The effects of interlocutor backchannels and L1 backchannel norms on the speech of L2 English learners

Alexander Flint

Thesis submitted to the University of Oxford for the degree of Doctor of Philosophy

St Hugh's College
Trinity Term 2016

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACTFL	American Council on the Teaching of Foreign Languages
ANOVA	Analysis of variance
AS–unit	Analysis of speech unit
BBC	British Broadcasting Corporation
BCs	Backchannels
CAF	Complexity, accuracy and fluency
CD	Compact disc
CI	Confidence intervals
CLperAS	Clauses per AS–unit
c–unit	Communication unit
DART	DART-Europe E-theses Portal
dB	Decibels
E-freeAS	Error-free AS–units
E-freeCL	Error-free clauses
ERIC	Educational Resources Information Center
F	Female
FBI	Federal Bureau of Investigation
HS	High school
Hz	Hertz
IBM	International Business Machines Corporation
IELTS	International English Language Testing System
L1	First language
L2	Second language
LLBA	Linguistics and Language Behavior Abstracts
M	Male
MANOVA	Multivariate analysis of variance
MLA Intl.	Modern Language Association International Bibliography
MLR	Mean length of run
msec(s)	Millisecond(s)
NHST	Null hypothesis statistical testing
NPTR	Non-phonation time ratio
PCM	Pulse-code modulation
ProQuest D&T	ProQuest Dissertations and Theses
PSR	Pruned speech rate
SD	Standard deviation
SEM	Standard Error of Measurement
TOEFL	Test of English as a Foreign Language
TOEIC	Test of English for International Communication
UK	United Kingdom
Uni	University
US	United States of America
USA	United States of America
WAV	Waveform audio file format
WperAS	Words per AS–unit
WperCL	Words per clause

ABSTRACT

Verbal backchannels – short responses such as 'uh-huh' and 'mhm' given by an interlocutor to the main speaker – have been studied extensively for several decades. The great majority of the research has been descriptive or based on backchannel uses. In contrast, little has been reported of their effects on spoken interaction and almost no research has examined their effects on second language (L2) speech. Given that first language (L1) backchannel norms vary, L2 speakers unaccustomed to different norms could be affected when exposed to such variation. This thesis investigated such effects through the use a quasi-experimental repeated measures design that compared the effects of two backchannel frequencies – one approximately a third of the other – on L2 English speech. The 37 L1 Japanese and 34 L1 Mandarin Chinese participants spoke in English to an interlocutor who varied the frequency of backchannels that they were given in different dyadic interactions. The resultant audio recordings were transcribed and analysed using common measures of speech complexity, accuracy and fluency. Multivariate analyses of variance and *t*-tests helped show that the fluency of each group was increased when the higher of the two frequencies was given and that, while the accuracy of the Japanese group did not alter, the Chinese group was less accurate in one set of interactions when receiving the higher frequency of backchannels. Effect sizes for these changes (d = 0.19–0.87) were comparable with other studies that used the same measures of fluency and accuracy. There were no statistically significant differences for measures of complexity. The findings show that the contribution of L1 norms to the effects of backchannels on L2 interactions is not as clear-cut as assumed by previous research. The implications of the findings extend into language testing, teaching, theory and research methods.

1 INTRODUCTION

1.1 Research Topic

In 1983, a Japanese newspaper reported on a corporate espionage trial that was taking place in the United States and involved a Japanese businessman accused of agreeing to buy trade secrets that would have to be stolen:

> In the trial, the prosecution argued that defendant Takaya Ishida of Mitsubishi had known the IBM information he obtained was stolen because when the FBI undercover agents earlier informed him that there was no choice but to steal it, he still wanted to get the information so he said 'yeah' and 'uh-huh' several times.
>
> The defense counsel, however, claimed that Ishida had had no knowledge of agreeing, saying Ishida used 'yeah' or 'uh-huh' just to mean he was listening.
>
> To support their argument, the defense council turned to [a] Japanologist […who] said that it is considered to be polite in Japan to constantly let others know that you are listening to them in conversations (Japan Times, 1983a).

Assuming that the defence's argument was a true account of events,[1] they were stating that the Japanese businessman's frequent use of 'yeah' and 'uh-huh' when listening was merely the second language (L2) manifestation of what was typical backchannel support in his first language (L1).

The difficulty of using backchannels appropriately is demonstrated by the fact that their usage develops slowly in an L1, "as part of children's pragmatic communicative skills and language development" (Heinz, 2003: 1117). Hess and Johnston found "a threefold increase in the frequency of back channel responses to adequate [readily understandable]

[1] This trial is returned to in the Discussion, where the outcome is also revealed.

messages between the years of 7 and 11" in children with L1 English (1988: 328). They suggested that "Back channel listener responses could be among the last conversational skills acquired" (1988: 319).

Although they may be developed late, backchannels are a very common feature of spoken interactions and appear to be used in all languages (Heinz, 2003: 1114; Rost, 2011: 92). As they are a fundamental part of conversation, which is a basic feature of human life, it is perhaps unsurprising that backchannels have been the subject of research in a diverse range of fields. The two main fields are linguistics and applied linguistics. Others include psychology, team dynamics, mental health, medicine and computer science. In team dynamics, backchannels have been researched in cohesion in design groups (e.g., Casakin et al., 2015). In mental health, they have been studied in psychiatry (e.g., DiNardo, Schober and Stuart, 2005) and counselling (e.g., Sharpley et al., 2000). In the field of medicine, research has included face-to-face doctor-patient interactions (e.g., Hall et al., 1994) and teleconsultation (e.g., Tachakra and Rajani, 2002). Research in computer science is a newer addition; one of its aims has been to create software that increases the naturalness of spoken human-machine interactions, from the perspective of the biological participants (e.g., Bevacqua et al., 2012).

Research in linguistics and applied linguistics has provided empirical evidence for what the Japan Times (1983a) story indicated: the details of backchannel behaviour in different languages vary (Heinz, 2003: 1114; Rost, 2011: 92). Generalising from the studies that have been done of L1 backchannels, there are language-specific norms and, because of these, there are expectations regarding an interlocutor's backchannel behaviour. Differences between these expectations of backchannel behaviour and what actually

occurs could affect an ongoing spoken interaction. Given their L1-specific nature, the effects of backchannels received not matching what a speaker expects are most likely to manifest themselves during interactions in which one or more of the people involved is speaking an L2.

Cutrone recently commented that "Research into the area of backchannel behavior, and particularly its effect on intercultural communication […], is in its infancy" (2014: 89). There is, however, already some understanding of backchannel differences. The impact of differences between L1-based backchannel expectations and reality has been described for those speaking their first language in L1-L2 interactions: Cutrone (2014) directly asked L1 English speakers about the backchannels that they had received when conversing in English with Japanese people; answers included assertions that the backchannels were too frequent, sometimes unnatural and often unvaried.

Some people are also able to identify changes in the backchannel behaviour of their interlocutor. One participant in a small-scale study (Flint, 2010) of the effects of verbal backchannels on L2 English fluency commented that:

> The first one [speaking task], you said 'mhm' or something like that, so it was very easy for me to talk, but the other one, you […didn't,] so suddenly I got nervous.

Another said that:

> I did feel comfortable when you […] gave me […] 'uh-huh' or whatever, but I don't think it made any difference to what I speak in terms of content or maybe the accuracy or the fluency.

These comments were unprompted and rare: the great majority of participants in that study did not mention noticing any differences in backchannels received, even when asked directly, and even though the difference had been between a high quantity of backchannels and none at all. Nevertheless, the comments demonstrate that some L2 speakers are consciously aware of backchannel differences and, importantly, that some also believe that those differences can have negative effects.

There is evidence, then, of miscommunication and communication breakdown as a result of backchannel differences, as well as frustration with those differences. The increasing prevalence worldwide of non-L1 interactions, especially in English, makes this an area of obvious importance for further research. There are several areas (besides mounting a legal defence in a foreign country) in which an improved understanding of the practical effects of backchannels on L1-L2 or lingua franca communication could be of benefit. One broad example is that the attitudes of L2 users towards the language and its speakers are associated with intentions to continue learning that language (R. C. Gardner, 1985), so greater awareness of backchannel differences and their possible effects could benefit both learners and teachers. One narrower example is language testing. For instance, IELTS is a high-stakes test that is taken by more than two million people a year (British Council, n.d.) and contains a speaking component in which just the candidate and examiner are present, while some Cambridge English exams have speaking components that require a candidate to interact with both an examiner and another candidate. Unfamiliarity with backchannel norms in different languages could potentially affect candidate performance in such tests.

The research topic of this thesis, arising from this brief introduction, is the effects of backchannels on L2 English speech. A lot more background, as well as the exact direction

of the research, is covered over the course of the next two chapters of literature review. Before embarking on this, the overall structure of the thesis will be outlined.

1.2 Thesis Overview

The structure of this thesis is largely conventional: Literature Review; Methods; Results; and Discussion and Conclusion. The combination of an abundance of literature on (some aspects of) backchannels and a paucity on how to investigate and measure their effects on speech means that each of the first two of these parts has been split into two, giving: Literature Review I – Backchannels; Literature Review II – Methodology; Methods – Data Collection; Methods – Data Analysis; Results; and Discussion and Conclusion.

Much of the considerable literature on backchannels is of attempts to categorise and describe based on their uses and has employed conversation analysis approaches. Literature Review I – Backchannels therefore summarises only those parts of this section of the literature that are germane to the current research. More detail is provided of the considerably smaller literature on backchannel effects. The theoretical framework of a model of speech is then introduced, to be referred to at other stages of the thesis. Backchannel features are then described, including through the use of an earlier systematic review of their frequency (Flint, 2012). The identifying of gaps in the literature builds towards the selection of the effects of verbal backchannel frequency on L2 English speech as the research focus and to the statement of the two general research questions.

Literature Review II – Methodology also builds towards a set of research questions; this time they are the two specific research questions. This is done in the first half by

considering and choosing a research design, interaction situation, participant L1s and other details of data collection. The second half of the chapter is on tools of data analysis. The decisions made feed in to the contents of the two methods chapters that follow.

The first of these is Methods – Data Collection. It details the whole data collection process, including two pilot studies, the main study itself and the checking of the methods. It describes, therefore, a quasi-experimental repeated measures design that used L1 Japanese and L1 Mandarin Chinese participants speaking English while receiving backchannels that varied in frequency across separate interactions. The next chapter is Methods – Data Analysis. It begins with transcription, continues with the chosen measures of complexity, accuracy and fluency (CAF), and ends with methods of statistical analysis and reliability checks. This chapter contains considerable detail, particularly on the calculating of CAF, as this is invariably missing in published studies.

The penultimate chapter is Results. This presents descriptive and inferential statistics for the CAF of the two L1 groups. The inferential statistics principally compare CAF within each L1 group when receiving the different backchannel frequencies. In keeping with recent trends in the field, 95% confidence intervals and Cohen's *d* effect sizes are reported, in addition to the more traditional *p* values and related numbers. The Discussion and Conclusion chapter continues from this by considering the results in relation to both the general and specific research questions. It goes on to consider implications, which are divided into those for people involved in the learning, teaching and testing of languages, and those that apply to theories of, and research into, speech itself. The final sections are future research directions and the conclusion.

2 LITERATURE REVIEW I – BACKCHANNELS

This is the first of two literature review chapters and seeks to support and build detail around the research topic – the effects of backchannels on L2 English speech. The literature that is wholly or partly on backchannels is voluminous, diverse and contains little agreement on terminology, categorisation or criteria for inclusion into any given category. An in-depth detailing of the literature would be sprawling and packed with dead ends, so the first three sections of this chapter both summarise and cut through the mélange with the purpose of setting the scene for the current research. 'Backchannels' is selected as the sole term to be used in this thesis, verbal forms are chosen as the most suitable focus of study and the nature and range of existing research into backchannel uses and effects are summarised. This is followed by the selection of an appropriate theoretical framework for the topic and for the research that will be done.

The remainder of the chapter concentrates on backchannel features, again with the goal of identifying what is best suited to addressing the research topic. Salient findings of an earlier systematic review of the literature on backchannel frequency (as measured by such things as the number of main speaker words per backchannel received) are presented, including the possible influence of factors such as the interaction situation and the age or gender of the speakers. The identifying of gaps in the literature leads to a narrowing of the research focus to the effects of backchannel frequency on L2 English speech, and the importance of comparing these across L1s, which then allows the general research questions to be stated.

2.1 Backchannel Terminology and Early Research

There are two people in archetypal spoken interaction; at a given point, one is doing most of the speaking and the other is mostly listening. These can be labelled as the 'main speaker' and the 'interlocutor', respectively (these labels will be used in the remainder of this thesis). Backchannels are short responses that are given by the interlocutor and that do not answer a question, interrupt or lead to a specific reaction from the main speaker.

Backchannels and their uses were mentioned in publications before they came to be labelled and studied more systematically. Covner, for instance, described counsellor-client interviews in which counsellor participation "would be limited to a single word such as 'Yes', or 'M-hm', which told the client that he was being listened to and should go on with his story" (1944: 197). Fries, analysing recordings of telephone conversations, is credited with putting backchannels together into a group: "the hearer, in some inconspicuous but conventional way, gives the speaker signals of his continued attention" (1952: 49), but the signals "do not interfere with the continuous flow of the utterances of the speaker" (1952: 50). Little research was then done on them for almost two decades. They were eventually incorporated into a description of spoken communication by Yngve, who contrasted the main speaker's channel of speech with the interlocutor's "back channel activity" (1970: 574). Yngve's neologism came to be used as a verb and, more commonly, as a label for the responses themselves.

Comparisons of backchannel use in different languages have been reported since the 1970s. Many of the early reports were largely anecdotal and could be regarded as Anglocentric cultural stereotypes. These included excessive backchannels from Japanese people (Lebra,

1976) and extended interlocutor silence among Finns (Lehtonen and Sajavaara, 1985). Subsequent research that is included in the remainder of this chapter was more empirical and encompassed backchannel use in L2 as well as in L1 interactions.

With the expansion of research came a proliferation of terminology. Fujimoto's (2007: 38) list of terms numbered 24, including, in alphabetical order, 'minimal feedback', 'minimal response', 'minimal response token', 'minimal token' and 'minimal vocalization'. Some of this was the product of investigating different categories and uses of backchannels, but the abundance of terms and variation in how they have been employed tend to obfuscate rather than enlighten. In the interests of clarity, only 'backchannels' will be used in this thesis, except in the next section, where some of the categories employed in the literature are described and the most appropriate ones for this study are selected.

2.2 Backchannel Categories and Uses

Backchannels can be split into verbal and non-verbal forms. Here, 'verbal' will refer to backchannels that belong to oral language production, irrespective of whether or not they consist of what is usually considered to be a lexical item. This does exclude coughing, laughter, sniffing and other such sounds.

2.2.1 Non-Verbal Backchannels

Verbal and non-verbal forms of backchannels began to attract more researcher interest at around the same time. However, limitations in technology caused practical difficulties in

recording and analysing non-verbal backchannels until comparatively recently and they were often studied as part of more general body language.

Much of the analysis of non-verbal backchannels has been of the most readily noticeable type – nods. Other actions that have been described as backchannels include shoulder shrugs (Dittmann and Llewellyn, 1968), smiles (Brunner, 1979) and gaze (Goodwin, 1981). Backchannels have also been studied in various sign languages, including tactile ones (Mesch, 2013).

Compared with verbal backchannels, which themselves are imperfectly defined and understood, current knowledge and understanding of non-verbal forms is limited. There have been a lot of descriptive studies, but considerably fewer on what non-verbal backchannels do: more basic research into the use, frequency and interdependence of different types of them is required before their effects can be studied.

An additional reason for preferring verbal backchannels as a focus of enquiry is that considerable practical difficulties remain for the study of non-verbal forms and those difficulties can undermine validity. For research that seeks to describe them, current video technology is likely to be adequate, but for investigations of effects, it is necessary to be confident not just that a backchannel has been given, but also that the main speaker has received it. This is very likely to be the case with verbal backchannels, because they will probably be undetectable only if they co-occur with a much louder utterance of the main speaker, which is rare. In contrast, non-verbal backchannels will not necessarily be seen and there is no way of knowing for sure which ones have been seen (using a video camera, with the same field of view as human eyes, built in to glasses worn by participants could

be a future and possibly necessary tool). For the reasons given in this section, then, verbal backchannels were chosen to be used in this research.

2.2.2 Verbal Backchannels

The basic description of backchannels given in Section 2.1 was that they are "short responses that are given by the interlocutor and that do not answer a question, interrupt or lead to a specific reaction from the main speaker". This excludes, for the purposes of this study, several things that some have claimed to be backchannels. These were influenced by Yngve's already mentioned distinction between the main channel and the back channel, as the examples of backchannels that he gave included "You've started writing it then – your dissertation?" and a 30-second aside that involved the giving of background information so that the main speaker could continue (1970: 574). Suggestions for inclusion as backchannels because of this early account include 'restatements' (briefly paraphrasing the main speaker), 'sentence completions' (finishing what the main speaker was saying) and 'clarification requests', all of which were suggested by Duncan (1974: 166–167). Clancy et al. (1996) were influential in analysing backchannels and other interlocutor behaviour in three languages, but also added to the confusion over nomenclature by renaming 'sentence completions' as 'collaborative finishes' and using 'repetitions' to describe what are another form of 'restatement'. For the reasons just given, none of these will be treated as backchannels in this thesis.

Schegloff (1982) introduced an important facet for conversation analysis-based study by describing some distinct uses of the short responses that previously had been assumed to have the same use. He categorised the archetypal backchannel such as 'uh-huh' or 'yeah' as

a 'continuer', which shows "on the part of its producer an understanding that an extended unit of talk is underway by another, and that it is not yet, or may not yet be […] complete" (1982: 81). He went on to introduce 'markers of surprise' such as 'really?' and 'assessments' such as 'wow' that put forward a judgement (1982: 85). This set the scene for further research, which predominantly started with transcripts of naturally occurring speech and employed conversation analysis as its investigative tool. Goodwin, for instance, went into further detail regarding 'continuers' and 'assessments', suggesting that the former are responses to the general fact that the main speaker has not finished, while the latter "deal with the specifics of the talk in progress as phenomena in their own right rather than as a prelude to further talk" (1986: 214).

There are innumerable further examples of research into the nuances of various forms of interlocutor response; just one more will suffice here. Beach (1993) studied the use of 'ok' and found that it could be used as a 'continuer', as an indicator of a possible topic transition, as an indicator of a possible change in speaker roles, and in several other ways. The key point is that a word or sound itself is not necessarily a backchannel (of course, 'ok', 'yeah', 'uh-huh' and so on can also answer a question), so its categorisation is based on the use to which it is put.

Uses that are mentioned in the literature, in addition to those already stated here, include the signifying of agreement (R. Gardner, 1998), involvement (Stenström, 1994), surprise (R. Gardner, 2001) or understanding (Tottie, 1991). The literature, or at least that published in English, has predominantly been on backchannels used by English speakers, but parts of it include uses in other languages, mostly Japanese. This has expanded on, and added detail to, the uses that have been described for backchannels among English

speakers. Their use in showing understanding may, in Japanese, go beyond that by also conveying "a sort of moral support for the speaker" (Maynard, 1997: 46). More generally, in Japanese, the use of backchannels can incorporate an empathetic and encouraging element, demonstrating "willingness to co-operate in the conversation and to show support of and attach value to the speaker" (LoCastro, 1987: 110).[2]

Distinguishing among some of the multitude of claimed uses can be very difficult, as Fujimoto (2007) and others point out. There are overlaps in uses that mean that more than one could be employed by one instance of a backchannel, and an identical backchannel employed in different contexts could have a different use. Drummond and Hopper provide an example of 'yeah', below, "which could be encouragement for D to continue, or it could be said in agreement with D's statement of her feelings" (1993: 207; the transcription has been simplified):

D: Death offends me it really bothers me
M: Yeah.
D: And I guess it's afraid of-

A further complication is one rarely mentioned: backchannels can be employed to pretend to show understanding, attention and so on (Cutrone, 2014). For the purposes of this research, all of these complications can be put to one side. 'Backchannels' will henceforth refer to the originally described, archetypal responses that are, or appear to be, used to show that the interlocutor is listening, that the main speaker need not stop, and which do not offer an assessment of the specific content of what the main speaker has said.

[2] Japanese has a non-specialist word – 'aizuchi' – for listener responses; Chinese and English do not (Deng, 2009: 115). Although this word has been employed in some of the literature on backchannels in English, it will not be used here.

Most of the research on backchannels has been descriptive (Cutrone, 2014: 89). Although, as just discussed, this has encompassed a wide range of *uses*, very little research has been done on the *effects* that backchannels can have on ongoing interaction. This is turned to next.

2.3 Backchannel Effects

Some research into the effects of backchannels has been based on having a third party rate a dyadic conversation that has been manipulated in some way. Bennett and Jarvis (1991) asked half of their participants about the interaction between two speakers in a recorded conversation; the other half did the same thing, but the recording that they listened to had had all of the backchannels removed. The participants completed a questionnaire to rate eight facets of the conversation: formality; conversational flow; interest shown; politeness; level of agreement; closeness of listening; how well the speakers knew each other; and how much the speakers liked each other. Only ratings of formality and level of agreement differed, with backchannels being regarded as lowering formality and raising the level of agreement perceived.

de Kok and Heylen (2011) used a more deceptive data collection process. They set up a live video feed of a main speaker to three separate would-be interlocutors and told each of the three that they could use backchannels and short responses while listening, but that they were not to ask questions. Each of the 'interlocutors' believed that the main speaker could see and hear him or her, but the main speaker received a video feed of only one of the three. All of the participants then completed a questionnaire on satisfaction with the conversations and, after being debriefed, were asked if they had noticed the deception;

their identification of when they were able to be seen and heard by the main speaker was no better than chance.

An attempt at combining conversation analysis with more experimental approaches for narratives was reported by Tolins and Fox Tree (2014). They first claimed that backchannels such as 'yeah' and 'uh-huh' lead to the main speaker continuing with a new event in a narrative, whereas more specific responses such as 'oh my gosh' lead to the giving of more information about what has already been mentioned. In the more experimental part, they showed partial transcripts of the same narratives to participants, who had to write down a possible continuation of the transcript. The required continuation was immediately after a backchannel or a more specific response. Half of the transcripts used exactly what had been said in the narratives, while the other half had had either the final backchannel swapped for a specific response, or the final specific response swapped for a backchannel. The main finding was that participants were more likely to write in a new narrative event after a backchannel than after a specific response, in keeping with the researchers' initial claim.

These three studies, and other similar ones, are of some interest, but they are concerned chiefly with the impressions of eavesdroppers and others who are exposed to speech offline, without participating in it directly. Of greater interest here is the effect of backchannels as interaction is ongoing.

Kraut, Lewis and Swezey (1982) investigated differing levels of interlocutor response on the main speaker's summary of a film. Their participants were in separate rooms and communicated via audio link. They found that restricting the responses led to the

interlocutors having a poorer understanding of the film. Their research, however, included one experimental condition in which questions could be asked and backchannels could be given, and the effect of giving only backchannels on the interlocutors' understanding is unclear.

Sannomiya et al. studied the effects of varying Japanese backchannel frequency on the number of ideas that a main speaker had when asked a question such as "What will happen if the number of aged people increases even more in Japan?" (2003: 43). The interlocutor was required either "to make backchannel utterances as frequently as possible" or "to be as silent as possible" while the ideas were being spoken (2003: 44). The number of ideas generated was higher when more backchannels were given, a finding they attributed to backchannels acting to "stimulate idea-generation itself at the cognitive and motivational level" (2003: 46).

Backchannel effects on interactions between human telephone users and computer-based call-handling systems have also been suggested. Edlund et al. (2008) and Gustafson, Heldner and Edlund (2008) commented on software that is designed to elicit information from people so that their telephone call can be routed to the most appropriate customer service department of a company. They suggest that a system that gives backchannels may cause the human callers to speak in a less command-like manner and possibly lead to the giving of more detail, allowing the system to identify the topic of their call. However, as with the Sannomiya et al. (2003) study, this effect is based on a dichotomous comparison of backchannels being given versus none being given.

The effects of backchannels have also been studied to a very limited extent in L2 interactions. Wolf (2008) followed up a series of L1 English experiments on counsellor behaviour (e.g., Kanfer and McBrearty, 1962; Siegman, 1976) that had produced contradictory results when investigating how long a main speaker spoke for when the interlocutor remained silent rather than showing interest and giving backchannels. Wolf's (2008) Japanese participants created an L2 English narrative based on a series of six cartoon drawings while their interlocutor gave backchannels that varied in number and type. He found that giving no backchannels led to lower fluency, as measured by things such as syllables per minute; no other facets of speech were reported on.

In summary, backchannel effects as opposed to uses have been studied very little. Only some of the research to date has been on speech that was ongoing and most of that utilised a crude comparison between a speaker receiving a high frequency of backchannels and receiving none at all. The near-total absence of research on L2 speech is also notable and represents a considerable gap in the literature.

Before moving on to describe the features of backchannels in Section 2.5, a further relation of backchannels to theory is required. This is a consideration of what form of theoretical framework for backchannels is appropriate for this study.

2.4 Theoretical Framework

There is no theoretical framework into which backchannels readily fit, but, as already discussed, there have been numerous attempts at classifying backchannels themselves. These attempts amount to taxonomies. Such taxonomies can be valuable in research that

has as its goal the description of language, and they have been incorporated into descriptions of turn-taking systems, for example. It would be possible, therefore, to present backchannels in relation to turn-taking or models of social interaction.

The first step, however, in identifying an appropriate theoretical framework for the purposes of this research is to consider what kind of research it will be. The type of research emerges from the research topic – the *effects* of backchannels on L2 English speech. Pre-empting the following chapter, which is on methodology, a study of effects requires an experimental (or quasi-experimental) design. Such a design differs fundamentally from those that have informed the production of taxonomies, accounts of turn-taking systems and models of social interaction, as they are based principally on descriptive analyses of naturally occurring conversation, as mentioned in Section 2.2.2. Employing an experimental design to examine the effects of backchannels, therefore, means that a taxonomy linked perhaps to a model of communication that is based on descriptive research is not an apposite framework for this study.

An additional step in identifying an appropriate framework is to consider the focus of the research topic. The focus of this experimental study is on the main speaker, rather than on the interlocutor. A framework that is both relevant to experimental research and centred on speech production itself is, therefore, justified. For more than two decades, a major model of speech production has been that of Levelt (1989, originally). Levelt's model was informed by existing research in a wide range of fields – he lists, in addition to the psycholinguistics literature, "conversational analysis, pragmatics, discourse semantics, artificial intelligence, syntax, phonology, speech communication, and phonetics" (1989: xiii) – but it was created to provide an account of the mental processing that occurs in a

speaker (1989: xiv). His model is for L1 production, but it has also been widely utilised in the L2 literature. It is appropriate for this research, then, as it is a model of speech production that is relevant to the experimental design that will be employed. Its main features will be described next.

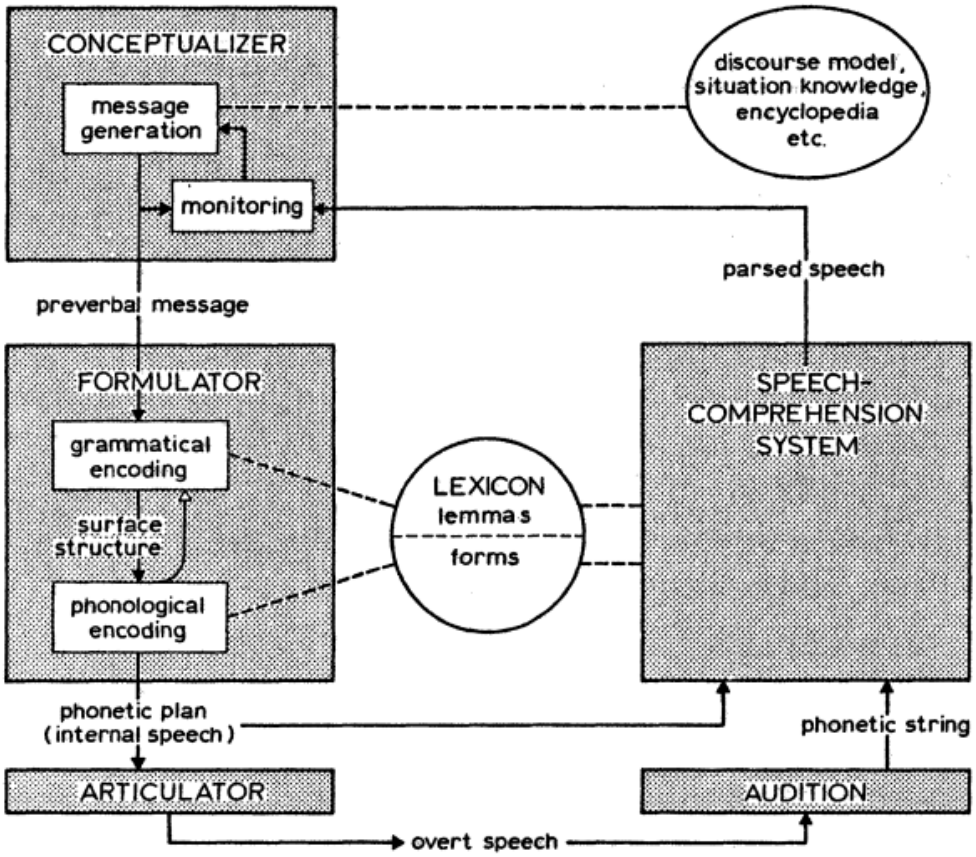

Figure 2.1 Levelt's "blueprint for the speaker" (1989: 9)

The main features of Levelt's model (reproduced in Figure 2.1)[3] are that three stages contribute: conceptualisation, formulation and articulation. Concepts are created, leading to a preverbal message being passed on for formulating through encoding for grammar, morphology and phonology, and phonetic content. The final articulatory score can be

[3] The account given here also incorporates aspects presented in Levelt (1999).

passed on for articulation. Feedback loops form part of self-monitoring and go to the initial stage of the system. One compares the preverbal message with the original conceptualisation prior to its being sent to the formulator; if the speaker finds, for example, that the message is inappropriate in content or for the situation, then it may be rejected. Another loop monitors the phonetic plan, for erroneously selected words, for instance. The third feedback loop involves checking overt speech after articulation.

Describing how backchannels could fit into this model of speech is speculative, but can be attempted, including by interpreting, in relation to the model, findings from some of the backchannel literature described earlier in this chapter. The findings of Sannomiya et al. (2003), Edlund et al. (2008) and Gustafson, Heldner and Edlund (2008), summarised in Section 2.3, indicate that backchannels might act to encourage an increase in the quantity of ideas expressed by a speaker, and expansion on an idea already partly expressed. This would occur at the conceptualisation stage, which, being the first stage in the model, is also the first point at which backchannels could act. Information that is available to the speaker and feeds in to the conceptualisation stage includes the "discourse record" of what the speaker and any interlocutor have said during an interaction (Levelt, 1989:10). This record is placed, in Levelt's (1989:10) account, with knowledge of the discourse situation, accumulated world knowledge, and other factors, in the category of "declarative knowledge", which is encircled in the top right corner of the schematic representation portrayed in Figure 2.1.

Conceivably, backchannels could be part of this discourse record. Their acting in this way would require a mechanism by which the ongoing backchannels being received could be processed while speaking. One potentially relevant piece of information is that some

people can comprehend two sets of aural inputs simultaneously (e.g., Buchweitz et al., 2012). This observation can be considered in combination with Levelt's (2001: 246) possibly rhetorical question – "Can you parse your interlocutor's utterance while you are speaking yourself?" – to conclude that, if the two aural inputs comprehended simultaneously can be from self and other, then interlocutor responses could be processed while speaking, at least by people with that ability. The nature of typical backchannels – their being very short and often non-lexical – is conducive to their being straightforward for the main speaker to process, so the idea that backchannels might act as feedback on overt speech, to the conceptualisation stage, is possible.

The extent to which the three stages of the speech production model operate in parallel is disputed, particularly for L2 speech. For proficient L2 speakers, formulation can be mostly automatic, permitting it to operate in parallel with the other stages, while conceptualisation and self-monitoring are more likely to depend on serial processing (Lambert, Kormos and Minn, 2016). Aspects of formulation, including the retrieval of lexical items and grammatical encoding, are likely to require attention in lower proficiency L2 speakers, meaning that the stages probably operate serially for them (Kormos, 2006; Lambert, Kormos and Minn, 2016). Greater serial operation and lower automatisation raises the possibility of backchannels acting at least in part at the formulation stage in such speakers. This might occur via backchannels influencing the second feedback loop, which monitors the phonetic plan before its articulation. In contrast, given that "the articulatory plan is relatively independent of context" (Levelt, 1989: 13), the articulation stage itself is unlikely to be directly affected by backchannels.

Positioning backchannels within the theoretical framework of Levelt's (1989; 1999) model of speech production has permitted an (albeit speculative) exposition of how they could affect speakers. This will help inform the choice of data analysis tools in this study, which is the topic of Section 3.2.1, as well as permit the study's findings to be more readily related to theory. Next, however, is a return to backchannels in themselves – their features.

2.5 Backchannel Features

The placement, sequencing, prosody and frequency of backchannels have been described in the literature to varying extents. Each of these is summarised in this section, again with the intention of identifying what is best suited to the research topic – the effects of backchannels on L2 English speech.

2.5.1 Placement

There are both similarities and differences in backchannel placement across languages. Backchannels in English typically occur at grammatical boundaries: at or after clause endings, for example (Oreström, 1983). Points of grammatical completion are also the most common backchannel location in Mandarin Chinese (Clancy et al., 1996) and Thai (Wannaruk, 1997): this is either the core location or one of the core locations in which backchannels occur.

It is also a core location for German speakers, who are likely to wait until a clause has ended and a pause begun before giving a backchannel. This is shown by the placement of 'man' in Heinz's (2003: 1123) example:

Egon: *Eigentlich hatte ich vierundzwanzig Stunden Dauerdienst*.
[Actually I was on continuous duty for 24 hours]
Und (**short pause**)
[And]
Ernst: *Mann*.
[Man]
Egon: *Das war echt stressig*.
[That was really stressful]

In Japanese, backchannels also occur at grammatical boundaries, but are "much more likely to occur 'midstream' during the primary speaker's talk, while s/he is still in the process of constructing a grammatical clause" (Clancy et al., 1996: 375). This can be seen in the following extract from an example given by Fujii (2008: 350–351; for clarity, the transcription has been simplified and combined with the translation here):

Masae: *e watashi wa*
[well I]
Eriko: *un*
[uh huh]
Masae: *baito shitete*
[was working part-time]
Eriko: *un*
[uh huh]
Masae: *de nijuuyokka ni*
[and on the twenty-fourth]
Eriko: *un*
[uh huh]
Masae: *honto wa kurisumasu dakara*
[because it was Christmas]
Eriko: *un*
[uh huh]
Masae: *kareshi n toko ni ikoo to omotta n dakedo*
[I thought I'd go to my boyfriend's place]
Eriko: *un*
[uh huh]

The perceived naturalness of backchannel placement in English was investigated by Poppe, Truong and Heylen (2011). They showed participants a screen that displayed a human main speaker next to a computer-generated animated interlocutor. The backchannel

behaviour of the main speaker's original human interlocutor was copied to the animated interlocutor for one experiment and altered in various ways for other experiments. The participants were asked to rate the naturalness of the backchannels. The findings were that backchannels artificially recreated in random positions were much more likely to be rated as unnatural, but that backchannels actually given by the human interlocutor were also rated as unnatural in an average of 15% of occurrences.

The importance of placement was also demonstrated by an early study of video conferencing technology by O'Conaill, Whittaker and Wilbur (1993). They compared interaction in face-to-face business meetings with similar meetings conducted over high bandwidth video and over low bandwidth video. They found that the frequency of backchannels was greatly reduced in the low bandwidth meetings, because the time delay between a backchannel being given and it being received meant that the main speakers hesitated or stopped speaking, as they believed that the backchannel was the start of an interruption. The participants in the low bandwidth meetings soon realised this problem and so almost eliminated their use of backchannels.

There has also been computer science-based research into backchannels being cued by prosodic features of the utterances of the main speaker. Using a corpus of spoken English, Gravano and Hirschberg (2011) identified five such features. After restricting their analysis to periods of speech containing a silence of more than 50 msecs and adding a part of speech indicator of grammatical boundaries, they found that, if all the features co-occurred, there was a 30% chance that a backchannel would follow (2011: 608, 631). The relationship between these proposed cues and a backchannel being given was quadratic, so the chance of a backchannel being given rapidly declined when the number of co-

occurring proposed cues fell: it was approximately 10% when two were not present (2011: 624). Similar attempts have been made for other languages, including forms of Arabic (Ward and Al Bayyari, 2010) and Spanish (Rivera and Ward, 2006), but with no greater predictive success. Research continues in this area: Dethlefs et al. recently added information density – "the distribution of information [content, rather than function words] across an utterance" – to the list of features of speech that may be associated with the positioning of backchannels (2016: 87), but it is also clear that a lot remains to be done to improve understanding of possible backchannel cues.

In short, backchannel placement is subject to a high degree of variation even in one language and an interlocutor has only the option, but not an obligation, to give a backchannel naturally in certain places. This is an area that requires more research to improve the naturalness of human-machine spoken interaction, but is not the best choice to investigate the effects of backchannels on L2 English speech.

2.5.2 Sequencing

There are two facets to the sequencing of backchannels. The first is what is said during a single backchannel occurrence. Tottie (1991) created a sub-classification of backchannels: simple (e.g., 'mhm'); double (repetition of the backchannel; e.g., 'mhm mhm'); and complex (more than one example of backchannels together; e.g., 'mhm yeah'). Cutrone subsequently suggested that combinations such as double and complex forms might indicate greater interlocutor involvement than those in the 'simple' category (2014: 86). This was based on Stubbe's (1998: 259) proposal for a "verbal feedback continuum", part of which included a gradation from low involvement, neutral affect to high involvement,

positive affect. Cutrone's (2014) suggestion could be accurate, but it is based more on supposition than on empirical evidence.

The second sequencing facet concerns the selection of backchannels over a period of talk. Schegloff (1982: 85) summarised the difference that can be made by varying the choice of backchannel used ("token"): "Use in four or five consecutive slots of the same token may then be used to hint incipient disinterest, while varying the tokens across the series, whatever tokens are employed, may mark a baseline of interest". He was writing about English, but this principle is likely to apply to other languages. In brief, however, sequencing is a relatively minor backchannel feature.

2.5.3 Prosody

A small quantity of research has gone beyond the examination of backchannels by category to look at some of their prosodic features. Müller (1996) used Italian radio phone-in recordings to study variation in several prosodic elements of backchannels, including intonation contours, loudness and length. R. Gardner also went into great detail about such things in English, contrasting, for instance, "a marked rise-falling tone or high pitch […indicating] encouragement or appreciation" with indifference shown by a backchannel that is "low and level in tone" (2001: 13). In summary, the core uses of a backchannel are associated with specific prosodic features, with variations in those features potentially altering the role that a backchannel plays (R. Gardner, 1998).

Prosody can, then, play an important part, although canonical prosodic features are similar for the most common uses of backchannels (R. Gardner, 1998). The final backchannel

feature to be discussed in this section is likely to be more pertinent and marked for L2 speakers. This is frequency, which was discussed in Section 2.3 as a feature of studies into backchannel effects.

2.5.4 Frequency

Backchannel frequency (or backchannel rate – the two terms will be used interchangeably) refers to the number of backchannels used in relation to some other unit. The other unit is usually either main speaker talking time (giving, for example, backchannels per minute) or the number of main speaker speech units (leading to, for instance, number of words per backchannel). A disadvantage of the former measures is that they depend on the speed of the main speaker, which can vary both between speakers and for the same speaker when doing different things. A disadvantage of the latter measures is that each speech unit, such as a word, is treated equally, even though there could be repetition, circumlocution and so on. Despite these limitations, they remain the standard measures. Alternative calculations have been suggested, but they often lack utility or validity: backchannels per minute of a person's own speaking time is one of the least appropriate, for the obvious reason that the giving of backchannels is done only while another person is speaking.

Frequency is often mentioned in the literature, especially that part of it which compares backchannel usage in different languages. Certain studies and their conclusions are often cited, but no methodical account of the literature on frequencies has been published. As part of the preparation for the research presented in this thesis, therefore, a systematic review was conducted of the literature published in English on verbal backchannel rates. (This was a review of the literature on verbal backchannel rates, so did not encompass

other backchannel features.) This review and its main findings (originally in Flint, 2012), are summarised in the remainder of Section 2.5.4. The full systematic map of 73 studies is reproduced in Appendix 1 and more than a third of the studies are referred to in the main text of this thesis.

2.5.4.1 Background to Systematic Review

No cut-off date was set for the material to be included in the systematic review. The main review question addressed was: 'What empirical research has been reported on verbal backchannel rates in L1 interaction, L2 interaction, and L1-L2 interaction?'

The following databases were searched: DART, ERIC, Google Scholar, Index to Theses, LLBA, MLA Intl., PapersFirst, ProQuest D&T and PsycINFO. These covered journal articles, books, book chapters, conference proceedings and doctoral theses in the fields in which backchannels have been studied. Also included were two websites – The Center for Applied Linguistics; Linguist List – and the bibliographies from the only two review articles of backchannels that were found (Fujimoto, 2007; Deng, 2009). Hand searching of journals was halted after all volumes of *Applied Linguistics* (1980 to 2012) and *Journal of Psycholinguistic Research* (1971 to 2012) produced only two studies that had not already been identified via the other means.

Search terms that were used to encompass the wide range employed to refer to backchannels were obtained from the two review articles (Fujimoto, 2007; Deng, 2009) that contain detailed discussions of backchannel terminology. This resulted in 29 search terms being used. The searches produced a total of 1,754 abstracts to screen, leading to

full-text screening of 357 studies, of which 73 were included in the systematic map. The inclusion and exclusion criteria are also listed in Appendix 1 of this thesis. In the systematic map, publications were dated from 1967 to 2010. Most of the studies were of backchannels in English. Japanese was the second most studied language, followed by Mandarin Chinese and then eight languages that appeared in only one or two studies each. Of the 86 cases identified,[4] 64 were of only L1 use, while 22 involved L2 use.

Where possible, only those backchannels identified in Section 2.2.2 as the archetypal ones were included in the backchannel rates obtained from the systematic map studies. Similarly, the rates included were, where possible, those just identified as standard: backchannels per minute and words per backchannel.

2.5.4.2 Findings of Systematic Review

2.5.4.2.1 Overview

Rates in dyadic speech, measured by backchannels per minute, ranged from 0.44 to 32.09. This is approximately once every two minutes to once every two seconds. Calculated as words per backchannel, the range was also very wide: from 8.69 to 833.33. Two things should be noted in relation to these values. First, each was obtained from a different study, so making additional calculations based on combining them is not appropriate. Second, all of them are means, so the range of rates across individuals was even greater.

[4] Some studies were of more than one language, setting, mode of interaction, etc. Where this occurred, the reporting on each of these was referred to as a 'case', so the number of cases exceeded the number of published studies.

These rates were for all dyadic interactions, regardless of any of a number of factors that could have influenced them. To help separate out some of these factors, findings on some of them will now be detailed, beginning with language.

2.5.4.2.2 Language

Direct comparisons could be made for several languages. These were possible when a study met three conditions: speakers of different L1s were used; participants were in the same interaction situation (for instance, a telephone conversation or a face-to-face informal conversation); and the same data analysis procedures were followed for all participants. The comparisons showed consistently that backchannel rates in Japanese were higher than in English and that rates in English were higher than those in Mandarin Chinese. Eight other languages were included in the systematic map: Arabic, Dutch, French, German, Korean, Spanish, Swedish and Thai (studies of others were identified but did not meet the review criteria). For these, there was insufficient evidence to reach more than the preliminary hypothesis that their typical backchannel rates are likely to be somewhere between those of Mandarin Chinese and Japanese.

2.5.4.2.3 Interaction Situation

Studies of numerous interaction situations were identified in the systematic review. These included telephone conversations, face-to-face conversations, counselling sessions, doctor-patient consultations, tutorials, and some researcher-created situations such as story narrations and collaborative tasks. Only the first two of these were reported in a sufficient number of cases and in sufficiently consistent units (backchannels per minute) to allow more detailed reporting.

Rates in telephone conversations varied from 1.43 to 15.60 backchannels per minute. In face-to-face conversations, the range was 2.13–17.65 and was typically higher than in the telephone interactions; the range was 3.58–9.31 in English. Even these observations, however, should be treated with caution, as the number of studies and variations in the methods they employed mean that comparing rates is problematic.

This leads to a broader point on the research that has been done to date: there are very few studies that compare people speaking in different situations. Furo (2002) compared telephone conversations with face-to-face conversations, but the participants were not the same. Only Plough and Gass (1993) both compared interaction situations (face-to-face discussions and spot-the-difference tasks) and used the same participants in the two situations. More such research would greatly aid the comparison of backchannels in different languages and different situations.

2.5.4.2.4 Gender

Comparing backchannel rates based on gender was possible in 19 cases. In L1 English, women's rates were higher than men's in seven of the 11 cases; four found the opposite. In L1 Japanese, all three cases were of women's rates being higher. The remaining five cases were of just one language each: in French, Mandarin Chinese, Spanish and Thai, men used backchannels more frequently; women employed a higher rate in Swedish. Again, these findings need to be treated with caution: they are based only on numerical values, not inferential statistics. The latter are almost entirely absent in the literature, partly because sample sizes are often very low, and not enough information is presented to calculate them.

Another common limitation of studies that report on backchannels by participant gender is that several other possible contributory factors are not controlled for. An example is Fellegy (1995). She collected data from three groups of men and three groups of women. The former were: "a dinner party of four gay males, a tutor-student meeting discussing upper-division physics problems, and a tutor-student meeting discussing freshman-level math problems" (1995: 187–188). The latter, in stark contrast, were: "an eleven-member book discussion group, a six-member hierarchically organized staff meeting of a displaced homemaker counselling agency, and a four-member, lesbian, power-sharing task force meeting in a women's shelter" (1995: 187). The additional factors of social setting, familiarity, power, number of speakers, subject matter of the talk, sexuality (an unusual inclusion) and others could be treated as a strength-in-diversity feature for interpreting results, as Fellegy claims (1995: 188). For the purpose of being able to make assertions about backchannel rates based on statistical analyses, however, such an approach to data collection is a hindrance.

There is also a legacy issue in the literature on the matter of gender and backchannels. Early investigations of links between the two led Coates (1986; 1989) to make the bold assertion that all of the evidence showed the same thing. "Research on the use of minimal responses is unanimous in showing that women use them more than men" (Coates, 1989: 68) was her claim and it has been repeated frequently since then (including in subsequent editions of Coates' original publication), despite having been contradicted by empirical evidence long ago. In summary, although gender is often claimed to be an influence on backchannel frequency, this may be based largely on the repeating of conclusions from early studies; considering all of the available evidence, the impact of gender is unclear.

2.5.4.2.5 Age and Status

Age was investigated in only one of the studies. This was Miyazaki (2005), who reported on L1 Japanese face-to-face conversations. Her three groups of participants, split by age, contained 5 dyads (aged 19–23), 4 dyads (aged 31–39) and 4 dyads (aged 54–60). A total of 20.5 minutes of recordings from informal conversations were analysed[5] and markers of surprise and assessments such as 'sugoi' ('great') and 'honto' ('really') were counted as backchannels. The youngest group averaged 9.62 backchannels per minute, those aged 31–39 produced 8.59 per minute and the oldest group averaged 17.65.

The existence of only one small study on age means that little beyond the confines of that study can be concluded on the influence of age on backchannel rates. Reports of the effects of other potential contributors to differences in status – workplace seniority, for instance – are also too few to allow any summarising conclusions to be reached.

2.5.4.2.6 L1 Variants

Only comparisons of English L1 variants were found in the systematic review. The earliest was Tottie (1991), who reported that speakers of US English had a backchannel rate 3.3 times greater than that of British English speakers. However, the US English data came from just one dyad (a pair of academics who discussed aesthetics in advertising), while the British English speakers formed two dyads (one reminisced about their university days and training; the other discussed a broad range of topics). Stubbe (1998) compared Pakeha (people of European descent) speakers of English in New Zealand with Maori counterparts.

[5] Further data were also analysed, but these were of each participant listening to the researcher reading instructions from a script prior to the main data collection process, not from interacting with another age-matched participant.

Her four Maori-Maori dyads and four Pakeha-Pakeha dyads were recorded in informal conversation and the former had a backchannel rate 1.4 times higher than the latter. Finally, O'Keeffe and Adolphs (2008) used two sub-corpora of approximately 20,000 words each to compare British English and Irish English face-to-face conversations among close friends. The speakers of British English were in two groups (one containing two people; the other having five), as were the Irish English speakers (one group had two participants; the other had four). On average, speakers of British English used backchannels 1.6 times more frequently than those speaking Irish English.

Overall, the number of participants in these studies means that little weight can be ascribed to their findings. In conclusion, there is only limited evidence of backchannel rate differences based on L1 variety.

