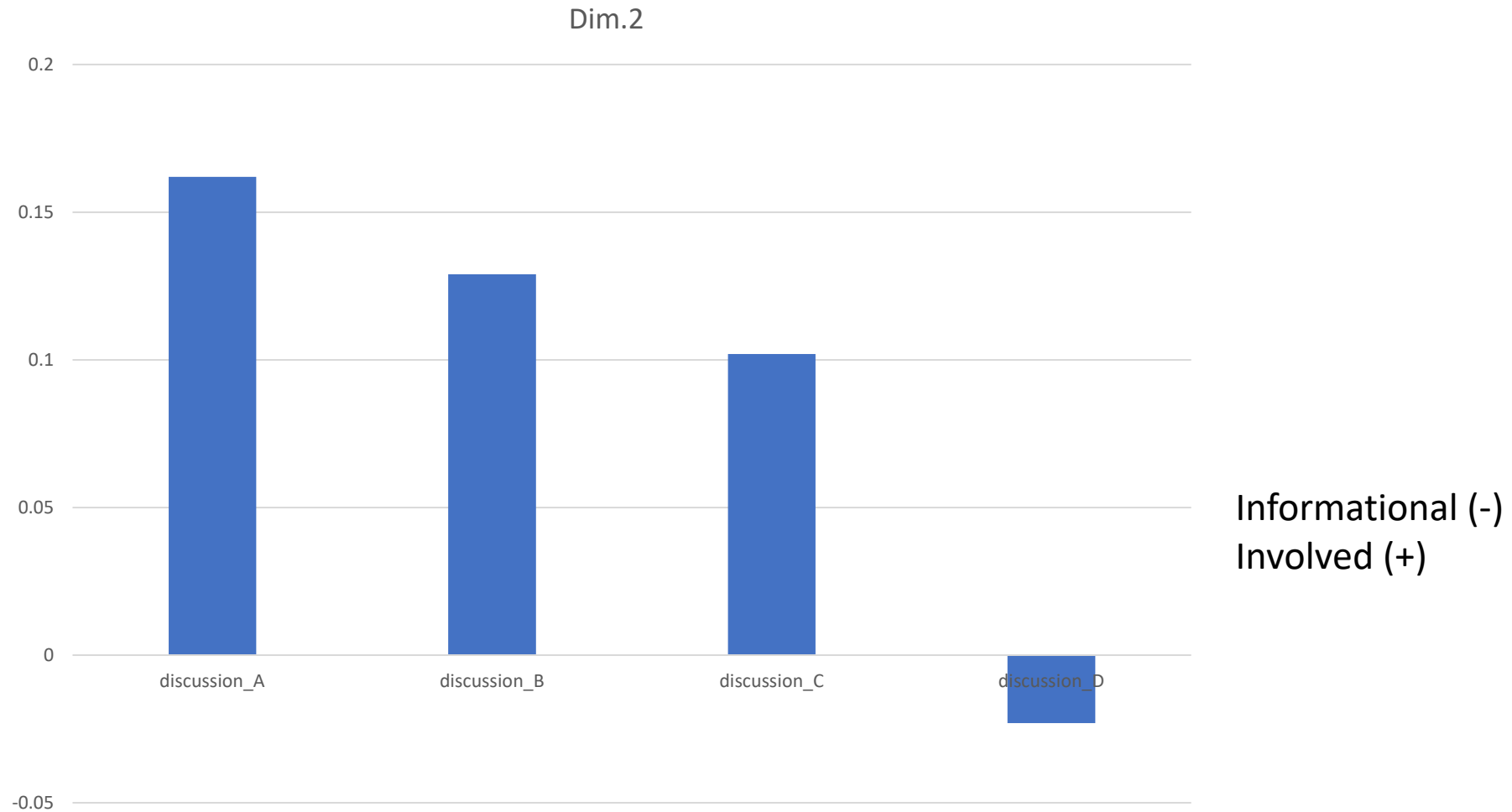
Negative: Informational

- Suasive Verb (*move*)
- Phrasal Verb (*find out*)
- Prediction Modal (*we'll*)
- Time Adverb (*now*)
- Stance Verb (*need*)
- Infinitive (*keep*)
- Nominalisation (*conversation, information*)
- Object Pronoun (*me*)
- Indefinite Pronoun (*something*)
- Public Verb (*ask*)
- Question Mark
- Second Person Pronoun (*you*)