2.5.5 Summary

Of the four backchannel features discussed, the strongest evidence of differences is found for their frequency. Backchannel rates were therefore the best choice as a target for this research into the effects of backchannels on L2 English speech. Evidence for differences in rates based on interaction situation, gender, age, status or L1 variant is very limited. Rates and other backchannel-related matters in L2 speech are discussed in the next section, which contains further detailed comments on selected studies from the systematic review (as does the following chapter), and then leads up to the stating of the general research questions.

2.6 Interactions Involving an L2

Backchannel rates in interactions involving an L2 were included in the systematic review (Flint, 2012) just reported. This determined that the evidence for differences between rates in L1 interactions and those involving an L2 was inconclusive, because of a combination of the limited quantity of the studies done and their quality. Some of these will now be described.

Ohira (1998) used recordings of conversations from five dinner parties containing four or five people each: one group spoke L1 Japanese; a second spoke L1 English; another contained Japanese-English bilinguals speaking Japanese; and the remaining two groups spoke in English and were made up of a mix of participants from the L1 English group and the Japanese-English bilingual group. One hour from each gathering was analysed and utterances such as 'uh-huh', 'yeah' and 'I see', as well as Japanese equivalents, were counted as backchannels.

Unusually, the L1 English group was reported to have a higher backchannel frequency than the L1 Japanese group (when calculated as main speaker words per backchannel): 31.91 words compared with 40.30. Ohira's (1998) posited explanation is that the L1 Japanese group were friends, whereas a combination of friends and strangers formed the L1 English group. In comparing the other groups of participants, she concluded that the bilingual participants used similar backchannel frequencies in their Japanese and mixed English conversations, while the L1 English speakers had a higher backchannel rate in the mixed interactions than in the L1 only one. Although, as just described, Ohira does discuss the possible relevance of not controlling for familiarity among participants, there is also an

undiscussed complicating factor: she included herself as a participant in all but the L1 Japanese dinner party. Being a source of some of the data to be used in her analyses is not a major limitation in itself, but the classification of one in five of the participants in most of the groups as both an L1 English (only) speaker and as a Japanese-English bilingual is considerably more problematic and undermines the reported backchannel data.

Different limitations were found in another study that compared L1 and L2 speech and was included in the systematic review: Hirokawa's (1995) study of conversation interaction management in Japanese and English, which included backchannels. She split her university student participants into three groups, with five dyads in each. Each had two face-to-face conversations based on topics that she suggested. For each dyad, one of the conversations was conducted in English and the other was conducted in Japanese. Her first group was of L1 Japanese speakers; the second contained L1 English speakers; and the third group paired L1 Japanese speakers with L1 English speakers. Eight minutes of each interaction were analysed and listener responses such as 'certainly', 'mhm', 'right' and 'sure', plus their equivalents in Japanese, were classed as backchannels.

The L1 Japanese group averaged 10.1 backchannels per minute when speaking in Japanese and 8.73 in English; this was found to be not a statistically significant difference. The L1 English group averaged 5.33 backchannels per minute in their L1. In the final group, when conversing in English, the L1 Japanese participants averaged 7.20 backchannels per minute and the L1 English participants averaged 4.15; Hirokawa (1995) indicates that this difference was not statistically significant because of high variation across individuals. The backchannel rates of the L1 English participants when speaking Japanese were not reported.

Hirokawa (1995) suggests that the Japanese participants may have transferred their L1 norms into the cross-cultural English conversations. She also indicates that accommodation, involving greater attempts to achieve cooperation and understanding, could have contributed to the apparently higher backchannel rates in those interactions. This study is an example of backchannel rates usually being higher in L1 Japanese than in L1 English, and serves to highlight some of the difficulties in conducting research into backchannels, including in L2 interactions. Small sample sizes can make the obtaining of meaningful findings difficult and using conversational data greatly restricts the confidence with which reasons underlying the findings can be given.

Other studies have speculated on the roles of transfer and accommodation involving backchannels during L2 speech. Some have suggested that L1 backchannel behaviour is mostly transferred to L2 interactions. For instance, Maynard (1997) was interested in what she suggested are differences in how interaction is managed across L1s and how these can influence the feelings of those involved in cross-cultural interactions. The feelings – she mentions (1997: 37) "disengagement", "estrangement", "sympathy" and "a disguised sense of superiority" – are assumed rather than empirically investigated, but her main interest is in contrasting L1 and L2 speech. One of her findings was that Japanese interlocutors, when conversing with an L1 English speaker who was a friend, placed backchannels at grammatical positions that would be natural in their L1 but not in English, indicating that they did not adapt to the style of their L1 English partners.

This has been contradicted by other studies of the same two languages. R. White's (1997) participants were experienced business people and took part in simulated sales negotiations in English. He also suggested that L1 differences could have negative effects on

perceptions of L2 English interactions, this time in business negotiations. His research indicated that the backchannel behaviour of the L1 Japanese and L1 English speakers was found to converge, with the Japanese participants adapting to the style of English in their backchannel use and the L1 English users also moving towards the Japanese style. S. White (1989) used both intercultural and intracultural English conversations that L1 English and L1 Japanese strangers took part in. She reported that, speaking in L2 English, the Japanese participants had a backchannel frequency that was 2.4 times that of the L1 English users speaking together. In L1-L2 English conversation, the L1 speakers were reported to have increased their backchannel rate substantially, while the L2 speakers reduced their rate slightly. S. White (1989) thus concluded that the L1 English interlocutors had accommodated to the Japanese style in intercultural exchanges.

As reported in Section 2.3, Wolf (2008) found that backchannel rates could influence L2 speech, but this appears to be the only example of such a study. Evidence for backchannel differences in L2 communication negatively influencing attitudes was reported by Cutrone for Japanese and British speakers of English (2005) and for Japanese and US speakers of English (2014). Cutrone's (2005; 2014) method was to ask participants about their dyadic conversations in English. Japanese participants were asked to explain their backchannel behaviour, which the L1 English speakers were asked to comment on; the Japanese participants were not asked for their impressions of the backchannels produced by the L1 English speakers, so the effects of backchannels on L2 learners or their speech were not assessed.

Evidence, then, for backchannel differences between L1 interactions and ones in an L2 are inconclusive. A combination of design problems, low sample sizes and a limited quantity of studies all contribute to the lack of clarity in the existing literature.

2.7 Summary and General Research Questions

At the end of Chapter 1, the research topic was stated to be the effects of backchannels on L2 English speech. Chapter 2 has selected verbal forms as the most appropriate for further investigation and considered several of their features. It has also identified some evidence that a difference between the backchannel frequency speakers expect to receive and what they actually receive can affect spoken interactions in a variety of ways. The limited extent of this evidence and for backchannel effects more broadly is a gap in the literature that the current research can address. The conclusion is that frequency is the most apposite backchannel feature to use to study the research topic and, as backchannel rates vary across L1s, a comparison across languages is also warranted.

The general research questions can now be stated:

1. What are the effects of interlocutor backchannel rates on spoken L2 English?
2. Do these vary with the speaker's L1?

Having stated the general research questions, it is necessary to decide on the best ways to investigate them. This is addressed in the next literature review chapter, which identifies an appropriate study design, selects participant L1s, details the required characteristics of those participants and chooses methods to analyse the data collected, all of which leads to the statement of the specific research questions.

3 LITERATURE REVIEW II – METHODOLOGY

This second literature review chapter is split into two main sections – data collection and data analysis. The data collection section addresses two key matters: the nature of the research design and the choice of participants. The research design parts begin with the selecting of a quasi-experimental design, continue with a discussion of the possible interaction situations that could be studied and end with the choosing of what backchannel frequencies to compare. The choice of participants – the L1 groups to use and their characteristics such as L2 proficiency – is based in part on studies that were introduced in the preceding chapter.

The data analysis section considers the possible means of analysing the samples of speech produced via the data collection procedure. Partly through reference to the study's theoretical framework, complexity, accuracy and fluency (CAF) are chosen, so, for each of these, the construct, some possible ways of measuring it, and the measures chosen for this research are described. This chapter, then, covers the main decisions taken on methodology; these inform the creation of the more specific research questions, which conclude the chapter. Descriptions and detailed features of the methods of data collection and data analysis that were actually employed are in the following two chapters.

3.1 Data Collection

3.1.1 Research Design

This section describes the main features of the research design for data collection purposes. A summary of the overall research design, incorporating data analysis as well as collection, is given in Section 3.3, towards the end of this chapter.

The required research design when evaluating the effects of something – of backchannel rates on speech, in this instance – is an experimental one. As the random sampling needed in a 'true experiment' was unrealistic in this research, 'quasi-experimental' was the more accurate label for the required design (Creswell, 2014: 168). The first decision to be made about the main details of the quasi-experimental design was whether the interactions studied should be of dyads or larger groups.

For quasi-experimental research the decision is straightforward. The great majority of research on spoken interaction is of dyads, as the complexities of dyadic interaction are already high and adding more people to the interaction would add considerably to that. Dyads were chosen for the current research for this reason and to allow for more ready comparisons to be made with existing research. The interactions were therefore between a main speaker and an interlocutor only (see Section 2.1 for the selection of the labels for these roles). Having a dyadic situation in which there is a high degree of control possible in the interaction, allowing the rate of backchannels to be experimentally varied, is also desirable. The next choice was therefore of what the interaction situation should be.

3.1.1.1 Interaction Situation

The chosen interaction situation preferably should be readily comparable with a real-world setting, so that any findings are applicable not only to theory development, but can be the starting point for further, practical research. Several such settings have been used in research into interaction that has included backchannels; examples are: doctor-patient consultations (Ohtaki, Ohtaki and Fetters, 2003); counselling sessions (Spicer, 2005); and tutor-student meetings (Farr, 2003). However, these have been descriptive rather than experimental studies: the ethical problems inherent in research activities that may have a negative effect on participants mean that only simulations of such encounters may be manipulated in the sort of way required in a quasi-experimental study.

Simulations of some such encounters have been used, including of doctor-patient consultations (Li, 2006) and counselling sessions (Sharpley et al., 2000). Doctors and counsellors are highly trained, however, so an adequate portrayal of one of these roles would require a real doctor or counsellor, to avoid a high level of pretence. While feasible, this would present clear practical difficulties, particularly in recruitment. Li (2006) used university students and gave the 'doctors' an information sheet on the use of codeine and the 'patients' an information sheet that contained their fictional case history. Sharpley et al. (2000) used trainees as 'counsellors' and two people trained to be consistent in their presentation of counselling problems as 'clients'. The second role in such dyadic interactions also needs to be considered: a healthy 'patient' or a psychologically untroubled 'client' could have difficulty in simulating the ascribed role, and ethical objections would resurface if participants did have physical or mental health problems.

A further drawback to using the interaction situations just discussed is that they would require participants to have a high level of L2 proficiency in order to play the roles assigned to them. For reasons detailed in Section 3.1.3.1, a lower level of L2 proficiency is preferable for the purposes of this study. Given the described problems of ethics, practicality and validity for these interaction situations, another situation was sought. Another dyadic encounter occurs in some tests of spoken language. Ethical considerations again prohibit the experimental manipulation of interlocutor backchannels in real language tests, but a data collection procedure based on one of them is practicable and requires little pretence for adult L2 speakers, who could participate without having a high level of proficiency. Dialogic tasks are common in high-stakes language tests. The speaking component of one of the most common ones, IELTS, involves one-to-one interaction between the interviewer / examiner and the candidate, and the testing procedure is somewhat controlled yet allows flexibility in some aspects of the examiner-candidate interaction, making it suitable for this research. Following the exact testing procedure is not necessary, as the focus here is on backchannels and their effects rather than testing, but it can be used as the starting point for developing an appropriate data collection procedure.

Backchannels in IELTS have been investigated to a limited extent by studying recordings from actual Speaking tests and there is some evidence that examiners are informed about backchannel use. Seedhouse and Egbert (2006: 22) quote *IELTS examiner training material* from 2001 (it is not publicly available): "Examiners should keep non-verbal interjections to a minimum. (Eg 'um', 'right', 'uh uh')", but observe that the frequency of examiner use of backchannels in their sample varied considerably. Given the known differences in L1 backchannel norms and examiner variation in backchannels produced, there is also clear real-world value of research that uses this interaction setting (albeit

where that research is initial and using this kind of language testing as the basis of a data collection procedure rather than its focus). Further details of both IELTS Speaking itself and those aspects of it chosen for the data collection procedure are given next.

3.1.1.1.1 Details of Chosen Interaction Situation

IELTS Speaking consists of three parts. In Part 1 ('introduction'), there are introductions and a check of the candidate's identity, then the candidate answers questions about himself or herself and topics related to himself or herself. In Part 2 ('individual long turn'), the candidate has one minute to prepare to speak for one to two minutes on a topic given by the examiner, who does not add comments or questions before the candidate has finished speaking. In Part 3 ('two-way discussion'), the two people discuss more conceptual aspects of the topic from Part 2; this usually involves the examiner asking questions that the candidate answers (Seedhouse and Harris, 2011: 4).

Part 1 is likely to lead to short answers that do not provide the opportunity for backchannels to be given, so would not be suitable for this research. In contrast, Parts 2 and 3 can readily allow the varying of interlocutor backchannels within a relatively controlled process and the responses are of sufficient length to warrant backchannels being used. Part 2 is monologue-like, highly controlled and thus potentially the most likely to reveal differences in speech based on backchannel rate variation; the nature of Part 3 is something approaching the conversational but still controlled. Using these two parts also confers the additional advantage of allowing comparisons to be made between different types of interaction (challenges to validity based on repetition of topic across the two parts are addressed in Section 4.3.5.3).

The length of speech samples obtained from participants in each of these two parts is also likely to be adequate for the purposes of analysis. In IELTS itself, Parts 2 and 3 are linked by the examiner asking "rounding-off questions" (Seedhouse and Harris, 2011: 4), but these need not be used for this research, to allow the two types of interaction to be kept largely separate. A final consideration is planning time. As mentioned, for Part 2 there is one minute to prepare to talk about a topic. Mehnert (1998) compared the effects of 0, 1, 5, and 10 minutes of planning time on the L2 speech of participants who were asked to leave messages on answering machines; in general, the longer the planning time, the more differences in the speech were found. While one minute may be an arbitrary planning length for Part 2, then, it is a reasonable period to choose for this research, provided that it is applied consistently for all participants. Having no planning time for Part 3 is also appropriate for the nature of the interaction.

The interaction situation was, therefore, based on the procedures of Parts 2 and 3 of IELTS, allowing samples of speech to be recorded with each participant as the main speaker in each part (similar to an exam candidate) and with the interlocutor (similar to an examiner) varying the frequency with which backchannels are given. In this thesis, these two parts will be referred to as the 'long-turn part' and the 'discursive part', respectively.[6] Before discussing backchannel frequencies, the next choice to be described is whether to use a between-subjects or repeated measures design.

[6] 'Discursive part' replaces 'discussion part' to minimise confusion with the Discussion chapter of this thesis.

3.1.1.2 Between-Subjects or Repeated Measures Design

Variation among individuals in between-subjects designs is uncontrolled, whereas each participant in a repeated measures design can act as his or her own control; this allows the latter design to have greater statistical power (Kim, 2012). Another advantage of repeated measures designs is that they enable individual differences, not just group differences, in performance to be examined (Kim, 2012). A repeated measures design is, therefore, preferable for this research. This means having each participant do the long-turn part and discursive part twice, with the backchannel rate received differing between the first and second iterations.

There are also several potential disadvantages to a repeated measures design and these need to be addressed. One drawback is the need to have two topics: if they vary in difficulty then the validity of the research will be compromised (Mackey and Gass, 2016: 167–168). How this was addressed is described in Section 4.3.4.1. A further challenge, that of sequencing effects, is also dealt with in the Methods chapter (Section 4.3.6). A final consideration is the potential effects of fatigue and boredom on validity (Mackey and Gass, 2016: 165–166). The expected data collection duration of 20–25 minutes per participant, with a variety of activities and topics, is likely to be comfortably within the normal experience of adults, so these potential effects would be minimal.

3.1.1.3 Backchannel Frequencies

Two contrasting backchannel frequencies were required. The small number of previous studies of backchannel effects, summarised in Section 2.3, mostly contrasted a standard rate (what the interlocutor gave naturally) with none at all. Using that contrast for this

research would be the most likely way to get differences in the speech of the main speakers, but the value of the findings would be very limited, as giving no responses at all is not something that is done in many real-world scenarios. The possibility of using a very high rate in contrast with a lower one was also a possibility. This was rejected in part because of the work of Ross (2007), who presents a discourse analysis from Japan of two L2 English oral proficiency interviews and suggests that a very high backchannel rate (one more similar to L1 Japanese than to L1 English) may be associated with relatively brief, phrase-like utterances from the main speaker and lower fluency. Extrapolating from an analysis of two samples of speech is unwarranted, but the dangers of applying an abnormally high backchannel rate are apparent: restricting the higher of the two planned backchannel frequencies to something that an L1 English speaker would typically provide is preferable.

Considering that the anticipated duration of the long-turn parts would be 1–2 minutes and that the participants' speech would be slow because of their L2 proficiency, the chosen contrast was between a typical rate (again, meaning what the interlocutor gave naturally) and a rate that was one-third of that. An even lower rate would mean that very few would be given in the long-turn part in particular, again reducing real-world applicability, and would increase the possibility of the purpose of the study being identified by the participants, an eventuality that was undesirable. One-third of the typical rate is unlikely to be abnormal in the context of IELTS Speaking: Seedhouse and Egbert (2006: 22–23) report from their analysis of actual test recordings that backchannel frequency varied considerably; and a full transcript (Seedhouse and Harris, 2011: 45–50) shows backchannels being used at highly variable rates in relation to the candidate's utterances even over the course of one test. As the exact difference between the typical and other rate

was unlikely to be exactly one-third, the latter will be referred to as the 'low' rate in the remainder of this thesis.

This concludes the section on the research design; again, further details on what was done will be presented in Chapter 4. The next section in this chapter identifies the L1 groups that would be used in this study.

3.1.2 Participant L1s

The general research questions require a comparison to be made between the effects of backchannel frequency on the L2 speech of people with different L1s. Comparing learners with L1s that differ maximally in their backchannel frequency would optimise the chances of finding differences when they speak in their L2. The findings of a systematic review (Flint, 2012) were reported in Section 2.5.4.2.2: the available evidence is that backchannel rates in L1 speech are Japanese > English > Mandarin Chinese, with insufficient evidence for other languages. Based on this, the best choice is to use L1 Japanese and L1 Mandarin Chinese speakers communicating in L2 English. Potentially important characteristics of the participants are described next.

3.1.3 Participant Characteristics

If participants do not greatly differ in certain characteristics, then the study's internal validity will be raised. Where a characteristic itself covers a large range of variation – L2 proficiency, for example – then which participants to select should be based on which will allow differences between the two populations to be revealed. Some of these

characteristics were introduced in Chapter 2; others are specific to L2 speakers. First is the one just mentioned: L2 proficiency.

3.1.3.1 L2 Proficiency

The chosen interaction situation permits a wide range of L2 proficiencies to be considered. Its discursive part is interactive in nature, so the participants need to be able to maintain a conversation in English; this excludes those with low English proficiency. Some research has been published on backchannel use at the upper end of the proficiency spectrum. Tao and Thompson studied L1 Mandarin Chinese speakers who had lived in the United States for a long time; they reported that "Mandarin speakers for whom American English has become their dominant language […] make extensive use of American English backchannel strategies" (1991: 209). In a larger study, Heinz (2003) reported similar findings for L1 German speakers: those for whom English had become their dominant language exhibited backchannel behaviour based on English norms rather than German ones even when speaking to monolingual Germans. This extreme familiarity with English backchannel norms is not desirable in the participants in this study either, so those of intermediate English ability appear to be most suitable.

The L2 proficiency of participants in Wolf's study of backchannel effects was judged to be intermediate-mid to intermediate-high on the ACTFL scale, indicating that "they could function competently in many routine communicative tasks and social situations" (2008: 285). Cutrone's first study that included attitudes to backchannel behaviour required participants to "be able to comfortably maintain a conversation in English with a native speaker of English" (2005: 250). His later study recruited Japanese people with a TOEIC

score of between 350 and 700 (Cutrone, 2014: 91). This represented a wide range: TOEIC scores can be from 10 (low) to 990.

Using TOEIC scores as an indicator of speaking proficiency is not appropriate, for two reasons: what TOEIC tests, and assumptions of links between what it tests and what it does not test. TOEIC was created in the 1970s, specifically for the Japanese market (IIBC, 2009b). It contains two sections – reading and listening – both of which contain only multiple-choice questions. Another product, TOEIC Speaking and Writing, has been created, but this is an entirely separate test and the assessment generally referred to as 'TOEIC' remains of reading and listening only.

The producers of TOEIC have asserted to users that the test of reading and listening might also be taken to be indicative of speaking ability. For example, they claimed that a TOEIC score of between 470 and 730 showed that a person "Can understand the gist of ordinary conversation and has no trouble forming responses. However, the individual shows some disparities in the ability to respond correctly and making himself/herself understood in more complicated situations" (IIBC, 2009a: 13). Given that an overall TOEIC score of 10 to 990 is formed by adding the reading score to the listening score, each of which can range from 5 to 495, it is possible for someone to obtain an overall score of 470 purely from the reading part (i.e. having no listening ability whatsoever and failing even to guess a listening part answer correctly). The claims that such a score can indicate anything about speaking require some very strong evidence.

These claims appear to have been removed from more recent TOEIC documents, but the most recent handbook for test-takers still maintains that correlation data of TOEIC scores

and test-taker self-reports of their abilities show "moderately strong correlations (.40s and .50s)" (ETS, 2015: 22) and implies that these extend to speaking (ETS, 2015: 23). This is very limited evidence for the claims made. In short, little, if anything, can be assumed about a person's oral proficiency based on TOEIC scores.

In contrast, IELTS does test speaking and reports separate scores for speaking, reading, listening and writing, in addition to an overall score. Each is reported on a scale of 1 (low) to 9, in whole bands or half bands. The most recent IELTS (Academic) data, from 2015, show that takers who reported their first language to be Japanese had an average speaking score of 5.6, a listening score of 5.9 and a reading score of 6.1 (IELTS, n.d. a). For those who reported that "Chinese" was their first language (different forms of Chinese were not distinguished), the average speaking score was 5.5 and the listening and reading scores were the same as those with L1 Japanese (IELTS, n.d. a). Speaking scores being lower than overall scores is a further indication that tests such as TOEIC, which do not assess speaking, should not be taken to be indicative of speaking proficiency.

The evidence, including that from previous practice, therefore provided little assistance on where within the broad category of 'intermediate' was the most appropriate L2 English proficiency for participants in this study. The relatively broad range of 'intermediate' was therefore retained as the starting point, and was to be evaluated in pilot studies, which are reported in Section 4.2. Subsequent amendments to the range were then put into operation via requests to gatekeepers in a form that they could readily understand and implement, as outlined in Section 4.3.2.2.1.

3.1.3.2 Time Spent Overseas

Time spent overseas can affect awareness of L2 backchannel norms (e.g., LoCastro, 1987), so this should also be taken into consideration when selecting participants. The three studies mentioned in the preceding section as precedents were also consulted on this characteristic. Cutrone (2005) does not mention it; Cutrone (2014: 91) set a lifetime limit of 100 days spent outside Japan; and Wolf's participants had lived in an English-speaking country for up to seven months (2008: 285). There is very little evidence informing the limit applied to time spent overseas and decisions appear to be arbitrary. Six months was set as the upper limit for this study. This was between the lengths applied in the studies just mentioned and included time spent in any other country, not only ones where English is the dominant language.

3.1.3.3 Gender, Age and Status

Section 2.5.4.2.4 concluded that, although gender has been suggested to be an influence on backchannel frequency, this reporting is based more on early, now-contradicted research conclusions, so the evidence for a gender effect is weak. There is, therefore, no compelling reason to investigate gender as a factor in this study. Nevertheless, having a mix of genders in the samples could allow for checks to be made on this and for the findings to be more readily generalised.

There is even less evidence of age- or status-based differences in backchannel rates (see Section 2.5.4.2.5). However, age effects on communication more generally have been reported (e.g., Yule, 2010), so it is a potential influence on backchannel rates.

In summary, with the exception of L2 proficiency, there are several participant characteristics that might be relevant to this study, but little to indicate that any of them is likely to be important. Given that little research has been done, the best option is to adopt the recommendation of Mackey and Gass (2016: 177) and report in the Methods chapter "major demographic characteristics such as gender, age, and race/ethnicity […] as well as information relevant to the study itself (e.g., the participants' first languages, previous academic experience, and level of L2 proficiency)".

A simple way to reduce the uncertainty over the possible impact of participant gender, age and status differences on the effects of backchannel rate variation was for their interlocutor to be the same person. The interlocutor is the final part of this chapter's data collection section.

3.1.4 Interlocutor

Ideally, the interlocutor would not also be the researcher, due to the possibility of researcher bias affecting the interaction (e.g., in psychology, Rosenthal, 1966). However, the aims of the study would be evident to anyone assuming the role, so potential levels of bias would be similar with another person as the interlocutor.

The interlocutor's familiarity with the participants' L1s or cultures also needs to be considered (Mackey and Gass, 2016: 167). As already mentioned in Sections 2.6 and 3.1.3.1, S. White (1989) found accommodation in backchannel behaviour during English conversations between L1 speakers of Japanese and L1 speakers of English who had little or no knowledge of Japanese language or culture, and, in contrast, Tao and Thompson

(1991) and Heinz (2003) found advanced L2 speakers transferred those backchannel norms to conversations in their (nominal) L1. If the interlocutor has some, but only partial, prior knowledge and experience of the participants' L1s, then backchannel behaviour can be more consistent over the duration of a study. The interlocutor, then, should have L1 English, and ideally have some experience of Japanese and Mandarin Chinese, but not of other languages.

A final validity challenge is the participants' familiarity with the interlocutor. O'Sullivan compared the L2 spoken performance of Japanese learners of English in tasks (personal information exchange, narrative, decision making) done with a stranger and done with a friend; he found "evidence of an 'acquaintanceship' effect, with subjects achieving higher scores when working with a friend" (2002: 277). In contrast, Plough and Gass (1993), comparing the effects of familiarity in spot-the-difference and discussion tasks, found that backchannels were not affected by interlocutor familiarity. Their study of L2 English speech used participants with seven L1s, however, so L1 backchannel norms were not accounted for. The conclusion, again, is that there is limited evidence, and that internal validity is best supported if familiarity between a participant and the interlocutor is the same for all. Details of how the chosen interlocutor fitted the desired criteria mentioned in this section are given in Section 4.1.2.

3.2 Data Analysis

The data collected would be recordings of L2 learners of English speaking with someone whose L1 is English. There are several ways in which the recordings could be analysed. Cutrone (2005; 2014) analysed his recordings for the influence of backchannels by using

stimulated recall combined with more general interview questions[7] to get the participants to comment on backchannel use. These studies, however, were of the effects of backchannels on attitudes and perceptions, not on speech itself, which is the focus of the current research.

More discourse-based analyses are another possibility. A pure form of conversation analysis would not be appropriate, given that it requires that recordings were of something naturally occurring and that the analytic process does not begin with categories or hypotheses (Richards, 2003: 26–27). Another discourse-based analysis is interactional analysis, described by Ellis and Barkhuizen (2005: 166) as studying "what kinds of functions learners perform when they interact with other learners or native speakers in different contexts and the structural properties of these conversations". In the current research, the nature of the interactions and their structural properties are largely pre-determined: the long-turn parts consist of one person as the main speaker throughout; and the discursive parts are based around a question-and-answer structure. This, of course, comes from the choice of a quasi-experimental design, which is required to control the backchannel rate. More suitable ways of analysing the data are those that produce numerical descriptors of speech, thereby permitting statistical analyses.

The numerical values need to be measures of performance; this was not to be a longitudinal study, so development did not need to be considered. Numerical values are produced by commonly used calculations of complexity, accuracy and fluency (CAF). These have been used together since the 1990s, permit various facets of speech to be

[7] Cutrone (2005; 2014) consistently refers to his method of analysis as 'interviews'; as it included participants being played parts of the recordings after completion of speaking tasks and then being asked questions about them by the researcher, 'stimulated recall' has also been listed here.

examined, allow the effects of other factors to be evaluated and have been shown empirically to be dimensions of L2 performance (Housen, Kuiken and Vedder, 2012b). Measures of CAF can be used to describe speech in both the long-turn and discursive parts of the data to be collected and can be used to evaluate the effects of backchannel frequency.

A critique of CAF as a framework is presented next (each of the individual components of CAF is discussed in Sections 3.2.2–3.2.4). This is principally on CAF in relation to L2 performance, because this is the focus of this research, rather than in relation to L2 acquisition, development, instruction, or task-based learning, although those things are among the major areas in which CAF is also applied (see, e.g., Norris and Ortega, 2009).

3.2.1 CAF

As mentioned in Section 2.4, this study's theoretical framework can aid in the selection of data analysis tools, so, before a critical examination of CAF as measures, the three components – complexity, accuracy, fluency – can be related to the theoretical framework of Levelt's (1989; 1999) model of speech. CAF is commonly used because the three parts are believed to represent dimensions of speech, which is multi-componential (Housen and Kuiken, 2009). These multi-componential and multi-dimensional features mean that the relationship between the components of CAF and elements of mental processing that occur in a speaker are not straightforward: there is unlikely to be a simple correspondence between the two (Housen, Kuiken and Vedder, 2012b). Nevertheless, as will be seen in Sections 3.2.2–3.2.4, fluency is linked to the speed and efficiency of access to, and use of, stored information, so operates mainly at the formulation and articulation stages of the model of speech, while L2 complexity and accuracy are regarded as operating chiefly in

relation to L2 knowledge representation, so are related to the conceptualisation and formulation stages (Housen, Kuiken and Vedder, 2012b). This differentiation, albeit partial, somewhat speculative and incompletely understood, supports the use of measures of CAF in this study of the effects of backchannels. This is because, as was suggested in Section 2.4, backchannels could act on the conceptualisation and/or formulation stage of speech production, so the possible differences in the CAF components' relationships to those stages could help to indicate at what stage, if any, backchannel rates affect speech in this study's participants.

While measures of CAF, then, appear to be well suited to be used in the data analysis parts of this study, they are not without controversy. A broad criticism of the CAF framework is from theory: there is only limited understanding of, and agreement on, the definitions of the terms used or the relationship of the three components. Housen and Kuiken (2009: 462) state the situation succinctly: "none of these three constructs is uncontroversial and many questions remain, including such fundamental questions as how CAF should be defined as constructs". The construct of each CAF component is also discussed in Sections 3.2.2–3.2.4, but, considering them together, an initial criticism is that CAF does not cover all facets of L2 performance. One facet was pointed out by Pallotti (2009): CAF does not encompass communicative adequacy, so a nonsensical yet grammatically and lexically accurate utterance that is smoothly and swiftly delivered would be evaluated highly using the CAF framework, despite being worthless communicatively. Equally, an attempt at communication that misuses a simple construction, employs a basic vocabulary and is delivered hesitantly and slowly could be effective at communicating a message, while being judged unfavourably by CAF criteria. Some work to address this gap has been done: Révész, Ekiert and Torgersen (2016), for instance, compared raters' assessments of

communicative adequacy in L2 English speaking tasks with CAF measures and found that at least one measure of each CAF component acted as a predictor of adequacy in their data.

While much remains to be done in this area, there are two points that make the topic of communicative adequacy in relation to CAF peripheral to the present research. First, this study used open-ended activities, not tasks that could be judged to be 'completed' or not, so that element of adequacy was of reduced relevance. In contrast, Révész, Ekiert and Torgersen give the example of their raters judging performance on one task – "refuse a teacher's suggestion" – based on participants doing three things: "(i) acknowledge receipt of phone message and/or professor's opinion, (ii) express disagreement with professor's position, and (iii) make case for own position and/or solution" (2016: 834). Communicative adequacy was also of limited importance in the present research because if a potential participant had, for any reason, been incapable of producing what would be a useable sample of speech for the purposes of data analysis, then the data collection procedure with that person would not have been started, or the person would have been excluded from the analysed sample. This point can also be applied more broadly: Pallotti's (2009: 596) example of a nonsensical utterance – "colorless green ideas sleep furiously on the justification where phonemes like to plead vessels for diminishing our temperature" – would score highly using CAF measures, but no researcher who was paying attention would choose to include it in data to be analysed.

Another concern is related to linking theory and practice: the lack of certainty over the constructs and how they are related has also led to the employing of a great number of ways of measuring them (e.g., Van Daele et al., 2007). Norris and Ortega (2009) report that there are CAF measures that have been used by some researchers as indicators of one

of the three CAF components and by other researchers as indicators of another of the three components. They advise taking into account existing empirical evidence and theory for which component a measure relates to, as well as considering how a value is calculated (its numerator and denominator, for instance) and how informative it is likely to be to the research being conducted. The plethora of available CAF measures also makes comparing results across published studies difficult, as pointed out by Ellis and Barkhuizen (2005: 163). Given these points, no attempt at measurement innovation was attempted in the current research. Instead, measures that are commonly reported and are justified by current (albeit incomplete) understanding of the underlying construct were used.

Pallotti (2009) also points out that there is not necessarily a linear relationship between a CAF value and proficiency or performance. For accuracy, he cites the example of a construction that is not correct from the perspective of prescriptive grammar, yet is commonly used by both L1 speakers and novice L2 users, but not by intermediate L2 speakers, who prefer the textbook form. For complexity, he mentions variety in stylistic preferences and variation in optimal levels of complexity depending on the nature of what is being done (he compares a narrative retelling with a telephone call opening). These challenges were dealt with in part in the present research by using in the data collection procedure topics that had been assessed to be of similar difficulty (see Section 4.3.4.1), as well as checking for topic-related differences and sequencing effects at the stage of statistical analysis. They were also addressed during data analysis by referring to the literature: for instance, the literature on the use of the historical present in narratives was consulted when judging the extent to which participants were making stylistic choices as opposed to selecting inappropriate tenses, as discussed in Section 5.8.1.1. Further examples and considerable detail exist throughout Chapter 5.

Linked to the relationship between CAF and proficiency or performance, it has been suggested that the CAF framework is inadequate for advanced L2 users. Ortega and Byrnes, for example, argue that the framework's focus on "isolatable domains such as lexis and grammar does not capture defining aspects of advanced levels of ability, particularly the textual oriented, socially embedded, and situationally motivated nature of language use" (2008: 282). This criticism, while being of obvious relevance to studies of those with advanced L2 skills, is not important for this investigation, which used participants of lower proficiency.

An often unmentioned factor in critiques of CAF is that the literature is dominated by a limited number and range of languages (Flint, 2013). Sakuragi, reporting her study of L2 Japanese, expressed this concisely: "although CAF measures have often been investigated in Indo-European languages, they have not been sufficiently investigated in other languages"; she goes on to mention the nature of Japanese morphology as an example of an element that may not be captured by current conceptualisations of CAF (2011: 158). The common, usually unstated, assumption that what is relevant and appropriate for L2 English research is also relevant and appropriate for research on other languages should, of course, be challenged, but the present study is of L2 English, so this objection can be put to one side here.

A final consideration here is individual variation. There can be considerable variation in CAF measurements across participants, even when those participants have been placed in a group based on shared characteristics (e.g., Housen, Kuiken and Vedder, 2012b). Comparing group means can help reduce the impact of such things on a study's results, but group means can also conceal individual differences. Reporting both group means and the

number of participants in a group who did and did not follow any pattern identified is a way of balancing these competing priorities. Another concern related to individual variation is that L2 speech may be influenced by L1 speech style. de Jong et al. (2015), for example, reported that an adjustment for L1 patterns in one fluency-related measure (syllable duration) made it a better predictor of L2 proficiency than the uncorrected measure was. This warrants further investigation, but for now the importance of individual L1 speech style on L2 style is little understood and, again, reporting both group means and how many participants fitted into any identified pattern is an adequate course of action.

Housen, Kuiken and Vedder (2012c: 299–300) conclude in the epilogue to their edited volume on complexity, accuracy and fluency in an L2 that CAF is:

> neither a type of analysis nor a method or methodological approach […] nor does it at the current state of affairs constitute a model or a theory of L2 learning, development or use. Rather, complexity, accuracy and fluency each represent heuristic dimensions for guiding systematic inquiry and observation of L2 performance, proficiency and development and they are most frequently used as dependent variables to assess variation in these areas with respect to independent variables[.]

This summarises the situation for the present research: despite problems with both the theory and practice of CAF considered collectively or as a loose triumvirate, CAF may be regarded as a tool or set of tools, and it is possible to address many of the inherent difficulties in selecting and implementing them by making appropriate decisions at the data collection and data analysis stages. As already mentioned, in the next three sections, each of the three parts of CAF is discussed separately, briefly covering the construct, some possible ways of measuring it, and the measures chosen for this research. First is fluency.

3.2.2 Choice of Fluency Measures

Lennon (1990) identified two ways in which the term 'fluency' is used: a broad sense, which refers to overall oral proficiency; and a narrow sense, relevant here, which is one part of oral proficiency. He later offered a draft definition: "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" (Lennon, 2000: 26).

Segalowitz (e.g., 2016) has gone further, describing three aspects of L2 fluency: cognitive (the mental processes underpinning speech); perceived (subjective ratings of speech fluency); and utterance (characteristics of the speech itself). Dimensions of utterance fluency are typically split into three: speed; breakdown; and repair (e.g., Skehan, 2009).[8] These can be compared with facets mentioned in Lennon's (2000: 26) definition: speed ("rapid"); breakdown ("smooth"); and repair ("accurate").

Innumerable measures of (utterance) fluency have been proposed and used, without a consensus having been reached on which are the most apposite or informative. Resisting the temptation to employ a large number of measures in the hope that at least one of them will reveal something of interest, the approach taken in this research was to choose measures that match the purposes of the research, have been empirically correlated with perceptions of fluency and are commonly used.

[8] Skehan (2003: 8) is often cited as an early instance of this tripartite description, even though four are stated: "(a) silence (breakdown fluency), (b) reformulation, replacement, false starts, and repetition (repair fluency), (c) speech rate (e.g., words/syllables per minute), and (d) automatisation, through measures of length of run". Studies that cite this one rarely mention automatisation.

Speech rate, often reported as syllables per minute, is one of the most commonly used measures. It is strongly associated with rater perceptions of fluency (Kormos and Dénes, 2004: 148) and is a combination measure, as it includes breakdown shown by pausing, and speed (Levkina and Gilabert, 2012: 182). Unpruned speech rate includes repairs such as false starts and repetitions; to avoid including three posited facets of fluency in one measure, an alternative is pruned speech rate, which excludes all such repairs.

Mean length of run is also strongly associated with rater perceptions of fluency (Kormos and Dénes, 2004: 148). It may also be linked to aspects of automatic speech production, the use of formulaic language and the use of language chunks (Kahng, 2014: 838–839). One way of calculating mean length of run is as "the mean number of syllables produced in utterances between pauses" (Towell, Hawkins and Bazergui, 1996: 91).

Phonation-time ratio – time spent phonating in proportion to total time spent on speaking – is also reported by Kormos and Dénes (2004: 148) to be a good predictor of fluency. Sometimes given as non-phonation-time ratio, it is a measure that combines pause frequency and duration (Wood, 2007: 215).

Including repair in measures of fluency has been challenged. The above linking of Skehan's (2009) fluency sub-dimension ("repair") with Lennon's (2000) definition ("accurate") indicates why. Housen, Kuiken and Vedder (2012c: 300–301) ask "Does repair fluency […] really constitute a distinct subdimension of the fluency construct, or is it perhaps more closely linked to accuracy?" Given this lack of clarity and the abundance of measures available that are more clearly of fluency (and others that are more clearly of

accuracy – see Section 3.2.4), direct measures of repair were not included among the chosen fluency measures.

The fluency measures chosen for this research were pruned speech rate, mean length of run and non-phonation-time ratio. These are commonly used (allowing comparisons with other studies to be readily made) and correlate with perceptions of fluency. They are also relatively general measures, which is in keeping with this research being of something not previously studied and therefore somewhat tentative.

3.2.3 Choice of Complexity Measures

Complexity as part of L2 speech tends to be operationalised and measured rather than conceptualised or defined. Most relevant here at the conceptual level is linguistic complexity,[9] which is divided principally into grammatical and lexical forms of complexity by Bulté and Housen (2012). Lexical forms were not measured in the current research, as the data collection process required the use of two topics (see Section 3.1.1.2) and lexical item choice can be subject to variation based on task content differences (e.g., Tavakoli and Foster, 2008).

Grammatical complexity can be split further into syntactic and morphological components; the latter are operationalised using specific features of grammar such as variety or frequency of certain tenses (Bulté and Housen, 2012). Selecting such specific units for analysis has a practical difficulty, as observed by Skehan (2014: 15): in a sample of speech,

[9] The terminology in this area can be as labyrinthine as the phenomena it attempts to describe: the terms here are taken selectively from Bulté and Housen, who present a figure (2012: 23) that contains a taxonomy of 24 L2 complexity constructs.

"there may not be enough tokens to work with". This was likely to be the case in the present research, with the long-turn parts of the data collection process expected to elicit samples of speech of 1–2 minutes. Measures of complexity used were therefore syntactic.

Measuring three parts of syntactic complexity is recommended by Norris and Ortega (2009): general complexity, subordination complexity and complexity via phrasal elaboration; these can be measured using mean length of syntactic unit, mean number of clauses per syntactic unit and mean length of clause, respectively. The most appropriate 'syntactic unit' to use in these measures has been the topic of some debate. The T–unit – a main clause plus its subordinate clauses – was created to assess the syntax of L1 writing, but came to be used to analyse L2 speech (Crookes, 1990). A variation on this is the c–unit, which encompasses some features of speech that would be excluded from analysis that is based on the T–unit (Crookes, 1990). While these continue to be used, the AS–unit, which is a further variation on the T–unit but was intended for L2 speech and can be applied more consistently than the other units, has become more common (Foster, Tonkyn and Wigglesworth, 2000; Norris and Ortega, 2009).

The AS–unit was selected as the syntactic unit of analysis in this study. Lengths are typically described in words, so the complexity measures chosen for this research were words per AS–unit, clauses per AS–unit and words per clause. As with the chosen fluency measures, these are all commonly used in L2 speech research. How they were calculated, and the difficulties in doing so, are covered in considerable detail in Chapter 5 (Methods – Data Analysis); this also applies to the fluency measures and the final CAF component: accuracy.

3.2.4 Choice of Accuracy Measures

Accuracy refers to how closely language conforms to a norm (Pallotti, 2009: 592). While this could encompass pronunciation, standard practice when using CAF is to apply norms of syntax and lexis only. The norms are invariably those of L1 speakers, with the usually unstated assumptions being that these are what learners aspire to and that L1 speakers have a shared sense of what is linguistically standard. This view of L2 deviation from a norm as being a deficit may be problematic and somewhat outdated (e.g., Jenkins, 2006), but two points can be made in defence of employing L1 norms for accuracy measures (and for some measures of complexity and fluency). One is that these 'norms' are, more precisely, 'idealised norms' of flawlessness that L1 speakers can also be compared with. The other is that there is a degree of subjectivity in the measurement of all CAF values and the key is to be as consistent as possible in calculating each one, so a set of L1 norms, reliably applied to analysing L2 accuracy, is no worse than any other norm.

For accuracy, there is a choice of specific or general measures. Specific measures home in on particular constructions and grammatical forms, such as articles and past tenses, and are particularly useful when such items are targeted by the research or are certain to appear in speech samples; when neither of these conditions applies, using general measures is more appropriate (Foster and Skehan, 1996: 304). Specific grammatical features were not targeted in this research and there is little certainty over which would be used by the participants, so general measures of accuracy were chosen.

As with complexity, a segmentation unit of analysis must be chosen. A base of 100 words is sometimes used, but T–units, AS–units and clauses have greater psycholinguistic

underpinnings and are commonly used (Foster and Wigglesworth, 2016: 102). AS–units and their clauses were selected in the previous section for complexity analysis, so using one or both for accuracy analysis is consistent. Using clauses increases the chances of variation in accuracy across participants and reduces the risk of a floor effect (Foster and Skehan, 1996: 304), while using the longer AS–units reduces the risk of a ceiling effect in accuracy scores. Although floor effects were more likely in the participants in this research, using both clauses and AS–units could be beneficial.

The next choice is between counting errors per unit and labelling whole units as either error-free or not. Identifying how many errors exist in a sample of speech is notoriously difficult and potentially unreliable (e.g., Ellis and Barkhuizen, 2005: 163), but it obviously allows the frequency of errors to be reported. Labelling whole units as error-free or not, in contrast, means that multiple errors within a clause, for instance, would be labelled in the same way as a single error within a clause. There is a similar limitation with counting errors, however: if they cluster together over a small number of clauses or AS–units, then accuracy over the whole sample would be distorted by using an error count rather than whole unit labelling. On balance, then, the greater simplicity, and hence achievable reliability, of labelling AS–units and clauses as error-free or not is preferable.

Applying a weighting to errors is sometimes proposed as a solution to some of the problems discussed in the previous paragraph. A recent suggestion (Foster and Wigglesworth, 2016) is to place each clause into one of four categories based on how seriously any errors it contains compromise meaning: error-free; minor errors; serious errors; or very serious errors. The final part of the suggestion entails counting the number of errors in each category and then multiplying each total by a weighting. Pallotti (2009:

592) criticises the weighting of errors, suggesting that it confounds two constructs: accuracy and comprehensibility. Foster and Wigglesworth (2016) counter this challenge, but, overall, the drawbacks to using such a measure in the current study are that it is not clearly superior to existing general measures of accuracy, either in clarity of the construct it purports to measure or in likely coding reliability, and that it is not commonly used.

Based on the above considerations, the accuracy measures chosen for this research were: error-free AS–units and error-free clauses. Having selected measures of fluency, complexity and accuracy, this chapter continues with a summary of the overall research design, then concludes with a brief summary that leads to a statement of the specific research questions.

3.3 Overall Research Design

This section summarises the overall repeated measures factorial research design in explicit terms, based on the contents of Chapter 2 and the earlier parts of this chapter. The independent variables were participant L1 (the between-subjects independent variable) and backchannel frequency. Each of these had two levels: L1 Japanese or Mandarin Chinese; and typical or low backchannel frequency. There was a possibility of an interaction between L1 and backchannel frequency, so this needed to be accounted for in the choice of statistical analyses, which is described in Section 5.9.

The quantity of other things that could potentially influence the effects of backchannel frequency is high. These are things that were not of primary research interest in this study (i.e. they were control variables) and they could be separated into two categories: those

that could lead to exclusion from the study's sample; and those that could not lead to such exclusion. In the former category were proficiency in a third language, L2 English proficiency, age, and time spent in another country. Principal variables in the latter category were gender and regional origin. It must be reiterated that there is little evidence to support most of these actually influencing the effects of backchannels, so the sampling used in this study was not designed or conducted to enable the inclusion of any of them as independent variables in statistical analyses.

Dependent variables numbered eight and were in a total of three categories: fluency, complexity, and accuracy. The three dependent variables for fluency were pruned speech rate, mean length of run and non-phonation time ratio. The three dependent variables for complexity were words per AS–unit, clauses per AS–unit and words per clause. The two dependent variables for accuracy were error-free AS–units and error-free clauses.

3.4 Summary and Specific Research Questions

At the end of Chapter 2 the general research questions were stated: What are the effects of interlocutor backchannel rates on spoken L2 English? Do these vary with the speaker's L1? Chapter 3 has allowed these to be narrowed down by choosing a quasi-experimental design, L1 groups to use and what forms of analysis to utilise. The two backchannel rates to employ during that procedure have also been selected. Comparisons across languages can be made by using two groups of participants whose L1 backchannel norms differ maximally – Japanese and Mandarin Chinese. The most appropriate way of analysing the resulting recordings was deemed to be measurements of CAF.

These key decisions allow the more specific research questions to be stated. They incorporate how the data were analysed and the two L1 groups that were used:

1. To what extent are the complexity, accuracy and fluency of the speech of adult L1 Japanese learners of English affected by variation in an interlocutor's backchannel frequency?
2. To what extent are the complexity, accuracy and fluency of the speech of adult L1 Mandarin Chinese learners of English affected by variation in an interlocutor's backchannel frequency?

As detailed previously, there is little evidence to support the addition of further specific research questions or sub-questions: age, gender, status and so on could conceivably have an influence on backchannel effects, but the currently available evidence did not justify their inclusion as independent variables in this study. Given that this is the first study of its kind, concentrating the samples on the characteristics most likely to be of relevance – L1 and backchannel rates – is prudent. This still allowed the possibility of further analyses of the data if effects of backchannel rates were found, with such analyses having the principal goal of helping to inform future research directions. This is taken up in Section 6.3.

The next chapter is on the methods of data collection, so describes pilot studies, populations, participant characteristics, the data collection procedure itself and many other pertinent things. It is based on the literature reviewed in Chapters 2 and 3, but also utilises more of the literature where it is required for the details of the methods employed.

4 METHODS – DATA COLLECTION

This chapter details the whole data collection process, from first pilot study to final checks of the data collection methods. Populations and the two groups of participants are described, as are the materials and the data collection process itself. Sections on the recording medium, counter-balancing of the design, checks of the methods, and the procedures and outcomes of ethical matters address the strength of the research design and threats to validity. First in this chapter is an overview of the data collection design.

4.1 Data Collection Overview

4.1.1 Design

This was a quasi-experimental study that used samples from two populations, distinguished principally by their L1: Japanese or Mandarin Chinese. A repeated measures design was employed. Each participant was audio recorded in two English interactions with an interlocutor, each of which was similar in nature to IELTS Speaking Parts 2 (long-turn) and 3 (discursive) combined and differed in topic and in the rate of verbal interlocutor backchannels given (one being approximately one-third of the rate given in the other). A questionnaire was used to check for conscious awareness of the backchannel differences and to check for differences in perceived topic difficulty and differences in difficulty between the long-turn and discursive parts.

Two pilot studies were done before the main study. These are described after one feature that was the same in all three studies – the interlocutor.

4.1.2 Interlocutor

Sections 3.1.3.3 and 3.1.4 discussed the desirable interlocutor characteristics in this study. They concluded that the same person should be used, and that the person should have L1 English and, preferably, some experience of Japanese and Mandarin Chinese, but not of other languages. The interlocutor being unfamiliar with the participants was also identified as being preferable on validity grounds.

The researcher, a male L1 speaker of British English in his thirties, was the interlocutor. I lived in Japan and mainland China for at least a year, but neglected to acquire more than a simple understanding of any language other than English. I have not been an IELTS examiner, but have done a large number of one-to-one assessments of L2 English, principally for class placement purposes. This combination put me, fortuitously, in a good position to be able to minimise accommodation in backchannel behaviour while also avoiding it being similar to that of the L1 of the participants. Participants in the main study were unknown to me, so possible acquaintanceship effects were also minimised (further details are in Section 4.3.2.2.1).

4.2 Pilot Studies

Two pilot studies are reported here. The first was intended to develop the data collection procedure, including assessing and improving my actions as the interlocutor. The second used participants from the target populations and was designed to fine tune the procedure further and to check if the participant criteria, particularly their English proficiency, that were assumed to be appropriate were indeed apposite for the main study. Both pilot studies

followed the basic procedural template: similar to IELTS Speaking Parts 2 and 3 combined; and done twice, allowing a different topic and different rate of interlocutor backchannels to be used.

The following sections on the pilot studies describe briefly what was done and state the principal conclusions reached. To avoid repetition, details of changes made to the procedure following these pilot studies are included only in Section 4.3, which describes the main study.

4.2.1 Pilot Study I

The first pilot study was with four advanced L2 English speakers who had lived in the UK for at least three years. People with these two characteristics – advanced rather than lower proficiency English; considerable experience of interacting with L1 speakers of English – were chosen for two reasons. First, needing less effort to concentrate on communicating in an L2, they would be more likely to be able to identify flaws in the procedure or the actions of the interlocutor. Second, their L2 ability would be sufficient to communicate clearly in English what they had noticed and they would be more likely to have the confidence to assert these things. Their L1s were Japanese (1), Mandarin Chinese (2) and Polish (1).

Three matters were assessed during this pilot study: the incidence of involuntary non-verbal behaviour that could act as backchannels; the visibility of my counting mechanism for the low backchannel rate; and my ability to use verbal backchannels with all non-verbal ones. These will now be described in turn.

For each of the interactions in this pilot study, a video camera was set up to record me as the interlocutor: the framing was such that I was fully visible but the participant did not appear at all. Each person was asked to view the recording of his or her interaction. The video was shown on a 15.6-inch laptop and without sound, so that the person who took part in each recording could concentrate on the image, without being led by linguistic content. The person was asked to state when any of my actions that could be interpreted as responding to him or her occurred. Before starting, they gave as examples such things as nods, eye contact and saying 'mm'. The resulting list of actions and timings was then compared with the equivalent list from the audio recording of what would be classed as backchannels in the main study. This revealed the potential extent of involuntary non-verbal behaviour that could act as backchannels, as perceived by the people who received it. One person identified occasional small head movements, but these were all immediately before a verbal backchannel. Another perceived one frown and one small nod (from a total of 70 responses). The other two participants did not report any involuntary non-verbal responses. Based on these reports by the people who took part in the whole data collection procedure, there was no evidence of widespread or systematic involuntary behaviour that could act as backchannels. There was evidence that some occurred, so there was a possibility that some would occur during the main data collection.

At the end of the nascent data collection procedure, each person was asked to comment on the procedure and my actions during it, including the relative rate of backchannels I had given during the two sets of interactions. One reported being aware of differences in my backchannel rate, but attributed this to the topic rather than the study's design; she also misidentified which interaction had contained the lower backchannel rate. Another correctly identified which interaction had contained the lower backchannel rate. The others

reported noticing no difference in the rate, although one rate was approximately three times that of the other. None reported noticing any conscious effort from me to alter the backchannel rate. Nothing in this regard was mentioned from watching the video recordings, either. The invisibility of my counting mechanism for the low backchannel rate was, therefore, provisionally confirmed.

Combining participant feedback from viewing the video recordings with my own separate viewing allowed the final assessment – of my ability to use verbal backchannels with all non-verbal ones – to be made. With the first participant, two of the 48 backchannels that I consciously gave were non-verbal only (one was a nod; the other was a nod with a mouthed but silent 'ok'). With the other three participants, none of the backchannels was non-verbal only. The evidence illustrated that I had a high degree of control in my efforts to use verbal backchannels with any non-verbal ones during the data collection procedure. This matter is returned to from the end of the next section.

4.2.2 Pilot Study II

The second pilot study was with 10 people who met the preliminary criteria for participation in the main study: intermediate speaking proficiency; no more than six months living in another country; and no more than basic proficiency in a third language. Audio recordings were made of each of the interactions. As in the first pilot study, the basic procedural template was followed, participants completed a questionnaire and they were asked for comments on the procedure.

The 10 pilot study participants were Japanese. Six were female and four were male. They were 19–21 years old and university students in Japan studying a variety of non-language subjects. None had taken IELTS. Two Chinese participants were also included, but they had been in the UK for more than six months, so, although no data collection differences were identified between them and the Japanese participants, they are not considered further. The most important conclusion from this pilot study was that those at the lower end of the initially targeted intermediate speaking proficiency were likely to be too limited to be used in the main study. The matter of L2 proficiency also needed to be put into more concrete terms for the main study. Based on the discussion of appropriate L2 proficiency for this research in Section 3.1.3.1, and the findings of this pilot study, the targeted proficiency was refined to IELTS Speaking 4.5–6, or equivalent. How this was put into operation for the main study is described in Section 4.3.2.2.1. This was still a relatively broad range, so was likely to lead to comparatively high variation within each group, reducing the study's statistical power. However, its chief advantages were that it was appropriate to this area of research being still nascent and it would allow for the findings to be more widely applicable than if a narrow range of proficiency had been selected.

Changes made to the data collection procedure itself and to the questionnaire as a result of this pilot study are described in the remainder of this chapter and are concentrated in Section 4.3.2.2 (Procedure). The next section discusses one matter that pertains to the basic nature of the data collection and emerged from the two pilot studies. This is the question of whether the recordings should be made using audio and video or only audio.

4.2.3 Recording Medium

Although there was good evidence suggesting that I could limit my non-verbal backchannels to almost none during the data collection procedure (see Section 4.2.1), a video recording of every interaction during the main study would have allowed this to be checked definitively. As two people in the first pilot study suggested that participants with more limited English proficiency might not feel comfortable if a video camera were in the room, some of the participants in the second pilot study were asked, after participating, about their attitude to the use of video recording. All responded that they would have felt considerably more nervous with the presence of a video camera, even if they knew that they would not appear in the recording at any time, and that this would have affected their performance. One person, for instance, said "I can't [couldn't] talk" when asked how she would feel if a video camera were present, and "I prefer this style", referring to the use of only audio recording. They also believed that their peers would have a similar reaction. Another person, for example, said "if other people saw that there was a video, they might feel nervous, like me".

Anxiety is known to affect L2 performance (e.g., MacIntyre and Gardner, 1991). Some anxiety was inevitable in participants in this study, as they were speaking individually for around 30 minutes to an L1 speaker of English who they did not know and who was studying their English use. It was clear from the comments of the pilot study participants who were from one of the target populations that the use of video, even when not of them, would heighten their anxiety to an extent that would adversely affect their language use.

There was thus a conflict between twin challenges to the validity of the main study: the participants being sufficiently relaxed that they could speak in a way that was representative of their ability (requiring audio recording only), versus being able to check the extent to which they received only verbal backchannels (requiring audio and video recording). Both theoretical and practical considerations pointed to the first of these concerns being prioritised. If the principal data – participants' speech – would be adversely affected by the circumstances in which it was collected, then maximising the validity of the checking of the data collection procedure would be futile. Practically, it would not be possible while conducting this study to monitor the extent to which participants' speech was affected by the presence of a video camera, but it would be possible to conduct checks on the interlocutor's use of non-verbal backchannels, by using video recording for a small number of interactions over the course of the main data collection period. On the basis of this, the main study's data collection was recorded via audio only, but some video recordings with participants from the target populations were also made during the main data collection phase, so that my actions with members from those populations could be checked. This checking of data collection methods is described in Section 4.4.2.

4.3 Main Study

This main study section describes, first, the two populations and samples, then the settings in which the data collection took place. What materials were used and how they were selected then precede details of the procedure itself. A sub-section on counter-balancing the design concludes.

4.3.1 Populations

There are two populations. One has L1 Japanese and is from Japan and the other has L1 Mandarin Chinese and is from mainland China. Both populations are adults (over 18 years of age), have no more than basic speaking proficiency in another language (excluding English) and have limited experience of being in another country.

4.3.2 Samples

4.3.2.1 Sample Sizes

The two known studies that have investigated the effects of backchannels on L2 speech using quantified measures of fluency – Wolf (2008) and Flint (2010) – found statistically significant differences in at least one measure of fluency using sample sizes of 14 and nine, respectively. Both reported, however, that these were small samples, limiting power. Neither study calculated measures of complexity or accuracy, so further precedents for sample sizes were sought in recently published studies of L2 speech in controlled settings that used such measures.

Two edited volumes (Robinson, 2011; Housen, Kuiken and Vedder, 2012a) that contain research of that kind were searched for studies that clearly present sample sizes of L2 speakers and several measures of the complexity, accuracy or fluency of their speech. The sample sizes used (per group in each study identified in these two volumes) were: 189 (de Jong et al., 2012); 44 (Kormos and Trebits, 2011); 44 (Kuiken and Vedder, 2011); 42 (Levkina and Gilabert, 2012); 42 and 40 (Gilabert, Barón and Levkina, 2011); 41 (Albert, 2011); 32 (Michel, 2011); 24 (Tonkyn, 2012); 24 (Ishikawa, 2011); and four (Ferrari,

2012). Most of these studies put participants with different L1s into the same group. Stated justifications for the sample sizes are rare. Excluding the highest and lowest, the range of sample sizes in these examples of published research was 24 to 44. The upper part of that range was targeted for this study. As detailed later in Section 4.3.2, the eventual sample sizes were 37 for the Japanese group and 34 for the Chinese. These were slightly below the target because a substantial number of people who took part were excluded, chiefly because they had spent more than the maximum permitted period overseas or because they spoke another language to more than a basic level of proficiency. Nevertheless, these sample sizes were within the typical range and likely to be adequate to identify differences that may exist in the observed values of the L2 speech measures used.

4.3.2.2 Participants

All participants were taking an English language course in the UK. The theory behind the required and other checked characteristics of the participants is mainly in Chapter 3 (Section 3.1.3). Here, these characteristics are briefly described, with the sampling and recruitment of the participants.

4.3.2.2.1 Sampling and Recruitment – Required Characteristics

Participants were recruited through language schools and universities in the southern half of England; the sampling type was purposive. All participants were unknown to me. In some instances, I met potential participants to arrange where and when they would take part. These congregations were brief and days in advance of the data collection meetings, so are unlikely to have had any meaningful effect on the later interaction.

Gatekeepers in the institutions that were approached were informed of the general nature of the research (that it was a study of L2 spoken English that compared L1 speakers of Japanese and Mandarin Chinese), but not of the detail (that backchannels were being studied, or that CAF measures would be calculated). Gatekeepers were also informed of the key required characteristics of potential participants and that participation was voluntary. The required characteristics of each participant are listed in Table 4.1.

Table 4.1 Summary of required characteristics of participants

Country of origin	Japan or China (mainland)
L1	Japanese or Mandarin Chinese
L2+	English; no more than basic in any other
English proficiency	IELTS 4.5–6
Age	18+
Time in English-speaking / other countries	≤ 6 months

Gatekeepers were asked to restrict the selection of participants to those who met the required characteristics. The broad category of 'intermediate' L2 English proficiency was selected as being the most appropriate for participants in this study in Section 3.1.3.1 and refined to IELTS 4.5–6 or equivalent following the second pilot study. 'Intermediate' would have been difficult for gatekeepers to interpret and implement, so they were given 'IELTS 4.5–6' as the target range, as they were likely to be familiar with IELTS and how their own institutions related internal or other assessments of proficiency to it. They reported using their databases to identify potential participants; these contained information on country of origin or nationality, L2 proficiency as measured by IELTS, TOEFL or other assessments, and age. Most of the characteristics were also checked as part of the data collection process.

4.3.2.2.2 Checking of Characteristics

English proficiency was checked by reference to the gatekeepers' information or, where no formal assessment was obtainable, by first comparing a participant's performance in the data collection process with publicly available descriptors of IELTS Speaking bands. The study did not have the financial resources to permit the formal assessment of participants who did not have a relevant major language test score. A combination of solutions was applied: providing experienced gatekeepers with a readily understandable guideline; comparing performance with publicly available descriptors; plus comparing performance with that of participants who did have a known major test score and retaining the option at the stage of data analysis of removing performances that led to outliers. While less than ideal, the wording '…, or equivalent' is a very common inclusion in descriptors of proficiency in the applied linguistics research literature; the proficiency of participants in this study was 'IELTS 4.5–6, or equivalent'.

The other required characteristics were checked by asking each participant directly. To minimise potential influences from other languages, participants should have spent a limited amount of time in other countries; which countries the person had visited, the length of each stay and how long the person had been in the UK at the time of participation in this study were recorded. All who started the data collection process continued to the end of it, but some were subsequently rejected because they had not met all of the required characteristics.

In addition to the required characteristics, others of possible relevance were also checked and are summarised here, as recommended by Mackey and Gass (2016: 177). Potential

influences from other languages were also controlled by limiting eligible participants' speaking proficiency in a third or further language to no more than basic. Also recorded were what part of Japan or China each participant was from; education level attained; subject studied if a university student; and knowledge and experience of taking IELTS. All of these were self-reported.

What part of the country each participant was from was collected even though there is very little evidence for this being important. Such evidence may appear at some point in the future, so being able to (re)assess the data based on this could be useful.

Ideally, participants would have approximately equivalent education backgrounds. This is based on Mulder and Hulstijn (2011), who found that education and profession affected performance in L1 Dutch speaking tasks to a greater extent than did age. In the target populations, the level of L2 proficiency required was likely to be found most readily among those who were studying at, or had graduated from, university.

4.3.2.2.3 Japanese Participants

Full details of the Japanese participants are given in Table 4.3, below. 26 were female and 11 were male. Their ages were very similar, ranging from 18 to 22. One had finished high school; the others were students at universities in Japan. Both time in the UK at the date of participation and time spent outside Japan varied widely: from two weeks to six months for each. Only one person had taken IELTS before; of the others, most knew nothing or very little about it, four stated that they knew something of its structure and two were preparing to take it. TOEFL scores were available for a large minority and could be

mapped to IELTS 5–6 (this range mainly comes not from the range of TOEFL scores, but from inconsistencies in mapping any TOEFL score to IELTS. ETS (2010) maps, for example, TOEFL 61 to IELTS 6, but many UK universities (e.g., Cardiff University, n.d.; University of Sheffield, n.d.) equate it to IELTS 5). Five participants stated that they had started studying English at elementary school; the others indicated having started at junior high school. They were from a high number of locations; the most common (nine participants) was Osaka. All but five of the participants had been overseas previously; most of them had visited several other countries, usually for short periods. 15 reported some knowledge of a third language; proficiency in these is reported as "basic" in Table 4.3, even when the person claimed to know only a few words of the language in question. Students at Japanese universities typically study two foreign languages, so the students who said that they knew no other language may have had the same proficiency in a third language as reported by the others, i.e. very limited. Four of the 37 said that they knew a basic amount of Chinese; all of these used the word "little" to describe their knowledge of Chinese – the example given by two was that they were limited to a self-introduction. None had spent more than a week in China. The university subjects that they were studying varied, with the most common being tourism and international studies.

4.3.2.2.4 Chinese Participants

Full details of the Chinese participants are given in Table 4.4, below. 15 were female and 19 were male. Their range of ages was 18–32, but all except four were aged 18–22. Most (24) were students at universities in China; of the others, six had graduated and four had completed high school. Time spent outside China was two weeks to five months; time in the UK at the date of participation was two weeks to four months. All but one were

familiar with IELTS; 30 had taken it at least twice. Their IELTS Speaking scores were in the range 5–6, with only two having a score of 6. Primary school was the most common starting point for their English studies (25 participants); others had started at middle school (six) or high school (three). They were from a high number of locations; the most common (eight participants) was Jiangsu Province. Only 13 had been overseas previously, most commonly to Asian countries. Nine reported some knowledge of one or more additional languages; proficiency in these is reported as "basic" in Table 4.4, even when the person claimed to know only a few words of the language in question. Five of the 34 said that they knew a basic amount of Japanese, being limited to greetings, a few words, or a few sentences, in each instance. None had spent more than two weeks in Japan. The university subjects that they were studying varied, with the most common being related to business or finance.

4.3.2.2.5 Comparison of Japanese and Chinese Participants

A summary of the numerical features of the separate tables for Japanese and Chinese participants is given in Table 4.2 for comparative purposes and includes means and standard deviations. The Chinese participants were slightly older than the Japanese ones, averaging 22.1 years versus 19.7. Even allowing for the four Chinese participants who were older than the typical 18–22 range, that group was older than the Japanese one and most were in their fourth year of university, rather than the Japanese students' second year. The gender balance was different: 70% of the Japanese group were female, compared with 44% of the Chinese group. While the shortest period spent outside their own country was the same across the two groups, the longest period and means varied. Experience of IELTS was perfectly opposite: all but one Chinese person had taken it, while all but one of the

Japanese people had not. The implication of this final point was that the participants would need to be given basic information on IELTS Speaking prior to their taking part; this is described in Section 4.3.5.

Table 4.2 Summary of participant characteristics

	Japanese	**Chinese**
Number	37	34
Gender	26 female; 11 male	15 female; 19 male
Age	18–22	18–32
	Mean 19.68; SD 0.88	Mean 22.09; SD 2.66
Time spent outside own country	2 weeks – 6 months	2 weeks – 5 months
	Mean 10.22 weeks; SD 5.99	Mean 6.94 weeks; SD 4.10
Length of stay when participated	2 weeks – 6 months	2 weeks – 4 months
	Mean 7.62 weeks; SD 4.99	Mean 5.76 weeks; SD 3.43

Table 4.3 Japanese participants

Participant	Gender	Age	Other languages spoken (proficiency)	Time spent outside Japan	Length of stay when participated	Education
J1	F	20	None	2 weeks	2 weeks	Uni 2nd year
J2	F	20	None	2 weeks	2 weeks	Uni 2nd year
J3	M	19	German (basic)	2 months	6 weeks	Uni 1st year
J4	F	19	Chinese (basic)	2 weeks	2 weeks	Uni 1st year
J5	M	20	Chinese (basic)	5 months	4 months	Uni 2nd year
J6	F	19	None	2 weeks	2 weeks	Uni 2nd year
J7	F	19	Spanish (basic)	2 months	6 weeks	Uni 2nd year
J8	F	20	None	4 months	4 months	Uni 2nd year
J9	F	20	None	1 month	2 weeks	Uni 2nd year
J10	F	19	German (basic)	5 weeks	1 month	Uni 2nd year
J11	F	18	Korean (basic)	3 months	2 months	Uni 1st year
J12	M	22	French (basic)	4 months	3 months	Uni 3rd year
J13	F	19	Korean (basic)	4 months	5 weeks	Uni 2nd year
J14	F	18	None	2 weeks	2 weeks	Uni 1st year
J15	M	19	None	1 month	1 month	Uni 2nd year
J16	M	21	French (basic)	4 months	3 months	Uni 2nd year
J17	M	21	None	5 weeks	5 weeks	Uni 3rd year
J18	F	20	Chinese (basic)	5 months	6 weeks	Uni 2nd year
J19	F	20	None	6 weeks	6 weeks	Uni 2nd year
J20	F	20	None	6 weeks	5 weeks	Uni 2nd year
J21	F	19	None	6 weeks	3 weeks	Uni 2nd year
J22	M	19	None	2 months	6 weeks	Uni 2nd year
J23	F	20	None	4 months	1 month	Uni 2nd year
J24	F	20	None	4 months	3 months	Uni 2nd year
J25	F	20	None	3 months	3 months	Uni 1st year
J26	F	19	None	3 months	2 months	HS finished
J27	F	19	Chinese (basic)	3 months	3 months	Uni 2nd year
J28	M	20	French (basic)	4 months	3 months	Uni 2nd year
J29	M	20	None	3 months	1 month	Uni 2nd year
J30	F	19	None	2 months	6 weeks	Uni 2nd year
J31	F	22	Spanish (basic)	6 months	6 months	Uni 3rd year
J32	F	20	French (basic)	1 month	1 month	Uni 2nd year
J33	M	20	None	4 months	3 months	Uni 2nd year
J34	F	19	None	2 months	2 months	Uni 2nd year
J35	F	20	German (basic)	3 months	3 months	Uni 2nd year
J36	M	20	None	4 months	3 months	Uni 2nd year
J37	F	19	None	2 months	2 months	Uni 2nd year

Note: Gender: F = female; M = male. Education: HS = high school; Uni = university, year of study at time of participation.

Table 4.4 Chinese participants

Participant	Gender	Age	Other languages spoken (proficiency)	Time spent outside China	Length of stay when participated	Education
C1	F	22	None	2 months	1 month	Bachelor's degree
C2	M	22	None	2 weeks	2 weeks	Uni 4th year
C3	M	30	Shanghai-hua (basic)	2 months	2 months	Bachelor's degree
C4	M	22	None	2 months	2 months	Uni 4th year
C5	M	22	None	2 months	2 months	Uni 4th year
C6	M	21	None	2 months	2 months	HS finished
C7	F	21	None	2 months	2 months	Uni 4th year
C8	M	21	None	1 month	1 month	Uni 4th year
C9	F	21	None	1 month	2 weeks	Uni 3rd year
C10	M	18	None	1 month	1 month	HS finished
C11	M	22	Japanese (basic), Russian (basic)	2 weeks	2 weeks	Uni 4th year
C12	M	22	None	2 weeks	2 weeks	Uni 4th year
C13	M	22	None	5 weeks	3 weeks	Uni 4th year
C14	F	24	None	2 months	2 months	Bachelor's degree
C15	M	22	None	2 weeks	2 weeks	Uni 4th year
C16	M	22	None	2 months	2 months	Uni 4th year
C17	M	22	None	2 months	2 weeks	Uni 4th year
C18	F	21	None	2 weeks	2 weeks	Uni 3rd year
C19	M	22	None	2 weeks	2 weeks	Uni 4th year
C20	F	21	None	3 weeks	3 weeks	Uni 3rd year
C21	F	22	None	1 month	1 month	Uni 4th year
C22	M	22	Japanese (basic)	2 months	2 months	Uni 4th year
C23	F	19	French (basic), Japanese (basic)	3 months	3 months	HS finished
C24	M	22	None	2 months	2 months	Uni 4th year
C25	F	22	Japanese (basic), Korean (basic)	2 months	2 months	Uni 4th year
C26	F	22	None	2 months	2 months	Uni 4th year
C27	F	22	French (basic)	1 month	2 weeks	Uni 4th year
C28	F	22	None	2 months	2 months	Uni 4th year
C29	F	19	German (basic)	5 months	2 months	HS finished
C30	M	32	Shanghai-hua (basic)	6 weeks	2 weeks	Bachelor's degree
C31	F	25	Japanese (basic)	4 months	4 months	Bachelor's degree
C32	M	22	None	3 months	2 months	Uni 4th year
C33	M	22	None	6 weeks	6 weeks	Uni 4th year
C34	F	18	None	3 months	2 months	HS finished

Note: Gender: F = female; M = male. Education: HS = high school; Uni = university, year of study at time of participation.

4.3.3 Settings

Features of data collection venues can affect psychological factors that may influence participant behaviour (Rosenthal, 1966), including in research into speech (Griffiths, 1991). Steps were therefore taken to make the participant experience as similar as possible across the period of data collection.

Adjustments were made to heating, ventilation and lighting to provide a comfortable environment in the room used. Practical matters meant that five locations were used; these were all small rooms with at least one window. Attempts to minimise environmental noise were also made, so that participants would not be disturbed and the audio recording would be as clean as possible. Some background noise was inevitable, however, given the locations used; this was most commonly from passers-by and traffic.

From the beginning of each meeting, no other people were permitted in the room. I wore very similar clothing to each session. A desk was positioned between two chairs that were behind but also slightly to one side of it, so that the desk could be utilised while not appearing to be a complete barrier. The participant's chair was placed so that the person would not be distracted by anything occurring outside any windows. The recording device was placed on the desk after consent had been obtained from the participant.

4.3.4 Materials

Materials were needed for the main data collection procedure and questionnaire. For the former, topics to be used in the equivalent of IELTS Part 2 (long-turn part) and questions

and prompts for the equivalent of Part 3 (discursive part) were needed. These are described in this section.

4.3.4.1 Main Procedure

For the long-turn part, two topics of equivalent difficulty were needed, as a difference in difficulty can affect research validity (Mackey and Gass, 2016: 167–168). Two topics from Weir, O'Sullivan and Horai (2006), which describes an analysis of nine Part 2 topics provided by IELTS and the identifying of four of them as being of equivalent difficulty, were used. The two from these four that appeared to be most appropriate for the participants in this study are shown in Figure 4.1, below, modified slightly to reduce the risk of some words being unknown (e.g., "explain why you particularly remember this event" was changed to "explain why you remember this event well").

Given the length of the process that Weir, O'Sullivan and Horai (2006) used to evaluate topic difficulty and related elements such as participants' perceptions of the topics (54 learners using a total of nine topics and completing a questionnaire), utilising topics already judged by others to be equivalent was preferable to attempting it independently. According to Google Scholar, this study had been cited four times prior to main data collection and seven times by its end. None of these used, reproduced or discussed any of the topics that were employed in the current research: although the topics were generic, there was no reason to believe that participants would have been exposed to them in this form previously.

Topic A:

Describe an enjoyable event that you experienced when you were at school.
You should say:
what the event was
when it happened
what was good about it
and explain why you remember this event well.

Topic B:

Describe a film or a TV programme which had a strong effect on you.
You should say:
what kind of film or TV programme it was (e.g., comedy)
when you saw it
what it was about
and explain why it had a strong effect on you.

Figure 4.1 IELTS Speaking Part 2 topics of equivalent difficulty (slightly modified from Weir, O'Sullivan and Horai, 2006: 11)

For the discursive part, questions and prompts for the topics shown in Figure 4.1 were prepared. These were a mix of questions on the participant's direct experience and more general topic-related issues, and were intended to allow the participants to use a variety of structures and tenses. For example, for topic A, in which a common theme, identified from the pilot studies, was participation in a school club or society:

What do other people say or remember about that event?

What were the good points and bad points about your school?

How important are clubs or societies at schools in [country]?

Do you think this will change in the future?

In the pilot study, some participants interpreted "school" in Topic A to be 'university'. This interpretation would not change the nature of the topic and in fact allowed a wider range of experiences to be drawn on, so the wording was not changed.

Pilot study feedback from two items in the questionnaire did not provide evidence for one topic being perceived of as being more difficult than the other: three participants indicated that Topic A was more difficult; four that Topic B was more difficult; and three that they were of about the same difficulty. When reporting on the questions asked during the two-way discursive parts, two stated that the questions asked when discussing Topic A were more difficult; three that those for Topic B were more difficult; and five that they were of about the same difficulty. This feedback indicated that there was no need to change the topics selected. Notwithstanding the evidence of equivalence between the topics, a final statistical check was made when analysing the data (see following chapter) and the same items were included in the final questionnaire.

4.3.4.2 Questionnaire

Several of the studies in the systematic review of backchannel rates (Flint, 2012) used a questionnaire to collect information on the participants' impressions of completed interactions. For the current study, a brief questionnaire (in Appendix 4) was used immediately after the main data collection procedure. This was intended to help determine whether or not there was conscious awareness of the backchannel differences during the procedure and to check for differences in perceived topic difficulty and differences in difficulty between the long-turn and discursive parts. Minor wording changes for clarity were made to the original questionnaire following feedback from the pilot studies.

Each of the five items was expressed as a question that compared two things (e.g., the difficulty of the long-turn parts versus the difficulty of the discursive parts) and could be answered by choosing one of four statements. These statements equated to: A

approximately equal to B; A more than B; B more than A; don't know. Space was left after each item for comments to be added. The key item – on backchannel differences – was positioned at number four. The other items therefore acted in part to conceal what was being studied, which is an important research consideration (Eckert, 2013: 15–16); also see Section 4.3.5.4 for more ways in which this was done. The questionnaire fitted on one page, so both looked and was short, as recommended by Dörnyei and Taguchi (2010: 12–14).

4.3.5 Procedure

After a short period of introductions and conversation about what they had done earlier that day, participants were given basic information about the research and asked to read and sign a consent form (see Section 4.6 for further research ethics details). Participants were then given a page to read (reproduced in Appendix 2) that contained a summary of the procedure, its timings and a brief outline of IELTS Speaking skills. Any questions that they had about these things were answered prior to beginning the procedure itself. This was intended to ensure that everyone knew what would happen and that everyone had at least a minimal background understanding of IELTS Speaking. This was important because instructions to participants need to be clear and adequately understood, as part of the actions taken to control threats to validity (Mackey and Gass, 2016: 169–170). The participants were also assured that the procedure was not a test and that the study was designed to compare the English of L1 speakers of Japanese and Mandarin Chinese, not to examine the English of any individual participant.

The data collection procedure with each participant consisted of four stages and was comparable to doing Parts 2 and 3 of IELTS Speaking twice. All stages were audio recorded and conducted only in English. These stages and their timings were:

1. Introductory stage	~5 minutes
2. First topic (long-turn and discursive parts)	7–9 minutes
3. Break	~1 minute
4. Second topic (long-turn and discursive parts)	7–9 minutes

The planned length of these stages totalled approximately 20–25 minutes. The four stages will now be described.

4.3.5.1 Stage 1: Introductory Stage

In the introductory stage, biographical information mentioned in Section 4.3.2.2.2 was collected verbally, via a question-and-answer format. This stage also served as a warm-up period, so some open-ended questions (e.g., "What do you think of your [university] subject?") were included to help the participants become accustomed to expanding on answers and voicing opinions, which would be required in the next stages.

Overall, during this stage not many backchannels were given, as the participants' responses were typically brief. The frequency with which they were given was targeted to be between the rates that the participant would receive when talking about the two topics, so that the frequency experienced during the first topic would not be greatly different from that encountered during this introductory stage.

The key information from the answers that the participants gave was written down on a prepared sheet of paper attached to a clipboard. This was also done in part to accustom the participants to their interlocutor looking at the clipboard and occasionally writing something down in the other stages of the data collection procedure. The scripts used during stages 2–4 were on a separate page, also attached to the clipboard, as was the digital watch used for timing. This positioning of the watch was a change from the first pilot study procedure, as one person who took part in that commented that I had several "distracted looks" down and to the left when listening to her speak. These looks were at the watch, which was on the table between us, so for subsequent interactions (including for the second pilot study), and at this person's suggestion, the (strapless) watch was fixed to the top of the clipboard.

4.3.5.2 Stage 3: Break

The break between stages 2 and 4 was approximately one minute in length. This was short to avoid distractions intruding and anxiety rising. Having a brief break would also help to offset any possible fatigue or boredom effects that were mentioned in Section 3.1.1.2. During this period, I prepared for the next stage by removing the pieces of paper used in the previous stage and bringing to hand those required for the next stage (these contained the topics and my questions), and wrote parts of the times for the following stage. These actions meant that continued interaction with the participant was minimised. During the pilot study, telling the participants that I was going to move some pieces of paper around (this was both required and symbolic of a change to a new topic) meant that there was no further speaking during this stage, so this was also done in the main study.

4.3.5.3 Stages 2 & 4: Main Data Collection

Each of these stages was similar to IELTS Speaking Parts 2 (long turn) and 3 (discursive) combined; the only difference between the stages was the topic. A script was used to introduce the long-turn part, to ensure that all participants received the same information (all scripts referred to in this section are in Appendix 3). There were two opportunities for the participants to ask questions so that the procedure was clear: when the information in Appendix 2 (on the procedure) was given to them; and after the reading of this scripted introduction to the first long-turn part. Some had questions, which were easily answered; none of the participants subsequently showed a lack of understanding of what was to happen.

After this scripted introduction, the participant was given a piece of paper containing the prompt for one topic from Figure 4.1. A separate piece of paper and pencil were available to the participant for the taking of notes during the 1-minute preparation period. This duration was based on IELTS timings and was strictly enforced, as increases in planning time can influence CAF values (e.g., Mehnert, 1998). After this, the person was asked to begin speaking; referring to any notes taken was permitted. This was followed immediately by a short, scripted section leading to the discursive part. In this part, prepared questions and prompts were adapted to the content of what had been said in the long-turn part (see Section 4.3.4.1 for examples). These typically began with requests for further detail on what had been described, progressed to questions on personal experiences and ended with broader or more abstract matters.

For each participant, I, as the interlocutor, used a typical backchannel rate during one of the topics and targeted giving one-third of the typical rate in the other. The typical rate was what I would typically give in that setting, but was not a constant. For the low backchannel rate, the pattern from the beginning of each long-turn and discursive part was: 'omit – use – omit – omit – use – omit – omit – use' etc. (i.e., omit the first one, then actually give every third backchannel that I would typically give). This was to standardise the procedure and to avoid a long interval before the first backchannel was given.

The backchannels given were limited to 'yeah', 'yes', 'ok', 'right', 'uh-huh', 'mm', 'mhm' and similar sounds. This avoided both assessments such as 'wow' and 'really' and non-scripted uses of words that could have been interpreted as feedback on performance, such as 'good' and 'fine'. In keeping with the observations of the literature in Sections 2.5.2 (Sequencing) and 2.5.3 (Prosody), several other things were done. First, a backchannel was given only singly, not in clusters: for instance, 'yeah', but not 'yeah, yeah' or 'ok, yeah', to avoid the possibility of indicating varying degrees of involvement. Second, the backchannel used was varied, to avoid serial repetition indicating disinterest. Finally, the prosodic features of the backchannels were restricted to those indicative of their core uses (indicating listening, that the main speaker need not stop, and not assessing the content – see Section 2.2.2). The first two of these were always achieved. A very small proportion of the backchannels given had prosodic features more likely to indicate surprise, amusement, or lack of surprise (in the sense that what was said was what had been anticipated). As these were rare occurrences that were natural responses to the content of the main speaker's talk, their impact on the study's results will have been negligible.

There were a few instances of a participant asking a question during a long-turn part. These were mostly of the form 'do you know _____?', asking about a person, place, or thing. They were responded to honestly, with reassurance, if ignorance had been professed, that there was no need to explain further.

The fact that the participants did the same thing twice, albeit with different topics, introduced the possibility of sequencing effects. As with checking for CAF differences based on topic (see end of Section 4.3.4.1), statistical checks for sequencing effects were made when analysing the data and are reported in the following chapter. Another possible repetition-related challenge to validity was the confounding of task topic and task mode (Tavakoli, 2016: 140): practice effects arising from each discursive part being based on the topic just spoken about in the long-turn part. As this study was designed to examine the effects of backchannel frequency on speech, not to examine directly CAF differences attributable to task mode, this was not a threat to validity, but did need to be considered when interpreting results.

4.3.5.4 Integrity of Procedure

People tend to become highly aware of the giving and receiving of backchannels when the phenomenon is pointed out to them.[10] When natural behaviour is required of participants in a study, concealing the research purpose may be necessary (Mackey and Gass, 2016: 168). This was the case here, as participants' knowing that backchannels were the focus of this research could have affected their behaviour and thus been a threat to validity, as could

[10] This is based on the researcher's observations of people's behaviour immediately after the topic of this research has been described to them. The effect appears to be short lived, but this may not be the case.

knowledge in advance of the topics used. These threats can never be eliminated, but some steps were taken to try to avoid them.

At the beginning of a recording session, the participants were asked what they knew about the nature of the study being conducted. This was to ensure that all of them would know the main points (if they did not know, they were told that it was a study that compared the spoken English of Japanese and Mandarin Chinese speakers) and to check that they had not been passed details of the procedure by others who had participated earlier. At the end of the procedure, each participant was asked not to discuss with other people what had happened until after the date on which the data collection process at that institution was due to end. No-one reported having been told details of the study and there were no clear signs of anyone being aware of, or specially prepared for, the topics that were discussed.

4.3.6 Counter-Balancing of Design

Sequencing can have an effect when all of the participants in a study are required to do the same thing but in different orders. The main way of dealing with this is through counter-balancing the design (Mackey and Gass, 2016: 197–198).

The sequences of topics and backchannel rates that were used are listed in Table 4.5 below. Each participant experienced one of these sequences; having approximately equal numbers of participants doing each of the sequences would result in a counter-balanced design. To minimise further any possible sequencing effects, this time related to interlocutor behaviour, which sequence was used first in each data collection session was varied. For instance, for the second pilot study, which involved meeting two participants a day, this

meant using the following sequences: 1, 2 (day 1); 3, 4 (day 2); 4, 3 (day 3); 2, 1 (day 4), then back to the day 1 sequence. This balanced which backchannel rate and which topic was used first over the period of data collection.

Table 4.5 The four sequences of topics and backchannel rates

Sequence	First	Second
1	Topic A / Typical	Topic B / Low
2	Topic A / Low	Topic B / Typical
3	Topic B / Typical	Topic A / Low
4	Topic B / Low	Topic A / Typical

The balancing of the four sequences was done separately for the Japanese and Chinese data collection, as they were to be analysed separately. This was usually straightforward, as data collection tended to occur in small clusters of each L1 at a time. Where balancing the four sequences clashed with which sequence was to be used first in each data collection session, the former was prioritised.

Table 4.6 Number of L1 participants who followed each sequence

Sequence	# Japanese	# Chinese
1	9	8
2	8	9
3	10	8
4	10	9

As can be seen in Table 4.6, the actual numbers of participants who followed each sequence were not as balanced as they could have been. This occurred because some people who took part were excluded because they did not meet all of the required criteria for inclusion in the study and the adjustments to the sequencing used towards the end of the data collection period were insufficient to overcome the imbalance introduced by their earlier exclusion. Nevertheless, the differences were small, as the most distant from the

ideal (25% of the total number of participants within each L1) was Sequence 2 for L1 Japanese, which, at 8/37, was 21.6%.

This ends Section 4.3, on the main study. Next are details of how backchannels produced during the data collection method were checked and what those checks revealed.

4.4 Checking of Data Collection Methods

Three checks of backchannel production were required. Checks for non-verbal backchannels were conducted before, during and after the data collection process. The first check of verbal backchannels was how close the difference between the typical rate and the low rate came to the intended ratio of 3:1. The second check of verbal backchannels was to compare the typical rate of backchannels that the participants received with backchannel rates reported in other studies, to ascertain how 'typical' they really were. These three checks are reported in the next two sections.

4.4.1 Verbal Backchannels

The mean ratios of backchannels given during the typical rate interactions in comparison with the low rate were calculated for words per backchannel and backchannels per minute and are listed in Table 4.7. In the long-turn parts, the ratio for the word-based measure was 2.49 and it was 2.48 for the time-based measure; for the discursive parts, the respective ratios were 2.65 and 2.72. All these ratios are lower than the intended value of 3. This is probably attributable to a combination of the short interactions and starting to give backchannels in the low rate interaction at the second rather than third possible instance

(as described in Section 4.3.5.3). This interpretation is supported by the ratios being closer to the target of 3 in the discursive parts, which were longer.

Table 4.7 Ratio of backchannels given at typical and low rates

	Long-turn parts		Discursive parts	
	Word ratio	**Time ratio**	**Word ratio**	**Time ratio**
Mean	2.49	2.48	2.65	2.72
SD	0.53	0.52	0.53	0.57
Range	1.70–3.88	1.69–3.87	1.71–3.77	1.72–4.00

While the ranges may appear to be wide, they must be considered in the context of short interactions. A small number of backchannels was often given during the low rate interactions, so a small change in this number could greatly affect the resultant ratio. For instance, the ratio of 4.00 shown as the maximum ratio in Table 4.7 would have dropped to 3.33 if just one more backchannel had been given in one of the interactions. Some interactions ended when a backchannel was about to be given, so the effect of such a circumstance on the calculated backchannel rate was not just hypothetical. The standard deviations were small, indicating that very high and low ratios were unusual. The consistency of the ratio of backchannels in the typical and low rates over the duration of the data collection process was also examined. This was done by taking the ratios from the first four participants, the last four participants, the four participants closest to the point at which one-third of the participants had taken part, and the four participants closest to the point at which two-thirds had taken part.[11] The ratios from each set of four were compared with the overall ratios presented in Table 4.7. As the ratios in each set were found to be spread across the overall range of ratios and not clustered in one part of the range, the ratio of backchannels in the typical and low rates over the duration of the data collection process was judged to be satisfactorily consistent.

[11] More accurately, these were selected from the participants who were included in the final study's data set, as the recordings of those who did not meet the required characteristics were not transcribed.

The second check involved comparing the typical rate of backchannels that the participants in this study received with backchannel rates presented in other studies. These rates were calculated using the same two measures as just reported. For words per backchannel, all words and abandoned words were included in the count, not just those classed as fluent, because backchannels can be given during disfluent speech; this is also the recommendation of Cutrone (2013: 47).

Words per backchannel averaged 18.41 in the long-turn parts and 18.49 in the discursive parts; backchannels per minute averaged 4.66 in the long-turn parts and 4.83 in the discursive parts. Judging how 'typical' these rates are is largely speculative, as very few data are available from a comparable combination of interaction setting and L1s of the people taking part. The systematic review of backchannel rates (Flint, 2012; reported in Section 2.5.4.2) found ranges of 8.69–833.33 for words per backchannel and 0.44–32.09 for backchannels per minute, including any interaction setting and any L1. The range for dyadic, face-to-face interactions was 2.13–17.65 backchannels per minute, with L1 English rates being 3.58–9.31 (words per backchannel was infrequently available). The study most comparable to the present one is that of O'Sullivan and Porter (1996). Their methods are not made clear, but they used an L2 English oral proficiency interview that elicited both short and long answers from Japanese participants, who received backchannels from L1 English interlocutors. The frequency they received was reported to average 3.27 backchannels per minute. However, this may have included talk time from both the main speaker and the interlocutor, not just the main speaker, so it should be taken to be a minimum value when comparing it with the 4.66 and 4.83 of the current research. In conclusion, the interlocutor backchannel rate labelled 'typical' in the current research is

likely to be comfortably within the range of what other interlocutors would produce in the same or similar circumstances.

4.4.2 Non-Verbal Backchannels

As stated in Section 4.2.3, some video recordings with participants from the target populations were made during the main data collection phase, so that the presence of non-verbal backchannels in my interactions with members from those populations could be checked. Four such recordings were made (these four participants were not included in the main study). The first and third were with Japanese speakers; the second and last were with Mandarin Chinese speakers. The first was made at the end of a week-long period in which the first seven participants took part; the second after a two-day period in which the first four Chinese participants took part; the third after the last of the recordings with Chinese participants; and the fourth one week after the final participant who was included, which was part of a batch of 13 Japanese participants spread over two weeks.

The procedure was that described for the main study, but with video recording as done for the first pilot study (see Section 4.2.1). The subsequent analysis of the recordings was similar to the procedure described for the first pilot study, but, this time, the video recordings were viewed not by the participants themselves, but by people with the two characteristics argued for in the same section: advanced rather than lower proficiency English; and considerable experience of interacting with L1 speakers of English. The Japanese recordings were viewed by someone with L1 Japanese and the Chinese recordings by someone with L1 Mandarin Chinese.

The viewer of the first recording reported one non-verbal response that was not associated with a verbal one; none were spotted from the second; two were pointed out in the third; and one was found in the last. This can be interpreted in relation to the number of verbal backchannels given and in relation to the number of interactions that were analysed statistically. The mean number of backchannels given in the main data collection procedure was 40 per participant. If the non-verbal response frequency found in the checking procedure were repeated in the main study, they would represent between 0 and 5% of the verbal backchannels given. There were 71 participants in the main study. Each took part in four interactions that were analysed independently (long-turn and discursive parts twice each), giving a total of 284 interactions. If the non-verbal response frequency found in the checking procedure were repeated in the main study, each participant would average one, meaning, on average, that three quarters of the interactions would be unaffected and one would potentially be influenced by an additional response that could be interpreted as a backchannel. Even assuming that every instance was seen by every participant, the influence of purely non-verbal responses on the data is likely to have been minimal.

4.5 Equipment

Comparing the audio from the recordings made in the first pilot study using a video camera and its internal microphone with the audio from a voice recorder used at the same time showed that the latter picked up all verbal backchannels, so this was used throughout the main study. This digital voice recorder was an Olympus DM–550. It is a small device, measuring 110 x 39 x 16 millimetres, making it unlikely to be off-putting in comparison with larger devices or ones that use an external microphone. Not using headset or lapel

microphones, which could have added the benefits of improved sound quality and separated inputs, also meant that any additional stress from the reality of recording was minimised, as hygiene concerns and physical contact with the other person were avoided. It was also simpler technically, as no synchronising of recording channels was required, a point that could be important for the subsequent analysis.

Even when something highly dependent on acoustic detail is not being studied, it is still best practice to make uncompressed, high-resolution audio recordings (Margetts and Margetts, 2012: 17). In keeping with this, the audio recordings were made in stereo in linear PCM format. This was at 44.1 or 48 kHz, which was the highest available on the device and equivalent to standard CD quality. The standard rate for linguistic research is 44 kHz (Podesva and Zsiga, 2013: 171). Each recording was saved as a WAV file. Video recordings for the first pilot study and checking for non-verbal backchannels were made in MPEG-2 format on a Sony DCR–SR35E Handycam or Sony DCR–SR68 Handycam.

4.6 Ethical Matters

Ethical matters are here divided into two sections. The first describes the processes and procedures followed; the second reports on participants' responses to these. Ethical matters pertinent to data analysis are presented in the following chapter.

4.6.1 Procedures

Approval from the University of Oxford's research ethics committee was obtained prior to data collection. A consent form, which included information about the research, was

presented to each participant prior to data collection. Given the English proficiency of the participants, offering this in their L1 was appropriate (Mackey and Gass, 2016: 37). The English version of the form (in Appendix 5) was therefore translated into Japanese and Chinese versions; one L1 speaker of each of those languages wrote an initial translation, which was then checked and amended slightly by a second. An L1 version was thus made available to all participants as an alternative to the English original.

The audio recordings were transferred to a password-protected computer and then deleted from the recording device. Video recordings were treated in the same way.

Participants were not told the precise purpose of the study, as this could have influenced the interactions. This need to avoid compromising the research aims met the first of Mackey and Gass' (2016: 35–36) recommended conditions for withholding information from participants; the other two are that no risks are undisclosed and that subsequent debriefing is offered. In keeping with these, no risks were identified and participants were given the opportunity to request more information about the purpose and findings of the study by ticking a box and providing an e-mail address at the end of the questionnaire.

4.6.2 Outcomes

Most participants chose to read both their L1 and English versions of the consent form. There were few questions about it and all who were offered the consent form signed it. No participant expressed any concerns about any aspect of their participation subsequently, or asked to withdraw from the study. The consent form stated that any comments written on

the questionnaire in Japanese or Chinese would be passed to a translator, but all who chose to add comments did so in English.

Approximately half of the participants requested further information on the research. This was provided by e-mail after all data collection had been completed.

4.7 Summary of Challenges and Limitations

This chapter and its predecessor have included some of the challenges for, and limitations of, this study's methods. The chief ones include facets of the research design, details of the procedure and characteristics of the participants and interlocutor. The repeated measures design had the challenges of requiring two topics of equivalent difficulty and involving possible sequencing effects. These were unavoidable, as they resulted from preferring the statistical power and possibility of examining individual differences that this design holds over a between-subjects one. They were dealt with by counter-balancing the design, using topics of demonstrated equivalence and then both checking their equivalence and checking for sequencing effects in the data collected. The participants varied in some characteristics and the importance of most of these to this research was unknown as well as being, based on the existing literature, unknowable. By removing from the samples participants who did not meet the required characteristics, while keeping those requirements relatively broad, variation was restricted but not curtailed to the point of greatly limiting the applicability of any findings. The main challenges to validity and reliability in the procedure itself were the contributions of the interlocutor. The choice of interlocutor met the L1 and L2 interlocutor characteristics that the literature revealed to be desirable. Checks of the production of non-verbal responses and the frequency of (verbal) backchannels given were

essential to address reliability and validity, so were performed in a combination of before, during and after the entire data collection process. The typical backchannel frequency being within the range of what can be described as typical for interlocutors was demonstrated using the entire data set, with reference to a systematic review of backchannel frequencies.