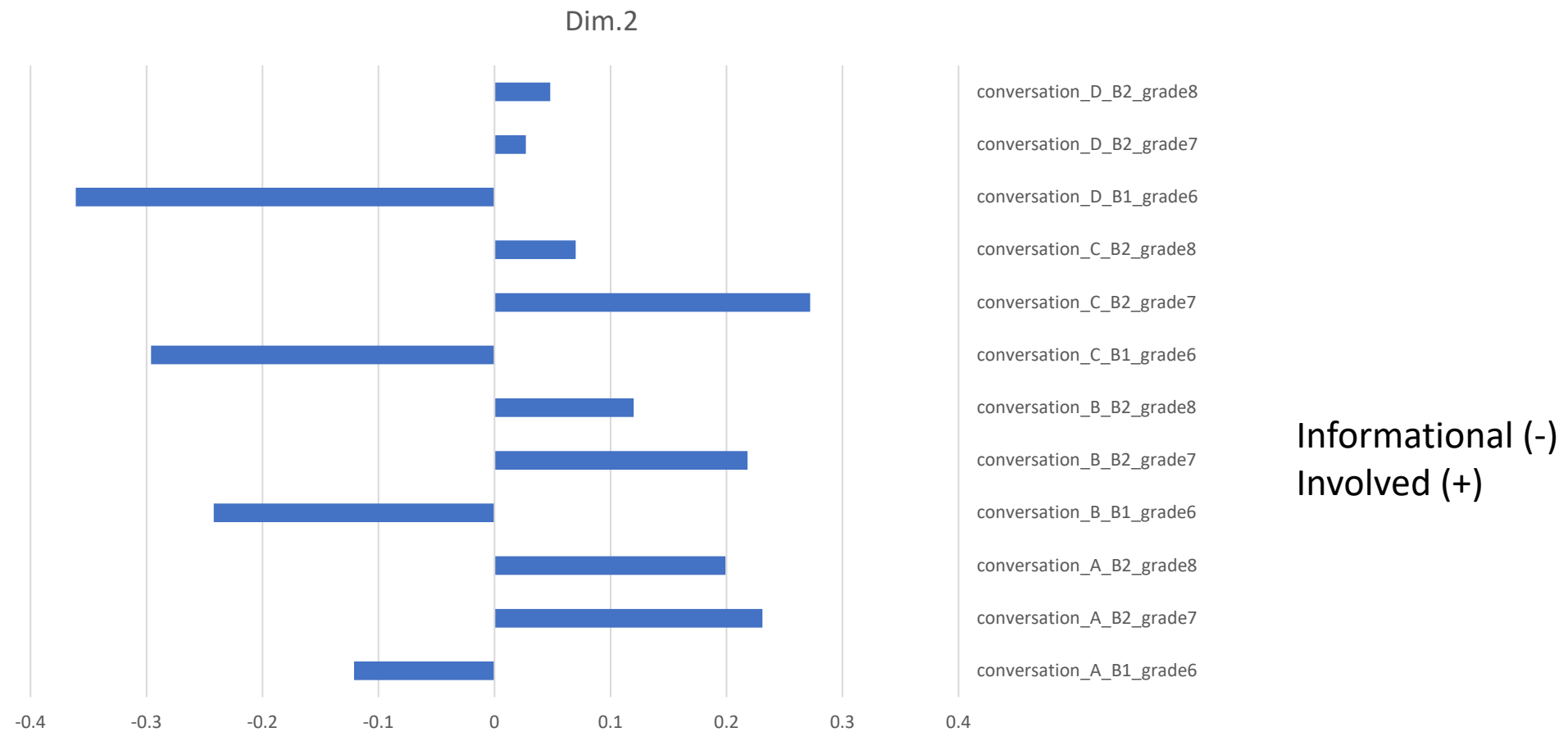
Dimension 2 – Learner Turns in Conversation



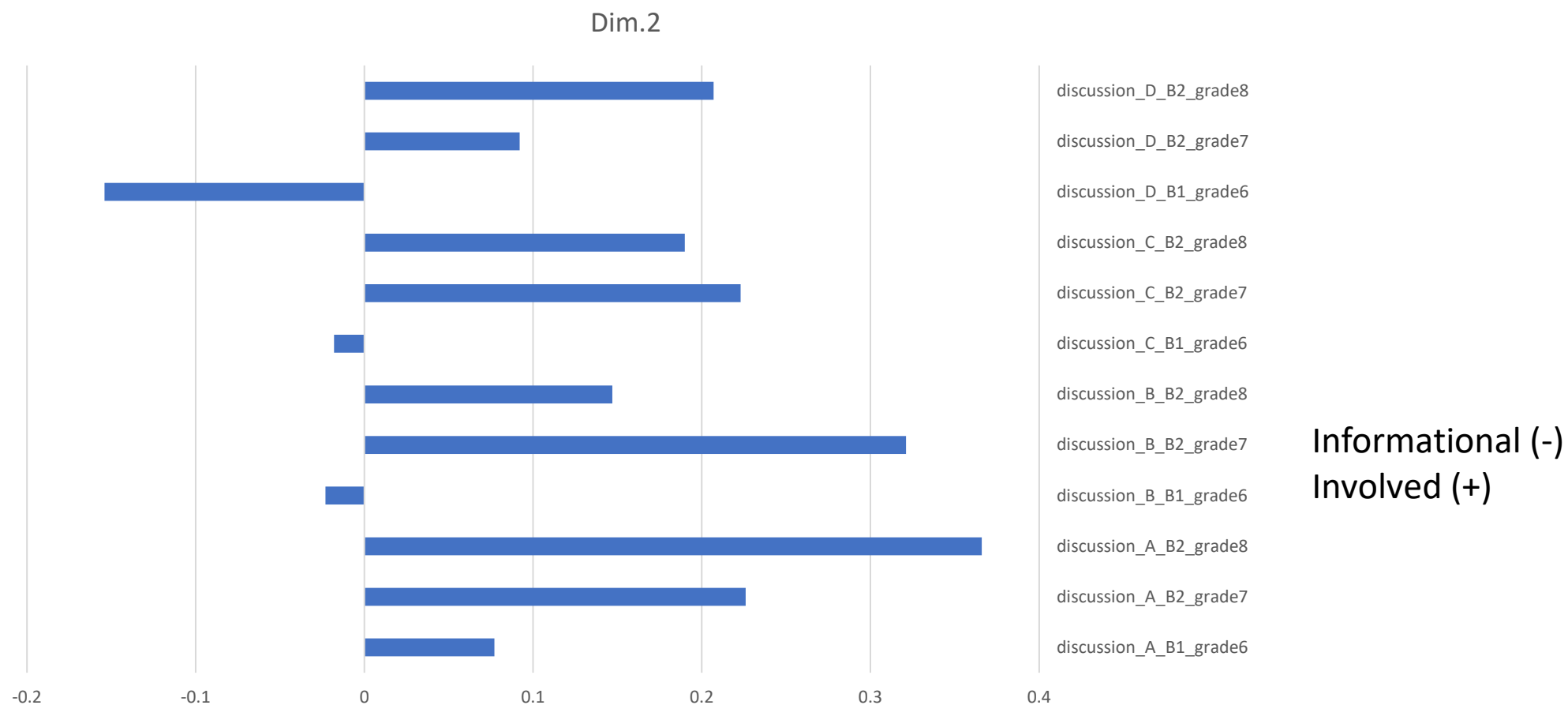
Dimension 2 – Learner Turns in Discussion



Learner Turns in Conversation



Learner Turns in Discussion



Consequences for Assessment, Notes for Teachers

- Key Takeaway 1 - more successful speakers ask more questions (both direct and indirect) while less successful candidates rely more on declarative sentences or chunks to move the conversation forward.
- So the students should seek information in an exam – the less passive they are, the better.

- Key Takeaway 2 - more successful speakers show more support as listeners than less successful speakers.
- The very same short sequences we saw early in the talk are precisely the same sequences that are hallmarks of success in learner speech.

- Key Takeaway 3 - more successful candidates used fewer hesitation markers ('er', 'erm' or unfilled pauses) than less successful candidates.
- But note that the pauses do not cause a conversation to collapse.
- Nor do they cause explicit correction.
- They evoke a *co-operative* response from the examiner

Examiner: to to be an actress wi-will you have to go to s... er theatre school to study?

Student: erm I'm like no I I w...

Examiner: mm

Student: not right now I'm in academy

Examiner: right

Student: of a of a famous actor

Examiner: okay

Student: singer

Examiner: mm

Student: and there but there I do musical comedy like I do it also> I sing and dance and act

Examiner: okay uhu

B1 exam, Spanish L1 speaker
Passed with a Merit score

- Key takeaway 4 - More successful speakers take a more active part in the conversation and assume responsibility for its continuation.
- The student should be co-operative too!