These main challenges and limitations were, then, addressed in a variety of ways. Some were checked and reported in this chapter; others were checked and are reported in the Results chapter; and some were checked and reported in this chapter and then re-checked and are reported again in the Results chapter. Analysing the data brought with it another set of challenges. These are included in the next chapter, which is on methods of data analysis.

5 METHODS – DATA ANALYSIS

This chapter describes all parts of the data analysis, from the preparation of transcripts and the calculating of complexity, accuracy and fluency (CAF) measures, to the running of statistical tests and the checking of intra-rater reliability. All steps, including transcription, were based on the need to conduct CAF analyses. The choosing of appropriate CAF measures was described in Chapter 3. How they are operationalised is rarely described or discussed in published research, so considerable detail is presented here, along with justifications for the decisions made. Principles, examples and problems are all stated. Where there was doubt over how to classify any part of the data, the primacy of the original audio recordings was enforced: clear intent evident from prosody was given more weight than what had been written in the transcript.

Most of the data analysis methods were the same for the long-turn and discursive parts of the recordings; any differences are described in the relevant sections of this chapter. All examples of speech given are from the data collected and analysed for this study, except where noted, and only one example is given per principle being explicated. A complete example of a long-turn and discursive part transcript marked up for analysis is in Appendix 6.

The structure of this chapter largely mirrors the sequencing of the data analysis process, so proceeds from transcription. Before this, ethical matters and the timing of the data analysis steps are described.

5.1 Ethical Matters

Transcripts were made from the audio recordings and subsequent analyses used those full transcripts, except where noted below. Extracts in this thesis are presented anonymously: no name or participant label is given. The extracts have also had identifying words such as the names of places, people and institutions changed. Following Jenks (2011: 23–24), country names were not changed, as they were of large areas. Before counting syllables using the automatic system (described in Section 5.6.2), the same identifying words were changed to non-identifying ones that contained a matching number of syllables.

5.2 Timing of Data Analysis Steps

Transcription was started after all data had been collected. This was to preclude the possibility of the ongoing data collection process being influenced by nascent findings. Similarly, other steps in the data analysis process were performed only after transcription had been completed. As the accuracy analysis depended on the complexity analysis, which depended on the fluency analysis, these were done in the order presented in this chapter: fluency, then complexity and then accuracy. More prosaically practical reasons meant that the long-turn parts were analysed before the discursive parts. The whole data analysis process was not entirely linear, however, because amendments to the transcripts were made as inaccuracies were identified; this occurred chiefly during the fluency analyses.

5.3 Transcription

The first step in data analysis was transcription. All transcribing was done by the researcher, in part because this can help to develop greater awareness of the content of recordings (Jenks, 2011: 90). Transcripts were written directly into Microsoft Word files and consisted of everything said between the end of instructions for each part being given and the participant's final utterance in each part.

5.3.1 Transcript Contents

Transcripts can vary considerably in the level of detail that they contain. The chief principle in selecting what detail to include is that it should be appropriate to how the transcripts will be used (Powers, 2005: 11; Jenks, 2011: 43). For the analyses required in this study, all syllables had to be counted and word-based assessments of complexity and accuracy done, so the transcripts produced were largely orthographic, with modifications to standard orthography made to account for non-words and unorthodox pronunciation of real words, especially where these altered the number of syllables. Pauses also had to be counted, so pause sounds were included, too.

Very brief sounds that were hard to identify or could not be represented using the English alphabet (clicks, for example), were not transcribed. Coughing and laughter, and paralinguistic features such as stress and intonation, were not transcribed, because they were not directly pertinent to the research questions.

5.3.2 Problems and Specific Instances

Where identifying words and sounds was difficult, the recording was slowed by 50% to attempt to elucidate what had been said. For instance, it could be very hard to distinguish 'ours' from the unfinished 'our s-', but doing so was easier with the recording slowed down. Where slowing did help, a final check was made by listening again at normal speed. This was done to counter the occasionally substantial difference in what appeared to have been said when listening at different speeds. For example, one participant's 'completely destress yourself' sounded like 'completed stress yourself' at half speed. Taking these different perspectives to listening is recommended by Jenks (2011: 91) as a means of reducing the risk of not considering alternative transcription possibilities.

Some utterances were almost impossible to distinguish. Examples were: 'uh' or 'a'; 'also' or 'all so'; 'um' or 'and' shortened to 'n'; and 'engaging' or 'engage in'. A two-step process was implemented to deal with such cases. The first was to compare the problematic instance with similar, clearer utterances from the same speaker. The problematic instance was then taken to be the same word or sound as the clearer one that it was most similar to, based on such things as their duration and intonation. If this was inconclusive or no similar utterances were available, the second step was to choose the transcription that would credit the speaker with greater semantic or syntactic control.

To reduce ambiguity when reading a transcript, a speaker's 'n' that was used instead of a full 'and' was transcribed as the latter. Other unambiguous shortenings such as 'coz' for 'because' were transcribed in their more abbreviated form. Spelling for pause sounds was

fixed as 'um', 'mm', 'uh', 'ah', or 'oh'. Similarly, backchannels were written as 'yeah', 'yes', 'ok', 'right', 'uh-huh', 'mm', or 'mhm'.

Vowel sounds that were clearly added unnecessarily to the ends of words were included in the transcripts and were almost all written as '-uh'. Such additions are often found in the L2 English of both Japanese and Chinese speakers, and varied in frequency among this study's participants from very common to not present at all. Even the pronunciation of the same word by the same participant could be inconsistent, as in the two versions of 'if' and 'help' in example 1:

(1) if-uh I have some problems they will help-uh me and if other students have problem they will help too

Participant use of an L1 was rare, apart from for the names of places and titles of films and television programmes. All L1 words were transliterated in standard ways, but without diacritics, and to allow for a ready understanding of the number of syllables they contained.

5.3.3 Punctuation and Layout

Punctuation was generally minimised in the transcripts, except for: hyphens, as described immediately below; conventional uses of apostrophes and question marks; some conventional capitalisation; and some commas or full stops to clarify meaning or to separate ideas. The following example contained conventional capitalisation (after a full stop; and for the name of an organisation) and a full stop to indicate that the second use of 'Ghibuli animation film' began a new idea:

(2) the film which has a strong effect on me is a Ghibuli animation film. Uh Ghibuli animation film is a uh Japanese famous animation film

A hyphen at the end of an abandoned word was used to indicate non-completion of that word:

(3) since I was a s- child

As described in the previous section, a hyphen also separated a superfluous vowel sound from the end of the word to which it was appended. The final reason for a hyphen being used was to indicate a pause within a word:

(4) the other agen-cy uh fight with this this girl

Square brackets were used to contain comments from the researcher on the transcript or audio recording. These included statements of the number of syllables in words that could be pronounced with a differing number of syllables, notes on rhetorical rather than disfluent repetition and a question mark to indicate uncertainty over the accuracy of the transcription of an utterance.

The start and end points of a participant's speech were not marked. No line breaks or numbering were used. To contrast with this, in the discursive parts, interviewer speech was enclosed within solidi and placed on a new line. Indenting was used in the discursive parts to indicate the positioning of any overlapping talk.

5.3.4 Backchannels

All backchannels were transcribed. A separate line in the transcript was not used to display only a backchannel, so they appeared in the transcripts with the ongoing speech and were distinguished from it by being placed between vertical arrows. These arrows pointed down when given by the interviewer:

(5) every [2 syllables] year we are looking forward to doing the school festival and-uh um uh uh in my high school I belonged to dance club ↓mhm↓ so uh the member of-uh the dance club uh show the dance ↓yeah↓

This allowed the backchannels to be identified in a transcript easily and to be counted automatically and with perfect reliability in Word by using the Find function. Backchannels given by a participant when listening to a question in the discursive part were placed between arrows that pointed up:

(6) /So about your school ↑mm↑ what were the good points and bad points about your school?/

Having transcribed the recordings, the next stage was to select which parts of the transcripts to include in the analyses. This is described in the next section.

5.4 Parts of Transcripts Included in Analyses

The full length of each long-turn part was transcribed and included in the analyses, even if a participant exceeded the allotted speaking time of 2 minutes. Imposing an arbitrary cut-off time would have simplified the analysis process, but this was deemed to be inappropriate, based on the following reasoning. The two topics talked about were an event at school and a film or television programme, so a narrative was frequently given. Narratives are known to have a particular structure and different grammatical constructions can be concentrated in different parts of that structure (e.g., Labov, 1972: 362–370, 384). There could, therefore, be varying CAF at different points in a narrative, so applying an arbitrary temporal cut-off could have resulted in CAF calculations being affected. The same principle could also apply to those participants who, in danger of going well beyond the allotted time, were stopped in a long-turn part before they had definitely finished what they intended to say. These were few in number, however, so the threat to validity was less than if an arbitrary cut-off had been applied after data collection.

For the long-turn parts, the beginning of what was analysed was taken to be the first fluent content syllable, so disfluencies (described in full in Section 5.6.1), pause sounds and discourse markers were never the first syllable. The last two were very common at the start of the long-turn parts, for example:

(7) Ok um I I watch a film many times

A few participants began speaking immediately after being asked to do so and some after a short silence, so including things such as the 'Ok um I' from the above example would have altered the fluency calculations in an inconsistent manner.

In a few instances, a participant asked a question after beginning a long turn. These were included in the analyses, but everything from the end of the question to the participant's resumption was not.

For the discursive parts, the interlocutor's speech was excluded from the analyses. The beginning of what was analysed was taken to be the start of the first non-pause sound of each participant utterance. This differed from the long-turn part starting point in that disfluent syllables were included. They were included so that the more interactive nature of the discursive parts would be featured in the data to be analysed. The concern over including in the analyses any initial pause sounds in the long-turn parts also applied to the discursive parts, because silent pauses and filled pauses were regarded as equivalent. The discursive parts contained an average of six participant responses, so there was likely to be some diluting of the effect of including initial disfluency in the analyses, in comparison with the long-turn parts, which contained only one start.[12]

If a participant's final idea in a long-turn part or a discursive part response was abandoned or disfluent, then it was excluded from the analyses; such instances were rare. This also

[12] As with many other methodological details in this area, the literature is almost entirely silent on where to begin and end analyses of speech, including interactive speech. An exception was published after the analyses reported here had been completed. This was Tavakoli (2016), who compared fluency measures calculated excluding 'between-turn' pauses with those calculated with such pauses divided equally between the two speakers. The findings were that fluency values differed but that, for her data, the statistical significance of most measures was not affected by the inclusion or exclusion of such pauses. Where to begin and end analyses warrants future investigation, then; a useful preliminary step would be for researchers to state clearly the decisions that they made.

applied to final non-content words and sounds, such as a self-directed 'yeah' or 'mm', but not to vowel sounds that were clearly added unnecessarily to the ends of words. Parts thus excluded were boxed:

(8) we can um discuss sports sports or other things after the w- after we played the basketball um and-uh mm

Further things were excluded from the analyses of the discursive parts. First, clarification requests and related things such as expressions of confusion at a question were excluded. These were removed because they were taken to be expressions of listening comprehension difficulty, including on the meaning of words used by the interlocutor, which was not being assessed in this study. In the following example, which came after a description of a school trip, the participant was confused ('school schools') by the change of topic, then explicitly asked for clarification:

(9) /Ok. And about your school what were the good points and bad points about your school?/

School schools

/Yeah your/

Not school trip?

/Yeah your high school/

In addition, very short responses to a question were excluded. This was operationalised as responses of three complete words or fewer. Pause sounds and incomplete words were not included in this count; contractions (e.g., 'don't') were counted as two words.

Backchannels were removed from the transcripts before subsequent analyses were performed. This was done to avoid the possibility of unconscious bias, based on knowledge of which backchannel frequency had been used, affecting any analysis. For ease of reading, they are not included in examples in the remainder of this chapter.

The parts of the recordings and, thus, transcripts that were used in CAF analyses were the same for all three CAF elements. These three are described in the following sections, after a brief delineation of the editing of the original audio files to prepare them for analysis.

5.5 Preparation of Audio Files

The editing of the original audio files was done using Audacity 2.1.1, a software package. This displays the waveform, allows parts of the recording to be excised with ease and permits the resultant edited version to be saved as a new file. The cutting was done in accordance with the principles described in the previous section, thereby creating, for each interaction, a new audio file that contained only those parts of a participant's speech that would be analysed for fluency.

5.6 Fluency

As described in Chapter 3, the three measures selected to evaluate fluency were pruned speech rate (PSR), mean length of run (MLR) and non-phonation time ratio (NPTR). These were defined as follows:[13]

[13] A 'correction' was made to the denominator in the MLR calculations: in the long-turn parts, one was added to the number of pauses to account for the final run being without a pause; in the discursive parts, one was added to the number of pauses for each interlocutor question answered, for the same reason.

PSR: total syllables minus disfluent syllables, divided by speaking time; expressed in syllables per minute

MLR: total syllables divided by number of pauses plus correction

NPTR: total pause time divided by speaking time; expressed as a percentage

MLR is usually calculated as described above, but variations exist. Ellis and Barkhuizen (2005: 157) suggest that disfluencies should be excluded from the pause count. As pauses occur among disfluent syllables as well as among fluent ones, this interpretation could lead to misleading figures, so all syllables and all pauses were included here.

The following were therefore required to perform these calculations: the total number of syllables; the number of disfluent syllables; the speaking time; the number of pauses; and the total pause time. These were obtained using a combination of software, adjustments to both the input and output of the software as required by flaws in the programs, plus manual checks to ensure consistency. These are all described in this fluency section, beginning with the identifying of disfluent syllables.

5.6.1 Disfluent Syllables

Disfluent syllables were disfluencies that appeared in the transcripts. They mainly consisted of self-interruptions and pause sounds (silent pauses, which were not included in the transcripts, are dealt with in Section 5.6.3). In the following consideration of disfluency types and principles and problems in identifying disfluency, there is considerable overlap with matters related to accuracy; those specific to accuracy are described in Section 5.8.

Although the starting point in identifying disfluent syllables was the transcripts, the audio recordings were also consulted when prosody was likely to help.

5.6.1.1 Disfluency Marking and Types

In the transcripts, disfluent syllables were marked in two ways: pause sounds were struck through, while other disfluencies were shaded. This left unmarked in the transcript those parts of speech that were deemed to be fluent. This mark-up is included in all examples for the remainder of Section 5.6, thus:[14]

(10) ~~uh~~ I think Chinese film ~~uh~~ films are very boring

In the literature, there are inconsistencies and overlaps in how disfluency terms are defined and used, and in how many types are distinguished. For the purposes of identifying disfluency in the transcripts here, three types need to be described: repetition; abandonment; and self-repair.

Repetition is simply "Words, phrases, or clauses that are repeated with no modification" (Foster and Skehan, 1996: 310). This typically occurs immediately, or following a pause. Example 11 was an extreme instance of repetition of one word:

(11) to do with my personality or or or or or or something not good

[14] Again, an example of a complete transcript marked up for analysis is in Appendix 6.

Abandonment involves a start being aborted and then a start on another construction or idea being made. (Abandonment was also deemed to have occurred when a participant aborted a start but did not continue speaking – i.e. at the end of a long-turn part or when the interlocutor took a turn to speak in a discursive part.) In the following, the 'school day is the' start is abandoned, then the 'sometimes I do' idea is started and completed:

(12) School school day is the. ~~Uh~~ sometimes I do really feel happy and sometimes I feel not happy in the university

Self-repair occurs when the speaker goes on to change and complete a started construction or idea. The frequently found constituent parts of a self-repair are: the reparandum; an editing term; and the repair (Husband, 2015: 21). These are shown in example 13:

(13) I had many works to do but the works are properly I mean are proper. They are

Reparandum *Editing term* *Repair*

The reparandum occurs first and is replaced by the repair, while an editing term is often found between the two. The editing term, where present, may take one of several forms, including a phrase such as "you know", or a pause (McKelvie, 1998: 4).

Self-repairs can be divided into error repairs and appropriateness repairs (Levelt, 1989: 459). An error repair is the attempted correction of one or more elements (chiefly lexis, syntax, morphology, or semantics) that have been said, as shown in the previous example.

An appropriateness repair addresses such things as specificity and ambiguity, and is again a change to what has been said. In 14, 'design' is changed to the more specific 'clothes design':

(14) it's not like the ~~um~~ university like design clothes design or

In addition, a superfluous vowel sound added to the end of a word (e.g., 'but-uh') was marked as disfluent.

5.6.1.2 Basic Principles in Identifying Disfluency

This section briefly describes the basic principles used to identify and mark disfluency. Exceptions are provided in the following section.

The starting principle in marking disfluency was that the final version of an attempt at saying something should be marked as the fluent one (this is implied but rarely stated explicitly in the literature, but is by Foster, Tonkyn and Wigglesworth, 2000: 368–369). The appropriateness of this principle can be seen in almost all of the examples provided here. A repair was deemed to have occurred and to be fluent if an attempt at it was completed. Some judgement was inevitably required in assessing this, but, in practice, intonation and pausing patterns helped to make the completeness of an attempt readily apparent, and there were few clear instances of abandonment in the data.

A second principle was that, if something in a transcript could be included as fluent, then it was. This was common with conjunctions: a conjunction followed by a disfluency and

then no subsequent conjunction was treated as fluent, if possible. This was in line with recent research on L1 English speakers and Chinese speakers of L2 English, who do not often start a repair with a repeated or altered conjunction (Fox, Maschler and Uhmann, 2010; Quan and Weisser, 2015). In 15, then, 'and' was marked as fluent, despite at least one attempt at repair occurring before fluent speech resumed:

(15) is-uh different and ameri- American movie is ~~um~~ is-uh the scale of American movies are bigger

A common limitation of studies of disfluency in L1s and L2s is that they fail to address errors that are not overtly corrected. These occurred in the data analysed for this study and were found especially in the use of non-content words, including articles and prepositions. This was perhaps more commonly a matter of accuracy, but the syllable count could also be affected. The principle applied was, similarly to the one just described, to mark as much as possible as fluent, taking into consideration the word order used. Thus, in 16, which was referring to the past and used correct word order, all words that could be inflected were inflected incorrectly, but, instead of marking 'practised' as disfluent to establish grammatical accuracy, all were kept as fluent.

(16) I ~~uh~~ we have to ~~uh~~ practised run and-uh climb

Another principle was that the participant was always attempting to say something that had meaning. Thus, even if part of what was said (and not then abandoned or repaired) appeared to make little sense at first inspection, a possible meaning was sought and that

formed the starting point of deciding what was disfluent. This principle was applied to the following:

(17) a. my university is a is n- is b- is normal university it's-uh um basical to um development the to some uh they the teacher most of teacher from the my university

The intended meaning was taken to be: 'my university, being a normal university, is fundamental to the development of its (student) teachers'. (A 'normal university' in China is a university at which trainee teachers are, or used to be, educated.) This was therefore marked as:

(17) b. my university is a is n- is b- is normal university it's-uh ~~um~~ basical to ~~um~~ development the to some ~~uh~~ they the teacher most of teacher from the my university

A final principle was that, where L1 Japanese- and Mandarin Chinese-specific editing terms occurred, they were treated in the same way as those that are found in L1 English. Such instances were rare in the data.

5.6.1.3 Problems and Specific Instances in Identifying Disfluency

This section provides additional detail to the principles stated in the previous section. It gives exceptions, contrasting examples and unusual instances. These are clustered by the principle on which they expand.

The principle of the final version being marked as the fluent one sometimes meant that a self-correction from a grammatically correct form to an incorrect form was retained (18), or that some detail was lost (in 19, 'ballet' was marked as disfluent because 'a' was repeated before 'dancer'):

(18) I love I loves Hollywood movie more than Japanese movie

(19) it's a film ~~uh~~ described-uh a ba- ballet ~~um~~ ~~uh~~ a dancer

In keeping with the same principle, if there was addition as modification or reformulation, then the first attempt was marked as disfluent. In 20, 'the test' was modified to 'the final test', so only the latter was kept as fluent:

(20) the points will added on the on the ~~uh~~ on the test the final test

If the addition was abandoned, as in 'girl f-' in 21, then it was also treated as disfluent:

(21) I could make many good friends many good f- girl f- ~~um~~ many good friends

In contrast, if there was addition of detail without modification or reformulation, then all parts were kept as fluent. In 22, the participant described a trip to New Zealand and added 'New Zealander' to the earlier 'local people':

(22) I stayed in the house the the the house ~~uh~~ local people ~~uh~~ New Zealander live in

Delayed correction of a word was not common, but sometimes required more than the final version to be classed as fluent. In the following example, marking either occurrence of 'divorced' as disfluent would create considerable loss of detail or semantic clarity, so both were retained as fluent:

(23) at that time I was just-uh divorced-uh from ~~uh~~ with my girlfriend. Divorced-uh separate or separate

Incomplete self-correction or repetition of an earlier, completed word meant that the original attempt was retained as fluent. Thus, in 24, 'daughter' was kept ('six' referred to her age):

(24) he had a daughter yeah dau- ah s- yeah six

Where left dislocation was judged to have occurred, the whole construction was marked as fluent. In 25, 'is' could have been marked as fluent and 'it's' as disfluent, but the construction was taken to be 'the business[,] it's well known':

(25) the business is very very we- well it's well known

In the next example, prosody indicated that a possible left dislocation – 'The music live[,] it was' – was a replacement of 'The music live' by 'it', which had an earlier antecedent, so only 'it' was kept as fluent:

(26) my enjoyable event was ~~uh~~ the music live I played. The music live asa [?] it it's it was in my high school yeah high school

Finally on retaining the final version as fluent, natural combinations of conjunctions were treated as fluent, unless they were separated by other content or prosody indicated abandonment. An example of a natural combination was 'and so'. The reverse – 'so and' – was unnatural, so was marked as 'so and'.

On the principle of a repair being fluent if an attempt at it was completed, an abandoned self-correction that turned into a self-confirmation was treated as disfluent. In 27, 'some teacher' was being changed ('some s-') but was then repeated:

(27) some some teacher ~~uh~~ or some ~~uh~~ some s- ~~uh~~ yeah some teacher w- can will random to catch ~~uh~~ tickets

Self-correction or repetition of part of an uncompleted word, without it being restarted, meant that the first attempt at any repeated or corrected syllable was marked as disfluent and the remainder as fluent:

(28) an enjoyable event that I exper-perience

The next six examples chiefly relate to the principle of something being treated as fluent if it could be. The first two are exceptions. First, editing terms (see Section 5.6.1.1) that occurred as part of a self-repair were marked as disfluent. Second, discourse markers were treated as fluent, but possible discourse markers that occurred as part of a disfluency (that

was not merely a pause) were classed as disfluent; instances of this can be seen in the treatment of 'yeah' in examples 26 and 27.

Summaries of what had been said could be regarded as repetitious in themselves, but they were not treated as disfluent, as they are a natural feature of speech (prosody helped identify such instances). The iterations of 'big difference' in the following were therefore not repetitions:

(29) Kyoto and Kobe are same area in Japan but they are they have-uh no di- they have big difference. I think ~~mm~~ I think big difference for me.

However, where something could be regarded as a summary but was part of an abandoned idea (having been evaluated as such based on prosody), it was treated as disfluent, as with the second 'comedy programme' below:

(30) Japanese comedy programme has ~~uh~~ many ~~mm~~ different types ~~uh~~ for example manzai or some kind of ~~um~~ yeah comedy programme

An unfinished word was kept if its full form appeared to be intended and there was no attempt at repair. Thus, in 31, 'univers-' was assumed to be an attempt at 'university' and marked as fluent:

(31) so I thinks ~~uh~~ the time of in univers- student is important

However, if another, less complete attempt was made and abandoned, then all were marked as disfluent, as with 'abou- ab-' here:

(32) so I was very surprise this abou- ab- this film

A variant of this principle – marking as much as possible as fluent, taking into consideration the word order used – requires further exemplification. Using this, an unnecessary article in a position where an article could be was kept, as in 33, which contains a location ('Tokyo'). However, the principle did not apply to duplication of meaning: 'a one room' (34) was said twice by this participant and would perhaps be regarded by him/her as not requiring correction, but it was taken to be self-repair here:[15]

(33) in Japan ~~um~~ now I living in the Tokyo

(34) four people in a one room in a school

There were also some occurrences that were largely outwith the stated principles. First, self-directed comments were counted as disfluencies, as in 'when was it' here:

(35) And-uh it-uh happens well when was it I start to do some exercises from my high school

[15] Inevitably, these fluency-based analysis choices could also influence accuracy measures. Both Japanese and Chinese learners of English typically have a poor understanding of article use. The choices made here meant that their syllable-based fluency measures were very slightly boosted, but this ensured that measures of their accuracy were not boosted when they had displayed no awareness of having made an error.

Second, repetition for rhetorical effect (adding emphasis) was classed as fluent. This was judged based on prosody. Finally, following Lennon (1990: 406) and others, comments on the task were treated as disfluent.

The transcripts were marked for disfluency based on the above criteria. The next step was to count the number of syllables, including those classed as disfluent.

5.6.2 Counting Syllables

Automatic systems exist to count syllables using only the speech signal. One that was tested using some L2 speech was described by de Jong and Wempe (2009). They used spurts (periods of speech between pauses) of at least 5 seconds and calculated speech rates based on the automatic detection of syllables by this means. They achieved a speech rate correlation with a transcript-based manual calculation of .71 (de Jong and Wempe, 2009: 387). Although they describe refinements that improved the correlation for their Dutch corpora and assert that the extent of their system's underestimation of the true speech rate was consistent across speakers, the validity of their and other purely signal-based calculations of syllables was deemed to be insufficient for the purposes of this study. Alternative, largely transcript-based, methods were therefore sought.

First, a manual approach was attempted for six transcripts (three Chinese and three Japanese). This simply involved adding up the number of syllables on each line of the transcript and then summing the total for each line. Where there was doubt over the number of syllables (e.g., 'different' can contain two or three syllables) the audio recording was listened to again (unusual numbers of syllables had already been noted during the

original transcription process). Second, to check the manual counts, two online tools that counted syllables in written texts automatically were tested.[16] The two reported different numbers of syllables for the same texts, so were compared to identify the reasons for the differences. The one at wordcalc.com was found to be considerably less accurate, as it introduced numerous errors for common words, such as assigning 'the' no syllables. The counter at poetrysoup.com was therefore continued with. This used a combination of two tools to calculate the number of syllables in a text that had been copied and pasted to it: a database of words and their syllables; and an algorithm to estimate the syllables in words not in the database.

The first finding from comparing the automatic syllable counter with the manual count was that, while the automatic version always returned the same count for a given text, the reliability of the manual count was considerably lower. This was caused by incorrectly adding up the number of syllables noted on each line of the transcript, not by misidentifying the number of syllables that words contained. The second finding was that the automatic counter did misidentify the number of syllables that some words contained. The superior reliability of the automatic system meant that it was selected to be the tool that would perform all of the syllable counts for this study. Before it could be used, however, problems with its accuracy had to be identified and addressed.

These problems were identified by comparing the automatic and manual counts from individual lines of the six transcripts just mentioned and, where they differed, comparing clusters of words from the line until the cause of the discrepancy was discovered. The most frequent inaccuracies of the chosen automatic system were: failure to identify when an /s/ or /ed/ inflection adds a syllable; miscounting the syllables in non-words containing

[16] One was at http://www.poetrysoup.com/poetry_resources/syllable_counter.aspx and the other was at http://www.wordcalc.com/

consecutive consonants; treating capitalised abbreviations as one syllable; and miscounting the syllables in some common words. These were corrected for by altering the transcripts before pasting them into the counter. Other problems resulted from limitations of, rather than flaws in, the automatic system: heteronymic variation and alternative pronunciations. These were readily accounted for by checking the original audio recording and, again, adjusting the transcripts where required so that the counter would report the apposite number of syllables.

The problems encountered are listed in Table 5.1. This also provides an example for each and how it was dealt with.

Table 5.1 Solutions to automatic syllable-counter problems

Problem	Example	Solution (to example)
/s/ inflection in some words	encourage (3) encourages (3; should be 4)	replaced 'encourages' with 'I can I can'
/ed/ inflection in some words	visit (2) visited (2; should be 3)	replaced 'visited' with 'I can I'
Consecutive consonants in non-words	tr- (2; should be 1)	changed 'tr-' to 't-'
Capitalised abbreviations	USA (1; should be 3)	changed 'USA' to 'U S A'
Some heteronyms	learned (2; could be 1 or 2)	listened to recording then replaced word if needed
Pronunciation variation	different (3; can be 2 or 3)	listened to recording then replaced word if needed
L1 words used in English	Tokyo (3; 2 in Japanese[17])	replaced 'Tokyo' with 'I can'
Syllable count for some common words	am (2; should be 1)	changed 'am' to 'a'
Mid-word hyphen in transcript	agen-cy (5; should be 3)	changed 'agen-cy' to 'agency'

Note: under 'Example', the number of syllables reported by the automatic counter is given in brackets.

After making these adjustments, the following procedure was followed for each transcript. The whole transcript was pasted into the counter to get the gross number of syllables. Where a transcript exceeded the maximum size permissible (1,500 characters), it was cut into smaller sections and the syllables in each of those sections were tallied separately. Disfluent syllables were then removed by cutting them in Word and the remaining

[17] Japanese sounds are split into morae, not syllables. 'Tokyo' contains four morae, but can be considered for the purposes here to contain two syllables when spoken in L1 Japanese.

syllables were pasted again. The final step was to remove pause sounds in Word and then paste the remaining syllables into the counter again. This gave the number of fluent syllables. All syllable counts were entered into a Microsoft Excel sheet. The built-in redundancy in this process helped to minimise the possibility of errors going undetected. For instance, Word reported the number of struck through items replaced and, as these were all monosyllabic pause sounds, this number could be compared with the difference in the number of syllables reported by the second and third counts.

At the start of each syllable-counting session (i.e. day), the first transcript part that had had its syllables counted in the previous session was pasted into the counter again, to check that the online counter had not changed. No evidence of it having changed was found.

5.6.3 Pauses

The next two things that were needed for the fluency calculations were the number of pauses and the total pause time. Before describing how these were obtained, more information on pause types and lengths is required.

5.6.3.1 Background on Pauses

Pauses in speech are often described as being 'filled' or 'unfilled'. The latter consist of silence, so are also known as 'silent pauses', while the former contain a vocalisation that is often written as 'uh', 'um', 'mm' and so on. Some researchers (e.g., Clark and Fox Tree, 2002: 73) have suggested that such vocalisations should be treated as words in English, but this is an unusual stance. The more common belief is that, as expressed by Lennon (1990: 406), "Filled pauses are, by definition, nonwords".

This is one part of a lack of consistency in the literature over what constitutes a pause, which ones should be included in analyses and how they should be measured. The confusion is sometimes highlighted by transcript extracts. Kormos and Dénes (2004: 162) provide a typical example of this in "by a car (4.074) err (0.620) the firemen". The figures in brackets are pause durations, which leads to several possible interpretations. The first figure could be the length of a silent pause and the second the length of the "err", which is treated as a filled pause, but that would require no silence between "err" and "the". Similarly, "err" could be linked without a silence to "car" and the second figure could be the length of a silent pause. Based on other examples from the same transcript, neither of these possible explanations is likely. Another possibility is that there could be two silent pauses separated by "err", which is regarded as a word, but the authors state that they counted both filled and silent pauses, so "err" would not have been a word. Finally, there could be two silent pauses separated by "err", the length of which was not measured, indicating that silent and filled pauses were regarded as different in some way.

This final (and most likely) interpretation raises a fundamental issue. In recent years, improvements in software have meant that silent pauses in speech can be identified and measured automatically. This has made pause-based analysis much less laborious, but has also arguably led to the spreading of another element of confusion: whether or not including only silent pauses in such analyses is justifiable. Briefly, it has been known for more than fifty years (e.g., Goldman-Eisler, 1961) that the number or length of filled pauses is not necessarily correlated positively with the number or length of silent pauses, and that the time occupied by filled pauses does not necessarily vary systematically with the time occupied by silent pauses, so considerably more validity can be achieved by including the number and length of both forms of pause in fluency analyses.

Turning to pause length, early research was conducted by Goldman-Eisler, who observed that many short pauses were the result of "the need to adjust the position of articulation" and went on to argue that applying a minimum pause length of 250 msecs in research "might mean some loss of data, but it ensures the clear separation of hesitation pauses from phonetic stoppages" (1968: 12). Research into oral fluency, including in L2s, has often used a cut-off of 250 msecs because of this (Towell, Hawkins and Bazergui, 1996: 91; de Jong and Bosker, 2013: 17). However, many studies (e.g., Foster and Skehan, 1996; Mehnert, 1998; Wolf, 2008) have used longer cut-off points – all three of these used 1,000 msecs, which is not justified by evidence or theory.

In contrast, some evidence for including pauses shorter than 250 msecs has been presented. Hieke, Kowal and O'Connell (1983) found a high number of pauses of 130–250 msecs in duration that were not caused by articulatory requirements. They used recordings of political speeches and poetry recitals, finding that, for poetry, most of the short pauses "occurred at the end of poetic lines or at punctuated positions" (1983: 211). These were, therefore, unlikely to be the result of the same psycholinguistic pressures that lead to pauses in spontaneous speech, so their findings were not relevant to the current research.

Recordings of spontaneous speech were used by Kirsner, Dunn and Hird (2003) to investigate relationships between short and long pauses. However, their choice to implement a minimum silence duration of just 20 msecs when identifying pauses meant that articulatory pauses were certainly included in their analyses; this cut-off point is simply too low.

More relevant to the current research is the work of de Jong and Bosker (2013), which aimed to identify an appropriate pause cut-off point for use in L2 research. They correlated pause lengths with L2 proficiency and pause-based fluency measures, investigated possible cut-off points from 20 to 1,000 msecs, and concluded that 250 msecs was an appropriate duration. One caveat is that their research was on silent pauses only, but the proportion of filled pauses shorter than 250 msecs is likely to be very low, so applying this cut-off to all pauses is reasonable.

The consequences of the above considerations for the research presented here were twofold. First, filled and silent pauses were not distinguished, so all were counted and measured. Second, 250 msecs was implemented as the minimum pause length. How the counting and measuring were done is described next.

5.6.3.2 Procedure for Identifying and Measuring Pauses

The manual identification and measurement of pauses in speech is very time-consuming.[18] To reduce the time required, software to semi-automate the process was used. This was the Annotate to Text Grid (Silences) function within Praat 5.4.17, which marked each part of the waveform as 'pause' (indicating silence) or 'speak' based on its amplitude. The following settings were used as defaults for this function:

Minimum pitch (Hz): 50

Time step: auto

Silence threshold (dB): -25

[18] Quantity is also a major contributor: more than 19,000 pauses were measured in this study.

Minimum silent interval duration (seconds): 0.25

Minimum sounding interval duration (seconds): 0.1

The silence threshold was the amplitude, relative to the peak amplitude in a recording, above which 'speak' could be marked as having occurred and below which 'pause' could be marked. The minimum silent interval duration implemented the 250 msecs threshold for pauses. The minimum sounding interval duration of 100 msecs meant that the software ignored very short sounds picked up by the microphone, as these were likely to be non-speech; typically, they were from a participant's hand or something held by a participant hitting the desk on which the recording device was placed.

In some instances, the silence threshold had to be altered to improve the consistency with which the software identified pauses across different recordings. For example, when there was additional background noise, the default setting meant that too few pauses were detected, so the threshold was changed from -25 dB to a smaller value. In one instance, the audio quality of a participant's recording was affected by the constant background noise of a computer fan. This made identifying pauses using Praat impossible, so the noise reduction tool in Audacity was employed to remove most of the unwanted noise before Praat was used. This was the only occasion on which the audio signal of any of the recordings was altered.

Although the use of Praat greatly accelerated the identification and measurement of pauses, a major problem was that it could identify only silences (unfilled pauses), so it labelled as speech both filled pauses and silences containing an interlocutor backchannel. These therefore had to be identified manually. This was done using the recordings, transcripts

and waveforms to find where such pauses occurred. How these and other problems with Praat were addressed will now be described.

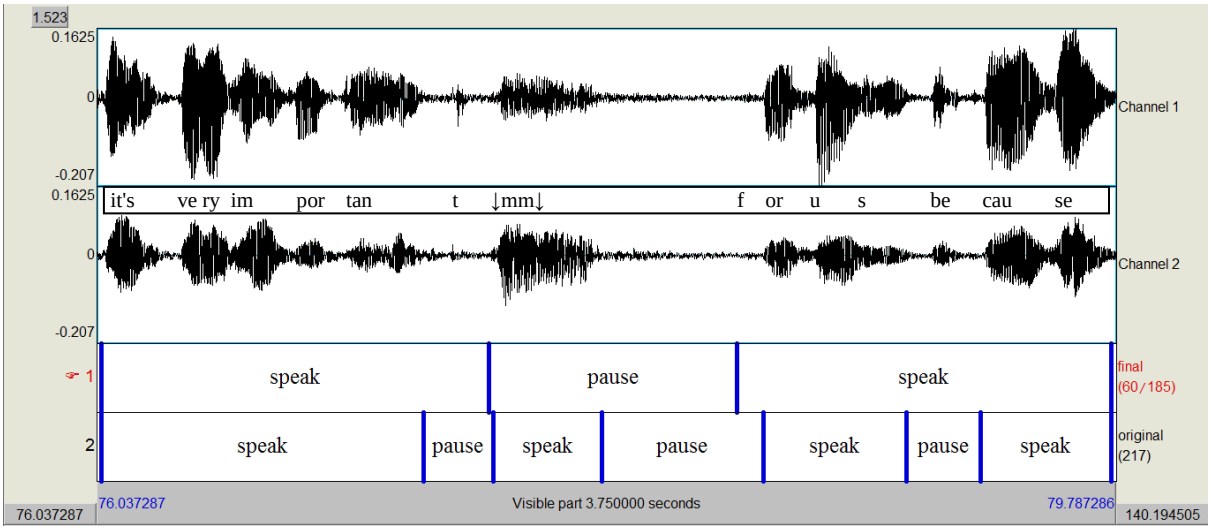

Figure 5.1 Screenshot from Praat. Waveforms plus original and final pause mark-ups are shown. Component parts of the words spoken have been added, in their approximate starting locations, between the waveforms.

Figure 5.1 is an example screenshot from Praat that shows waveforms and the 'speak' and 'pause' labels. The speech represented has been added to the screenshot for illustrative purposes; it is: "it's very important ↓mm↓ for us because". The approximate locations of the start of component parts of this are represented, where relevant, by the positioning of the letters between the two waveforms. The bottom row of 'speak' and 'pause' labels is the product of the automated analysis done by Praat. The vertical bars separating those labels are temporal boundaries. The upper row of labels is the final, manually adjusted analysis. The manual adjustments illustrated in this example are as follows. The first (farthest left) 'pause' part in the automated analysis contained the transient of a /t/ sound, so was manually included in the preceding 'speak' part. The second 'speak' in the automated analysis corresponded to an interlocutor backchannel that coincided with part of a pause, so this was changed to 'pause'. The last 'speak pause speak' sequence in the automated analysis had two changes made to it. First, the beginning of the 'speak' part was moved

forward (as depicted by the vertical bar having been moved to the left), to include the start of the /f/ sound of 'for'. Second, the 'pause' was removed, as it contained the /b/ at the start of 'because'.

Other problems with the automated analysis that required manual adjustments included inhalations, artefacts and extraneous noises being labelled as speech. These were dealt with by adding, removing or adjusting the placement of 'pause' and 'speak' labels in the manner just described. Identifying the start and end points of speech that co-occurred with other sounds could also be difficult; on occasion, the option to view in Praat the signal from separate audio channels was useful in tackling this.

The screenshot in Figure 5.1 is of a 3.75-second segment of a recording. This was the standard level of zoom used when checking and amending the automated identification of pauses. This was chosen as it displayed an appropriate level of detail in the waveform: sufficient to be able to identify and correct inaccuracies, but not so great that a trivial discrepancy would appear to warrant adjustment (a greater level of zoom was occasionally used when the waveform was messier, but this was not used routinely, as adjustments made would have been only approximately 1 millisecond).

The temporal position of the vertical bars separating the 'speak' and 'pause' labels is recorded by Praat. The manual adjustment to their positions just described was therefore the final stage in the pause measurement process. The remaining step was to extract this information so that the total pause time could be calculated.

5.6.3.3 Total Pause Time

For each interaction, the whole of the pertinent text grid in Praat (the part containing the manually adjusted 'speak' and 'pause' labels in Figure 5.1) was saved as a text file and exported to Excel. This created an Excel file with start and finish times for every section of speech and pausing in the recording. The duration of each pause was then calculated in Excel by deducting the finish time of each interval marked as 'pause' from its start time. Total pause time was calculated simply by summing these durations. Praat presented times to six decimal places of a second. Although this level of precision was unrealistic, it was retained in subsequent calculations for reasons of simplicity.

As a safeguard, the 'include empty intervals' option in Praat was selected when transferring the data to Excel. This added a question mark to an otherwise blank cell instead of a 'pause' or 'speak' label. On the few occasions that this happened, the blank part of the Praat text grid was filled as appropriate and the corrected text file was then exported to Excel.

5.6.4 Total Speaking Time

The final thing required to calculate fluency values was total speaking time. As each recording had been edited so that it began with the first utterance to be analysed and ended with the last, total speaking time was simply the length of the recording analysed in Praat.

5.7 Complexity

The three measures selected in the previous chapter to evaluate complexity were words per AS–unit, clauses per AS–unit and words per clause. These were defined as follows:

Words per AS–unit:	number of fluent words divided by number of AS–units
Clauses per AS–unit:	number of clauses divided by number of AS–units
Words per clause:	number of fluent words divided by number of clauses

The following were therefore required to perform these calculations: the number of fluent words; the number of AS–units; and the number of (subordinate) clauses. These are all described in this complexity section, beginning with the number of words.

5.7.1 Number of Words

As with pauses, there is often a lack of clarity in the literature over the counting of words. The specific issue for words is straightforward: whether or not to include in the count words that are classed as disfluent. Foster, Tonkyn and Wigglesworth, who devised AS–units, the other units to be calculated for complexity here, commented that "If the length of AS–units is being counted, false starts, or items which are replaced for grammatical or lexical reasons, will typically not be included in the count" (2000: 374). The simple starting position taken for the current research was that only what was said could be analysed; what was not said could not be analysed. Therefore, just as an 'uh' or 'um' could not and should not be taken to be an expression of an idea, a series of disfluent (as judged by the contents of Section 5.6.1) words could not and should not be taken to be an addition to the complexity of an already or soon-to-be expressed idea. Words counted for measures of complexity were, therefore, only of those classed as fluent.

The necessary counts were done automatically in Word. A shortening such as 'BBC' was counted as one word, as were words that had been spelled out. In keeping with the

selection of transcript parts to include in CAF analyses (Section 5.4), contractions such as 'don't' were treated as two words.

5.7.2 AS–Units and Subordinate Clauses

The remaining things that were needed for the complexity calculations were the number of AS–units and the number of subordinate clauses. AS–units were introduced by Foster, Tonkyn and Wigglesworth (2000) and most of their recommended procedures were followed. However, there are omissions, ambiguities and apparent contradictions in their explication, so both straightforward and problematic instances are discussed and exemplified here, after some relevant definitions.

5.7.2.1 Definitions and Transcript Marking

Foster, Tonkyn and Wigglesworth (2000: 365) state that an AS–unit "is a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either". They also define each of these elements: an independent clause is "minimally a clause including a finite verb" (2000: 365); an independent sub-clausal unit consists of "either one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation […] or a minor utterance" such as "Thank you very much"; and a subordinate clause consists "minimally of a finite or non-finite verb element plus at least one other clause element (subject, object, complement or adverbial)" (2000: 366).

They do not define "clause", "phrase", or several other terms employed. They do, however, provide examples, the interpretation of which formed the basis of much of what is described here.

The transcripts were marked in Word during the complexity analyses. Double vertical lines (||) were used to indicate AS–unit boundaries; a double colon (::) indicated a subordinate clause boundary. These are included in all of the examples given in the remainder of Section 5.7. In these examples, pause sounds (e.g., '~~um~~') and vowel sounds added to the end of words (e.g., 'because-uh') have been removed, for ease of reading.

5.7.2.2 Basic Principles, Problems and Specific Instances

Foster, Tonkyn and Wigglesworth appear to support the idea, adopted from T–unit analysis mainly of written texts, that "coordinated main clauses should be treated as separate units" (2000: 363). This is not stated explicitly or clearly exemplified, but repetition of the subject appears in an example in their appendix (2000: 372):

A: || I bought this cassette recorder here last weekend ||

B: || yes ||

A: || and when I took it home :: I couldn't record with it ||

B's 'yes' is a complicating factor, but could be accounted for as being an "interruption" or instance of "scaffolding" (Foster, Tonkyn and Wigglesworth, 2000: 369). Examples given in another relatively rare instance of published explanations in this area – Ellis and Barkhuizen (2005: 146) – support the separation of coordinated independent clauses, so

the principle applied to the current analysis was that coordinated independent clauses featuring a repeated or different subject were treated as belonging to separate AS–units:

(36) || we can do part time job || and we can belong to club ||

There is a lack of clarity concerning the distinction between "clause" and "verb phrase" in Foster, Tonkyn and Wigglesworth (2000). They contrast the above treatment of coordinated independent clauses with the treatment of coordinated verb phrases by asserting that, "where coordination of verb phrases occurs, the coordinated phrases will normally be considered to belong to the same AS–unit" (2000: 369). This was interpreted as stating that, where coordination occurs in the absence of a repeated or different subject, one AS–unit of two clauses should normally be recorded.[19] In the following example, 'we can' became a subject-free 'and can', so two clauses were marked:

(37) || through this kind of activities we can s- improve our connections between t-
between teachers :: and can understand more about the teachers ||

This was described as being "normally" the case because Foster, Tonkyn and Wigglesworth also have criteria to identify a new start when coordination occurs in the absence of a repeated or different subject: if there is an intonation change and a pause of at least 500 msecs, two AS–units are recorded (2000: 367). They chose the pause length for reasons of convenience only, but it was retained here in the absence of justification for an

[19] This classification in Foster, Tonkyn and Wigglesworth's (2000) system is not often commented on in the literature. In effect, some forms of coordination are classed as subordination. The obvious problem is that what would be considered to be a series of subject-free coordinated clauses (Beaman, 1984) becomes conflated with genuine subordination. An analysis of the impact of this on numerical measures of complexity is worthy of investigation, but was beyond the scope of this study.

alternative. The combination of such a pause with a change in intonation was not common, but did occur; for instance:

(38) || the boy the boy didn't like the girl [falling intonation; 611 msec pause] || and and experience experienced something ||

Turning to subordination, the definition of a subordinate clause allowed very short combinations to be regarded as clauses:

(39) || I want :: to go back ||

This highlights the fact that the AS–unit is a highly verb-centred unit of analysis. Of the three possible components – sub-clausal unit; independent clause; subordinate clause – only the first does not have to contain a verb in Foster, Tonkyn and Wigglesworth's (2000) system. It is not, however, a simple matter of one verb per AS–unit or subordinated clause, as a subordinated clause must also contain a subject, object, complement or adverbial, thus 40a, not 40b or other possible variants, was recorded:

(40) a. || because losing will allow me :: to do what I want to do ||

b. || because losing will allow me :: to do what :: I want to do ||

The extent of ellipsis that Foster, Tonkyn and Wigglesworth permit in a sub-clausal unit is unclear: the very brief utterances that they give in an extended example (2000: 372–373) are from a transactional interaction, so do not transfer well to the current study's data. A response of three words or fewer was already excluded from analysis in this study (see

Section 5.4), so, to maximise consistency, the absence of a verb meant that a sub-clausal unit was not established. A consequence of this is in accordance with what is implied by Foster, Tonkyn and Wigglesworth (2000): ellipsis is not permitted in a subordinate clause. Thus:

(41) || I saw the news when I high school student ||

Here, subordination using 'when' is clearly attempted, but the omission of a second verb (the obligatory 'was') meant that the attempt could not be treated as successfully creating a subordinate clause. More broadly, if a word in a relative clause was optional for grammatical completeness, it was assumed to be present, but if such a word was obligatory, its absence meant that a relative clause – and subordination – was not assumed. This has the disadvantage of giving no credit for an attempt at subordination, but the advantage of consistency in applying a principle of analysis.

'That'-clauses were typically treated as subordinate. The null form is possible in some constructions, which created some ambiguity in the data as to when a 'that'-clause was being used. The principle applied was that, where such ambiguity existed and the null form was not required for semantic reasons, the ambiguous part started a new AS–unit, where possible. Thus, in 42, the television programme could have mentioned two things (the machine can be made anywhere; people do not have secrets), but the second, being without an explicit 'that', was incorporated into a new AS–unit:

(42) || because this tv programme also mentioned :: that this machine can maybe made at everywhere || and you have n- you do not have your personal private s- private secret ||

The remaining instances are all of things not discussed by Foster, Tonkyn and Wigglesworth (2000) and not readily deducible from the examples that they give. First, where use of a non-word led to doubt over the intended word class, the attempted word was treated as not being a verb, to avoid giving credit for added complexity where it was not definitely attempted. Thus, in the following example, the non-word 'coperate' was treated as being an attempt at 'cooperation', meaning that there was no subordination, while the word 'encourage' in the next AS–unit was sufficient to establish subordination.

(43) || I learned about importance of coperate || and I learned about I learned about :: encourage each peoples ||

Second is the treatment of 'be' forms such as 'be going to' and 'be able to'. In this analysis, the components of such forms were regarded as integrated, so they were not split into separate clauses:

(44) || I'm going to tell you about school trip ||

(45) || but I I was ab- able to be to know about the history ||

Finally, the use of 'so' and 'and' at the start of clauses, without a clear coordinating intent, was common. As Leech (2000: 705) suggests, these were treated as discourse connectives rather than as grammatical conjunctions.

5.7.2.3 Counting

The required counts were done in Word. This could be done automatically by identifying the number of double vertical lines (‖) and double colons (::) in the transcripts.

5.8 Accuracy

The measures selected in the previous chapter to evaluate accuracy were error-free AS–units and error-free clauses. These were calculated as a percentage of the total number of AS–units, and as a percentage of the total number of clauses, respectively.

The following were required to calculate these values: the number of AS–units, the number of clauses, the number of error-free AS–units and the number of error-free clauses. Much of the preparation was done in the identification of disfluencies for fluency calculations, while the counting of AS–units and clauses was completed for complexity calculations. These are detailed earlier in this chapter, so this section describes only those principles, problems and examples that were specific to accuracy calculations. As with the complexity examples, pause sounds and vowel sounds added to the end of words have been removed, for ease of reading.

5.8.1 Basic Principles, Problems and Specific Instances

5.8.1.1 AS–units

Only fluent parts of speech were considered when judging accuracy: those parts marked for fluency analysis as disfluent were disregarded. This followed Vercellotti (2015: 7) and

the recommendations of Ellis and Barkhuizen (2005: 151) and was done because the final attempt at communicating an idea was taken to be the definitive one. Accordingly, the basic principle was that any uncorrected error of syntax, morphology or lexis meant that an AS–unit was not classed as error-free, whereas accuracy in pronunciation was not required.

Where there was ambiguity about the intended meaning and, thus, about the accuracy of an AS–unit, it was recorded as being accurate. The following example concluded a description of one participant's leadership role in a school trip. An article could have been required either before or after 'really', but the meaning conveyed without one was also possible, so this was marked as accurate:

(46) || so it was really good experience for me ||

Syntax alone was insufficient to establish accuracy; the meaning required by the context also had to match the syntax employed. In the first AS–unit of the following example, 'now I'm studying the UK' was syntactically accurate, but the context showed that 'the UK' was the location, not the topic of study, so should have been 'in the UK'. This was, therefore, not classed as an error-free AS–unit:

(47) || now I'm studying the u- UK || so I want :: to talk about the my s- school experience in st- Saint Bede st- Saint Bede ||

Similarly, 'and they killed many people' in the next example was syntactically accurate, but, in the context of discussing television programmes, the chosen construction, giving the dramas agency, was not appropriate:

(48) || because because the dr- there was a lot there were a lot of murder drama || and they killed many people || and the the reason was ver- a lot the reason was many ||

Word type choices that could be considered errors from a highly prescriptive position but which are in common use among those who speak little but English were treated as accurate. Example 49 is of adverb versus adjective choice; 50 is of present perfect versus past simple tense choice:[20]

(49) || and my partners also did very good ||

(50) || yeah but when I but recently I didn't watch tv ||

Doubts over whether a lexical choice or pronunciation error had occurred were resolved by referring to the remainder of the recording and transcript. In the following example, 'low' was treated as a lexical error because there were numerous examples of accurate pronunciation of /ɔː/ in the participant's speech and the context required 'law', 'rules' or a similar word.

(51) || he she was tired of the low [should be 'law'?] in in the palace ||

Incorrect morphology that created a non-word was inaccurate. The 'importance of coperate' from 43, above, is an example.

[20] For flummoxed readers in the future: 2016 standards are that 'good' is commonly used as an adverb (e.g., 'I am good', once an adjectival proclamation of virtue, is a largely phatic phrase masquerading as an expression of robust health) and that 'recently' has, of late, increasingly come to be used with past simple rather than present perfect constructions.

Use of the historical present – referring to past events using the present tense – is common in English, especially in narratives (Wolfson, 1979; Schiffrin, 1981). The learners in this study, however, were unlikely to have had the control required to regulate changes between the past continuous or past simple and the historical present, to take the most straightforward instances. The principle applied in this analysis was that, if the use of a present tense when referring to the past was not preceded by the use of a past tense, it was assumed to be not a deliberate use of the historical present. This is supported by evidence that L1 English narratives rarely employ only the historical present and rarely begin with that tense (Schiffrin, 1981). If this condition was not met, the AS–unit in question was not considered to be free of error.

Using the appropriate pronoun – 'he' or 'she' – when giving a narrative is difficult for both Japanese and Chinese L1 speakers. Where the wrong one was selected, this was regarded as an error.

If the title of things such as organisations, films and books was inaccurate, this was not regarded as an error. The following was therefore accurate, despite the absent article:

(52) || My favourite film is Sound Of Music ||

Use of an L1 word that is also used in English or which is not conventionally translated was not treated as an error. Syntax related to this was also treated as being flexible. The next example contains the Japanese word 'yukata' (an item of clothing), which is normally used unchanged in English and could take or not take an article and could be used uninflected as a plural, so was accurate:

(53) || and they have wear yukata ||

Finally, in the discursive parts, a response that was accurate but demonstrated that the question had been misunderstood was marked as error-free. In this example, the topic was changes in television programmes over time, but the participant took the new question to be about his own future:

(54) /Ok. So do you think this will change in the future?/

|| I I don't think I don't think :: it will change my future ||

These principles and specific instances for assessing the accuracy of AS–units also applied when assessing the accuracy of clauses within AS–units. Further clause-specific examples are given next.

5.8.1.2 Clauses

The overarching principle for clause accuracy was that a clause should not be judged in isolation. One instantiation of this was that all of the clauses in an AS–unit could not be marked as accurate if the whole AS–unit was not marked as error-free. In the following example, both clauses could be taken to be accurate, but 'it is a' was needed for the whole to be error-free, so only one was marked as accurate:[21]

(55) || And I think :: really relaxing study experience and I I like it ||

[21] In this example, the first clause was marked as accurate. For the purposes of the analyses done here, however, such choices made no difference.

In contrast, an inaccurate clause in one part of an AS–unit did not condemn all subsequent clauses to being inaccurate. The most common instance of this was the use of tenses. In 56, the participant was talking about the past, so the first two clauses were classed as inaccurate, despite being syntactically correct, but the last was marked as error-free because it could be referring to the past:

(56) || and and i- it's it it is very important for me :: because it helps me :: to be more confident ||

Syntax alone being insufficient to establish accuracy was also relevant to some responses to questions in the discursive parts. These were marked as inaccurate when the chosen structure was not appropriate to the question asked. In 57, therefore, 'it is' was not accurate (the participant's continuation could have made it accurate, but failed to do so):

(57) /Do you think people watch too much television?/
|| I think :: it is. ||

5.8.2 Counting

AS–units that were error-free were highlighted in the Word-based transcripts. Clauses that were error-free were also highlighted, but a different colour was used. Instances of each were then counted.

5.9 Statistical Analyses

The statistical analyses performed were related to the overall research design, which was summarised in Section 3.3. To compare across L1s, the starting point for each part – long-turn and discursive – was a repeated measures multivariate analysis of variance (MANOVA). This tested for an interaction effect of L1 and backchannel frequency, as well as for main effects of L1 and, separately, of backchannel frequency. Finding an interaction effect would indicate that the L1 groups differed in some way in relation to the effects of backchannel frequency. The dependent variables included in a MANOVA should be justified by theory, meaning that they need to be related conceptually (Dattalo, 2013: 2–3). As measures of complexity, accuracy and fluency are regarded as being inter-related facets of speech (see Section 3.2.1), including all eight of the dependent variables in the MANOVA was justified.

There are several assumptions underlying a MANOVA. The literature is inconsistent on what they are and how important each is, but the checking of the following assumptions is either mentioned by all sources consulted or is stated to be very important by a high proportion of sources: multicollinearity; normality; outliers; and homogeneity of covariance matrices (e.g., Dattalo, 2013: 2–3; Field, 2013: 642–643; Verma, 2015: 189–238). Multicollinearity can be checked using correlation coefficients; values exceeding .900 indicate the presence of multicollinearity, which should be avoided (Verma, 2015: 206). The normality assumption is multivariate, and univariate normality is a necessary, but not sufficient, condition for multivariate normality (Dattalo, 2013: 16). The data were therefore first checked for univariate normality of distribution using a combination of visual inspection of Q-Q plots and the running of Kolmogorov-Smirnov

goodness-of-fit tests. MANOVA is generally reported as being robust to violations of normality, and violations are likely to lead to a loss of power, rather than an increased risk of Type I errors (Bray and Maxwell, 1985; Larson-Hall and Herrington, 2009). Outliers were first checked for using stem-and-leaf plots and any potential strong outliers were then compared against the mean plus or minus three standard deviations. The labelling as an outlier any point that is beyond ±3 standard deviations is commonly used and was employed recently for CAF speech data by Révész, Ekiert and Torgersen (2016). Homogeneity of covariance matrices can be judged using Box's M test, but this is known to be highly sensitive to deviations from normality (Bray and Maxwell, 1985). As with normality, MANOVA is regarded as being reasonably robust to violations of homogeneity of covariance matrices (Dattalo, 2013: 50), although the situation is less clear when sample sizes are not equal (Bray and Maxwell, 1985). Where sample sizes are unequal, an alternative approach is to compare the sizes of the variances and covariances across the two groups (which are the two L1s in this instance): if the larger sample has greater variances and covariances, then a Type I error is unlikely; if the opposite is true, then there is a raised risk of such an error (Tabachnick and Fidell, 2012). Where sample sizes are nearly equal, robustness to violations of homogeneity of covariance matrices may be similar to when they are equal (O'Brien and Kaiser, 1985); in this research, L1 group sizes were close to equal (34 and 37) and group sizes for the independent variable backchannels were equal, so violations of this assumption might not have serious implications for the test's robustness.

The repeated measures MANOVAs indicated whether or not there were statistically significant differences across the CAF measures, based on L1, backchannel frequency, or an interaction between L1 and backchannel frequency. Finding an interaction effect or a

main effect of backchannels meant that further examination, using *t*-tests, could then investigate where any differences were located. The study was interested principally in these direct comparisons of the speech of the L1 Japanese group when receiving the typical backchannel rate with their speech when receiving the low rate, and, separately, of the speech of the L1 Mandarin Chinese group when receiving the typical backchannel rate with their speech when receiving the low rate. The research design that permitted this had the considerable merit of simplicity and was also based on an acknowledgement that the two groups of participants varied in some characteristics (see Section 4.3.2.2.5).[22] In addition, *t*-tests were done to check that the two topics were of equivalent difficulty and to check for sequencing effects.

Paired samples *t*-tests were used, as was appropriate for the repeated measures design. All *t*-tests were two-tailed paired samples with an alpha level of .05. Where the previously described checks of normality indicated deviation from a normal distribution, comparisons based on 1,000 bias-corrected accelerated bootstrap samples were done. The bootstrapping tool is a comparatively recent one to applied linguistics research and involves resampling repeatedly, making the statistical analysis "less sensitive to irregularities such as outliers, thus providing descriptive and test statistics that are robust to deviations from normality in the original sample" (Plonsky, Egbert and LaFlair, 2015: 593).

Another relatively recent movement concerning the use of statistics in applied linguistics is the encouraging of a change away from null hypothesis significance testing (NHST) and towards the greater reporting of confidence intervals (CI) and effect sizes (Larson-Hall,

[22] One characteristic that could vary was proficiency. The units of analysis to be used were CAF ones, not language test bands, and there were likely to be differences in these between the groups, which might not all consistently point in the same direction of either higher or lower proficiency.

2015: 128–167; Norris, 2015). As the current situation is likely to lie between these two visions, both *p* values from NHST and 95% CI with effect sizes were calculated. All inferential statistics were done using IBM SPSS Statistics 23.

Effect sizes (Cohen's *d*) were calculated using the formula presented below. This is the recommendation of Cumming (2012: 290–291) as being the most appropriate for repeated measures designs.

$$d = \text{Mean}_1 - \text{Mean}_2 / \text{SD}_{pooled}$$

$$\text{where } \text{SD}_{pooled} = \sqrt{[(\text{SD}_1^2 + \text{SD}_2^2) / 2]}$$

There is debate among statisticians over the need to make adjustments to the alpha level when conducting multiple comparisons of pairs of means. Opinions vary, from recommending Bonferroni adjustments (dividing the alpha level by the number of comparisons) to encouraging more comparisons, albeit with a conservative interpretation of the outcomes (Cumming, 2012: 421). This problem is reduced, if not wholly eliminated, by using CI (Cumming, 2012: 422; Larson-Hall, 2015: 287). The position taken here follows Cumming, who argues that a reasonable number of comparisons, if planned for rather than post hoc, and with an underpinning rationale, means "we can reach reasonably confident conclusions" (2012: 422–423). Therefore, no adjustments were made to the alpha level and, by examining 95% CI in addition to *p* values, caution was applied in interpreting the outcomes of statistical comparisons.

5.10 Reliability Checks

5.10.1 Background

Reporting estimates of reliability is important and can be done in multiple ways (Plonsky and Derrick, 2016). Mackey and Gass observe that there are no widely accepted guidelines for applied linguistics research and recommend that "researchers should state which measure was used to calculate […] reliability, what the score was, and, if there is space in the report, briefly explain why that particular measure was chosen. Some researchers also explain how data about which disagreements arose were dealt with" (2016: 141).

To try to identify the best ways of reporting reliability in this area of study, the ten recent publications mentioned in Section 4.4.2.1 were referred to, as they used several CAF measures for L2 speech. No consistency was found: some did not report any reliability estimates; either intra-rater or inter-rater checks were done; reliability data were given for all CAF measures collectively or for each measure separately; and some reported on final CAF measures, while others reported on only a component used in calculating those measures. As no set of standards was apparent, what is reported next is what was done in the current research and the reasons that those things were considered appropriate.

5.10.2 Choices in Procedure

The training of raters for inter-rater reliability checks often has to go through several iterations before acceptable consistency is obtained. Although reasons for this are rarely or never given in the literature, the common statement that 'differences were resolved via discussion' indicates that it is often attributable to the new raters not acting in ways that are

similar to the original rater. Having only one rater and unmodified protocols should allow reliability checks to indicate flaws in those protocols rather than combining this with variation attributable to using different raters. For the current study, the amount of detail in this chapter on how the CAF measures were calculated illustrates that preparing an additional person to the standard required would take considerable training. As this would add little to a suitably time-delayed and careful reliability check done by one rater, and because checking the reliability of the protocols was considered to be of greater importance than being able to claim consistency across raters, intra-rater checks were chosen. (The detail in this chapter on the calculating of CAF measures also serves to show the care taken to improve the reliability of the calculations.) The intra-rater checks were done at least three months after the initial data analyses. A minimum of three months was not sufficient to remove all familiarity with the recordings and transcripts, but was likely to have been more than sufficient to be regarded as a fresh start for the CAF analyses of them.

A random number generator was used to select four Japanese and four Chinese participants for the fluency calculations and the same number for the accuracy and complexity calculations combined. This was a random sample of 11% of the data (8/71). The percentage of data re-analysed varies considerably in the literature, but 10% is common. In this research, data for the long-turn and discursive parts are reported separately, so reliability statistics were calculated separately for them, too.

An important consideration was the point from which the second analyses should be done. For this research, the starting point was the transcripts. This excluded transcription from the reliability checks, but the final version of the transcripts was the product of listening to the recordings numerous times, both to produce the initial transcript and then to conduct

CAF analyses. As amendments to the original transcripts were made based on listening during the CAF analyses, the final versions were as close to definitive for the purposes of this research as was likely to be possible. Conducting reliability checks only on the remaining parts of the data analysis process – those that had not already been subject to multiple checks – was therefore judged to be sufficient. Put into practice, this meant that the following were measured again for the randomly selected 11% of the data: for fluency, number of disfluent syllables, number of fluent syllables, number of pauses, total speaking time and total pausing time; for complexity, number of words, number of AS–units and number of clauses; and for accuracy, number of error-free AS–units and number of error-free clauses. These were then used to calculate each of the CAF measures again.

For the current research, the final CAF measures, not components of them, were considered to be the most appropriate units to analyse for reliability, as these were what were included in subsequent statistical tests. Although some studies merge the reliability statistics from several CAF measures, they are given individually here, to avoid masking differences. Choosing the most appropriate reliability measure is considered next.

Numerous measures of reliability have been reported, sometimes with no consideration for the nature of the data – nominal versus ordinal, for instance – that are being checked (Feng, 2014). The current research generated data of a continuous nature (in some instances expressed as a percentage), so percentage agreement was appropriate and had the added merit of simplicity (Mackey and Gass, 2016: 140).

Calculating percentage agreement should not be done by adding up the values obtained for a CAF measure in the second analysis and comparing the resulting total with that obtained

for the same measure in the original analysis, as this may allow differences to cancel out. Instead of this, percentage differences between the two analyses can first be calculated for each participant. In the reliability checks done for this study, the first step was to subtract the second analysis' value from the original analysis' value. This difference was then expressed as a percentage of the original analysis' value, any negative percentages were changed to positive, and then the mean of the percentages across the sample was calculated.

Some imprecision is inherent in measures of reliability that do not account for another way in which differences can cancel out. A simple, hypothetical example illustrates this. If there are three AS–units – a, b, c – and the original analysis reports that only b is error-free, while the second analysis reports only c as error-free, then reliability is 100% (because both reported 1/3 as error-free). This problem was relevant to the measures used here, so was recognised by noting, for syllables (fluent and disfluent), AS–units and clauses, when such complete or partial cancelling out occurred. No action was taken to counter the effects of this, as such instances were not common, but the occurrences are reported along with the reliability percentages in the next section.

5.10.3 Outcomes

The reliability percentages are in Table 5.2. Reliability was very high (at least 96%) for all fluency and complexity measures. It was lower but still acceptable for accuracy measures (91 to 94%). Long-turn and discursive part reliability did not differ markedly.

Reliability for measures of accuracy being lower may be attributed to a combination of two factors. One is simply that accuracy was low among the participants, so even a

difference of one error-free AS–unit or clause between the original and second analyses could lead to a large percentage difference in accuracy values. There is little doubt, however, that calculating the accuracy values used here is also subject to a greater element of subjectivity than is calculating the fluency and complexity values.

Table 5.2 Intra-rater reliability

Measure	Long-turn reliability (%)	Discursive reliability (%)
PSR	98.61	98.59
MLR	96.20	98.73
NPTR	98.82	99.27
CLperAS	99.04	96.79
WperAS	99.67	96.70
WperCL	99.36	99.45
E-freeAS	93.94	90.57
E-freeCL	93.29	93.80

Note: PSR = pruned speech rate; MLR = mean length of run; NPTR = non-phonation time ratio; CLperAS = clauses per AS–unit; WperAS = words per AS–unit; WperCL = words per clause; E-freeAS = error-free AS–units; E-freeCL = error-free clauses

As mentioned at the end of the previous section, instances of differences between the original and second analyses of syllables (fluent and disfluent), AS–units and clauses completely or partially cancelling out were recorded. These are listed in Appendix 7. Each of the eight participants used for the reliability checks did two long-turn parts and two discursive parts, so the potential maximum number of instances of cancelling in one of those parts was 16. The highest number of instances found was six (for AS–units and clauses in the discursive parts), the next highest was three, while half of the analyses contained no instances. The cancelling out phenomenon was therefore rare for syllables, AS–units and clauses.

Finally in this section and chapter is a brief statement of what was done when there were differences between the original and second analyses. If there was a clear error in the original analysis, it was changed and the reported CAF calculations were updated. In all other cases, the original version was retained.

6 RESULTS

The specific research questions call for the complexity, accuracy and fluency of the participants' speech to be measured. This chapter reports the results for all of these, separated into two main sections by the two parts of the data collection process – long-turn parts and discursive parts. Within these main sections, the MANOVA findings and each of the complexity, accuracy and fluency results are reported separately. For each CAF component, descriptive statistics (including means and standard deviations for all measures, for ease of comparability) and *t*-test results are given separately for each of the L1 groups. An additional section considers age, gender, L1 varieties and L2 proficiency. The final main section in the chapter describes relevant questionnaire responses.

Although the two topics used in the data collection process were selected based on evidence that they were of equivalent difficulty (see Methods Section 4.3.4.1), statistical checks were done to establish if this held for the participants in this study. The same checks were made for sequencing effects, to see if there were CAF differences in participants doing the procedure the first time versus the second time. No L1-backchannel frequency interaction effects were expected for the data when comparing CAF measures based on topic or sequence, so *t*-tests were done without an initial MANOVA for these checks of the data collection procedure.

For the CAF *t*-tests, most of the recommendations of Larson-Hall and Plonsky (2015) on improving the reporting of L2 research findings via descriptive statistics and effect sizes have been followed. Although, as they argue (2015: 138–139), including confidence intervals (CI) makes the reporting of *p* values redundant, they have been included here for

reader convenience and both are commented on. All *t*-tests reported in this chapter were two-tailed paired samples and had a .05 alpha level. Degrees of freedom for the *t*-tests were 36 for the Japanese group and 33 for the Chinese group; to avoid repetition, these are not included in the tables presented in this chapter. The bootstrapping of non-normally distributed data procedure does not return *t* values, so some blank spaces appear in the tables. The full form of abbreviations are given beneath the table in which they first occur in this chapter and are all listed on page 10.

6.1 Long-Turn Parts

6.1.1 MANOVA

Assumptions for the repeated measures MANOVA that needed to be tested were of normality, outliers, homogeneity of covariance matrices, and multicollinearity, as described in Section 5.9. Several of the dependent variables were non-normally distributed: NPTR, clauses per AS–unit and error-free AS–units for the Japanese group; and error-free AS–units and error-free clauses for the Chinese group. The MANOVA assumption of multivariate normality was thus not met. Stem-and-leaf plots indicated one potential strong outlier in the Japanese MLR data. This was a MLR of 5.84 compared with the group mean of 3.71; as this was within three standard deviations of the mean, it was retained. For the same reason, one potential outlier in the Japanese clauses per AS–unit data was also retained in the main data set. For homogeneity of covariance matrices, Box's M test was not appropriate because of the non-normal distribution. Although the Japanese and Chinese groups were of similar size (37 and 34, respectively), they were not perfectly equal, so comparisons were made of variance and covariance values. Differences were

inconsistent in which group had the larger variance and covariance, so the evidence was inconclusive on how much trust could be placed in the MANOVA test statistic based on homogeneity of variance covariance matrices. Finally, for multicollinearity, the highest correlation among the dependent variables was .862, below the preferred maximum of .900.

Noting that these deviations from assumptions would be likely to reduce power, a MANOVA for the long-turn parts was conducted. This indicated a statistically significant effect of backchannel frequency (Wilks' Lambda = .64; $F = 4.32$; $p < .001$; $\eta^2 = .36$), of L1 (Wilks' Lambda = .29; $F = 18.58$; $p < .001$; $\eta^2 = .71$) and of the interaction between backchannel frequency and L1 (Wilks' Lambda = .75; $F = 2.59$; $p = .017$; $\eta^2 = .25$). The presence of an interaction effect did not mean that it existed for all of the CAF measures; univariate tests indicated that it was found only for the measures of accuracy: error-free AS–units ($F = 6.27$; $p = .015$) and error-free clauses ($F = 5.56$; $p = .021$). Graphical representations of the data for these two measures were used to look at the nature of the interactions; these are presented in Section 6.1.4.2. The primary area of interest in this study was the effect of backchannels, so *t*-tests were also performed to look for differences within each of the L1 groups based on backchannel rates.

6.1.2 Fluency

6.1.2.1 Japanese Group

Descriptive statistics of fluency measures for the Japanese group when doing the long-turn parts are in Table 6.1. The means and standard deviations indicate that there was higher variance in the PSR values than the others. The means were very similar and the standard deviations were similar for the topic- and sequence-based data sets. The results of the

associated *t*-tests are in Table 6.2. For the topic comparison, the *t*-tests indicate that none of the fluency measures was affected to a statistically significant extent (*p* values ranged from .260 to .994). Similarly, the *t*-tests indicate that sequence did not have a statistically significant effect (*p* values ranged from .370 to .927).

Table 6.1 Long-turn parts: Japanese group descriptive statistics for fluency

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PSR	77.18	21.29	77.19	24.48	77.34	23.10	77.03	22.77	79.84	24.21	74.53	21.26
MLR	3.62	0.73	3.64	0.82	3.64	0.75	3.64	0.80	3.71	0.84	3.56	0.69
NPTR	52.09	8.84	53.08	9.29	52.94	9.31	52.23	8.83	52.12	9.15	53.05	8.98

Note: BCs = backchannels; PSR = pruned speech rate (syllables per minute); MLR = mean length of run (syllables); NPTR = non-phonation time ratio (percentage).

When receiving the typical backchannel frequency, the Japanese group had a higher PSR than when receiving the low rate. This difference, 5.31, 95% CI [0.93, 9.69], was significant, $p = .019$ and had an effect size of 0.23. Differences in MLR ($p = .110$) and NPTR ($p = .258$) were not statistically significant, although the CI for the former crosses zero by only a short distance, suggesting that there was close to being a difference.

Table 6.2 Long-turn parts: Japanese group inferential statistics for fluency

			95% CI	95% CI			
Measure		**Mean**	**Lower**	**Upper**	***t***	***p***	***d***
Topic	PSR	-0.02	-4.75	4.72	-0.007	.994	< 0.01
	MLR	-0.02	-0.21	0.18	-0.184	.855	0.03
	NPTR	-0.98	-2.85	0.68		.260	0.11
Sequence	PSR	0.30	-4.43	5.04	0.130	.897	0.01
	MLR	0.01	-0.19	0.20	0.093	.927	0.01
	NPTR	0.71	-0.84	2.25		.370	0.08
BCs	PSR	5.31	0.93	9.69	2.456	.019	0.23
	MLR	0.15	-0.04	0.34	1.638	.110	0.20
	NPTR	-0.92	-2.48	0.75		.258	0.10

Note: Blank spaces appear because the bootstrapping of non-normally distributed data procedure does not return *t* values.

Most of the Japanese participants – 22 of the 37 – had a higher PSR when receiving the typical backchannel rate than when receiving the low rate. For 16 of these, the difference between PSR values when receiving the different backchannel rates exceeded 10%. For 11

of the 15 others whose PSR was not higher when receiving the typical backchannel rate, the difference was less than 10%. The clear trend towards a higher backchannel rate leading to a higher PSR resulted in only a low effect size, probably reflecting the high standard deviations, which are the basis of the denominator of the *d* calculation.

6.1.2.2 Chinese Group

Descriptive statistics of long-turn part fluency measures for the Chinese group are in Table 6.3. Means were similar across the topic- and sequence-based data sets, as were standard deviations. The results of the associated *t*-tests are in Table 6.4. As with the Japanese group, none of the fluency measures was affected by the topic to a statistically significant extent (*p* values ranged from .274 to .981), or by sequence (*p* values of .339 to .517).

Table 6.3 Long-turn parts: Chinese group descriptive statistics for fluency

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PSR	103.04	29.46	99.96	27.54	102.85	30.32	100.15	26.63	105.19	29.51	97.81	27.07
MLR	5.93	1.69	5.79	1.69	5.92	1.75	5.80	1.63	6.06	1.84	5.65	1.50
NPTR	36.78	10.15	36.76	10.26	37.09	10.55	36.45	9.84	35.87	10.60	37.68	9.71

For the Chinese group, participants receiving the typical backchannel frequency had a higher PSR than when receiving the low rate. This difference, 7.38, 95% CI [2.28, 12.47], was significant, $p = .006$ and had an effect size of 0.26. When receiving the typical backchannel frequency they also had a higher MLR than when receiving the low rate. This difference, 0.41, 95% CI [0.11, 0.70], was significant, $p = .008$ and had an effect size of 0.24. The difference in their NPTR approached, but did not reach, statistical significance, with the CI crossing zero by only a small amount.

Table 6.4 Long-turn parts: Chinese group inferential statistics for fluency

Measure		Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	PSR	3.08	-2.55	8.70	1.113	.274	0.11
	MLR	0.14	-0.18	0.47	0.902	.374	0.08
	NPTR	0.02	-1.97	2.02	0.024	.981	< 0.01
Sequence	PSR	2.69	-2.96	8.34	0.969	.339	0.09
	MLR	0.12	-0.21	0.44	0.739	.465	0.07
	NPTR	0.64	-1.34	2.62	0.655	.517	0.06
BCs	PSR	7.38	2.28	12.47	2.944	.006	0.26
	MLR	0.41	0.11	0.70	2.819	.008	0.24
	NPTR	-1.81	-3.70	0.08	-1.951	.060	0.18

Of the 34 Chinese participants, when receiving the typical backchannel frequency, 22 had a higher PSR, 25 had a higher MLR and 24 had a lower NPTR (all indicating higher fluency). 19 participants displayed higher fluency across all three measures when receiving the typical backchannel rate; six had lower fluency across all three measures. The PSR and MLR results, being linked with speed and automatisation of speech, respectively, in addition to pausing, indicate that these were positively affected by a higher backchannel frequency. As with the Japanese group, the effect sizes were partly a reflection of the high variation within the group, as shown by the standard deviation values.

6.1.3 Complexity

6.1.3.1 Japanese Group

Descriptive statistics of long turn complexity measures for the Japanese group are in Table 6.5. Topic and sequence means did not vary greatly, nor did their standard deviations. The associated *t*-tests are in Table 6.6. For the topic comparison, none of the complexity measures was affected by the topic to a statistically significant extent (*p* values ranged from .290 to .677). The tests also indicate that sequence did not have a statistically significant effect (*p* values ranged from .447 to .797).

Table 6.5 Long-turn parts: Japanese group descriptive statistics for complexity

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
CLperAS	1.40	0.24	1.42	0.20	1.41	0.21	1.42	0.23	1.38	0.21	1.45	0.22
WperAS	9.30	1.36	9.56	1.23	9.34	1.33	9.53	1.27	9.38	1.09	9.48	1.48
WperCL	6.71	1.01	6.80	1.06	6.70	0.98	6.81	1.09	6.92	1.01	6.59	1.04

Note: CLperAS = clauses per AS–unit; WperAS = words per AS–unit; WperCL = words per clause.

Although no statistically significant differences were found for the Japanese group for backchannel frequency received, two measures – clauses per AS–unit and words per clause – approached significance. The trend for these was in opposite directions, however: clauses per AS–unit tended to be lower when receiving the typical backchannel rate, whereas words per clause tended to be higher. As these were weak trends, little can be read into them.

Table 6.6 Long-turn parts: Japanese group inferential statistics for complexity

	Measure	Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
	CLperAS	-0.02	-0.11	0.08		.677	0.09
Topic	WperAS	-0.26	-0.77	0.24	-1.074	.290	0.20
	WperCL	-0.09	-0.55	0.36		.673	0.09
	CLperAS	-0.01	-0.10	0.08		.797	0.05
Sequence	WperAS	-0.19	-0.70	0.31	-0.77	.447	0.15
	WperCL	-0.12	-0.55	0.31	-0.56	.583	0.11
	CLperAS	-0.08	-0.18	0.02		.097	0.33
BCs	WperAS	-0.10	-0.61	0.40	-0.416	.680	0.08
	WperCL	0.33	-0.09	0.74	1.605	.117	0.32

6.1.3.2 Chinese Group

Descriptive statistics of complexity measures for the Chinese group are in Table 6.7. Means were similar across the topic- and sequence-based data sets, as were standard deviations. Table 6.8 contains the results of the associated *t*-tests. For the topic comparison, none of the complexity measures was affected by the topic to a statistically significant extent (*p* values were between .313 and .804). The tests also indicate that sequence did not have a statistically significant effect (*p* values were between .715 and .996).

Table 6.7 Long-turn parts: Chinese group descriptive statistics for complexity

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
CLperAS	1.74	0.31	1.78	0.28	1.76	0.30	1.75	0.30	1.78	0.31	1.74	0.28
WperAS	10.87	1.98	10.78	1.57	10.82	1.87	10.83	1.71	11.10	1.93	10.55	1.59
WperCL	6.31	0.85	6.12	0.77	6.18	0.69	6.25	0.93	6.32	0.94	6.11	0.66

As with the Japanese group, no statistically significant differences were found for complexity measures for the Chinese group based on backchannel frequency received (*p* values were from .123 to .579).

Table 6.8 Long-turn parts: Chinese group inferential statistics for complexity

			95% CI	95% CI			
	Measure	**Mean**	**Lower**	**Upper**	***t***	***p***	***d***
	CLperAS	-0.04	-0.18	0.09	-0.654	.518	0.14
Topic	WperAS	0.09	-0.65	0.83	0.250	.804	0.05
	WperCL	0.20	-0.19	0.59	1.026	.313	0.23
	CLperAS	0.01	-0.13	0.14	0.117	.907	0.03
Sequence	WperAS	> -0.01	-0.74	0.74	-0.005	.996	0.01
	WperCL	-0.07	-0.47	0.32	-0.368	.715	0.09
	CLperAS	0.04	-0.10	0.17	0.560	.579	0.14
BCs	WperAS	0.56	-0.16	1.27	1.581	.123	0.31
	WperCL	0.20	-0.19	0.59	1.066	.294	0.26

6.1.4 Accuracy

6.1.4.1 Japanese Group

Descriptive statistics of long-turn part accuracy for the Japanese group are in Table 6.9. Standard deviations were particularly high, being more than half of the mean for the error-free AS–units topic and sequence data and close to half for the error-free clauses data. The results of the associated *t*-tests are in Table 6.10. Neither topic nor sequence affected accuracy to a statistically significant extent.

Table 6.9 Long-turn parts: Japanese group descriptive statistics for accuracy

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
E-freeAS	25.25	14.49	21.36	13.83	22.80	14.54	23.81	14.03	22.55	10.20	24.06	17.43
E-freeCL	34.04	15.62	34.55	14.68	34.39	15.31	33.63	15.99	32.32	13.44	35.70	17.44

Note: E-freeAS = error-free AS–units (percentage); E-freeCL = error-free clauses (percentage).

For backchannel frequency, the standard deviation was much higher in the low backchannel rate data set than in the typical rate one. This was because the low backchannel rate data set contained most of the lowest and most of the highest values for both measures. The means, however, were similar and no statistically significant difference was found for either measure of accuracy (p = .623 and .308).

Table 6.10 Long-turn parts: Japanese group inferential statistics for accuracy

	Measure	Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	E-freeAS	3.89	-2.50	9.73		.188	0.27
	E-freeCL	-0.51	-6.66	5.64	-0.168	.867	0.03
Sequence	E-freeAS	-1.01	-7.22	5.23		.738	0.07
	E-freeCL	0.77	-5.93	7.47	0.233	.817	0.05
BCs	E-freeAS	-1.52	-7.36	4.21		.623	0.11
	E-freeCL	-3.37	-9.98	3.24	-1.035	.308	0.22

6.1.4.2 Chinese Group

The Chinese group's accuracy descriptive statistics are in Table 6.11. Standard deviations were high, particularly for error-free AS–units. Neither topic nor sequence affected accuracy to a statistically significant extent for either measure (p values varied from .331 to .962, as listed in Table 6.12).

Table 6.11 Long-turn parts: Chinese group descriptive statistics for accuracy

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
E-freeAS	21.13	15.32	18.28	13.44	19.78	16.97	19.62	11.48	14.48	10.58	24.92	15.87
E-freeCL	34.82	14.59	33.33	14.29	35.58	15.12	32.57	13.60	28.30	11.13	39.85	15.01

In the Chinese group, receiving fewer backchannels was associated with greater accuracy. For error-free AS–units, the difference, -10.44, 95% CI [-16.31, -4.88], was significant, p = .004 and had an effect size of 0.77. For error-free clauses, the difference, -11.55, 95% CI

[-16.51, -7.15], was also significant, p = .001 and had an effect size of 0.87, the largest found in this study.

Table 6.12 Long-turn parts: Chinese group inferential statistics for accuracy

Measure		Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	E-freeAS	2.85	-3.07	8.35		.390	0.20
	E-freeCL	1.49	-4.39	7.08		.634	0.10
Sequence	E-freeAS	0.16	-6.50	6.82	0.048	.962	0.01
	E-freeCL	3.02	-3.21	9.24	0.986	.331	0.21
BCs	E-freeAS	-10.44	-16.31	-4.88		.004	0.77
	E-freeCL	-11.55	-16.51	-7.15		.001	0.87

It was reported in Section 6.1.1 that an interaction effect of L1 and backchannel frequency was found for the two measures of accuracy. These are displayed graphically in Figure 6.1. The non-parallel, intersecting lines indicate the interactions. The error-free AS–units interaction (displayed on the left of the figure) is the more interesting, as it shows the similarity of each L1 group's percentage when receiving the low backchannel rate (there was not a statistically significant difference between the L1 groups for this), that the Japanese group was minimally if at all affected by backchannel rates and that the Chinese group was much less accurate based on this measure when receiving the typical rate.

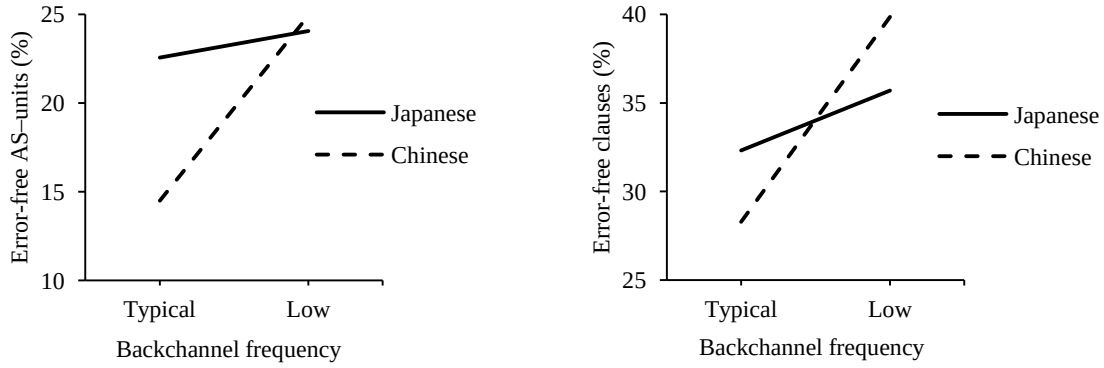

Figure 6.1 L1 and backchannel frequency interaction effects for long-turn part accuracy

Of the 34 Chinese participants, 28 followed the pattern of the low backchannel rate being associated with greater accuracy, as measured by error-free clauses, and 24 followed the pattern when judged by error-free AS–units. Improvements in error-free clauses between receiving a typical backchannel rate and a low rate varied from 1 to 48% (calculated by subtracting one error-free clause percentage from the other); for error-free AS–units the range was 3 to 53%. The participants who did not follow the pattern had an error-free clause difference of between 1 and 24% and an error-free AS–unit difference of between 1 and 14%.

One possible explanation is that more backchannels may have encouraged expansion on ideas in this group, leading to more errors, or to errors that were more difficult to self-repair, as participants sought to say more. Linked to this, the backchannels could also have been taken as signals of interlocutor comprehension, discouraging the self-repair that would have increased accuracy. An alternative explanation is that the higher backchannel frequency led to more backchannels acting as interruptions or distractions, disturbing ideas and leading to more errors.

The former explanation is favoured when the long-turn fluency results for this group (see Section 6.1.2.2) are also considered. Fluency was greater when the backchannel rate was higher. The suggestion that more backchannels encouraged the participants to continue speaking is congruent with fluency being relatively high while accuracy was relatively low. Conversely, the suggestion that more backchannels acted to interrupt the participants would mean that both accuracy and fluency should be reduced, as the disruption of the memory of the syntax of what had just been said would reduce accuracy by restricting the

ability to self-correct, while the disruption of an idea would reduce fluency by interrupting the planning of what was about to be said.

This explanation matches the data, but is largely conjecture. To help support or reject it, evidence for the mechanisms underlying the difference in accuracy was sought by examining the data in more detail. The first more detailed analysis that was done was to look for differences in accuracy caused directly and immediately by a backchannel being given. This was done by examining and comparing the long-turn part transcripts of examples from three sets of Chinese group participants: those that followed the pattern of elevated accuracy when receiving the low backchannel rate; those that followed the pattern and were extreme examples of it (i.e. their accuracy was greatly increased); and those that did not follow the pattern (i.e. they were less accurate when receiving the low rate). As already mentioned, lower accuracy when receiving the typical backchannel rate could have been because of an increase in errors (in other words, a decrease in perfectly formed clauses) or a decrease in self-repair. The comparison of the different sets of Chinese group participants revealed no trend in or between any of the sets for backchannels immediately or soon leading to errors in clauses, to reduced self-repair, or to self-repair resulting in error-free clauses. The conclusion was that the effect of backchannels on accuracy for this group in the long-turn parts was not a simple matter of a backchannel directly influencing the speech that immediately followed it.

Another more detailed set of analyses of the data was therefore performed. These included all of the Chinese participants during the long-turn parts. The initial step was to calculate the percentage of their clauses that were both error free and contained no self-repair. This was very similar across the two backchannel frequencies: 21.24% when the typical

frequency was received and 21.86% when the low frequency was received. This demonstrated that the Chinese group's greater accuracy when receiving the low backchannel rate was not attributable to a greater proportion of perfectly formed clauses being produced at the first attempt. The difference must, instead, be attributable to self-repair.

The quantity of self-repair (i.e. the number of self-repairs in relation to the number of clauses) was also almost identical when receiving the two backchannel rates, so the difference in accuracy appeared to lie in the proportion of self-repairs that resulted in error-free clauses. Confirmation of this was sought by making a comparison between self-repair that led to an error-free clause being formed and self-repair that did not lead to an error-free clause being formed. (The identifying of self-repairs is described in Section 5.6.1.1.[23] Using error-free clauses was important as this was one of the two measures of accuracy for which a difference was found. Error-free clauses were preferred to error-free AS–units because there were more instances of them.) As shown in Figure 6.2, there was a considerable increase in the percentage of self-repairs that led to an error-free clause being formed when the low frequency of backchannels was received compared with when the typical rate was received: 31.74% compared with 19.63%. Including repetitions and abandonments (again, as described in Section 5.6.1.1) in the calculations made very little difference to these percentages. This supported the idea that the difference in accuracy for the Chinese group in the long-turn parts was attributable to self-repair, specifically, more

[23] Briefly, self-repairs had already been marked, as a type of disfluency, and an instance of self-repair was counted here only when it was followed by at least one word that was not marked as disfluent. The following was therefore marked as containing two self-repairs: || and-uh she ask ~~uh~~ she asked her class classmates || If it had been a little different, one self-repair would have been recorded: || and-uh she ask her class ~~uh~~ she asked her classmates ||

successful self-repair, when the low rate of backchannels was received. The long-turn part results end here; the next section presents results from the discursive parts.

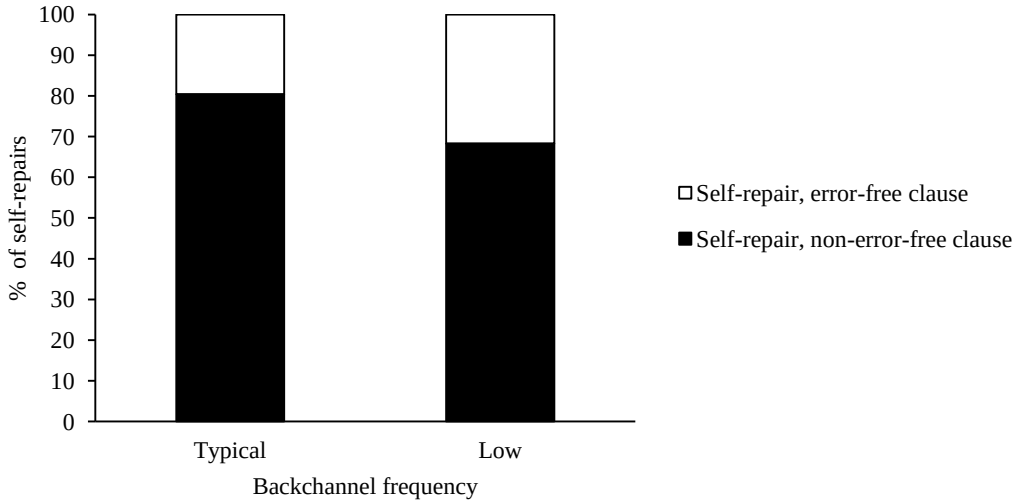

Figure 6.2 Chinese long-turn part self-repairs leading to error-free / non-error-free clauses

6.2 Discursive Parts

This section reports the descriptive and inferential statistics for the discursive parts of the recordings. As in the long-turn section, each CAF component is reported on separately for each of the L1 groups, after the initial MANOVA.

6.2.1 MANOVA

As for the long-turn parts, the assumptions that needed to be tested were those of normality, outliers, homogeneity of covariance matrices, and multicollinearity. One of the dependent variables was non-normally distributed: MLR for the Chinese group. The MANOVA assumption of multivariate normality was thus not met. Stem-and-leaf plots did not indicate any strong outliers. For homogeneity of variance covariance matrices, Box's M

test was not appropriate because of the non-normal distribution. Although the two groups (Japanese and Chinese) were of similar size, they were not perfectly equal, so comparisons were made of variance and covariance values. Differences were inconsistent in which group had the larger variance and covariance, so the evidence was inconclusive on how much trust could be placed in the MANOVA test statistic based on homogeneity of variance covariance matrices. Finally, for multicollinearity, the highest correlation among the dependent variables was .839, below the preferred maximum of .900.

Noting that these deviations from assumptions would be likely to reduce power, a MANOVA for the discursive parts was conducted. This indicated a statistically significant effect of backchannel frequency (Wilks' Lambda = .69; $F = 3.51$; $p = .002$; $\eta^2 = .31$) and of L1 (Wilks' Lambda = .31; $F = 17.65$; $p < .001$; $\eta^2 = .69$), but not of the interaction between backchannel frequency and L1 (Wilks' Lambda = .93; $F = 0.59$; $p = .781$; $\eta^2 = .07$). It can be stated with confidence that this last finding was not attributable to a reduction in power, given the *p* value of .781. The primary area of interest in this study was the effect of backchannels, so finding a main effect of L1 without an interaction effect meant that differences within each L1 group, rather than between them, became the focus of further investigation: the identifying of a main effect of backchannel frequency meant that *t*-tests were performed to look for differences within each of the L1 groups based on backchannel rates.

6.2.2 Fluency

6.2.2.1 Japanese Group

Descriptive statistics of fluency measures for the Japanese group in the discursive parts are in Table 6.13. For Topic A compared with Topic B and for sequence in the procedure (first or second), the means and standard deviations show that there was higher variance in the PSR values than the others. The results of the associated *t*-tests are in Table 6.14. For the topic comparison, none of the fluency measures was affected to a statistically significant extent (*p* values ranged from .234 to .463). Similarly, sequence did not have a statistically significant effect (*p* values ranged from .417 to .923).

Table 6.13 Discursive parts: Japanese group descriptive statistics for fluency

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PSR	78.33	20.14	75.49	19.09	76.75	18.61	77.07	20.68	81.18	19.99	72.64	18.37
MLR	3.49	0.62	3.55	0.67	3.49	0.65	3.55	0.65	3.63	0.66	3.41	0.62
NPTR	53.09	6.99	53.88	7.90	53.53	7.12	53.45	7.80	52.52	7.44	54.46	7.37

When receiving the typical backchannel frequency, the Japanese group had a higher PSR than when receiving the low rate. This difference, 8.54, 95% CI [4.64, 12.44], was significant, $p < .001$ and had an effect size of 0.44. MLR was also higher when receiving the typical backchannel frequency. The difference, 0.22, 95% CI [0.08, 0.35], was significant, $p = .002$ and had an effect size of 0.34. Finally, NPTR also differed significantly with backchannel frequency. The difference, -1.94, 95% CI [-3.39, -0.49], was significant, $p = .010$ and had an effect size of 0.26. The values for NPTR show that it was lower when receiving the typical backchannel frequency, which, as with the other two measures, indicates higher fluency.

Table 6.14 Discursive parts: Japanese group inferential statistics for fluency

Measure		Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	PSR	2.84	-1.92	7.59	1.211	.234	0.14
	MLR	-0.05	-0.20	0.09	-0.742	.463	0.09
	NPTR	-0.79	-2.36	0.79	-1.015	.317	0.11
Sequence	PSR	-0.32	-5.17	4.53	-0.133	.895	0.02
	MLR	-0.06	-0.21	0.09	-0.822	.417	0.09
	NPTR	0.08	-1.52	1.67	0.098	.923	0.01
BCs	PSR	8.54	4.64	12.44	4.445	< .001	0.44
	MLR	0.22	0.08	0.35	3.341	.002	0.34
	NPTR	-1.94	-3.39	-0.49	-2.704	.010	0.26

Of the 37 participants, 29 had a higher PSR when receiving the typical backchannel frequency. For MLR, 27 followed this pattern; for NPTR, 24 of the 37 followed it. Just over half – 19 of the 37 – were more fluent in all three measures when receiving the typical backchannel frequency. Only five were more fluent in all three measures when receiving the low backchannel frequency. In comparison with the solo speaking style of the long-turn parts, the more interactional, conversational nature of the discursive parts is likely to have been more familiar to, and comfortable for, the Japanese participants. This is supported by the questionnaire findings (reported in more detail in Section 6.4) that the long-turn parts were regarded as being more difficult by 15 of the Japanese participants, while eight reported that they were easier than the discursive parts (the remainder stated that the two parts were of similar difficulty). This could explain why all three fluency measures, rather than only the speed-based PSR in the long-turn parts, indicated higher fluency when the higher backchannel rate was received. A statistically significant difference in MLR and NPTR in the discursive parts could have occurred via a greater frequency of backchannels acting as reassurance for the participants, reducing self-doubt and, thus, hesitation when speaking. It is important to highlight, however, that the differences in these values when receiving different backchannel rates was small to moderate.