Student: erm have you ever seen visit Paris?

Examiner: yes I have

Student: yes

Examiner: yes I have yeah

Student: do you vis=

Examiner: a long time ago

Student: <laugh/>

Examiner: but yeah

Student: did you visit the Eiffel Tower?

Examiner: yes I did yes I did we went up not to the very very top but maybe

Student: the middle

Examiner: halfway

Student: mm

B1 Exam, Spanish L1 Speaker
Passed with grade Merit

- Key take away 5 - less successful speakers did not fully understand the prompt and did not seek sufficient clarification.
- Again – be cooperative. This happens in L1 conversation, it is not abnormal, it is not a sign of failure to ask for clarification.

Resources

- B1/B2/C1 Worksheets for students and Teacher Notes



Successful speakers... are clear about their opinions

At Trinity College London, we have studied the scripts from hundreds of our speaking exams. We wanted to find out what makes some test takers more successful than others. One thing that test takers who perform better do is make their opinions clear.

In any discussion, being clear about what you think helps the other speaker to understand what you are saying and continue the conversation.

You can be clear about a positive opinion: *Oh yes – I agree!*

You can be clear about a negative opinion: *No, I can't agree with you.*

You can be clear about being uncertain: *I'm really not sure.*

Being clear does **not** mean that you have to know exactly what you think. It **does** mean that you can say that you are certain or uncertain about your opinion.

Activity 1

Read these two exam extracts of a discussion between a test taker <TT> and an examiner <E> about people using mobile phones on trains.

Corpus extract

- | | |
|--|---|
| <p>A</p> <p><E> Do you think that mobile phones should be banned on trains? Should they be forbidden?</p> <p><TT> Well, I think mobile phones are very useful.</p> <p><E> Yes, they can be. But what about people talking on them on trains? Do you think that is sometimes annoying?</p> <p><TT> It can be annoying, yes.</p> <p><E> Some people think they should be forbidden. Do you think they are right?</p> <p><TT> I think that people shouldn't talk too loudly.</p> <p><E> I see.</p> | <p>B</p> <p><E> Do you think that mobile phones should be banned on trains? Should they be forbidden?</p> <p><TT> Because people speak too loudly?</p> <p><E> Yes. Some people find that very annoying.</p> <p><TT> I can see their point. It is sometimes annoying. But I don't agree that they should be banned.</p> |
|--|---|

Trinity tip

The examiner is **not** testing your opinion. He or she is testing how well you can **express** your opinion, whatever it might be.

Which test taker is clearer about his or her opinion? _____

Which test taker actually answers the question? _____

Activity 2

Read these phrases aloud. Then write A (agreement), D (disagreement) or U (uncertainty)

- | | | | |
|--------------------------------------|-------|--|-------|
| a) I disagree | _____ | g) I can't agree | _____ |
| b) I agree | _____ | h) I'm sorry but I can't agree with that | _____ |
| c) I think that's right | _____ | i) I'm not sure | _____ |
| d) I don't agree | _____ | j) I completely disagree with this | _____ |
| e) I really don't know what to think | _____ | k) Absolutely! | _____ |
| f) I just can't decide | _____ | l) That seems wrong to me | _____ |

Teacher Notes

Successful speakers... are clear about their opinions

These teacher notes give you guidance on how you could manage the activities with your class. We encourage you to consider follow-up activities where students can reflect on their language production and to create checklists for them to consider their own performance. You may want to offer extra support on some activities, eg by doing an example with the whole class first and/or having an observation checklist for when they listen to each other. You may also prefer to present the activities in other ways.

This lesson is about helping students to state their opinions clearly, whether they agree, disagree or are ambivalent about the topic.

When we are clear about what we mean, conversations flow more easily and, in an exam situation, the examiner needs to check the test taker's meaning less often.

Feedback

Remember to give students feedback on how they are performing throughout. You may choose to cover different aspects of speaking in single parts, eg activity 1 = pronunciation, activity 2 = fluency, activity 3 = accuracy etc, then give whole feedback on the final activity.

Conclusion

- The results shown today would never have been produced without AI helping researchers to process the data
- Yet without the data, the AI would have had nothing to work on
- AI can be a useful tool in allowing us unprecedented insight into previously rarely studied areas of research
- Yet AI is not a magic wand – we can use it to gain insights, but human experts need to identify the question, match the technique to the question and interpret the findings.
- Human experts also need to determine when and how to develop the right datasets to get the insights
- There is much more to show and say so read the book! 😊