Several other factors could have contributed to the effect of increased backchannel frequency in the discursive parts being stronger than in the long-turn parts. A comparison of the means and standard deviations reveals that a narrowing of the variability is likely to have been one underlying element. Mean PSR, for instance, was 79.84 (SD 24.21) for the typical backchannel rate in the long-turn parts and 81.18 (SD 19.99) in the discursive parts. The SD narrowing in the discursive parts, while the mean was little changed, shows that these participants became more similar in their fluency in the discursive parts. Elements of the discursive parts that could have contributed to this include: shadowing the interlocutor's question structure when responding; the use of simple, fixed beginnings such as "I think"; using discourse markers; and, at the end of responses, summarising what had already been said.

6.2.2.2 Chinese Group

The Chinese group's descriptive statistics for fluency measures are in Table 6.15. Means were similar across the topic- and sequence-based data sets, as were standard deviations. The results of the associated *t*-tests are in Table 6.16. For the topic comparison, none of the fluency measures was affected by the topic to a statistically significant extent (*p* values ranged from .699 to .759). The *t*-tests also indicate that sequence did not have a statistically significant effect (*p* values ranged from .628 to .817).

Table 6.15 Discursive parts: Chinese group descriptive statistics for fluency

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
PSR	112.51	27.43	113.43	24.77	113.47	25.40	112.46	26.84	116.27	27.47	109.67	24.28
MLR	5.86	1.59	5.80	1.54	5.79	1.59	5.86	1.54	5.97	1.64	5.69	1.47
NPTR	33.67	8.85	33.45	8.62	33.64	9.09	33.47	8.36	32.37	8.51	34.75	8.79

Chinese participants receiving the typical backchannel frequency had higher fluency values in all three measures than when receiving the low rate. For PSR, the difference, 6.60, 95% CI [2.37, 10.83], was significant, $p = .003$ and had an effect size of 0.25. For NPTR, the difference, -2.39, 95% CI [-3.61, -1.17], was significant, $p < .001$ and had the largest effect size, 0.28. Although the difference in MLR, 0.29, 95% CI [0.03, 0.53], was significant at an alpha level of .05 ($p = .041$; effect size 0.19), the use of multiple statistical tests makes this more likely to be a Type I error. Nevertheless, the trend across the three fluency measures was consistent.

Table 6.16 Discursive parts: Chinese group inferential statistics for fluency

Measure		Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	PSR	-0.93	-5.75	3.89	-0.390	.699	0.04
	MLR	0.05	-0.22	0.29		.741	0.03
	NPTR	0.23	-1.26	1.71	0.309	.759	0.03
Sequence	PSR	1.01	-3.81	5.83	0.426	.673	0.04
	MLR	-0.07	-0.36	0.24		.628	0.04
	NPTR	0.17	-1.31	1.65	0.233	.817	0.02
BCs	PSR	6.60	2.37	10.83	3.176	.003	0.25
	MLR	0.29	0.03	0.53		.041	0.19
	NPTR	-2.39	-3.61	-1.17	-3.984	< .001	0.28

Note: Blank spaces appear because the bootstrapping of non-normally distributed data procedure does not return *t* values.

Of the 34 Chinese participants, when receiving the typical backchannel frequency, 23 had a higher PSR, 21 had a higher MLR and 22 had a lower NPTR (all indicating higher fluency). Under half – 14 of the 34 – were more fluent in all three measures when receiving the typical backchannel frequency; three of the participants were less fluent in all three measures.

In comparison with the long-turn figures for this group, PSR and NPTR indicated substantially higher fluency in the discursive parts, while MLR was similar. Standard

deviations for PSR also narrowed as the means rose in the discursive parts, as reported for the Japanese group, but the differences were less marked.

6.2.3 Complexity

6.2.3.1 Japanese Group

Table 6.17 contains descriptive statistics of complexity measures for the Japanese group in the discursive parts. Means did not vary greatly; standards deviations varied to a more than typical extent for these data between Topics A and B. Table 6.18 contains the results of the associated *t*-tests. None of the complexity measures was affected by the topic to a statistically significant extent (*p* values of .345 to .814). The *t*-tests also indicate that sequence did not have a statistically significant effect (*p* values of .202 to .590).

Table 6.17 Discursive parts: Japanese group descriptive statistics for complexity

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
CLperAS	1.47	0.17	1.51	0.25	1.49	0.22	1.47	0.21	1.48	0.24	1.49	0.19
WperAS	9.29	1.33	9.55	1.61	9.32	1.49	9.52	1.47	9.52	1.72	9.33	1.19
WperCL	6.37	0.71	6.41	1.13	6.26	0.86	6.52	1.01	6.48	0.87	6.29	1.01

For backchannel frequency, no statistically significant differences were found (*p* values were between .375 and .639). Complexity of the Japanese group's speech in the discursive parts was not affected by backchannels, topic or sequence.

Table 6.18 Discursive parts: Japanese group inferential statistics for complexity

	Measure	Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	CLperAS	-0.04	-0.15	0.06		.360	0.20
	WperAS	-0.25	-0.78	0.28	-0.956	.345	0.17
	WperCL	-0.05	-0.46	0.36	-0.236	.814	0.05
Sequence	CLperAS	0.02	-0.06	0.11		.590	0.12
	WperAS	-0.19	-0.73	0.33	-0.759	.453	0.13
	WperCL	-0.26	-0.66	0.14	-1.300	.202	0.28
BCs	CLperAS	-0.02	-0.12	0.07	-0.473	.639	0.10
	WperAS	0.19	-0.34	0.72	0.722	.475	0.13
	WperCL	0.18	-0.23	0.59	0.898	.375	0.19

6.2.3.2 Chinese Group

The Chinese group's descriptive statistics for complexity measures are in Table 6.19. The results of the associated *t*-tests are in Table 6.20. For the topic comparison, none of the complexity measures was affected by the topic to a statistically significant extent (*p* values ranged from .207 to .462). The tests also indicate that sequence did not have a statistically significant effect (*p* values ranged from .273 to .796).

Table 6.19 Discursive parts: Chinese group descriptive statistics for complexity

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
CLperAS	1.69	0.23	1.63	0.19	1.67	0.19	1.65	0.23	1.66	0.20	1.66	0.22
WperAS	10.61	1.40	10.34	1.22	10.65	1.32	10.29	1.29	10.48	1.43	10.46	1.21
WperCL	6.29	0.63	6.39	0.67	6.41	0.61	6.27	0.68	6.33	0.63	6.35	0.67

For backchannel frequency, no *p* value was lower than .916. The Chinese group's complexity of speech in the discursive parts was not affected by backchannels, topic or sequence.

Table 6.20 Discursive parts: Chinese group inferential statistics for complexity

	Measure	Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	CLperAS	0.07	-0.04	0.17	1.288	.207	0.28
	WperAS	0.27	-0.36	0.84		.394	0.20
	WperCL	-0.09	-0.37	0.17	-0.744	.462	0.15
Sequence	CLperAS	0.01	-0.09	0.12	0.261	.796	0.07
	WperAS	0.35	-0.29	0.99	1.115	.273	0.27
	WperCL	0.13	-0.14	0.40	1.006	.322	0.20
BCs	CLperAS	< 0.00	-0.11	0.11	0.007	.995	< 0.01
	WperAS	0.03	-0.62	0.68	0.085	.933	0.02
	WperCL	-0.01	-0.29	0.26	-0.106	.916	0.02

6.2.4 Accuracy

6.2.4.1 Japanese Group

Descriptive statistics of accuracy for the Japanese group are in Table 6.21. Standard deviations were high, being more than half of the mean for two of the four error-free AS–units topic and sequence data sets. The results of the associated *t*-tests are in Table 6.22. Neither topic nor sequence affected accuracy to a statistically significant extent.

Table 6.21 Discursive parts: Japanese group descriptive statistics for accuracy

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
E-freeAS	23.99	12.09	26.19	12.16	25.08	13.09	25.11	11.19	25.28	13.16	24.91	11.10
E-freeCL	35.93	12.15	38.30	12.06	38.42	12.34	35.81	11.84	36.73	12.34	37.50	11.98

For backchannel frequency, standard deviations were also high, particularly for the error-free AS–units data. No statistically significant differences were found (p = .856 and .686).

Table 6.22 Discursive parts: Japanese group inferential statistics for accuracy

	Measure	Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	E-freeAS	-2.21	-6.24	1.82	-1.111	.274	0.18
	E-freeCL	-2.37	-6.13	1.39	-1.279	.209	0.20
Sequence	E-freeAS	-0.03	-4.13	4.07	-0.016	.988	< 0.01
	E-freeCL	2.61	-1.13	6.35	1.415	.166	0.22
BCs	E-freeAS	0.37	-3.73	4.47	0.183	.856	0.03
	E-freeCL	-0.77	-4.60	3.06	-0.408	.686	0.06

6.2.4.2 Chinese Group

Descriptive statistics of accuracy for the Chinese group are in Table 6.23. Similarly to the Japanese figures, standard deviations were high. As with the Japanese data, neither topic nor sequence affected accuracy to a statistically significant extent (*p* values ranged from .168 to .797, as listed in Table 6.24).

Table 6.23 Discursive parts: Chinese group descriptive statistics for accuracy

	Topic A		Topic B		First		Second		Typical BCs		Low BCs	
Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
E-freeAS	24.58	11.20	25.25	11.79	26.08	10.44	23.76	12.38	25.94	12.39	23.89	10.46
E-freeCL	41.39	12.10	37.41	11.55	40.90	11.18	37.90	12.59	40.54	12.45	38.36	11.48

For backchannels, standard deviations were again high. No statistically significant differences were found (p = .432 and .455). This contrasts with the long-turn parts, where greater accuracy was associated with the lower backchannel rate. The explanation for the long-turn parts was that the typical backchannel rate was associated with less effective self-repair and could have encouraged more expansion on ideas (see Section 6.1.4.2). Perhaps the change in the discursive parts was that expansion was not required to the same extent: participants knew that they were expected to speak on one topic for approximately 2 minutes in the long-turn parts, whereas in the discursive parts they knew that the interlocutor would ask another question once they had finished a response. This could have reduced the pressure to expand on what had been said. In addition, knowing that the interlocutor could ask a clarification question if lack of accuracy led to loss of understanding could have reduced the effect of backchannels on the sense of need to self-repair, in comparison with the long-turn parts.

Accuracy when receiving the low rate was similar (24.92 in the long-turn parts; 23.89 in the discursive parts, using error-free AS–units), but varied substantially for the typical rate (14.48 in the long-turn parts; 25.94 in the discursive parts). This suggests that the accuracy data from the discursive parts are more likely to have been close to a 'normal' or 'baseline' level, making the long-turn part accuracy data the finding that needs to be explained. Most (20 out of 34; see Section 6.4) Chinese participants reported that the long-turn parts were more difficult, so maintaining accuracy was probably more challenging in those parts.

Table 6.24 Discursive parts: Chinese group inferential statistics for accuracy

Measure		Mean	95% CI Lower	95% CI Upper	*t*	*p*	*d*
Topic	E-freeAS	-0.67	-5.95	4.61	-0.259	.797	0.06
	E-freeCL	3.98	-1.49	9.69		.168	0.34
Sequence	E-freeAS	2.31	-2.91	7.53	0.902	.374	0.20
	E-freeCL	3.01	-2.53	8.10		.300	0.25
BCs	E-freeAS	2.04	-3.19	7.28	0.795	.432	0.18
	E-freeCL	2.18	-3.69	8.05	0.757	.455	0.18

This concludes the main presentation of the results from the discursive and long-turn parts for the Japanese and Chinese groups. The next section is also on the CAF results, but on differences that could be attributable to something other than participant L1.

6.3 Age, Regional Differences, Gender and L2 Proficiency

As discussed in the literature review chapters, there is only limited evidence for differences in L1 backchannel frequency based on age, gender, or which form of an L1 a person speaks. The lack of evidence meant that they were not included as sub-questions with the other research questions, but the possibility was left open of investigating them for the purposes of potentially informing future research directions. These things, and L2 proficiency, were therefore considered for further analysis of the data collected for this

research. The finding of an interaction effect of L1 and backchannel frequency for the long-turn parts, reported in Section 6.1.1, meant that the L1 groups were not combined in any such analysis.

6.3.1 Age

There was little difference in participant ages within either group in this study: as summarised in Section 4.3.2.2.5, all but four of the 71 participants were aged 18–22. This precluded the possibility of investigating possible age effects.

6.3.2 Regional L1 Differences

Numerical checks were made for the most common areas of origin for the participants in each group, with area of origin being taken to be a possible proxy for the form of an L1 spoken. These were Osaka in Japan (nine participants) and Jiangsu Province in China (eight participants). The effects of differences in backchannels received were compared for the members of each of these sub-groups. No indication was found of effects being systematically stronger or weaker than for the whole group: for instance, some were more fluent with more backchannels, but some were less fluent. This being the case, no statistical tests were done. A further observation is that the possibility (and relevance) of investigating the influence of regional L1 differences on L2 English interactions is likely to be fading. Although area of origin was recorded for this study, the reality is that a single place of origin for a person is becoming increasingly rare: even in China, where practical limits on geographic mobility exist because of government policies, moving from one area of the country to another during childhood is not unusual, as reported by some of the participants in this research.

6.3.3 Gender

The smallest sample size for gender was 11, for males in the Japanese group. This was low for inferential statistics, reducing power, but with the purpose of informing future research, comparisons based on gender were done in each of the L1 groups, using the long-turn and discursive parts data separately. Only those measures for which a statistically significant difference based on backchannel frequency was found in the original analyses were included.

There was only one dependent variable – PSR – that needed to be analysed in the Japanese group, so this was done using a mixed design ANOVA. The data met the assumptions (normality was assessed using Q-Q plots and Kolmogorov-Smirnov goodness-of-fit tests; homogeneity of variance was checked using Levene's test). There was no interaction effect of gender and backchannel frequency (F = 0.10; p = .755). The four relevant dependent variables in the Chinese group – PSR, MLR, error-free AS–units and error-free clauses – were analysed using a repeated measures MANOVA. First, the MANOVA assumptions detailed in Section 5.9 were checked. Not all of the data were normally distributed, there were no outliers, the evidence was again inconclusive on how much trust could be placed in the MANOVA test statistic based on homogeneity of variance covariance matrices, and multicollinearity was absent. Noting these caveats, the MANOVA was performed and it showed no interaction effect of gender and backchannel frequency (Wilks' Lambda = .86; F = 1.22; p = .325; η^2 = .14) and no main effect of gender (Wilks' Lambda = .96; F = 0.30; p = .875; η^2 = .04). It was therefore concluded that there was no evidence of gender influencing the effects of backchannel rates in the long-turn parts.

The same MANOVA process was followed for the discursive parts data. For the Japanese group, all three measures of fluency were included. Again, not all of the data were normally distributed, there were no outliers, the evidence was again inconclusive on how much trust could be placed in the MANOVA test statistic based on homogeneity of variance covariance matrices, and multicollinearity was absent. The MANOVA performed showed no interaction effect of gender and backchannel frequency (Wilks' Lambda = .97; $F = 0.36$; $p = .782$; $\eta^2 = .03$) and no main effect of gender (Wilks' Lambda = .89; $F = 1.33$; $p = .281$; $\eta^2 = .11$). For the Chinese group, PSR and NPTR were included in a MANOVA after the same observations on the assumptions had been made as for the Japanese group. Again, no interaction effect of gender and backchannel frequency was found (Wilks' Lambda = .98; $F = 0.26$; $p = .775$; $\eta^2 = .02$) and no main effect of gender was observed (Wilks' Lambda = .99; $F = 0.06$; $p = .939$; $\eta^2 = .01$). It was therefore concluded that there was also no evidence of gender influencing the effects of backchannel rates in the discursive parts.

An important final point on gender is that not all possible combinations could be compared. There was one male interlocutor and the main speakers were male or female, but the combination of female interlocutor and male or female main speaker was not included. Again, this was a matter of research design choices, and not finding evidence of a gender effect in this research should not lead to the conclusion that it does not have an effect in different combinations of speakers or other factors.

6.3.4 L2 Proficiency

One way to examine the relationship between L2 proficiency and the effects of backchannel frequency would be to conduct correlation analyses. These, however, would need to be planned for at the stage of research design, in part because a sizeable range of proficiencies would be needed (available IELTS Speaking scores for the Chinese group were 5, 5.5 or 6, which is a range of only three values), and this was not the case here. An alternative method would be to split the participants into two or more groups, based on their L2 proficiency. IELTS Speaking scores were available for almost all of the Chinese participants and TOEFL scores were available for some of the Japanese participants, so the potential for such a split could be considered.

A key consideration when contemplating the advisability of doing further analyses based on splitting an L1 group into two (the simplest possibility, in comparison with splitting into three, or more) based on L2 proficiency is the need to be confident that such a split really would separate participants according to that criterion, so that all those with relatively low proficiency would be in a 'low' group and all those with relatively high proficiency would be in a 'high' group. Without being sure of this, the outcomes of any statistical analyses would be impossible to interpret with confidence. With this in mind, it was possible to use the IELTS Speaking scores of the Chinese participants, together with data from the producers of IELTS, to evaluate whether or not a split was advisable.

The score that a person obtains in a test is classed as the observed score, which consists of the person's true score plus an error part. Reducing the error part increases confidence that the observed score is close to the true score, and the size of the error part, which is related

to the reliability of the test, can be expressed as the standard error of measurement (SEM), given in the same units as the test score (Brown, 2012). It is possible to state with 68% confidence that the true score is within one SEM of the observed score; if this is extended to two SEM, then confidence increases to 95% (Brown, 2012).

As mentioned, the IELTS Speaking scores of the Chinese participants for whom such data were available were 5, 5.5 or 6. The producers of IELTS report an SEM for the Listening component and one for the Reading component; in the most recent data, for 2015, these were 0.37 and 0.38 respectively (IELTS, n.d. c). They state that producing an equivalent value for the Speaking or Writing component is not possible, because they are not item-based (IELTS, n.d. c). They do, however, report that studies have given the Speaking component reliability coefficients in the range 0.83–0.86, which are lower than those for Listening (0.92) or Reading (0.90) (IELTS, n.d. c). Therefore, although a Speaking SEM is not available, it is possible to evaluate what the lowest SEM of an IELTS component – that for Listening – means in relation to the separability of scores in the range 5–6. With 68% confidence, an observed score of 5.5 reflects a true score of between 5.13 (5.5 minus the SEM of 0.37) and 5.87 (5.5 plus the SEM of 0.37); such values would be rounded for reporting purposes to 5 and 6, respectively. Even for the most reliable IELTS component, then, a reported score of 5.5 cannot be said with much confidence to be different from a true score of 5 or 6. The Speaking component is less reliable than this, so, again, the conclusion is that those participants who held an IELTS Speaking score of 5.5 could have had a true score of 5, 5.5 or 6, while those with 5 or 6 could have had a true score of 5.5 (this is assuming 68% confidence; obviously, increasing to 95% confidence would make the range of true scores much wider). This consideration of the participants' reported L2 proficiencies thus led to the conclusion that even a simple split into 'low' and 'high' could

not be done with confidence, so examining the relationship between proficiency and the effects of backchannel frequency would not be appropriate with this data set.

A final possibility was to use median splits based on CAF values instead of test scores. However, checks of where the splits would have been made showed that the CAF values were very close around that point, so a division into 'low' and 'high' proficiency would have been too artificial. Further, a person could have been classed as 'low' based on one measure of, for instance, fluency, while being classed as 'high' based on another fluency measure. A further possible reality, highlighting the artificiality of such a CAF median split, is that a person could be classed as 'high' based on speech when receiving one backchannel frequency but 'low' when receiving the other, introducing a paradoxical element to the process.

Considerations of possible further research into some of the matters raised in this section are made in the following, Discussion chapter. Next, however, is a report of the questionnaire data.

6.4 Questionnaires

In addition to the CAF analyses, participant responses to the questionnaire items (reproduced in Appendix 4) were also examined. This was done to check the extent of conscious participant awareness of backchannel differences and for differences in perceived task part and topic difficulty.

The questionnaire contained five items. All of the participants answered all of the items. For the item on backchannel frequency received – "Comparing the interaction with the other person… were his backchannels ('uh-huh', 'mhm', etc. while listening)" – the most common response (28 of 71 participants) was that the frequency had been "about the same" during each of the topics; 18 answered "don't know". Of the 25 who replied that the frequency had differed, 16 correctly identified which set of interactions had featured a higher backchannel frequency and nine were incorrect. The correct identification of backchannel frequency differences was thus no better than chance, calculated either in relation to the number of available responses (22.5% compared with a chance value of 25%) or in relation to the number of positive responses (ignoring "don't know", there were 53 answers; 16 correct represents 30.2%, compared with a chance value of 33.3%).

Looking in more detail, the Japanese participants were much better at identifying frequency differences than their Chinese counterparts were: of the 11 who chose one way or the other, nine were correct; the Chinese were split, with seven being correct and seven incorrect. Even when considered as a separate group, however, the Japanese participants were still no better than chance at spotting the difference: nine from 37 is below 25%. One Japanese participant, though, the only one who added a comment when responding to the item on backchannel frequency, was certainly aware of the difference that they can make. She wrote, "The backchannels made me relax to speak, thank you", and correctly identified with which topic a higher frequency had been given. Overall, however, a simple check for a link between participants identifying backchannel frequency differences and their CAF values when receiving those different frequencies revealed no clear connection. The results described earlier in this chapter show that backchannel frequency did affect some aspects of the participants' speech. As they appear to have been unaware of the

differences in backchannel frequency, the mechanisms by which this occurred are likely to have been subconscious too.

Another Japanese participant who correctly identified the difference in frequency also stated that the long-turn parts were more difficult than the two-way discursive parts: in response to questionnaire item 2 – "Comparing the individual long turn parts and the two-way discussion [discursive] parts… which was more difficult?" – she wrote that "I like talking. But when the person who I talk with don't answer anything, I feel sad." Approximately half of the participants felt that the long-turn parts were more difficult (35 of 71; 15 Japanese and 20 Chinese). Those who stated the opposite numbered 16 (eight Japanese; eight Chinese). The remainder reported little difference in difficulty. The possible implications of these questionnaire responses were discussed earlier in this chapter.

Opinions of relative topic difficulty differed similarly. 37 participants reported that Topic B was the more difficult (21 Japanese; 16 Chinese) and 13 that Topic A was more difficult (four Japanese; nine Chinese). The remainder reported little difference in topic difficulty. The checks reported above showed that there were no statistically significant differences based on topic for any of the CAF measures. Nevertheless, these questionnaire findings differed from those of the pilot study (Section 4.3.4.1), where there was a more even split of opinions on topic difficulty. All of these matters are returned to in the Discussion chapter, which follows a brief summary of the findings.

6.5 Summary of Findings

The CAF of the two L1 groups were considered for the long-turn parts and for the discursive parts. Consistent findings, for each of the L1 groups and for each part of the procedure, were that increased fluency was associated with the higher frequency of backchannels and that complexity was not affected by backchannel rate. The one finding that did vary was for accuracy: for only the Chinese group in the long-turn part, a higher frequency of backchannels was linked to lower accuracy; the interaction effects identified in the MANOVA indicated that this was a difference between the L1 groups. This was associated principally with self-repair. For the Japanese group and for the Chinese group in the discursive part, backchannel rate did not affect accuracy. With a small number of exceptions, the participants were not consciously aware of differences in the backchannel rates that they received.

7 DISCUSSION AND CONCLUSION

This study compared the complexity, accuracy and fluency of the L2 English speech of participants when they were receiving different frequencies of backchannels. Two L1 groups were chosen because their backchannel norms differed widely. Studies of the effects of backchannels are rare, and this one investigated those with Japanese or Mandarin Chinese as their L1, in two types of interaction – long-turn and discursive parts. Numerous outcomes from this combination of factors were possible and this chapter discusses those that did occur and queries why some others did not. A consideration of the contribution of the L1 to backchannel effects is the central theme, but, regardless of how important L1 norms may be, the implications of the findings extend into language testing, teaching, theory and research methods.

The first main section of the chapter directly addresses the research questions. Findings related to the two specific research questions are discussed in relation to CAF effect sizes reported in other studies. For the subsequent discussion of the two general research questions, the similarities and differences between the Japanese and Mandarin Chinese groups are considered, to identify the extent to which the findings may be broadened to other languages. The second main section covers implications. They are split into those that apply to people involved in the learning, teaching and testing of languages, and those that apply to theories of, and research into, speech itself. The third main section is on future research directions. Limitations are mentioned in these main sections as the need arises, but of course include those summarised in Section 4.7 and the fact that this was a study only of verbal backchannels in dyadic settings.

7.1 Discussion of the Findings

7.1.1 Specific Research Questions

The specific research questions were:

1. To what extent are the complexity, accuracy and fluency of the speech of adult L1 Japanese learners of English affected by variation in an interlocutor's backchannel frequency?
2. To what extent are the complexity, accuracy and fluency of the speech of adult L1 Mandarin Chinese learners of English affected by variation in an interlocutor's backchannel frequency?

The main findings for the Japanese group were that their fluency, as measured by PSR and, less consistently, by MLR and NPTR, was higher when receiving a greater frequency of backchannels. The complexity and accuracy of their speech were not affected by backchannel rate. The main findings for the Mandarin Chinese group were similar to those of the Japanese group for fluency and complexity. For them, a higher backchannel rate was also associated with lowered accuracy, although this occurred only during the long-turn parts.

Part of the discussion of these findings entails comparing them with those of other published studies. The extent to which CAF were affected by backchannel frequency can be evaluated by two sets of comparisons. The first is the L2 CAF values reported for other studies involving backchannels; the second is the L2 CAF values reported for studies that did not involve backchannels. The second is required because there is only one study of

the first type. This is Wolf (2008), who reported a PSR of 65.73 (SD 18.33) when a typical backchannel rate was given to Japanese participants during L2 English narrative tasks and 56.76 (SD 16.07) when none were given, leading to an effect size (*d*) of 0.52; CI were not reported and no statistically significant effects were found for MLR or NPTR.

A particularly useful metric for comparisons is effect sizes. Plonsky and Oswald (2014) conducted a meta-analysis of effect sizes in L2 research; their sample was predominantly from the journals *Language Learning* and *Studies in Second Language Acquisition*, which, as their titles indicate, are concerned mainly with L2 acquisition. The current research, however, concerned L2 use and did not have an acquisition element, beyond considering possible implications for L2 teaching and learning. The suggestions made by Plonsky and Oswald (2014) on interpreting effect sizes appear to be principally for L2 acquisition research, so are not directly applicable to the present research, but their recommendation to compare a study's effect sizes with those presented in similar studies is pertinent. Before summarising effect sizes of CAF values reported for comparable studies that did not involve backchannels, it is necessary to make some comments about how they were calculated. Cumming (2012: 290–294) states that there are many ways of calculating Cohen's *d* and that they can produce widely differing figures, so it is necessary to compare like with like. The individual studies mentioned here did not report *d*, so means and standard deviations were used in the same formula used to calculate *d* in the present study (given in Section 5.9).

The 10 studies from two edited volumes (Robinson, 2011; Housen, Kuiken and Vedder, 2012a) that were used in Section 4.3.2.1 to identify sample size precedents in CAF research were also used to find effect size precedents. Only three studies that permitted

direct comparisons of measures with the present research were identified. This was because of a combination of some of the studies using different CAF measures and their discovering no statistically significant differences. The three are included in Table 7.1, which shows the effect sizes for four measures in the current and other research, all of which studied L2 English. None of them was included in Jackson and Suethanapornkul's (2013) meta-analysis of nine cognition hypothesis research studies that reported CAF measures. All of those nine were quasi-experimental studies that "used repeated-measures designs, measured oral performance, and adopted monologic tasks" (2013: 349). For the purposes of comparison with the current research, they were checked individually[24] and the two of them that used a CAF measure that was the same as in the current research were added to Table 7.1.

Table 7.1 Effect size (*d*) comparisons

Study	PSR	MLR	E-freeAS	E-freeCL	Notes
Current	0.23; 0.44	0.34	-	-	L1 Japanese.
Current	0.25; 0.26	0.19; 0.24	0.77	0.87	L1 Mandarin Chinese.
Wolf (2008)	0.52	-	-	-	Compared typical backchannel rate with no backchannels in picture-story tasks. L1 Japanese.
Levkina and Gilabert (2012)	0.26	-	-	-	Compared planned, few elements with unplanned, many elements tasks. L1 Russian or Spanish.
Tonkyn (2012)	-	0.61	0.36	-	Compared before and after nine weeks of a pre-sessional programme. Various L1s.
Gilabert, Barón and Levkina (2011)	0.17; 0.37; 0.46	-	-	-	Compared task types and task complexity. L1 Spanish and Catalan.
Iwashita, McNamara and Elder (2001)	-	-	-	0.41	Compared here and now with there and then tasks. Various L1s.
Gilabert (2007)	0.28; 0.31; 0.45; 0.49	-	-	-	Compared planned and unplanned, here and now or there and then tasks. L1 unclear.

Note: PSR = pruned speech rate (syllables per minute); MLR = mean length of run; E-freeAS = error-free AS–units; E-freeCL = error-free clauses. A blank space indicates that the measure was not used or no statistically significant difference was found.

Three studies could have been included in Table 7.1, but were not. Michel's (2011) study (L2 Dutch; L1s unclear) had PSR effect sizes of 0.48 and 0.94, but these are misleading if contrasted directly with the current study, as they compare a 'simple' task in a 'monologue'

[24] Jackson and Suethanapornkul (2013: 349) state that nine studies were included in their meta-analysis, but they highlight only eight in their list of references, so only those could be checked.

format with a 'complex' one in a 'dialogue' format.[25] Similarly, Michel, Kuiken and Vedder (2007) compared 'monologic' and 'dialogic' tasks (L2 Dutch; L1s unclear) and had PSR effect sizes of 0.91 and 1.21. While the impact of task difficulty/complexity is open to question, it is known that fluency tends to be higher in 'dialogic' tasks than 'monologic' ones (Tavakoli, 2016). Effect sizes would also be higher in the current study if PSR in the long-turn parts were compared with PSR in the discursive parts, but such comparisons were not part of the research focus. The third study, Rahimpour (1999), used error-free T–units rather than AS–units, with effect sizes of 0.60 and 0.79 (using 'here and now', 'there and then' tasks; L2 English; L1 unclear). A final omission from the table is NPTR, which was significantly different in some of the analyses in the current research, but was used only by Wolf (2008), who found no statistically significant differences.

The relevant aggregated *d* effect sizes in Jackson and Suethanapornkul's (2013) just-mentioned meta-analysis were: fluency (mean -0.16; SD 0.17; 95% CI -0.25, -0.08); general measures of complexity (mean -0.11; SD 0.17; 95% CI -0.22, 0.00); and general measures of accuracy (mean 0.28; SD 0.41; 95% CI 0.14, 0.42). It is important to note that these are based not just on the measures included in Table 7.1, but on any measure of fluency, or general measure of complexity or accuracy. The effect size formula they used to calculate these values is unclear, as they do not give a formula for pooled standard deviation, the denominator. Polarity in the values is a reflection of the predictions of the cognition hypothesis, so effect sizes of opposing polarity in different studies may have had a cancelling effect in the meta-analysis. Notwithstanding these caveats, the mean and CI values can be compared with the *d* values from the current study. These comparisons, and ones with the effect size values in Table 7.1, are contained in the following two paragraphs.

[25] See Conclusion on the use of scare quotes for the two -logues.

For the Japanese group, PSR effect sizes were within the range of those reported in other studies, but a little below that of Wolf (2008), who studied backchannels and Japanese learners of English. The two studies are not equivalent, as there are substantial differences in methods. For example, in Wolf (2008), the tasks entailed telling a story from cartoon strips, planning time was 4 minutes, the main comparison was between a typical backchannel rate and no backchannels at all, and the sample size was 14. What can be said is that both studies show that receiving a higher frequency of backchannels is associated with a higher speaking rate for Japanese learners of English. The most likely contributor to the difference in effect sizes is that the current study used a lowered backchannel rate while the other used a no backchannels rate. The MLR effect size for the Japanese group was below that of Tonkyn (2012), yet higher than the upper CI for fluency in Jackson and Suethanapornkul's (2013) meta-analysis. The conclusion for the Japanese group is that their fluency is affected by variation in an interlocutor's backchannel frequency to a moderate extent.

For the Chinese group, PSR effect sizes were also within the range of those reported in other studies. Compared with Jackson and Suethanapornkul's (2013) figures, the fluency effect sizes were clustered around the upper CI of the meta-analysis. Considered together with the MLR effect sizes, the fluency of the Chinese group can also be said to be affected by variation in an interlocutor's backchannel frequency to a moderate extent. Effect sizes for accuracy, measured by error-free AS–units and error-free clauses, were considerably higher than those reported in other studies, but there were only two comparisons possible. They were also comfortably greater than the upper CI of the meta-analysis for accuracy. This applies only to the long-turn parts of the current research, as no significant differences for accuracy were found in the discursive parts. The conclusion for the Chinese group is

that variation in an interlocutor's backchannel frequency can affect their accuracy, possibly to a substantial extent, but that this may be highly dependent on the nature of the interaction.

The important question of how meaningful the CAF differences are is addressed in the Implications section. That is preceded by the next section, which moves on from a focus on the research questions for two specific languages to discussing the more general research questions.

7.1.2 General Research Questions

The two general research questions were:

1. What are the effects of interlocutor backchannel rates on spoken L2 English?
2. Do these vary with the speaker's L1?

The consideration of the general research questions is split into two sections. First are broader matters, which concentrate on discussing the influence of the main speaker's L1. These are followed by narrower considerations of individual differences, complexity and the backchannel frequencies chosen for this research.

7.1.2.1 Broad Considerations

Answering these questions requires stating the similarities and differences between the Japanese and Mandarin Chinese groups in this study and considering whether or not the findings can be broadened to other languages.

The basis of choosing Japanese and Mandarin Chinese learners of English for this study was that the available evidence indicated that they were maximally different in their L1 backchannel norms and that these norms could be carried over into their L2 speech. This idea was supported by the findings for the Japanese group, for whom the literature overwhelmingly reports a high backchannel frequency: fluency was higher when the backchannel rate was greater. How the findings for the Chinese group, for whom the small quantity of available literature consistently reports a lower backchannel frequency, fit in with this idea is less clear. For them, fluency was also higher when the backchannel rate was higher, but accuracy fell in one of the two parts – the long-turn ones. The presence of interaction effects of L1 and backchannel frequency in the MANOVA conducted for the long-turn part, and the identifying of these as occurring in the accuracy measures, supported the conclusion that the L1, and not just the backchannel rates, contributed to the accuracy differences found. Without this difference in accuracy, the preliminary conclusion would have been that the effects of backchannels do not vary with the speaker's L1.

The reported differing effects of backchannel frequency on accuracy have several possible explanations. The first thing to consider is that it was an artefact of the study's methods, with practice effects from using the same topic in the long-turn and discursive parts manifesting themselves in the latter. Accuracy appeared to increase slightly from long-turn parts to discursive parts among the Chinese group: mean error-free AS–unit values based on topic and sequence were between 18 and 21% in the former and were 24–26% in the latter. Having already spoken about the topic in the long-turn part may have aided their accuracy by allowing more consistent selection of topic-specific words and phrases, for instance. However, such things cannot have been sufficient to explain why the

backchannel rate difference in the long-turn parts of 14 (low rate) to 25% (typical rate) became the very similar 26 and 24% in the discursive parts. Experience in taking IELTS could also have been responsible: all but one of the Chinese participants had taken it, most several times, whereas only one of the Japanese participants had taken it. There could have been surprise in the Chinese group at a relatively high backchannel rate during the procedure. This is countered, however, by two pieces of evidence. First, their fluency was higher, when such a sense of surprise would surely have lowered it. Second, the questionnaire data showed that there was no conscious awareness of backchannel differences in the procedure, but a comparison with a backchannel norm that was as highly situation-specific as occurring in a language test would probably need to be made consciously, so would have been reported in the questionnaire. Another thing that can be queried is the strength of the evidence for L1 Mandarin Chinese backchannel frequencies being low in comparison with other languages. The evidence comes from only a few studies, so is not as well developed as that for Japanese or English. Nevertheless, it is consistently reported (Deng, 2008) and not queried in the literature. So, perhaps differences in L1-norm-based backchannel expectations did play a part in producing the differences in accuracy between the two L1 groups in the current study.

The next explanation to consider for the differences in the findings between the two L1 groups is based on why the accuracy of the Chinese group varied with backchannel frequency in the long-turn parts but not in the discursive parts. The suggestion in the Results chapter was that the typical backchannel rate in the long-turn parts may have encouraged the expansion of ideas while reducing effective self-repair, and thus caused an increase in fluency and a decrease in accuracy. Idea generation being enhanced by backchannels was reported by Sannomiya et al. for Japanese people; they suggested that

they "stimulate idea-generation itself at the cognitive and motivational level" (2003: 46). In the discursive parts, in contrast, the amount of expansion required could have been regarded as less because the participants knew that the interlocutor would ask more questions. This reduction in the need to generate ideas (meaning the need to continue speaking) could have permitted more consistent concentration on accuracy in the discursive parts, regardless of backchannel frequency.

This leads, of course, to the consideration of why the accuracy differences were not also found in the Japanese group. As the Sannomiya et al. (2003) research was on Japanese people, it could be argued that the same idea-generation effects should have occurred in this group, too. Their accuracy was, however, comparable to that of the Chinese group in the discursive parts and in the topic- and sequence-based calculations for the long-turn parts. There is, then, no obvious explanation for the combination of similar backchannel effects on fluency and differing effects on accuracy between the two L1 groups.

For now, there is insufficient evidence to show in what direction the answer to the second general research question may lie. There could be an L1-independent association of raised backchannel rates and increased fluency, with L1-dependent variation in accuracy and complexity. There could be the same L1-independent association of raised backchannel rates and increased fluency, but with interaction type leading to variation in accuracy and complexity. Alternatively, there could be an L1-dependent association between raised backchannel rates and all parts of CAF, with the effects varying with interaction type. What can be stated is that this study added to the evidence of effects of backchannels on Japanese speakers and contributed to the preliminary understanding of the same in Mandarin Chinese speakers. It can also be stated that more evidence is required before

reaching any conclusions on the centrality or contribution of elements of the interaction situation, or of L1 norms as opposed to universals. The assumption (in, e.g., Cutrone, 2005; Wolf, 2008; and, for its research design, the current study) that L1 norms are key is based almost entirely on Japanese and English, has been weakened by the current study, and needs to be challenged further by future research.

7.1.2.2 Narrow Considerations

As reported in the Results chapter, even for those CAF measures for which statistically significant differences based on backchannel frequencies were found, not all of the participants followed the same pattern. For example, six of the 34 Chinese participants in the long-turn parts were more fluent, as judged by all three fluency measures, when receiving the low rate of backchannels. Vercellotti (2015) recommends examining individual differences rather than aggregated data for longitudinal studies of change in CAF. The current study was not longitudinal, but a numerical check was made of the Chinese participants to see how many followed the pattern of being both more fluent and less accurate when receiving a higher backchannel rate in the long-turn parts. A participant was taken to have followed the pattern when at least two of the three fluency measures and at least one of the two accuracy measures were in accordance with the trend. This occurred in 22 of the 34. Therefore, and it is close to axiomatic to state it, individual differences also clearly play a part in the extent to which backchannel rates affect L2 speech.

Complexity has been mentioned very little in the Results and Discussion chapters, as no statistically significant differences were found for any of the three measures of it in either of the L1 groups. A point to note, then, is that the complexity values reported here were

not abnormal for L2 speech. Skehan (2014: 16) has observed that clauses per AS–unit values are typically "above 1.2, generally but not always below 2, with group means often in the range 1.4 to 1.6". In the current research, group means for clauses per AS–unit ranged from 1.38 to 1.78. Complexity may be the least likely of the three CAF components to change in experimental studies: Jackson and Suethanapornkul's (2013) meta-analysis reported a mean effect size of 0.11 for general measures of complexity, the lowest of the three, and the CI crossed zero, indicating no statistically significant difference. Whether the absence of any spoken complexity differences attributable to backchannel rates in the current research is just another example of this trend, or if backchannels are simply unlikely to affect complexity, cannot be known from the data available.

A final consideration in this section is the methodological appropriateness of employing the backchannel rate that was labelled 'typical', and of targeting a low rate of one-third of that typical rate. As described in Section 4.4.1, the attained difference in rates was 2.48–2.72 times, rather than the targeted 3, and the typical rate was within the range of what the literature reports for dyadic, face-to-face interactions in L1 English. Perhaps pushing the 'typical' rate to a higher one would have highlighted greater or other differences than the ones found in this study. Perhaps targeting a low rate of one-quarter instead of one-third might have had similar effects. Even in retrospect, the decisions taken appear to be apposite, but, again, an obvious point is to be stated: a different interlocutor, with a different 'typical' starting point, could influence the extent to which backchannel rates affect L2 speech.

7.2 Implications

The implications are split into two parts: those that apply to L2 speakers and those that pertain to theory and research methods. A later section goes on to discuss future research directions.

7.2.1 L2 Speakers

The L2 speakers to which this section applies are certainly those of the intermediate proficiency, L1 Japanese or Mandarin Chinese learners who participated in this research. As has been seen from the earlier discussions, the degree to which it could be extended to apply to others in unclear. The two parts of this section are on language testing and teaching. They relate existing literature and practice to the findings of this research.

7.2.1.1 Testing

The current study was not of language testing, but the data collection process was similar to that used in some tests of L2 speaking, so there are implications for several aspects of such tests. These tests vary in their import, but some are high stakes, as they are used for things such as university admissions, employment promotions or overseas placements, immigration, and acquiring permanent residency status in a country.

A lot of research has been done on the impact of characteristics of the examiner-interlocutor in tests of L2 speech. Test standardisation – "making sure that all test takers have the same experience, in so far as any variation in the administration may impact negatively upon their performance or scores" (Fulcher and Davidson, 2007: 258) – is the

major concern. Stated simply, "the presence of an interlocutor in the test setting introduces an immediate and overt social context" (McNamara and Roever, 2006: 45). This can be based on fixed qualities such as gender, but also on more mutable matters such as examiner behaviour leading to varying levels of rapport (Brown and Hill, 2007). Morton, Wigglesworth and Williams (1997) associated backchannels with rapport and found that differences in backchannel rates were linked to judgements of how well examiners had assisted candidates to perform to their full capability in one type of speaking test. There is also evidence that such differences can be noticed: teacher-raters were asked by Ducasse and Brown to comment on interaction in L2 Spanish peer-peer discursive tasks; among other things mentioned as salient were backchannels, which were "deemed to support the interaction and were attended to by the raters" (2009: 436).

Backchannels can, then, contribute to interaction in language tests and the contributions can be attended to by raters. This alone is sufficient to warrant attention from test makers who are concerned about standardisation. The current research adds to this by showing that differences in backchannel frequency can change aspects of speech that are likely to be included in the criteria that lead to test scores being awarded. Differences in frequency could be from the examiner in examiner-candidate tests such as IELTS or a candidate in paired candidate tests such as some Cambridge English exams. The finding that a higher frequency was associated with greater fluency in both L1 Japanese and L1 Mandarin Chinese speakers, but that accuracy was affected differently between L1 groups, indicates that the relationship between backchannels and test or task performance is likely to be a complex one.

Very little information on training and guidelines for examiners is publicly available. As mentioned in Section 3.1.1.1, Seedhouse and Egbert (2006: 22) state that official IELTS guidelines are that "Examiners should keep non-verbal interjections to a minimum. (Eg 'um', 'right', 'uh uh')", but that the frequency of examiner use of backchannels varied considerably in their sample of recordings of actual tests. Three issues arise from this in relation to the current research. First is the guidelines to examiners in this particular test. If the IELTS guidelines are strictly followed and a minimal rate of backchannels is given (this is not defined, but can be taken to be much closer to the low rate than to the typical rate in the current study), then underperformance in fluency is likely to occur, at least among some L1 groups. Second is the possible effects of an elevated backchannel rate resulting from an examiner-interlocutor's familiarity with another L1 or the L2 speech characteristics of that group. Ross (2007) suggests that a very high backchannel rate in oral proficiency interviews may be associated with relatively brief, phrase-like utterances from the candidate and that these may reduce fluency. Third is the possibility of different performance based on L1. The Chinese group in the current study had differing accuracy in the long-turn parts when receiving different backchannel rates, but the Japanese group did not. Such variation cannot be countered by a simple guideline that is to be applied to all. What examiner guidelines should contain is, therefore, unclear. The extent to which this really matters is turned to next.

Whether or not the differences reported here for fluency and (for some) accuracy would be sufficient to lead to a candidate being given a different test score cannot be ascertained from the research done to date. Numerical measures of some aspects of speech have been shown to correlate with raters' scores awarded, including for speech rate and mean length of run (e.g., Ginther, Dimova and Yang, 2010). In the current research, the largest effect of

backchannels on PSR was a *d* value of 0.44 and had CI of 4.64–12.44 on an overall PSR of approximately 80 syllables per minute. The likelihood of this level of backchannel-induced differences leading to test score variation is low, unless the test assigns a strong weighting to fluency.

Nevertheless, another matter remains – that of test fairness. Standard conceptualisations of test fairness include the idea of "equitable treatment of all test takers in the testing process" (Xi, 2010: 147). How to interpret "equitable" is open to question: it could mean 'the same', 'not biased against certain groups', or something else. Adopting the perhaps naïve view that language tests should allow candidates to perform as well as they are capable of, then having interlocutor-examiners take into account interactional differences, including that of backchannel rates, would improve test fairness. The difficulties in doing this would be high, however, so test-makers may prefer to gloss over these things by maintaining that candidates' ability to adapt to a different style of interaction from that of their L1 is part of the construct being assessed.

An alternative test format that can avoid these issues is computer-based testing that requires a candidate only to respond to prompts and be recorded by a computer. In this setting, there would be no expectation of receiving backchannels. Authenticity and construct validity are ongoing research concerns with such settings (e.g., Sawaki, 2012), but this is straying beyond the scope of this backchannel-based discussion, so Swain's words on treating everyone in the same way when testing can conclude: "[if we] do that, […] we are washing out all of that variability which is what human nature and language is all about" (Swain interviewed in Fox, 2004: 240).

Fixing a backchannel rate for all candidates in a test is problematic; prescribed altering of the rate based on a candidate's L1 is problematic; and ignoring the issue is problematic. Instead of changing a test, the participants in it could be changed: building awareness and understanding of backchannel norms in different languages could aid both candidates and examiners. Building candidates' tolerance of differences could also help their performance. These are matters of teaching, which is addressed next.

7.2.1.2 Teaching

Three teaching-related matters are considered in this section. The first is the training of L2 speakers in backchannel use, which blends into the second, on the raising of awareness of differences in L1 backchannels. The section concludes with the third – teachers' use of backchannels in classrooms – and a revisit to the trial of the Japanese businessman who may have overused 'uh-huh'.

Recommendations and proposals for the inclusion of backchannels as part of L2 pedagogy are longstanding: R. Gardner (1998: 220–221), for example, suggested "exposing learners to […backchannels], and providing opportunities for practice and feedback". Some attempts have been made. Ward et al. (2007) created software, for learners of Arabic, that gave examples of backchannels, demonstrated where they could be given and provided feedback on the learners' attempts at appropriate backchannel behaviour when listening to pre-recorded speech. They reported that their participants were interested in the topic and that most improved the naturalness of their backchannels. However, the study was small in scale and did not use delayed post-tests.

A different approach to teaching was attempted by Sardegna and Molle (2010). They used video-conferencing to teach English backchannels to L2 learners in Japan and reported a positive effect on backchannel production over the two-hour period. Again, however, there was no follow-up to check if this would last beyond immediate production.

A more thorough study of the teaching of backchannels was done by Cutrone (2013), who also compared the effectiveness of explicit versus implicit instruction on Japanese learners of English. The explicit instruction group was given instruction in differences in interlocutor behaviour across languages, analysed examples and practised giving backchannels in English. The implicit instruction group was exposed to example conversations and discussed differences across languages; they also had conversations with L1 English speakers which were followed by periods of reflection. At the end of the eight-week process, the backchannel use of the explicit instruction group was closer to that of L1 English than was that of the implicit instruction group or that of a control group which had received no instruction in backchannel use. An interesting continuation of these teaching attempts would be to check if the speech of those who had been instructed in backchannel *production* was less sensitive to the *receiving* of different backchannel rates than were the participants in the current study.

These attempts at teaching have been based on the idea of having learners use backchannels in ways that are similar to those in L1 use. They necessarily entail building awareness of the norms of that L1. With so much L2 use being with speakers who are also using English as an L2, rather than as their L1, the element of awareness-raising in any teaching may need to be given added weight, so that understanding and tolerance of different backchannel behaviours is developed, not just those of the targeted L1.

Awareness and tolerance were mentioned at the end of the previous section, on testing. While being able to produce backchannels in a certain way – frequency, placement and so on – could be of benefit, it would not necessarily develop a tolerance to differences in backchannels produced by others. For language testing and other formal situations such as job interviews, it might also be beneficial for learners to practise the main speaker role while an interlocutor produces varying backchannel frequencies. Conversely, examiners and interviewers could well attain a better understanding of the dynamics of the interactions that they lead if they, too, are made aware of backchannel differences.

The discussion of teachers' use of backchannels in the classroom must be restricted to one-to-one teaching and possibly very small groups, as the dynamics of larger classes will be very different. The main thing to consider is learner affect. Comments reported in the Introduction and Section 6.4 – including, from this study, "The backchannels made me relax to speak, thank you" and "when the person who I talk with don't answer anything, I feel sad" – show that, when some learners of English consciously notice backchannel differences, they can also be aware of affective consequences. For them, the differences in backchannel rates were meaningful, irrespective of any effects on their speech as subsequently measured by a researcher. Aspects of one-to-one and small group teaching can be similar to what was done in the discursive parts in particular of this study, and both Japanese and Chinese learners of English are often worried about making mistakes when speaking, so this study's findings are likely to be of relevance to the learning and teaching of English for these L1 groups.

L2 learning and performance can be affected by anxiety and negative beliefs (Dörnyei, 2005), so the teacher's providing a frequency of backchannels that will minimise these is

desirable. As the continuation of language learning is also associated with learner attitudes to speakers of that L2 and to the teachers of it (R. C. Gardner, 1985), teachers with L1 English who are new to instructing Japanese learners should probably use at least the backchannel rate that they are comfortable with. Experimenting with using a higher rate for learners of no more than intermediate proficiency could also be tried, as the classroom atmosphere and rapport might improve. For teaching those with other L1s, observing the behaviour of more experienced teachers is advisable. Part of the goal would be to use backchannels as part of "conversational form […not just linguistic content, to] shape perceptions of good collaboration (experienced as a feeling of conversational flow) and this in turn can feed perceptions of group solidarity" (Koudenburg, Postmes and Gordijn, 2016: 4). As reported by others (e.g., Cutrone, 2005), the backchannel behaviour of teachers who have been in a country with a different L1 for a sustained period of time is likely to have become more like the norms of that language, so the matter will be less important for them.

Both the teaching of backchannels and the teacher's use of them in classrooms could be informed by their inclusion in textbooks. Thonus (2007) reported around a decade ago that they were not, and little appears to have changed since then. Their inclusion could be extended to classroom listening materials, which would also be a relatively straightforward means of improving the authenticity of those materials.

Would any of these suggestions have helped the Japanese businessman who was reported on in the Introduction? He was arrested for corporate espionage and argued that his frequent use of 'yeah' and 'uh-huh' was misunderstood as meaning that he was agreeing to buy stolen information, but that it actually showed, as in his L1, only that he was listening

supportively (Japan Times, 1983a). Perhaps the situation would not have arisen if he had been taught about backchannel differences. Perhaps he would not have been able to mount that defence if he had been. The truth was known only by the businessman himself (and possibly his English teachers): at a later hearing, his plea was changed to 'no contest', thereby avoiding further revelations in court, but admitting no guilt (Japan Times, 1983b).

7.2.2 Theory and Research

There are two parts to this implications section. The second, on research methods, deals with specific matters arising from the methods employed during this research, as well as broader issues that emerge from considering the findings. The first is more theoretical and is on this study's theoretical framework and other models of speech.

7.2.2.1 Theoretical Framework and Models of Speech-Comprehension

Levelt's (1989; 1999) model of speech was selected in Section 2.4 as the theoretical framework for this study. The consideration of how backchannels could influence stages of speech in that model and how CAF measures might be linked to the stages was described as being somewhat speculative, so the following relating of the study's findings on backchannels and CAF measures to the model must also be prefaced with the same caveat.

A sensible place to start in relating the findings to the theoretical framework is with the relatively consistent finding that fluency was greater when a higher backchannel frequency was received. Although the correspondence between CAF components and speaking stages is unlikely to be straightforward, fluency can be regarded as being related mainly to the formulation and articulation stages of the model of speech (Housen, Kuiken and Vedder,

2012b). Articulation itself, being the "execution of the articulatory score by the laryngeal and supra-laryngeal apparatus" (Levelt, 1999: 88), is unlikely to be a location of backchannel influence, so formulation can be focused on as the possible process influenced by backchannels, as revealed by measures of fluency in this study. Rather than being a single process, of course, formulation involves lexical retrieval, grammatical encoding, morpho-phonological encoding and phonetic encoding related to a syllabary (Levelt, 1999). These processes require attention in non-advanced L2 speakers – they are not as automatic as in L1 speech. Using only the fluency data from this study, it is not possible to pick from among these processes which is or are more likely to be involved in backchannel effects.

Incorporating the findings for accuracy in the Chinese group during the long-turn parts allows for some narrowing of the list of formulation processes that are likely to be involved. An objection to this incorporation is possible: this accuracy finding was for only one of the two L1 groups and was observed in only one of the two data collection parts. The counter to this objection is that, in relating the findings to the theoretical framework, the goal is not (and cannot be at this point in time, given the absence of other relevant studies on the matter) to suggest ways in which backchannel rates *will* act, but to suggest ways in which they *could* act. It is therefore justified to consider all instances in this study in which there is evidence for backchannels having had an effect, as this indicates where they could act in non-advanced L2 speakers. An explanation posited in the Results chapter for the Chinese long-turn accuracy finding, then, was that the effectiveness of self-repair was greater when receiving a lower backchannel rate. This is unlikely to map to phonetic encoding, because the calculations of accuracy and self-repair did not include pronunciation. It is also unlikely to map to the process of preverbal message creation that

occurs at the conceptualisation stage (which is also linked with accuracy). However, self-repair for grammatical and lexical accuracy does fit the lexical retrieval, grammatical encoding, or morpho-phonological encoding parts of formulation. These are thus the most likely processes to be have been influenced by backchannel rates in this study. It should, of course, be noted that the evidence collected in this study may not be exhaustive of the possible ways in which backchannel frequencies could influence L2 speech. In particular, previously mentioned L1 studies of idea generation, and the earlier interpretation of the findings from this study, indicate that backchannels could also act at the conceptualisation stage.

The means by which backchannels are perceived and are thus able to have an effect on non-advanced L2 speakers might be theorised or conceived of as being an addition to the feedback loop which involves the self-monitoring of overt speech in Levelt's model. (All of the feedback loops in the model go to the conceptualisation stage, but this does not mean that feedback more relevant to later stages cannot occur – self-correction that is overtly pronunciation-based exists, for instance.) In this vision of a mechanism, instead of overt speech being monitored by the main speaker (self-monitoring) and feeding back to the conceptualisation stage, overt speech could, in effect, be monitored by an interlocutor and lead to feedback to the conceptualisation stage through backchannels. This idea ties in well with the fact that self-monitoring is regarded as involving the same mechanisms as for the comprehension of others (see, e.g., Levelt, 1999). An extension of this idea for non-advanced L2 speakers is indicated by the questionnaire responses showing that almost all of the participants had no conscious awareness of what had been a substantial difference in backchannel frequencies. Self-monitoring (as well as the comprehension of others) is regarded as requiring attention in L2 speakers (Kormos, 2006: 173), whereas the apparent

lack of conscious awareness of changes in backchannel rates suggests that they require little if any attention: backchannels being short and often non-lexical allows them to be straightforward to process, hence their potential to affect L2 speakers who have more limited free processing capacity.

This research, then, has helped to add to the case for a greater role of external factors in Levelt's model of speech, stressing that those that originate with an interlocutor do not need to have lexical content. It has also pointed towards those parts of the model that the evidence indicates are most likely to be influenced by backchannel rates. The apparent lack of conscious awareness of backchannel changes also suggests that the discourse record that contributes to the conceptualisation stage of speech in the model contains more than declarative knowledge.

Recently, theories that attempt a greater integration of aspects of speaking and comprehension have been proposed, mostly based on the idea of prediction involving the system of speech production (Meyer, Huettig and Levelt, 2016: 3). In the introduction to a special volume of *Journal of Memory and Language*, Meyer, Huettig and Levelt concluded that "the authors of the volume appear to be in good agreement that speech production and comprehension engage skills and representations that are distinct but tightly linked" (2016: 6) and they commented that "a particularly pressing issue is [still] how people shape each other's language in actual conversation" (2016: 5). Research on this idea of models of speech-comprehension has also been conducted in the natural sciences. One line of research has led to the suggestion of a direct link between speech production and comprehension to the extent that coupling of neural activity between interactants occurs. This overlap of activity in brain areas used by a speaker and listener does not occur

merely because they are either producing or hearing the same thing; it appears to require that the interactants are able to decode and understand what is being said (Schoot, Hagoort and Segaert, 2016). A consequence of this could be that "between-pair differences in the extent of between-brain neural coupling may be explained by the level of alignment between speaker and listener at multiple levels of linguistic and extra-linguistic representations" (Schoot, Hagoort and Segaert, 2016: 456). If, as argued above, backchannels can be easily processed, then they could help promote this neural coupling or synchrony between main speaker and interlocutor, thereby aiding the interaction. This ease of processing is, potentially, a key difference between backchannels and speech with linguistic meaning, and may validate the conceptual separation of backchannels from other, meaning-based, speech. Being the main speaker while also receiving feedback on that speech avoids the main speaker becoming the interlocutor: the roles typically labelled 'speaker' and 'listener' become blurred, which may become manifest in the proposed coupling of neural activity between interactants.

7.2.2.2 Research Methods

Implications for research methods fall into two groups. First are those arising from the methods employed during this research; these are narrow and specific. Second are implications that emerge from considering the research findings; these are broader but perhaps of more fundamental importance.

The narrow implications are based not on the results of this study, but on observations of the literature that uses CAF measures. Strict limits on space in publications mean that details of methods are often given a low priority and therefore simply omitted or glossed

over. As a consequence, presumably, it is common to find a short statement such as 'AS–units were used', possibly backed up by a definition taken from the original paper, or that 'disfluent syllables were not included in the count', with no further elaboration. If units such as these were unambiguous and straightforward to identify, then this would not represent a problem. This is clearly not the case.

An alternative is to establish one or more detailed accounts or protocols that can be referred to with similar brevity but which provide a greater level of detail than existing sources. It is hoped that the accounts presented in this thesis, particularly on disfluency, can be the basis of such a protocol, remembering that, in their current form, they were established and adapted for use with particular samples of language. The need for more detailed standards could be assessed readily. A check of the consistency of interpretation of things such as 'AS–unit', 'disfluency' and 'error-free clause' could be performed by asking multiple researchers to code the same text according to their normal methods. Instead of receiving training to improve inter-rater reliability, they would be asked explicitly not to consult others. A comparison of the choices made at the level of the text and the impact of them on CAF measures could be revealing and highlight the need – or lack thereof – for more standardised protocols in the field.

The broader implications for research methods come from a consideration of the research findings. Interpreted widely, they show that an interlocutor merely giving a different rate of backchannels can affect the main speaker. This needs to be considered when interpreting research findings, so it also needs to be considered in research designs. An example of the potential problem is in Ferrari's (2012) study: she compares so-called 'monologic' tasks and so-called 'interactive' tasks, but the former involved an interlocutor

who is labelled the "interviewer". This particular example could be a matter of mislabelling, but it is symptomatic of the ignoring in research of the interlocutor who could be considered to be merely present: "there is an unstated assumption that an interlocutor will have no effect on the CAF of the study's participants in any of the activities" (Flint, 2013: 271–272).

This also applies to any research in which one person's speech is the focus. If it is intended to be a study of solo speech, then there should be no-one else present when it is being recorded and, if the presence of another person is essential, then that person should not be visible to participants during the data collection procedure and should not respond in any way to what is said. The extent of differences brought about by backchannels may be small to moderate, but differences reported in studies of L2 speech typically are, so the risks of backchannels having a confounding influence on the data collected could be considerable.

This extends into research on language testing. Research into interaction in IELTS Speaking has typically used data from people with a wide range of L1s: O'Sullivan and Lu (2006) looked at examiner variation in interaction and used participants of at least 15 L1 backgrounds; Seedhouse and Egbert (2006) studied interaction using recordings from test centres in 19 countries; and Seedhouse and Harris' (2011) examination of topics and their development used recordings from 25 countries. Such sampling designs are very likely to obscure any L1 differences that may have an effect on the test interaction. This has been commented on by the researchers: "it may be found that candidates from particular regions of the world repeatedly run into trouble in relation to a particular interactional sequence, topic or question in the Speaking Test" (Seedhouse and Egbert, 2006: 35).

This section and the ones earlier in the chapter have raised several matters that could be addressed in subsequent research. Possibilities for future research are outlined in the following section, along with comments on the current and future status of speech-related research in different disciplines.

7.3 Future Research Directions

Future research could initially use different interaction situations from the one used in this study. Using the same L1 groups, with approximately intermediate proficiency in L2 English, would enable ready comparisons with the findings of this study. As stated in the literature review, there is still the need for a comparison of, say, telephone and face-to-face interaction using the same participants, so using just one L1 group and examining the effects of backchannel frequency on the same people in different interaction situations would also be worthwhile. In a more educational context, the applicability of the current research beyond the one-to-one teaching and very small groups mentioned in Section 7.2.1.2 could be evaluated via studies of larger classes, perhaps in an English as a Medium of Instruction or Content and Language Integrated Learning setting.

A simple continuation of the current study would be to use recordings of actual language tests to establish the range of backchannel frequencies that occur in such a setting. The extent of the relevance of this research to that setting could then be understood: if frequencies vary little, then candidate experiences would be consistent in this regard, but greater variation would indicate that the matter could be of more importance. A further continuation for language testing would be to see if the CAF differences identified in this

study led to varying test scores being given by raters. This would need to be done using research that followed a test procedure more closely than was the case in this study.

The relationship between backchannel frequency and L2 proficiency is also something that could be investigated. There is a wide range of L2 proficiencies that could be affected by variation in backchannel frequency (see Section 3.1.3.1), but they might not be affected in the same way. Comparing two groups within the broad 'intermediate' category could provide insights into the similarities and differences of backchannel effects based on L2 proficiency. Careful consideration would have to be given to how the difference in proficiency of the two groups would be established, as basing it on their CAF measures could be problematic, given that a clear separation of groups in one measure may well coincide with overlaps of the two in another measure. A further consideration should be the topics used: although no statistically significant differences were found when comparing CAF for the two topics used in this research, their equivalence for other L2 proficiencies for L1 groups should not be assumed.

Similarly to the choice of topic, evaluating the effects of backchannels using different CAF measures from the ones employed in this study could also be done. This study used general measures of both complexity and accuracy, and relatively general measures of fluency, in keeping with it being of something not previously studied and employing common topics. Using topics or interaction settings that could encourage the use of particular structures or patterns of speech could allow different, more specific measures to be used. This might help identify more details of the mechanisms underpinning any effects of differences in backchannel rates.

An obvious future continuation would be to compare other L1 groups using the same methods that were used in this study. This would help to clarify the role of the L1, by separating it out from other participant characteristics. Unfortunately, however, the two L1s chosen for this research on the basis of the available evidence showing that they were maximally different in their backchannel norms are also the ones on which most evidence is available. There is too little known about other languages to make an informed choice of which ones to choose for further comparisons. One or more could be chosen speculatively, but there would be little theory available to interpret any findings based on L1 differences. This comment overlooks another possibility, hidden in the Anglocentric clouds: an L2 other than English could be studied. L1 Japanese learners of Mandarin Chinese, or L1 Mandarin Chinese learners of Japanese, could be revealing to study and doing so would be justified by the existing literature.

If or when some of this further research is done, it should be easier to evaluate what individual characteristics contribute to the effects of backchannel frequency differences. At that point, things such as gender and age differences could be studied more systematically. Targeting these things before some of the further research just mentioned has been done would, based on the evidence from this study, be the wrong sequence of actions.

This section is entitled 'Future Research Directions' rather than a mere 'Future Research' because a final point needs to be made. The Literature Review chapters of this thesis, combined with this section, have illustrated both the wide range of fields in which backchannels have been studied and the lack of anything resembling a systematic approach to their investigation. A combination of journals such as *Counselling Psychology Quarterly, Journal of Telemedicine and Telecare,* and *Artificial Intelligence for*

Engineering Design, Analysis and Manufacturing is an unusual set for one list of references and can be taken as a positive indicator of interest and variety of methods utilised. The relevance of this topic, then, extends beyond backchannels and into the broader arena of speech, to which it belongs.

Future studies of speech eventually will have to involve multiple disciplines that up to now have been largely separate. Studies of interaction are typically conversation analysis-based and ignore the fact that speech involves the brain. Psycholinguistic studies in applied linguistics are usually statistics-based and downplay the fact that speech is social. Psycholinguistic studies in linguistics often use just words or short sentences and display no interest in the social. Physiological studies employ imaging tools but are largely limited to isolated words. Conversation analysis methods are settled, while psycholinguistic methods and tools are so numerous that it is rare to find two studies that use the same set. The one certainty is that imaging tools for physiological research, originating in medicine, will improve. This is likely to force studies in this area towards the biological. A recent article in the journal *Neuron* commented that "research on the brain basis of social cognition and interaction should move from passive spectator science to studies including engaged participants and simultaneous recordings from the brains of the interacting persons" (Hari et al., 2015: 181). It is also not an abstract notion, as it is already underway, using imaging tools: hyperscanning – "a technique for measuring brain activity simultaneously from two people" – "allows us to use inter-brain effects as neuromarkers of the properties of social interaction in daily life" (Koike, Tanabe and Sadato, 2015: 25). Therefore, if the speech-interaction-psycholinguistics field(s) is not to be swept aside by the physiologic, then ways of linking research strands need to be found and research directions need to be set out and followed. Currently, the field(s) resembles Leacock's

(1912: 75) hero, who "flung himself from the room, flung himself upon his horse and rode madly off in all directions".

7.4 Conclusion

This ending contains a triumvirate of observations. They very briefly cover the essentials of the research findings, the import of the research findings for interactions, and fundamental considerations of what the role of backchannels may reveal.

This study found both similarities and differences in the effects of backchannel frequency on the L2 English speech of Japanese and Chinese groups of participants. In doing so, the previous assumption that it is L1 norms that dictate backchannel effects has been challenged. The possible and relative contributions of the L1, the interaction situation, the type of interaction, the personal characteristics of the interactants, and several other factors, all remain to be identified in this nascent area of research.

The magnitude of the effects found is small to moderate, so how meaningful they may be in any specific set of circumstances, even for the types of learners and interaction situations used in this research, should be judged on a case-by-case basis. Backchannel frequency can have a measurable effect, however, even without conscious awareness of differences, so should be considered in dyadic situations, particularly in those that are important to one or both of the interactants.

Finally, and more fundamentally, is the consideration of backchannels in relation to prevailing conceptualisations and categorisations of spoken communication. Backchannels

are not used as content-containing words, so are not produced by those traditionally categorised as the 'speaker'. They are spoken, however, so cannot be attributed to a pure 'listener'. Speaking listening, production comprehension, self-monitoring monitoring by other – these concepts that are traditionally separated are increasingly blurring and fusing. Backchannels could act as part of a feedback loop that is analogous to the self-monitoring of overt speech; a logical extension from this is that the distinction between the concepts of 'monologue' and 'dialogue' also becomes fuzzy and then hard to maintain. The effects of backchannels on speech – L1 or L2 – may be considerable.

Appendix 1 Systematic Review

The inclusion and exclusion criteria for the systematic review of verbal backchannel rates are as below (from Flint, 2012: 12–13):

> A study was included in the systematic map if it met all of the following criteria:
>
> 1. It reports on verbal backchannels.
> 2. It presents a verbal backchannel rate, or sufficient information for the reader to calculate such a rate.
> 3. It presents original data and/or analysis.
> 4. The participants' L1 is stated, or can reasonably be surmised.
> 5. It is available in English.
> 6. It is published (in a peer-reviewed journal, a book or conference proceedings), or is a doctoral thesis.
>
> In addition, a study was excluded if it met any of the following criteria:
>
> 7. Only one or two types of verbal backchannel are included.
> 8. It reports on selected parts of speech only.
> 9. It reports data mainly from children (i.e., those under 18 years old).
> 10. The speech of one or more participants was scripted.
> 11. It reports on interactions that were for public broadcast (i.e., on television or radio).
> 12. The participants had a known speech/hearing/mental health problem.

The resulting systematic map of 73 studies is on the 21 pages that follow this summary guide to reading it (from Flint, 2012: 54–75):

Information added from another source is given in square brackets and the source is stated in a footnote. Also in a footnote is the name of any more widely used corpus that the data a study used came from.

The following is a gloss of the terms that are used in the table:

(Assumed): The preceding information is not stated explicitly, but can reasonably be surmised.

Backchannels included: This describes and/or gives examples of the backchannels that were included by the researcher in rate calculations. Parts of speech mentioned in brackets as things that were *also reported* have been excluded from the rate calculations by the current author.

Mixed (language of interaction): Some spoke their L1, which for other interlocutors was an L2.

Naturally occurring: The interaction would have occurred even if it was not being recorded.

Researcher devised: The interaction would not have occurred if it was not going to be recorded.

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Alberts et al. (2005) - Journal article - PsycINFO	- Not stated - Phoenix, U.S.A.	English - L1 (assumed)	- 10 M, 10 F - Couples (8 heterosexual; 1 gay; 1 lesbian), together for 2-33 years - Early 20s - late 50s	n/a	Daily conversations in each couple's home - Naturally occurring	(~29% of 2 days) x 10 recordings - One weekday, one weekend day for each couple	"those comments that functioned to indicate the speaker was listening and/or involved in the conversation, including remarks such as 'uh huh,' 'I see,' and 'yeah' (when not used as a response to a question)" (p309)	Not available	12.99 thought units of speech per backchannel (A thought unit is an "utterance segment that expresses a complete and autonomous idea" (p307))
Alcón and Guzman (1995) - Journal article - LLBA	- Not stated - Not stated	English - L2 (L1 Spanish)	- 40 M (20 dyads) - Strangers - In the range 18-34	- At least conversational; 1 in each dyad stronger - Not stated	(High L2 - lower L2, equal subject knowledge) "present a multinational network for a certain company" (p23) - Researcher devised	15m x 20 recordings = 5h - First 15m of each interaction	"back channels give listener's feedback to the speaker" (p24)	High L2: 1.17 Lower L2: 0.33	Partner's words per backchannel (As backchanneller): High L2: 28.28 Lower L2: 129.89
	As row above	As row above	- 40 M (20 dyads) - Strangers - In the range 18-34	As row above	(High L2 but subject non-expert - lower L2 but subject expert) "present a multinational network for a certain company" (p23) - Researcher devised	As row above	As row above	High L2, non-expert: 0.32 Lower L2, expert: 0.12	Partner's words per backchannel (As backchanneller): High L2, non-expert: 135.51 Lower L2, expert: 263.81
	As row above	As row above	- 40 M (20 dyads) - Strangers - In the range 18-34	As row above	(High L2 and subject expert - lower L2 and subject non-expert) "present a multinational network for a certain company" (p23) - Researcher devised	As row above	As row above	High L2, expert: 1.30 Lower L2, non-expert: 1.91	Partner's words per backchannel (As backchanneller): High L2, expert: 21.03 Lower L2, non-expert: 26.26
Benus, Gravano and Hirschberg (2007)[26] - Conference proceedings - Google Scholar	- Not stated [2004] - Not stated [Columbia, U.S.A.]	English - L1 (Standard U.S.)	- 7 M, 6 F (12 dyads: 11 participants took part twice, each time with a different partner) - Not stated - ~20 - ~50	n/a	2 collaborative games using computers Partner could not be seen - Researcher devised	9h 8m in total [8h 55m in total] - Whole recordings	"single affirmative words {*alright, mmhm, okay, right, uhhuh, yeah, yep, yes, yup*}" (p1066) labelled as 'backchannels' ["Indicates only "*I hear you and please* continue", in response to another speaker's utterance" (p97) Affirmative words: *alright, mm-hm, okay, right, uh-huh, yeah, gotcha, huh, yep, yes, yup*]	1.39 [1.41]	- [Words per backchannel: 92.81]

[26] This uses the Columbia Games Corpus. The same corpus was used by Gravano (2009), which is the source of information presented in square brackets.

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
Bjørge (2010) - Journal article - MLA Intl.	A: - 2006, 2007 - Not stated B: - 1995, 1997, 2006 - Not stated	English A: - L2 (L1 not stated; 14 nationalities) B: - L2 (L1 not stated; 9 nationalities)	A: - 10 M, 17 F - Master's degree classmates (assumed) - 20-25 B: - 14 M, 10 F - Master's degree classmates (assumed) - 20-25	A: - Upper-intermediate to advanced - Not stated B: - Advanced - Not stated	A: Role-played business bargaining without specific roles. 6 negotiations with 4-6 people in each - Naturally occurring B: Role-played business bargaining with specific roles. 7 negotiations with 3 or 4 people in each - Naturally occurring	A: 2h 7m in total B: 1h 39m in total - Not stated	"Non-lexical items: mhm, ah, oh"; Lexical: "absolutely, brilliant, certainly, definitely, exactly, excellent, fine, good, great, I see, of course, ok, perfect, quite, really, right, so, sure, that's nice / right / not bad, true, yes / yeah, yes I know"; "Repetition of other speaker's utterance" (p196)	A: 2.47 B: 2.01	A: 60.30 words per backchannel B: 75.42 words per backchannel
Boyle, Anderson and Newlands (1994)[27] - Journal article - PsycINFO	- Not stated - Glasgow, U.K.	English - L1 [(Scottish English)]	- 32 M, 32 F (128 dyads) - Half of dyads were friends, half strangers - 17-30	n/a	A: Gave/received map route with unseen partner B: Gave/received map route with seen partner - Researcher devised	Not stated ["six minutes on average" (p53) ≈ 12h 48m] - Not stated [Whole recordings]	"an ″uhuh" or ″mhm" standing alone or repeated" (p10) [*right*, *ok*, *mm-hmm*, *uh*, *right*, *okay*, *yeah*, *uh right*, *mm*, *oh*, *ok right*, *aye* (these made up <77% of the total backchannels)]	(Using recording length in Cathcart, Carletta and Klein (2003)): A: 2.69 B: 1.52 [> 4.13 (using only those listed in column to left) ~5.37 (taking those in column to left as 77% of total)]	A: 78.18 words per backchannel B: 115.15 words per backchannel
Caspers (2000) - Conference proceedings - Google Scholar	- Not stated - Not stated	Dutch - L1 (assumed)	- Not stated (8 dyads) - Not stated - Not stated	n/a	Gave/received map route - Researcher devised	"over 40 minutes" (no page number) - Whole recordings	""ja" ('yes'); […] ″oké" ('okay') […] ""hmhmm" ('uh-huh'), ″oh" ('oh') or ″goed" ('good')" (no page number)	< 4.73	8.21 inter pausal units (one person's speech bounded by pauses of >100 milliseconds) per backchannel
Clancy et al. (1996) - Journal article - LLBA	- Not stated - Not stated	Mandarin Chinese - L1	- Not stated (8 groups of 2 or 3 people) - Friends - Not stated	n/a	"face-to-face ordinary, non-argumentative conversations" (p357) - Naturally occurring	23m in total - Not stated	'Backchannel': "a non-lexical vocalic form, and serves as a 'continuer' […] display of interest, or claim of understanding" (p359) 'Reactive expression': "a short non-floor-taking lexical phrase or word" (p359) (Collaborative finishes, repetitions and resumptive openers are also reported)	'Backchannels': 0.61 'Reactive expressions': 0.39 = 1.00 (Note: reactive expressions are based on the table on p370, which is not completely compatible with the specific backchannel data)	56.61 intonation units ("speech uttered under a single coherent intonation contour" p365) per backchannel or reactive expression

[27] This uses the HCRC Map Task Corpus. The same corpus was used by Cathcart, Carletta and Klein (2003), which is the source of information presented in square brackets.

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
… **Clancy et al. (1996)** cntd.	- Not stated - Not stated	Japanese - L1 (assumed)	- Not stated (8 dyads: 2 M-M, 2 F-F, 4 M-F) - Friends - Not stated	n/a	"face-to-face ordinary, non-argumentative conversations" (p357) - Naturally occurring	23m in total - Not stated	As row above	'Backchannels': 4.57 'Reactive expressions': 1.13 = 5.70 (See note in row above)	9.73 intonation units per backchannel or reactive expression
	- Not stated - Not stated	English - L1 (U.S.) (assumed)	- Not stated (8 groups of 2-5 people) - Friends - Not stated	n/a	As row above	44m in total - Not stated	As row above	'Backchannels': 1.55 'Reactive expressions': 1.39 = 2.94 (See note two rows above)	9.54 intonation units per backchannel or reactive expression
Cutrone (2005) - Journal article - LLBA	- Not stated - Japan	English - Mixed (L1 U.K. English with L1 Japanese)	- 8 M, 8 F (8 dyads: 4 M-M, 4 F-F) - 4 dyads were strangers, 4 were not - 20-41	- "able to comfortably maintain a conversation in English with a native speaker of English" (p250) - Not stated	Face to face; any topic - Researcher devised	3m x 8 recordings = 24m in total - Middle 3m, as likely to be most natural	"yeah (u)n oo m ah uhum uhuh no yes really right oh yep o(k)" (p260)	British interlocutor: M: 4.92 F: 2.25 Overall: 3.58 Japanese interlocutor: M: 8.17 F: 9.42 Overall: 8.79 (Note: p260 data do not match p259 British backchannel data. The latter are used here)	Speaker words per interlocutor backchannel: British interlocutor: M: 8.69 F: 13.85 Overall: 10.31 Japanese interlocutor: M: 9.18 F: 9.60 Overall: 9.41
Demo (2006) - Doctoral thesis - ProQuest D&T	- Not stated - Washington, D.C., U.S.A.	English - A: L1 B: Mixed (L1 U.S. English with L1 Latin American Spanish)	A: - 8 M, 7 F - Not stated - 19-21 B: - 8 M, 7 F - Not stated - 19-37 A&B: With 1 M, 37 years old ('interviewer')	A: n/a B: - Advanced - 10 months - 7 years	One-to-one simulated employment interview with 'interviewer' - Researcher devised	Averaged "13-14 minutes" (p60) x 30 recordings = 6h 30m - 7h - Whole recordings	"a non-floor taking turn with the purpose of supporting the primary speaker discourse"; e.g., *wow*, *uh-huh*, *mm-hmm* (p67)	A: 0.60 - 0.65 B: 1.23 - 1.33 (Note: a range is given as the recording length used is also given as a range)	Speaker words per interlocutor backchannel: A: Interviewer as interlocutor: 119.95 Interviewee as interlocutor: 225.00 Overall: 132.41 B: Interviewer as interlocutor: 28.30 Interviewee as interlocutor: 202.05 Overall: 42.08
Deng (2008) - Journal article - LLBA	- Not stated - Not stated	Mandarin Chinese - L1	- 15 M, 15 F (15 dyads: 5 M-M, 5 F-F, 5 M-F) - Friends - Mean 20.8	n/a	"participants were given two conversation topics of general interest, but it was emphasised to them that they should feel free to	10m x15 recordings = 2h 30m - Randomly selected from	'Backchannels': "non-lexical forms […] which are used by the interlocutors to show their recipiency" (p309)	'Backchannels': 2.32 'Reactive expressions': 1.52 = 3.84	-

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
… **Deng (2008)** cntd.					talk about anything they liked" (p308) - Researcher devised	after 1m of each	'Reactive expressions': "A short free-standing or turn-incipient lexical phrase or word, or an assessment-type non-lexical form, produced by a recipient in reaction to the speaker's talk" (p310) (Collaborative finishes, repetitions and resumptive openers are also reported)		
	- Not stated - Not stated	English - L1 (Australian)	- 15 M, 15 F (15 dyads: 5 M-M, 5 F-F, 5 M-F) - Friends - Mean 21.2	n/a	As row above	As row above	As row above	'Backchannels': 1.92 'Reactive expressions': 5.55 = 7.47	-
Deutschmann and Panichi (2009) - Journal article - ERIC	- 2008 - Online ('Second Life')	English - Mixed (L1 U.K. English, Australian English with other nationalities - L1s not stated)	- 4 M, 5 F - 2 teachers of English, 7 students (but 1 'student' had L1 English) - 25+ - 45+	- From limited to near-native - Not stated	Online in 'Second Life': self-introduction to the group, plus Q&A after individual presentations to the group - Naturally occurring	48m 28s in total - Whole recordings of two situations stated in column to left	"short utterances (such as *uhu*, *hmm*, *okay*, *yeah*, *I see*) that are produced by the listener while a speaker is holding the floor" (p316); plus "longer responses such as brief statements of interest" (p317)	3.59	-
DiNardo, Schober and Stuart (2005) - Journal article - LLBA	- Late 1960s to 2005 - U.S.A.	English - L1 (assumed)	- Analysts: not stated, but all M Patients: 2 M, 8 F - Analyst and patient - Not stated	n/a	Analyst - patient sessions - Naturally occurring	10m x 10 recordings for each pair = 16h 40m - Typically, 10m after start	"suggest attention to one's conversational partner" (p213); specifically, "*absolutely*, *alright*, *exactly*, *fine*, *hmm*, *I know*, *I see*, *oh*, *okay*, *mm-hm*, *right*, *sure*, *uh-huh*, *yeah*, *yes*" (p222)	Not available	Analyst words per patient backchannel: 833.33
Dittmann and Llewellyn (1967) - Journal article - PsycINFO	- Not stated - U.S.A.	English - L1 (assumed)	- 4 M, 2 F - Acquaintances - Not stated	n/a	Simulated telephone conversation on everyday topics - Researcher devised	~4m x 6 recordings ≈ 24m - Whole recordings	Words: serving "some purpose for the listener like seeking clarification"; Brief vocal sounds, e.g., *yes*, *um-hmm*, *I see* (p344)	Not available	3.06 phonemic clauses per listener response
Dixon and Foster (1998) - Journal article - PsycINFO	- Not stated - South Africa (assumed)	English - L1 (South African)	50 M, 54 F - Not stated (assume strangers) - Not stated	n/a	Half of dyads got to know one another, half debated a controversial issue - Researcher devised	~8m x 52 recordings ≈ 6h 56m - Whole recordings	"e.g., "hmm," "yeh"" (p135)	Not available	M-F dyads: 43.67 words from F per M interlocutor backchannel F-F dyads: 65.36 words from partner per interlocutor backchannel

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Duncan and Fiske (1977) - Book - Bibliographies	- Not stated - Chicago, U.S.A.	English - L1	- 2 M, 1 F (dyads: 1 M-M, 1 M-F; same M in each) - M-M: Friends M-F: Therapist-client - M: ~40; F: early 20s	n/a	M-M: Client discussion M-F: First therapist-client meeting - Naturally occurring	19m x 2 recordings = 38m - M-M: To match M-F M-F: First 19m, up to "a major turning point in the conversation" (p 146)	e.g., *m-hm*, *yeah*, *right*, *I see* (Sentence completions, clarification requests and brief restatements are also reported)	3.05	-
	As row above	As row above	- 6 M, 6 F (dyads: 1 M-M, 4 M-F, 1 F-F) - 5 dyads: strangers; 1 dyad: friends - Not stated	n/a	Casual conversation - Researcher devised	40m in total (6 recordings) - Not stated	As row above	5.65	-
Dungan (1993) - Conference proceedings - ERIC	- Not stated - U.S.A. (assumed)	English - L1 (assumed)	- 4 M, 10 F (dyads: 1 M-M; 2 M-F; 4 F-F) - Supervisor-teacher - Not stated	n/a	Instructional supervisors in post-observation conference with teachers (one-to-one) - Researcher devised	Not stated	"encourager or cooperative simultaneity (words such as "uhhuh," "yeah", or "right") that facilitate continued discussion" (p8)	Not available	168.51 teacher words per supervisor's backchannel; 125.46 supervisor words per teacher's backchannel
Farr (2003) - Journal article - ERIC	- Not stated - Limerick, Ireland	English - L1 (Irish English) (assumed)	- 2 F tutors; 9 master's students (3 M, 6 F) (10 dyads in total) - Tutor-student - Not stated	n/a	One-to-one master's degree student post-observation meetings - Naturally occurring	30-40m x 10 recordings - Whole recordings	Minimal response tokens ("yeah, mm hm, mm, yes, okay and no") (p74); Non-minimal response tokens (single words; assessment-oriented)	Not available	54.86 words per minimal response token; 472.97 words per non-minimal response token Overall: 49.16 words per response token
Feke (2003) - Conference proceedings - Google Scholar	- Not stated - Pittsburgh, U.S.A.	A: Spanish - L1 (From: Chile; Argentina) B: English - L1 (Canadian English; U.S. English)	A: - 2 M, 2 F (dyads: 1 M-M, 1 F-F; then 2 M-F) - Friends - In the range 19-40 B: - 2 M, 2 F (dyads: 1 M-M, 1 F-F; then 2 M-F) - Strangers - In the range 19-40	n/a	A&B: Discussion: Same-sex dyads - "What do you like or dislike about the University of Pittsburgh?" Mixed-sex dyads - "Where have you traveled in the past? If you could travel anywhere in the world, where would you travel and why?" (p98) - Researcher devised	A: 4m x 4 recordings = 16m B: 4m x 4 recordings = 16m A&B: - Typically, 2m when one interlocutor held the floor; 2m when the other interlocutor did	"nonverbal (i.e.- "Mmm-hmm", "Uh-huh", "Ah-hah"), short comments (i.e.- "*No me digas"(No way)*, "That's awesome."), sentence fragments (i.e.- "So, your own personal advanced Quechua", *"con su carrera"(with his career)*, "right in the middle of"), and short questions" (p99)	A: M-M: 11.5 M-F: 14 F-F: 9 B: M-M: 11 M-F: 14.13 F-F: 9.5	-

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Fellegy (1995) - Journal article - LLBA	- Not stated - U.S.A. (assumed)	English - L1 (U.S.) (assumed)	- 8 M, 21 F (3 all-M groups; 3 all-F) - M groups: 2 x tutor-student; other not stated F groups: 2 x business groups; 1 social group - Not stated	n/a	M groups: one-to-one tutor student meeting (x 2); gay dinner party (4 people) F groups: counselling staff meeting (6 people); lesbian task force meeting (4 people); book discussion club (11 people) - Naturally occurring	20m x 6 recordings = 2h - Not stated	"utterances produced by listeners that clearly were not turns, generally in the form of one- or two-word responses, such as mmhmm, right, uh-huh, and yeah" (p188)	M: 1.5 F: 3.1	-
Freedman and Sperling (1983) - Conference proceedings - ERIC	- Not stated - San Francisco, U.S.A.	English - L1	- 1 M, 4 F (4 dyads with the same F teacher in each) - Teacher-student - Not stated	n/a	College writing tutorial A: "high achieving Caucasian" student B: "high achieving Asian-American" student C: "low achieving Caucasian" student D: "low achieving Asian-American" student (all p8) - Naturally occurring	A, B & C: 7m 20s each D: 7m 30s - Test discussion segment	e.g., *ok*, *right*, *aha*, *uh huh*	(Student only) A: 1.09 B: 1.09 C: 1.64 D: 4.00	-
Fujimura-Wilson (2005) - Doctoral thesis - Index to theses	A: - 1999-2000 - Japan and U.K. B: - 1999-2000 - U.K.	A: Japanese - L1 B: English - L1 (U.K.)	A: - Not stated - Family and/or friends - 19+ B: - Not stated - Family and/or friends - 19+	n/a	A: Face-to-face informal conversations (3 or 4 people in each); 17 in Japan, 16 in U.K. B: 12 face-to-face informal conversations (3 or 4 people in each) A&B: - Naturally occurring	Not stated	English: e.g., *mm*, *yeah*, *indeed*, *ok*, *sure*, *I see* Japanese: (in the style of the English ones)	Not available	A: 2.32 utterances per backchannel (In Japan: ~2.38 In U.K.: ~2.27) B: 2.78 utterances per backchannel
Furo (2000)[28] - Book chapter - LLBA	- Not stated - Not stated	Japanese - L1 (assumed)	- 3 M, 3 F (2 same-sex groups) - Friends - Late 20s - ~30s	n/a	Face-to-face conversation on ordinary topics - Naturally occurring	20m x 2 recordings = 40m - Whole recordings	After Clancy et al. (1996) Backchannels, reactive expressions (Collaborative finishes, repetitions and laughter are also reported)	M group: 7.25 F group: 16.55	-
	- Not stated - Not stated	English - L1 (U.S.) (assumed)	- 3 M, 3 F (2 same-sex groups) - Friends - Late 20s - ~30s	n/a	As row above	20m x 2 recordings = 40m - Whole recordings	As row above	M group: 4.60 F group: 6.85	-

[28] Furo (2001) presents the same data as the female English group and reports slightly different information but an identical per minute rate for the female Japanese group.

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Furo (2002) - Conference proceedings - Specialist website	- Not stated - Not stated	Japanese - L1	- Not stated (all F) (dyads) - Friends - Not stated	n/a	A: Telephone conversations - Researcher devised B: Face-to-face conversations - Researcher devised	A: 4m x 5 recordings B: 4m x 5 recordings = 40m - After greetings	As row above	A: 15.60 B: 11.25	Words per backchannel: A: 14.35 B: 16.25
	As row above	English - L1 (U.S. English)	- Not stated (all F) (dyads) - Friends - Not stated	n/a	A: Telephone conversations - Researcher devised B: Face-to-face conversations - Researcher devised	A: 4m x 5 recordings B: 4m x 5 recordings = 40m - After greetings	As row above	A: 6.35 B: 5.00	Words per backchannel: A: 47.17 B: 49.71
Hall et al. (1994) - Journal article - PsycINFO	- Not stated - Massachusetts, U.S.A	English - L1 (assumed)	- 75 M, 75 F (dyads: 25M doctor-M patient; 25M doctor-F patient; 25F doctor-M patient; 25F doctor-F patient) - Doctor-patient - Doctors: not stated; patients: 23-88	n/a	One-to-one doctor-patient consultations - Naturally occurring	41h 11m in total (100 recordings) - Whole recordings	"verbal indicators of sustained attention and encouragement emitted by a speaker who does not hold the speaking floor. The behaviors coded as back-channel responses included such responses as *mm-hmm*, *yeah*, *okay*, and *right*" (p386)	M doctor-M patient: 2.32 from doctor 0.34 from patient M doctor-F patient: 2.33 from doctor 0.46 from patient F doctor-M patient: 2.98 from doctor 0.43 from patient F doctor-F patient: 3.21 from doctor 0.29 from patient	Speaker utterances per interlocutor backchannel: M doctor-M patient: 29.03 with doctor as speaker 3.38 (patient) M doctor-F patient: 20.45 with doctor as speaker 3.35 (patient) F doctor-M patient: 23.71 from doctor as speaker 3.00 (patient) F doctor-F patient: 30.69 from doctor as speaker 2.66 (patient) Utterance: "smallest meaningful string of words" (p386)
Hannah and Murachver (1999) - Journal article - PsycINFO	- Not stated - Dunedin, New Zealand (assumed)	English - L1 (New Zealand English) (assumed)	- 36 M, 36 F (dyads made up of 1 of 4 M or 4 F confederates of the researcher with each of 32 M and 32 F participants) - Strangers - Confederates: mean of 23; participants: mean of 20 (M), 19 (F)	n/a	Discuss Dunedin and holiday planning. Half of dyads had a facilitative confederate (frequent backchannels; eye contact; no interruptions when listening); half had a non-facilitative confederate (interrupted; fewer or delayed backchannels; some looking away) - Researcher devised	10m x 64 recordings = 10h 40m - After the first 2m of each conversation	"any audible response that was uttered by the listener while the other person was speaking. This did not include response tokens used as a precursor to speech" (p161)	Participants with facilitative confederate: 3.29 Participants with non- facilitative confederate: 3.47	-

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
Harrigan (1980) - Book chapter - LLBA	- Not stated - Not stated	English - L1 (assumed)	- 2 M, 3 F (1 group) - Each knew at least 3 of the others - 27-33	n/a	A social gathering - Naturally occurring	13m 30s - Whole recording	Acknowledgements, exclamations (Restatements, requests for clarification and apologies are also reported)	6.22	-
Heinz (2003) - Journal article - MLA Intl.	- Not stated - Germany	German - L1	- 10 (gender not stated) - Friends - Not stated	n/a	Phone call with friend - Naturally occurring	5m x 5 recordings = 25m - Not stated	"brief vocal responses […] which do not constitute an attempt to take the conversational floor" (p1120)	6.24	-
	- Not stated - U.S.A.	English - L1 (U.S.)	- 10 (gender not stated) - Friends - Not stated	n/a	Phone call with friend - Naturally occurring	5m x 5 recordings = 25m - Not stated	As row above	8.92	-
	- Not stated - Between Germany and U.S.A. (phone calls)	German - L1 (L1 monolingual German with balanced German-English bilinguals)	- 10 (gender not stated) - Not stated (assume strangers) - Not stated	n/a	Phone call - Researcher devised	5m x 5 recordings = 25m - Not stated	As row above	Monolingual speakers: 2.64 Bilingual speakers: 5.64 Overall: 8.28	-
Heitman (1999) - Journal article - LLBA	- Not stated - Not stated	English - Mixed (L1 U.S. English with 5 L1 Japanese, 3 L1 Taiwanese, 1 L1 French, 1 L1 Arabic)	- 17 (7 L1 English, 10 other L1); (gender not stated); (dyads) - Strangers - "university age" (p77)	- Not stated - Not stated	Face-to-face casual conversation - Researcher devised	A: 5m x 10 recordings = 50m - Not stated B (after backchannelling training and 6 weeks after A): 5m x 10 recordings = 50m - Not stated	"*yeah*, *uh-huh*, *ohh*, *oh really* and so on" (p76) (Repetitions are also reported)	A: L1 English as backchanneller: 5.46 L2 English as backchanneller: 2.56 B: L1 English as backchanneller: 4.58 L2 English as backchanneller: 3.16	-
Hirokawa (1995) - Doctoral thesis - ProQuest D&T	- Not stated - Michigan, U.S.A.	Japanese - L1 and English - L2 (L1 Japanese)	- 10 M (5 dyads) - Strangers / acquaintances - In the range 23-35	- Not stated - 2-18 months	Face-to-face casual conversation - Researcher devised	8m x 5 recordings per language = 1h 20m - After the first 2m of each recording	e.g., *mhm*, *sure*, *right*, *certainly*, *wow* (and similar in Japanese)	Speaking Japanese: 10.1 Speaking English: 8.73	-
	As row above	English - L1 (U.S.) and Japanese - L2 (L1 English)	- 10 M (5 dyads) - Strangers / acquaintances - In the range 18-34	- Not stated - 6 weeks-5months	As row above	8m x 5 recordings (English only) = 40m - After the first 2m of each recording	As row above	Speaking English: 5.33	-

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
... **Hirokawa (1995)** cntd.	As row above	English - Mixed (L1 U.S. English - L1 Japanese) and Japanese - Mixed (L1 U.S. English - L1 Japanese)	- 10 M (5 dyads: L1 English - L1 Japanese) - Strangers - In the range 18-35	- Not stated - L1 English: 6 weeks-5 months L1 Japanese: 2-18 months	As row above	8m x 5 recordings (English only) = 40m - After the first 2m of each recording	As row above	Speaking English: Japanese people: 7.20 U.S. people: 4.15 = 11.35	-
Jurafsky et al. (1998)[29] - Conference proceedings - Google Scholar	- Not stated - [U.S.A.]	English - [L1 (broadly representative of U.S. English)]	- [From 292 M, 239 F] - [Not stated] - [20-69]	n/a	[One-to-one telephone discussions on various topics] - [Researcher devised]	Not stated (1,155 conversations)	Continuers (passive recipiency); not agreements, answers or incipient speakership	Not available	19% of utterances were backchannels ("an utterance roughly corresponds to a sentence") (p2)
Kjellmer (2009)[30] - Journal article - MLA Intl.	- Not stated - Not stated	English - L1 (U.K.)	- Not stated - Not stated - Not stated	n/a	Not stated - Not stated	Not stated	"*Mhm*, *Mm*, *Right*, *Uh huh*, *Yeah*, *Yes*" (p85)	Not available	46.44 words per backchannel (turn-external and turn-internal backchannels only); this assumes that the sample used is representative of the whole corpus module
Knight (2009)[31] - Doctoral thesis - DART	- Not stated - Not stated	English - L1 (U.K.)	- 1 M, 1 F - Supervisor (M) - Master's degree student (F) - Not stated	n/a	Master's degree supervision session - Naturally occurring	10m - Randomly from the corpus; from the middle of the session	By function: Continuers: "floor-yielding tokens signalling that the addressee is listening" (p47) Convergence tokens: "have a 'higher relational value' than continuers" (p47) Engaged response tokens Information receipt tokens	M: 4.5 F: 2.8 = 7.3	- (Number of words per participant is reported inconsistently) [31.71 words per backchannel, based on data showing 6.8 backchannels per minute]
	As row above	As row above	- 8 M, 4 F - Supervisor - master's degree student (dyads: 3 M-M, 1 M-F, 1 F-M, 1 F-F) - Not stated	n/a	As row above	5h 7m in total - Whole recordings	As row above	7.92	23.13 words per backchannel

[29] This uses some of the Switchboard corpus. Information in square brackets comes from University of Pennsylvania (2002), which describes the procedure followed to collect the data which were used by Jurafsky et al. (1998).
[30] This uses "the spoken British English module of the CobuildDirect Corpus" (Kjellmer, 2009: 85).
[31] This appears to use the same data as Knight and Adolphs (2008), which is the source of information presented in square brackets.

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Kogure (2003) - Doctoral thesis - ProQuest D&T	- Not stated - Not stated	Japanese - L1	- 15 M, 15 F (dyads: 5 M-M, 5 M-F, 5 F-F) - Friends / acquaintances (3months - 6 years) - 18-25	n/a	Face to face; talk about "memory of school lunches" (p68) - Researcher devised	5m x 15 recordings = 1h 15m - After the first 3m of each conversation	"short verbal utterances which are given by the listener as responses to what the speaker has just said" (p104). Classed similarly to Clancy et al. (1996), except 'hai' and 'soo' are classed as continuers here, not reactive expressions (Collaborative finishes, repetitions and resumptive openers are also reported)	M-M: Continuers: 6.56 Reactive exps.: 1.96 = 8.52 (per dyad) F-F: Continuers: 11.96 Reactive exps.: 2.08 = 14.04 (per dyad) M in M-F: Continuers: 4.44 Reactive exps.: 1.00 = 5.44 (per person) F in M-F: Continuers: 4.40 Reactive exps.: 1.00 = 5.40 (per person)	(Coninuers and reactive expressions are combined here) Moras per backchannel: M-M: 44.78 F-F: 28.75 M backchanneler in M-F: 38.61 F backchanneler in M-F: 36.84
Kubota (1991) - Doctoral thesis - ProQuest D&T	- Not stated - U.S.A.	Japanese - L1 and English - L2 (L1 Japanese)	- 5 M, 11 F (dyads: 1 M-M, 3 M-F, 4 F-F) - Acquaintances (5 dyads), strangers (3 dyads) - 20s - 60s	- Minimum ACTFL Intermediate - 1.5 - 39 years (in U.S.A.)	Description of story after watching a drama on video - Researcher devised	3m x 29 recordings = 1h 27m - First 3m of each interaction, to include story descriptions	"a signal that the person holding the floor receives from the floor supporter" (p33). Includes 'short backchannels' ('long backchannels': sentence completions, restatements, requests for clarification, other turns are also reported)	In Japanese: 8.26 In English: 7.26	-
	As row above	English - L1 (U.S.) and Japanese - L2 (L1 U.S. English)	- 9 M, 9 F (dyads: 4 M-M, 1 M-F, 4 F-F) - Acquaintances (3 dyads), strangers (6 dyads) - 20s - 60s	- Minimum ACTFL Intermediate - 9 months - 29 years (in Japan)	As row above		As row above	In Japanese: 3.19 In English: 3.19	-
Lee and Mukai (1998) - Journal article - LLBA	- 1997 - Canberra, Australia	Japanese - L1	- 10 (gender not stated) (5 dyads) - Not stated - 20-26	n/a	Face-to-face conversation on any topic - Researcher devised	15m x 5 recordings = 1h 15m - From 5m after start	Backchannel expressions (e.g., *honto*, *n*, *ee - really*, *uh-huh*, *yeah*) (Repetitions, cooperative completions and nods are also reported)	14.20	Syllables per backchannel: 13.8 Note: this includes all of the backchannel forms listed two columns to the left
	- 1997 - Canberra, Australia	Japanese - Mixed (L1 Japanese with L1 English)	- 10 (gender not stated) (5 dyads of Japanese L1-English L1) - Not stated - 20-26	- Advanced - ~1year	As row above	15m x 5 recordings = 1h 15m - From 5m after start	As row above	L1 Japanese backchanneller: 8.05 L1 English backchanneller: 5.05 Overall: 13.10	Syllables per backchannel: Japanese backchanneller: 11.6 English backchanneller: 14.0 See note above

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Levow, Duncan and King (2010) - Conference proceedings - Google Scholar	- Not stated - U.S.A.	English - L1 (U.S.) (assumed)	- 31 dyads (mixed and same gender) - Friends/family - Not stated	n/a	Relate a story watched on video - Researcher devised	Not stated	"Since the speaker controls the floor throughout the retelling, we treat all listener utterances as instances of vocal feedback in our setting" (no page number)	(As stated times may be for the whole corpus or the section reported on here, backchannels per minute cannot be calculated reliably)	"ratio of listener feedback phrases to total utterance phrases" (no page number): 0.2
	As row above	Spanish - L1 (Mexican Spanish) (assumed)	25 dyads (mixed and same gender) - Friends/family - Not stated	n/a	As row above	As row above	As row above	As row above	0.15
	- Not stated - U.S.A.; Jordan; U.A.E.	Arabic - L1 (Iraqi or Emirati Arabic) (assumed)	23 dyads (mixed and same gender) - Friends/family - Not stated	n/a	As row above	As row above	As row above	As row above	0.29
Li (2006) - Journal article - LLBA	- Not stated - Canada	English - L2 (L1 Mandarin Chinese)	- 10 M, 10 F (same sex dyads) - Not stated - Not stated	- TOEFL 575+ - 2 weeks - 5 years	Simulated doctor-patient interview: 'patient' describes problem, then 'doctor' describes treatment - Researcher devised	1h 38m 50s - Whole recordings	"any verbal […] act occurring during the conversation in a non-intrusive manner (not interrupting the speech turn of the current speaker)" (p103) *ok, uhm, yeah, I see, right, oh,* (repeats)*, shi ma*	5.58	-
	As row above	English - L1 (Canadian)	- 10 M, 10 F (same sex dyads) - Not stated - Not stated	n/a	As row above	1h 17m 50s - Whole recordings	As row above	4.28	-
	As row above	English - Mixed (L1 Mandarin Chinese with L1 Canadian English)	A: - 10 M, 10 F (same sex dyads with partner from B) - Not stated - Not stated B: 10 M, 10 F - Not stated - Not stated	A (L1 Mandarin Chinese): - TOEFL 575+ - 2 weeks - 5 years B (L1 Canadian English): n/a	As row above	A as 'doctor', B as 'patient': 1h 28m 50s B as 'doctor', A as 'patient': 1h 28m 40s - Whole recordings	As row above	A as 'doctor', B as 'patient': A: 2.21 B: 2.56 = 4.77 B as 'doctor', A as 'patient': A: 2.04 B: 2.22 = 4.26	-
Macaulay (2005) - Book - Specialist website	- 1997 - Glasgow, U.K.	English - L1 (Scottish English)	- 8 M, 9 F (same sex and class dyads) - Friends or acquaintances - 40+	n/a	Face-to-face conversation on any topic - Researcher devised	Not stated	Minimal responses; e.g., *mhm*, *uhuh*	Not available	Words per backchannel: M working-class: 23.42 M middle-class: 23.42 F working-class: 28.49 F middle-class: 19.69

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
Maynard (1989) - Book - Bibliographies	- 1985 - Tokyo, Japan	Japanese - L1	- 20 M, 20 F (dyads: 10 M-M, 10 F-F) - Friends - Not stated	n/a	Face-to-face conversation on any topic - Researcher devised	3m x 20 recordings = 1h - After the first 2m of each conversation	"brief utterances such as un 'uh-huh', honto 'really', and soo 'I see'" (p168)	10.23	-
	- 1985 - New Jersey, U.S.A.	English - L1 (U.S.)	- 20 M, 20 F (dyads: 10 M-M, 10 F-F) - Friends - 18-32	n/a	As row above	As row above	English versions of those in row above	3.58	-
McLachlan (1991) - Journal article - LLBA	- Not stated - U.K. (assumed)	English - L1	- 22 M, 22 F - Friends (dyads: 6 M-M, 5 F-F) or strangers (dyads: 5 M-M, 6 F-F) - 18 - ~45	n/a	Face-to-face discussion of hypothetical problems where there is (researcher-known) agreement or disagreement - Researcher devised	~1h 31m 58s - Whole recordings	Acknowledgement / attention signal which does not stop the speaker	Agreeing: M-M: 4.00 F-F: 6.17 Friends: 4.74 Strangers: 5.43 Overall: 5.08 Disagreeing: M-M: 3.47 F-F: 3.79 Friends: 2.90 Strangers: 4.37 Overall: 3.63	Words per backchannel: Agreeing: M-M: 43.05 F-F: 29.51 Friends: 38.37 Strangers: 31.49 Disagreeing: M-M: 50.78 F-F: 44.14 Friends: 60.66 Strangers: 38.35
Miyazaki (2005) - Doctoral thesis - LLBA	- Not stated - Japan (assumed)	Japanese - L1	- 30 F (15 dyads) - Friends - 19-61 (In final data: A: 5 dyads, listeners aged 19-23 B: 4 dyads, listeners 31-39 C: 4 dyads, listeners 54-60)	n/a	Face-to-face conversation on memory of a school trip / other trip - Researcher devised	20m 26s in total (13 recordings) - Single topic; one main speaker; speaker-listener from different workplace and age difference <5y. Plus similar total length for A, B and C	After Clancy et al. (1996) Backchannels (hai, ee, un, aa, haa, hoo, huun, hee); reactive expressions (e.g., *sugoi* (*great*), *honto* (*really*)) (Repetition, collaborative finishes, resumptive openers, short comments and paraphrasing are also reported)	A: 9.62 B: 8.59 C: 17.65	-
Mott and Petrie (1995) - Journal article - LLBA	- Not stated - U.K.	English - L1 (assumed)	- 9 F + unknown number of others - Recruitment consultant (one of the 9F)-client or -employee - Not stated	n/a	Telephone conversations at workplace - Naturally occurring	Not stated (78 recordings) - Interlocutors: employees (27 F, 13 M), clients (22 F, 16 M). Max. 3 of each per consultant; where >3 available, selected by recording clarity and length	"the speaker did not intend to take control of the floor but made an utterance relevant to the interlocutor (e.g., "mm," "yes," etc.)." (p328)	Not available	Backchannels per speaker's 100 words: To employee: 1.59 To client: 2.64

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Nordenstam (1992) - Journal article - LLBA	- 1979 - Gothenburg, Sweden	Swedish - L1 (assumed)	- 18 M, 18 F (dyads: 6 M-M, 6 M-F, 6 F-F) - Close friends / relations - Not stated	n/a	Face-to-face conversation on any topic - Researcher devised	Not stated	"short supports like *hm, ja* 'yes', etc., exclamations of surprise: *åh*, and exclamatory questions: *vad säger du* 'aha, you don't say'" (p92)	Not available	Words per backchannel: M-M: 136.95 M-F (combined): 227.61 F-F: 82.98
O'Conaill, Whittaker and Wilbur (1993) - Journal article - PsycINFO	- Not stated - U.K.	English - L1 (assumed)	A: - Not stated - Work colleagues - Not stated B: - Not stated - Officials at different colleges - Not stated C: - Not stated - Work colleagues - Not stated	n/a	Business meetings; mainly information exchange, appraisal, idea-generation A: low bandwidth video conferences (5 recordings; 4-7 people in each) B: higher bandwidth video conferences (4 recordings; 7-9 people in each) C: face-to-face meetings (5 recordings; 4-7 people in each) - Naturally occurring	20m x 14 recordings = 4h 40m - 20m after start	"short feedback utterances, produced by the listener to indicate functions such as attention, support, or acceptance of the speaker's message"; e.g., *mm, uhu, ok* (p400)	A: 0.35 B: 1.53 C: 3.04	Words per backchannel: A: 458.86 B: 115.72 C: 55.70
Ohira (1998) - Doctoral thesis - ProQuest D&T	- Not stated - Japan	Japanese - L1	- 4 F - Friends - 20	n/a	Dinner party - Researcher devised	1h - When participants were used to the setting	"short utterances […] that do not influence the flow of the speaker's talk" (p75) e.g., *uh-huh, yeah, I see*; and similar in Japanese (Repetitions are also reported)	4.45	Utterances per backchannel: 9.63 Words per backchannel: 40.30 Note: Utterances = intonation units; this includes backchannels in word counts
	- Not stated - U.S.A.	Japanese - L1 (described as Japanese-English bilingual)	- 5 F (1 group, including researcher) - Some friends, some strangers - 25-29	- High enough to participate actively - 6 months - 6.5 years in U.S.A.	As row above	As row above	As row above	5.63	Utterances per backchannel: 6.84 Words per backchannel: 26.95 See note above
	- Not stated - U.S.A.	English - L1 (U.S. English)	- 5 F (1 group, including researcher) - Some friends, some strangers - 23-27	n/a	As row above	As row above	As row above	4.28	Utterances per backchannel: 7.19 Words per backchannel: 31.91 See note above
	- Not stated - U.S.A.	English - Mixed (L1 Japanese with L1 U.S. English)	(Combinations from the 2 groups above) - 9 F (2 groups of 5, each including researcher)	As 2 rows above	As row above	As row above	As row above	Americans (2 people in each): 2.82 Japanese (3 people in each): 4.18 Overall: 6.99	(Overall) Utterances per backchannel: 5.59 (Overall) Words per backchannel: 23.35 See note above

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
Ohtaki, Ohtaki and Fetters (2003) - Journal article - PsycINFO	- 2000 - U.S.A.	English - L1 (assumed)	- Doctors: 5 M Patients: 11 M, 9 F - Doctor-patient - 39-52	n/a	One-to-one doctor-patient consultation - Naturally occurring	3h 42m 54s (20 recordings) total - Whole recordings	"markers of continued attention uttered by the listener"; e.g., (in English) *hmm*, *ok*, *right* (p278)	3.90	-
	- 2000 - Japan	Japanese - L1 (assumed)	- Doctors: 4 M Patients: 11 M, 9 F - Doctor-patient - 40-59	n/a	As row above	2h 48m 20s (20 recordings) total - Whole recordings	As row above	6.74	-
Okamoto and Sato (2008) - Conference proceedings - Papers First	- Not stated - Not stated	Japanese - L1	- 8 F (7 dyads) - Dyads: Mother-daughter (4); 'daughter'-friend (3) - 47-56 (mothers) 17-23 (daughters and their friends)	n/a	Face-to-face conversation on any topic - Researcher devised	15m x 7 recordings = 1h 45m - First 15m of each recording	After Clancy et al. (1996) Backchannels and reactive expressions (Repetitions and collaborative finishes are also reported)	Mother-daughter: Mothers: 1.58 Daughters: 0.55 = 2.13 'Daughter'-friend: 9.58	-
O'Keeffe and Adolphs (2008)[32] - Book chapter - MLA Intl.	- Not stated - U.K.	English - L1 (U.K.)	- 7 F (1 group of 2, 1 group of 5) - Close friends - ~20	n/a	Face-to-face chat - Naturally occurring	Not stated - 20,000-word sub-corpus	Continuers, convergence tokens, engagement tokens, information receipt tokens	Not available	Words per backchannel: 65.79
	- Not stated - Ireland	English - L1 (Irish English)	- 6 F (1 group of 2, 1 group of 4) - Close friends - ~20	n/a	Face-to-face chat - Naturally occurring	Not stated - 20,000-word sub-corpus	As row above	As row above	Words per backchannel: 104.71
Oreström (1983)[33] - Book - Bibliographies	- Not stated - Not stated	English - L1	- 13 M, 7 F (dyads: 4 M-M, 5 M-F, 1 F-F) - Relatives - Not stated	n/a	Face-to-face conversations - At least 4 were naturally occurring	4h 35m 30s - Not stated	Supports (e.g., *m*, *mhm*, *yes*); exclamations (e.g., *oh*, *gosh*); exclamatory questions (e.g., *what*, *really*); sentence completions and restatements	Supports: 2.63 Exclamations: 0.09 Exclamatory questions: 0.15 Sentence completions and restatements: 0.12 Total: 2.99	Words per backchannel: 56.12
O'Sullivan and Porter (1996) - Conference proceedings - ERIC	- 1995 - Okayama, Japan	English - Mixed (L1 Japanese with L1 English)	- Interviewers: 3 M, 3 F Interviewees: 6 M, 6 F (dyads: each interviewee with 1 M and, separately, 1 F interviewer) - Not stated - (Means): Interviewers: 29.6 Interviewees: ~20	Interviewers: n/a Interviewees: - Not stated - Not stated	'Interview' in 2 parts: short answers, then longer responses - Researcher devised	Not stated x 16 recordings (8 interviewees, each with 1 M and 1 F interviewer) - 10% or more difference between score with F versus M interviewer	"utterances such as 'yeah', 'mmmm', 'uh-huh' etc." (p5)	Interviewers only: M: 4.24 F: 2.30	-

[32] This uses parts of the CANCODE corpus (U.K. English data) and parts of the Limerick Corpus of Irish English (Irish English data).
[33] This uses parts of the London-Lund Corpus.

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
Pillon, Degauquier and Duquesne (1992) - Journal article - Hand search	- Not stated - Not stated	French - L1 (assumed)	- 20 M, 20 F (dyads: 20 M-F) - Strangers - 19-26	n/a	Discuss a given topic - Researcher devised	11m x 20 recordings = 3h 40m - Not stated	Non-turn minimal responses "*oui*, *huhum*, *c'est ça*, etc." (p154)	M: 1.48 F: 1.48 Overall: 2.96	Speaker words per interlocutor backchannel: M backchanneller: 59.82 F backchanneller: 66.61
Plough and Gass (1993) - Book chapter - LLBA	- Not stated - U.S.A.	English - L2 (L1s: 6 Japanese, 1 Arabic, 1 Chinese, 1 Spanish, 1 Turkish)	- 10 M (5 dyads) - Known for 4-7 months - 20-29	- "low intermediate to high intermediate" (p38) - Not stated	Two tasks: spot the difference; discussion - Researcher devised	24m x 5 = 2h - First 12m of recording of each task	"responses such as 'Uh huh', 'Mmm', and 'Yeah' made by one speaker during the other speaker's utterance" (p40)	Spot the difference task: 2.83 Discussion: 2.48	-
	As row above	English - L2 (L1s: 4 Japanese, 2 Arabic, 1 Chinese, 1 Korean, 1 Gourmantdi, 1 Spanish)	- 10 M (5 dyads) - Strangers - 22-43	As row above	As row above	As row above	As row above	Spot the difference task: 3.25 Discussion: 3.45	-
Reid (1980) - Doctoral thesis - ProQuest D&T	- 1979 - Arizona, U.S.A.	English - L1 (assumed)	- 28 M, 28 F (dyads: 28 M-F) - Not stated - 19-46	n/a	Simulated counselling interviews (participants were counselling students, each played each role separately with the same partner) - Researcher devised	10m x 56 recordings = 9h 20m - Not stated	1 or 2 words that help the interviewee continue; e.g., *mm*, *hmm*, *really*	As counsellor: M: 30.28 F: 32.09	-
Robey, Canary and Burggraf (1998) - Book chapter - PsycINFO	- Not stated - Not stated	English - L1 (assumed)	- 20 M, 20 F (dyads: 20 M-F) - Husband-wife - Mean age 33y	n/a	Talk about any topic, face to face - Researcher devised	"approximately 7 minutes" (p381) x 20 recordings ≈ 2h 20m - Whole recordings	e.g., *um-hmm*, *uh-huh*, *really*	Not available	Speaker words per interlocutor backchannel: M backchaneller: 66.23 F backchanneller: 90.09
Schleef (2009) - Journal article - LLBA	A: - 2002-2003 - Germany	German - L1 (for students: assumed)	- 18 M, 18 F lecturers; 125 students - Lecturer-student - Not stated	n/a	Interactional classes and lectures at university (1 lecturer per session; natural sciences plus humanities) - Naturally occurring	Not stated - Not lectures, because of low frequency of backchannels	Indicating participation in the interaction, encouraging the speaker to continue; e.g. *ok*, *yeah* (English), *gut*, *ja* (German)	Not available	Student words per lecturer backchannel: 67.11
	B: - Not stated - Not stated (assume U.S.A.)	English - L1 (U.S. English) (for students: assumed)	- 16 M, 16 F lecturers; 181 students - Lecturer-student - Not stated	As row above	As row above	As row above	As row above	As row above	Student words per lecturer backchannel: 136.99

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Sharpley et al. (2000) - Journal article - PsycINFO	- Not stated - Melbourne, Australia	English - L1 (assumed)	- Trained 'clients': 2 F; 'counsellors': 12 M, 47 F (59 dyads of 1 'client' to 1 'counsellor') - Not stated - 'Clients': 24, 45; 'counsellors': 21-60	n/a	Simulated face-to-face counselling interviews ('counsellors' were trainee counsellors) - Researcher devised	46h 13m in total - Whole interviews	Minimal encouragers	5.22	-
Spicer (2005) - Doctoral thesis - ProQuest D&T	- 2004 - Florida, U.S.A.	English - L1 (assumed)	- Counsellors: 2 M, 2 F; clients: 4 M, 4 F (1 counsellor per M-F partner) - Couple with marriage counsellor - Over 21	n/a	Marriage counselling - Naturally occurring	37m 46s; 35m 34s; 33m 46s; 36m 34s = 2h 23m 20s - Last part, minus final summary	Minimal responses	Counsellor to M client: 0.39 Counsellor to F client: 0.27 M client to counsellor: 0.08 F client to counsellor: 0.07 M client to partner: 0.06 F client to partner: 0.01	Client words per counsellor backchannel M client: 102.07 F client: 144.90
Stubbe (1998)[34] - Journal article - Bibliographies	- Not stated - Not stated	English - L1 (New Zealand English; some balanced bilingual Maori)	- 8 M, 8 F (dyads: 2 M-M Maori, 2 M-M Pakeha, 2 F-F Maori, 2 F-F Pakeha) (*Pakeha*: people of European descent) - Not stated - 40-60	n/a	Informal conversation - Naturally occurring	Not stated - Approximate length and natural topic boundaries	Minimal responses: neutral (e.g., *mm*, *uhuh*, *yeah*) and supportive (e.g., *oh gosh*) (Cooperative overlaps are also reported)	Not available	Words per backchannel: Maori: 28.03 Pakeha: 19.43
Tachakra and Rajani (2002) - Journal article - PsycINFO	- Not stated - U.K.	English - L1 (assumed)	Patients: 60 Doctors and nurses: not stated (gender not stated) (60 groups of 1 patient, 1 doctor, 1 nurse) - Doctor-nurse-patient - Not stated	n/a	Minor accident consultations: A: Teleconsultations - patient and nurse together, consulting doctor via video link. B: Face-to-face consultations - patient, doctor and nurse all together - Naturally occurring	A: 7h 55m 30s in total (30 recordings) B: 2h 3m 30s in total (30 recordings) - Whole recordings	"comments such as 'mhm' or 'uh-huh'" (p227)	A: 2.46 B: 4.37	Words per backchannel: A: 74.33 B: 83.33

[34] This uses part of the Wellington Corpus of Spoken New Zealand English.

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Tanaka (2007) - Doctoral thesis - ProQuest D&T	- 2003 - Japan	Japanese - L1	- 82 M, 82 F (82 same-sex dyads) - Friends - 18-22	n/a	Conversation on any topic - Researcher devised	6m x 82 = 8h 12m - At random	After Clancy et al. (1996): backchannels, reactive expressions (Collaborative finishes, repetitions and resumptive openers are also reported)	Backchannels: M: 2.16 F: 3.75 Reactive expressions: M: 1.17 F: 1.49 Total: M: 3.33 F: 5.24	-
Tao and Thompson (1991) - Journal article - PsycINFO	- Not stated - Not stated	Mandarin Chinese - L1	- 2 or 3 (gender not stated) - Friends - Not stated	n/a	Casual conversation - Researcher devised	"about 5 minutes" (p211) - Not stated	"short, non-lexical utterances produced by an interlocutor who is playing primarily a listener's role" (p210) (But some words are included)	~2	-
	- Not stated - Not stated	English - L1 (U.S.)	As row above	n/a	As row above	As row above	As row above	~12.6	-
	- 1978 - Taiwan	Mandarin Chinese - L1	- 2 M - Strangers - 28, 38	n/a	Casual conversation (For one speaker, "His English was as fluent as his native Mandarin" (p212) and he had used English almost exclusively for 20 years) - Researcher devised	5m - Not stated	As row above	Bilingual: 5 Monolingual: 2.8	-
	- Early 1970s - U.S.A.	Mandarin Chinese - L1	- 1 M, 1 F - Not stated - M: ~35 F: Mid-20s	n/a	"eight interview-style conversations about Chinese culture" (p212) M had similar Mandarin and English ability and had used English almost exclusively for 20y; F was Chinese-English balanced bilingual, but had used Mandarin primarily during her life - Researcher devised	5m and (possibly separate) 30m recordings - Not stated	As row above	From 5m recording: M: 10.6 F: 0.4 From 30m recording: M: 10.2 F: 0.17	-
Thonus (1998) - Doctoral thesis - ProQuest D&T	- 1997 - Indiana, U.S.A.	English - Mixed (L1 English tutors with L1 not stated tutees)	- Tutors: 1 M, 5 F; tutees: 2 M, 4 F (one-to-one dyads) - Tutor-tutee - Tutors: 22-37; tutees: 20-27	- Not stated - Not stated	Academic writing tutorials - Naturally occurring	5h 18m in total (6 recordings) - Whole recordings	""off-line" hearer continuers not constituting a taking of the floor", excluding "continuers that fill turn slots and are thus main channel" (p87)	Tutor: 1.56 Tutee: 3.15	Partner's words per interlocutor backchannel: Tutor as backchanneller: 23.08 Tutee as backchanneller: 21.90

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
… **Thonus (1998)** cntd.	As row above	English - L1	- Tutors: 2 M, 4 F; tutees: 1 M, 5 F (one-to-one dyads) - Tutor-tutee - Tutors: 25-50; tutees: 18-29	n/a	As row above	5h 30m in total (6 recordings) - Whole recordings	As row above	Tutor: 1.98 Tutee: 2.48	Partner's words per interlocutor backchannel: Tutor as backchanneller: 29.31 Tutee as backchanneller: 30.66
Tottie (1991)[35] - Book chapter - MLA Intl.	- Not stated - Not stated	English - L1 (U.K.) (assumed)	- 3 M, 1 F (dyads: 1 M-M, 1 M-F) - Not stated - M-M dyad: ~30 M-F dyad: 22, 23	n/a	M-M dyad: discussion of serious topics M-F dyad: talking about work and study experiences - Not stated	53m 30s in total - Not stated	Supportive sounds/words, not responded to by the main speaker. Includes laughter	5.05	-
	- Not stated - Not stated	English - L1 (U.S.) (assumed)	- 1 M, 1 F - Academic colleagues - 30s, 40s	n/a	Academic discussion - Not stated	19m - Not stated	As row above	16.74	-
Truong, Poppe and Heylen (2010)[36] - Conference proceedings - Google Scholar	- Not stated - Not stated	Dutch - [L1]	- [10 M, 24 F] (20 dyads) - ["good friends, relatives, or long-time colleagues" (p501)] - [12-72; 2 were minors]	n/a	Face-to-face conversation on any topic - Researcher devised	15m x 20 recordings = 5h - Whole recordings	"short responses such as 'yeah' and 'hm' that do not contribute contentful information to the conversation" (p3058) (Laughter is also reported)	7.09	[Words per backchannel: 32.53]
Wannaruk (1997) - Doctoral thesis - ProQuest D&T	- Not stated - Thailand	Thai - L1	- 30 M, 30 F (single-sex dyads) - Friends - 17-22	n/a	Telephone conversation on any topic - Researcher devised	10m x 30 recordings = 5h - Minutes 5-15 of 20-m recordings	Short utterances: supports, exclamations, questions (e.g., *ok*, *mhm*, *oh*, *really*) (Sentence completions, short questions, restatements and laughter are also reported)	M: 2.37 F: 2.26	-
	- Not stated - Illinois, U.S.A.	English - L1 (U.S.)	- 30 M, 30 F (single-sex dyads) - Friends - 18-23	n/a	As row above	As row above	As row above	M: 1.43 F: 2.24	-
Ward (1996) - Conference proceedings - Papers First	- Not stated - Not stated	Japanese - L1 (assumed)	- Not stated (17 dyads) - Not stated - Not stated	n/a	Conversation on any topic (In most, participants could not make eye contact, because of how they were seated) - Researcher devised	1h 20m in total (17 recordings) - Whole recordings	"responds directly to the content of an utterance […] is optional, and […] does not require acknowledgement by the speaker" (no page number)	9.86	-

[35] This uses parts of the London-Lund Corpus (U.K. data) and parts of the Santa Barbara corpus (U.S. data).
[36] This uses the IFADV corpus. Details on the corpus, taken from van Son et al. (2008), are in square brackets.

Item - Publication type - Found via	Data collected - Year - Place	Language of interaction - L1/L2/mixed (with L1 details)	Participants - Number; gender - Relationship - Age (years)	L2 participants - Level - Time in L2 countries	Interaction situation - Naturally occurring or researcher devised?	Recording length used - How selected	Backchannels included	Backchannels per minute (per dyad or group, unless otherwise stated)	Other rates (per dyad or group, unless otherwise stated)
Ward and Al Bayyari (2007)[37] - Book chapter - MLA Intl.	- Not stated - Telephone (North America to - typically - Egypt)	Egyptian colloquial Arabic - L1 (2 speakers "not truly fluent in Arabic" (p190))	- Not stated (15 conversations, mostly dyads; up to 4 people in some) - Mostly family / close friends - Mostly adults	n/a	Telephone calls on everyday topics - Naturally occurring	2h 48m in total (15 recordings) - First 5 recordings from each of 3 CDs	Same definition as row above (Laughter is also reported)	3.48	-
Ward and Tsukahara (2000) - Journal article - LLBA	- Not stated - Tokyo, Japan	Japanese - L1	- 15 M, 9 F (18 dyads) - Not stated - 20s	n/a	Conversation (In most, participants could not make eye contact, because of how they were seated) - Researcher devised	1h 20m in total (18 recordings) - Whole recordings	Same definition as row above	10.91	-
	- Not stated - U.S.A. (one recording in Tokyo, Japan)	English - L1 (U.S., except 1, "who had lived in the U.S. since her early teens" (p1190)	- 10 M, 2 F (8 dyads) - Not stated - Not stated	n/a	As row above	1h 8m in total (8 recordings) - Whole recordings	As row above	5.28	-
White (1989)[38] - Journal article - LLBA	- Not stated - Hawaii, U.S.A.	English - L2 (L1 Japanese)	- 10 F (5 dyads) - Strangers - In the range 18-37	- [16 were level '4'; others were '3' or '5' (on 1-6 scale) on Foreign Service Institute's language proficiency interview test] - Less than 6 months in Hawaii	"get to know one another for 30 minutes in English", face to face (p61) - Researcher devised	10m x 5 recordings = 50m - Middle 10m, to avoid beginnings and ends	*mmhm*, *yeah*, *uh-huh*, *oh* and *hmm* (These made up 74% of the total; the other 26% were ignored)	6.83 (per person)	14 words per backchannel
	As row above	English - L1 (U.S.)	- 10 F (5 dyads) - Strangers - [18-37]	n/a	As row above	10m x 5 recordings = 50m - Middle 10m, to avoid beginnings and ends	As row above	2.84 (per person)	37 words per backchannel
	As row above	English - Mixed (L1 Japanese with L1 U.S. English)	(10 dyads, formed from 1 Japanese and 1 American from the above 2 groups)	L1 Japanese: as 2 rows above L1 English: n/a	As row above	10m x 10 recordings = 1h 40m - Middle 10m, to avoid beginnings and ends	As row above	Japanese listener: 6.01 American listener: 4.61 = 10.62	-

[37] This uses parts of the CallHome Corpus of Egyptian Arabic Speech.
[38] This presents the same data as White (1986), which is the source of information presented in square brackets.

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
Yamada (1992) - Book - LLBA	- 1988 - San Francisco, U.S.A.	Japanese - L1	- 3 M (1 group) - Colleagues - 37-40	n/a	Bank officers' weekly business meeting - Naturally occurring	20m - Whole recording	e.g., *un*, *m*, *ee*, *hai*	8.95	Tone units (pause-bounded chunks of speech) per backchannel: 5.01
	As row above	English - L1 (U.S.)	- 1 M, 2 F (1 group) - Colleagues - 25-39	n/a	As row above	27m - Whole recording	e.g., *mhm*, *uhuh*, *yeah*, *ok*	6.04	Tone units per backchannel: 6.20
	As row above	English - Mixed (L1 Japanese with L1 English)	- 3 M, 3 F (1 M Japanese-1 F American in one group; 2 M Japanese-2 American F in other) - Colleagues - 23-59	- Not stated - Not stated	As row above	1h 14m in total (2 recordings) - Whole recordings	As row above	L1 Japanese: 6.24 L1 English: 3.30 Overall: 9.54	Tone units per backchannel: L1 Japanese backchanneller: 2.81 L1 English backchanneller: 4.89
Young and Lee (2004) - Journal article - LLBA	- Not stated - Not stated	Korean - L1	- 4 F (2 dyads) - Strangers - 20s	n/a	Discussion of a film and more general discussion - Researcher devised	13m x 2 recordings = 26m - Not stated	Continuers and assessments; e.g., *yey*, *ney*, *ah*	11.0	-
	As row above	English - L1 (U.S.)	- 4 F (2 dyads) - Strangers - 20s	n/a	As row above	13m x 2 recordings = 26m - Not stated	Continuers and assessments; e.g., *uhuh*, *yeah*, *wow*	9.31	-
	As row above	English - Mixed (L1 Korean with L1 U.S. English)	- 8 F (4 dyads; same people as in above 2 rows) - Strangers - 20s	- Advanced - Not stated	As row above	13m x 4 recordings = 52m - Not stated	As row above	L1 Korean: 6.04 L1 English: 4.89 = 10.93	-
Zuengler (1993) - Journal article - LLBA	- Not stated - Not stated	English - Mixed (L1 English with L1 various: 1 of 12 L1s)	- 90 M (45 dyads: L1 English - L2 English) - Strangers - L1 English: 19-44 L2 English: 20-52	- ~Intermediate - 2m - 5y (U.S.A.)	Discuss major (equal major knowledge) - Researcher devised	10m x 15 recordings = 2h 30m - Whole recordings	"Listeners can actively signal to the speaker that they are listening. Such examples as *right*, *mhmm*, *yeah*, and others provide this function without serving as an attempt to take the speaker's turn away" (p432)	L1 English: 2.88 L2 English: 4.04	Speaker words per interlocutor backchannel: L1 English backchanneller: 21.96 L2 English backchanneller: 19.13
	As row above	As row above	As row above	As row above	Discuss major (L1 English but major non-expert - L2 English but major expert) - Researcher devised	10m x 15 recordings = 2h 30m - Whole recordings	As row above	L1 English: 4.07 L2 English: 3.09	Speaker words per interlocutor backchannel: L1 English backchanneller: 18.20 L2 English backchanneller: 20.49

Item **- Publication type** **- Found via**	**Data collected** **- Year** **- Place**	**Language of interaction** **- L1/L2/mixed (with L1 details)**	**Participants** **- Number; gender** **- Relationship** **- Age (years)**	**L2 participants** **- Level** **- Time in L2 countries**	**Interaction situation** **- Naturally occurring or researcher devised?**	**Recording length used** **- How selected**	**Backchannels included**	**Backchannels per minute (per dyad or group, unless otherwise stated)**	**Other rates (per dyad or group, unless otherwise stated)**
… **Zuengler (1993)** cntd.	As row above	As row above	As row above	As row above	Discuss major (L1 English and major expert - L2 English and major non-expert) - Researcher devised	10m x 15 recordings = 2h 30m - Whole recordings	As row above	L1 English: 3.31 L2 English: 4.21	Speaker words per interlocutor backchannel: L1 English backchanneller: 16.89 L2 English backchanneller: 22.81
Zuengler and Bent (1991) - Journal article - Hand search	- Not stated - Not stated	English - Mixed (L1 English with L1 various: 1 of 12 L1s)	- 90 M (45 dyads: L1 English - L2 English) - Strangers / not well acquainted - L1 English: 19-37 L2 English: 18-43	- Not stated - 4m - 11y (U.S.A.)	Discuss major (equal major knowledge) - Researcher devised	10m x 15 recordings = 2h 30m - Whole recordings	"indicate to the speaker that she or he is saying something of interest to the listener, and should continue talking" (p400); e.g., *uhhuh*, *mhmm*, *I see*	L1 English: 3.16 L2 English: 4.13	Speaker words per interlocutor backchannel: L1 English backchanneller: 22.45 L2 English backchanneller: 21.78
	As row above	As row above	As row above	As row above	Discuss major (L1 English but major non-expert - L2 English but major expert) - Researcher devised	10m x 15 recordings = 2h 30m - Whole recordings	As row above	L1 English: 2.33 L2 English: 3.37	Speaker words per interlocutor backchannel: L1 English backchanneller: 36.30 L2 English backchanneller: 20.14
	As row above	As row above	As row above	As row above	Discuss major (L1 English and major expert - L2 English and major non-expert) - Researcher devised	10m x 15 recordings = 2h 30m - Whole recordings	As row above	L1 English: 2.42 L2 English: 3.95	Speaker words per interlocutor backchannel: L1 English backchanneller: 26.78 L2 English backchanneller: 26.48

Appendix 2 Description of Procedure for Participants

There are four sections.

1. Introduction. I will ask you some background questions about yourself. This is not part of the test.
2. First topic. This is like IELTS Parts 2 ('individual long turn') and 3 ('two-way discussion'), using one topic.
3. Break (1 minute).
4. Second topic. This is like IELTS Parts 2 and 3, using one topic, which is different from the first topic.

For IELTS Part 2 ('individual long turn') (3–4 minutes)
I will give you a card which asks you to talk about a topic and which includes points which you can cover in your talk. You will have 1 minute to prepare your talk, and a pencil and paper to make notes. You will then talk for 1–2 minutes on the topic.
I will not ask questions or give hints when you are speaking.

For IELTS Part 3 ('two-way discussion') (4–5 minutes)
I will ask questions which are connected to the topic of Part 2. These questions give you the chance to discuss more general issues and ideas.

Skills assessed
A wide range of speaking skills is assessed. These include: the ability to communicate opinions and information on everyday topics and common experiences and situations by answering a range of questions; the ability to speak at length on a given topic using appropriate language and organising ideas coherently; and the ability to express and justify opinions and to analyse, discuss and speculate about issues. [Adapted from IELTS, n.d. b]

Appendix 3 Scripts Used in Procedure

Introduction to the long-turn part: I will give you a topic. You will have one minute to prepare to talk about the topic. You can make notes using the pencil and paper if you want to. After one minute, you will talk about your topic for up to two minutes. I will not ask questions or give hints when you are speaking… Here is your topic. You have one minute to prepare, from now.

Between the long-turn part and discursive part: Thank you. We've been talking about an enjoyable event that you experienced when you were at school. I'd like to discuss with you some more general questions about this topic. First…

After the discursive part: Thank you. Time is up, so that is the end of this section.

For the break: In one minute, we will start the next section, with a different topic.

Appendix 4 Questionnaire

Please circle (○) one answer for each item and write any comments in the space.

1. Comparing the enjoyable event topic and the film or TV topic that you spoke about… which was more difficult?

A. They were about the same
B. The enjoyable event topic was more difficult
C. The film or TV programme topic was more difficult
D. Don't know

Comments:_____

_____

2. Comparing the individual long turn parts and the two-way discussion parts… which was more difficult?

A. They were about the same
B. The individual long turn parts were more difficult
C. The two-way discussion parts were more difficult
D. Don't know

Comments:_____

_____

3. Comparing the interaction with the other person… were his questions

A. About the same difficulty in the enjoyable event tasks and the film or TV programme tasks
B. Easier in the enjoyable event tasks
C. Easier in the film or TV programme tasks
D. Don't know

Comments:_____

_____

4. Comparing the interaction with the other person… were his backchannels ('uh-huh', 'mhm', etc. while listening)

A. About the same frequency in the enjoyable event tasks and the film or TV programme tasks
B. More frequent in the enjoyable event tasks
C. More frequent in the film or TV programme tasks
D. Don't know

Comments:_____

_____

5. Comparing the interaction with the other person… was his attention to what you said

A. About the same in the enjoyable event tasks and the film or TV programme tasks
B. Higher in the enjoyable event tasks
C. Higher in the film or TV programme tasks
D. Don't know

Comments:_____

_____

Thank you!

☐ Please tick (✓) this box if you would like more information about the purpose and results of this research.

If so, please write your e-mail address: _____

Appendix 5 Consent Form

[The header is not reproduced in this electronic version.]

Research Consent Form

I am a graduate student at the University of Oxford conducting research that compares the spoken English of people with different first languages. I am doing this by recording people individually as they complete some short speaking tasks and then analyzing the audio recordings. There is also a short questionnaire after the tasks. I would like you to take part in my study and therefore ask that you read the information below, ask me any questions that you may have and, if you agree to take part, sign the final section.

The total time required will be less than 35 minutes. The recording of you speaking will be stored on my password-protected computer and will not be made available to anyone else, except another researcher or researchers who will have signed a research ethics agreement and will only check my analyses. If the questionnaire is written in your first language, it will be passed to a translator. However, only I will have access to your name and any report of research findings will also preserve your anonymity.

This research has been reviewed by, and received ethics clearance through, the University of Oxford Central University Research Ethics Committee. If you wish to make a formal complaint, you may contact them via e-mail (ethics@socsci.ox.ac.uk) or telephone (UK 01865 614871). Your participation is voluntary, so you may withdraw from the study without penalty at any time. To do this, please contact me.

If you have no more questions about the research and agree to participate, please complete the section under the line below, confirming that you have read and understood this research consent form.

Thank you,

Alexander Flint
alexander.flint@education.ox.ac.uk

_____	_____	_____
Name	Signature	Date

Appendix 6 Example Transcript and Mark Up

The full transcript and mark up of an example long-turn and discursive part interaction is below. Full details of how all recordings were transcribed and prepared for CAF analysis are in Chapter 5. Backchannels were removed from transcripts as a first step prior to CAF analysis, so are not present in this example.

Key to main features

/ /	Contains interviewer's speech
[ ]	Contains transcriber's comments
||	AS–unit boundary
::	Clause boundary
abc	Start of error-free AS–unit
abc	Start of error-free clause
abc	Disfluency
abc	Excluded from analysis
~~mm~~	Pause sound

|| The opera [2syllables] in my high high school ~~um~~ is the biggest event in my ~~um~~ in my life til now || ~~um~~ because ~~uh~~ we must ~~um~~ we must practise by ourselves :: and dress up make up for ourselves ~~mm~~. || Do all the things :: ~~uh~~ we need ~~uh~~ we need practise ~~um~~ every day || and we we have two parts of ~~uh~~ team || the one part is :: to show to audience || and the other parts should ~~uh~~ should prepare the words || yeah and ~~um~~ and I find ~~uh~~ good

teacher yeah [?] not in our school ~~uh~~ just ~~uh~~ my parents' friend || and I find a teacher :: to teach us :: how to ~~um~~ how to ~~um~~ show a good opera to audience || and ~~uh~~ in this in in in this period ~~uh~~ I find out :: ~~um~~ I can ~~um~~ I can organise a team || and ~~uh~~ I have ability :: to ~~um~~ to talk ~~uh~~ to talk my team :: how to ~~um~~ how to get ~~uh~~ how to ~~uh~~ how to say how to get the improvement in in our practice || and at last we get the champion in the c- ~~um~~ competition. ||

/So first what did the other people in your group say about the the opera?/

Um. || At first ~~uh~~ the [?] ~~uh~~ we can't ~~uh~~ we can confirm :: who is the major actor :: and who should be the ~~um~~ second party :: to prepare the words || ~~um~~ but ~~um~~ we just ~~uh~~ set a set a test || ~~uh~~ because ~~uh~~ the teacher ~~uh~~ will help us || and ~~uh~~ he test all the people || ~~um~~ the maybe someone voice is better || and ~~uh~~ they should ~~uh~~ take part in the ~~um~~ second team || and maybe someone ~~uh~~ have the good action || ~~uh~~ or they have ability :: to show the audience || and they ~~uh~~ they weren't nervous :: to face to audience || ~~uh~~ they could be the actor. || Yeah b- ~~uh~~ so the test i- is a good method :: to separate two parts of people. ||

/So how important are these kinds of activities at schools in China?/

Um. Many many [repetition] high school have-uh less less activities || but ~~uh~~ in my high school ~~uh~~ we have many activities like this || ~~um~~ maybe ~~um~~ it's a art festival || yeah the opera [2 syllables] is ~~uh~~ one of the art festival. || So ~~um~~ many people enjoying [?] the activities || because ~~uh~~ the pressure in in study ~~um~~ is uncomfortable [5 syllables] for us ~~um~~ [?] in normal life || and ~~uh~~ the something different [2 syllables] like ~~um~~ art festival ~~uh~~ PE festival yeah and [?] many many [repetition] activities could help :: us ~~uh~~ learn :: to how to make a team :: to ~~um~~ achieve a goal :: or achieve something ~~um~~ for ourselves || and do it for ourselves. ||

/And about your school what were the good points and bad points about your school do you think?/

Mm. In my high school I my high school is the best ~~uh~~ high school in our province || and in in whole China is fay- ~~uh~~ is famous. || Because many ~~um~~ I don't ~~uh~~ I don't think there are ma- I don't think :: there has any bad ~~um~~ bad poi- ~~uh~~ bad parts in in our high schools. || I just have a good memory for for there || yeah.

/So I want you to imagine. If you have children in the future would you send them to the same kind of school that you went to?/

Yeah. Because I think :: the ability ~~uh~~ is most important in our life || and if we just ~~uh~~ pay the time ~~um~~ on study or test :: ~~uh~~ it's n- it's not ~~um~~ it's not good for a person's development. || Yeah. And I hope :: he or she or the ~~uh~~ have ~~uh~~ many opportunities :: to choose what kind of life ~~uh~~ :: they need || yeah ~~uh~~ so I hope :: I can offer ~~uh~~ ~~um~~ complete ~~uh~~ environment for them. ||

/And some people say that school days are the best days of your life. What do you think?/

Um. Before I came here :: I think :: oh school school life is awful || I need ~~uh~~ assignment ~~uh~~ or kinds of lessons || I need get up early. || ~~Uh~~ [?] but ~~um~~ after maybe [?] w- when I came here :: ~~um~~ I find :: ~~um~~ school life is really is really better than the job life || ~~uh~~ maybe if I am a student :: ~~um~~ my f- f- ~~uh~~ firstly I ~~uh~~ I ca- I haven't consider :: ~~um~~ I have a ~~uh~~ i- if I haven't ~~uh~~ salary :: what should I do. || I have I don't need :: ~~uh~~ consider many things || and ~~um~~ ~~uh~~ I have a family || ~~uh~~ I have a parents || I have many things :: could ~~uh~~ rely on || and ~~um~~ I don't ~~uh~~ I just need study || and ~~uh~~ just ~~uh~~ pass the test || it's ok || but ~~um~~ if ~~um~~ if when we go to the society :: we must find a job :: maybe the salary in one month ~~uh~~ is so more ~~uh~~ ~~uh~~ i- i- i- i- is is so less || and maybe ~~uh~~ it's less than the costs in ~~um~~ at now || yeah so maybe ~~um~~ at that time w- I will feel upset :: because maybe the the life is so hard || yeah I have many things should consider. ||

Appendix 7 Cancelling Out in Reliability Checks

As stated in Section 5.10.3, instances during the reliability checks of differences between the original and second analyses of syllables (fluent and disfluent), AS–units and clauses completely or partially cancelling out were recorded. These are listed in the table below.

In the table, each intersection of column and row is for a separate participant, so the first row shows that a cancelling effect occurred in the reliability analyses of six participants for AS–units and clauses in the discursive parts. The most extreme example, listed in the first row as "+1 AS, -1 clause, (x6)", was of six clauses marked in the original analysis of a discursive part of one participant each being marked as an AS–unit in the second analysis.

Table A.1 Instances of cancelling out in reliability checks

Units (part)	**Instances in different participants' analyses**					
AS–units, clauses (discursive)	+1 AS -1 AS +1 clause -1 clause	-2 AS +1 AS -1 clause	+1 AS -1 clause (x6)	+2 AS -2 clauses	-1 AS +1 clause	+1 clause -1 clause
fluent syllables, disfluent syllables (discursive)	16 fluent→dis 2 dis→fluent	+4 fluent +2 disfluent	+1 fluent +1 disfluent			
error-free AS–units, error-free clauses (discursive)	-1 AS +1 clause (x2)					
AS–units, clauses (long-turn)	+1 clause -1 clause					

Note: 'x2' and 'x6' indicates the number of times that cancelling out occurred.

REFERENCES

Albert, A. (2011) When individual differences come into play: the effect of learner creativity on simple and complex task performance. In: P. Robinson (ed.). *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, 239–266.

Alberts, J. K., Yoshimura, C. G., Rabby, M. and Loschiavo, R. (2005) Mapping the topography of couples' daily conversation. *Journal of Social and Personal Relationships*, 22(3), 299–322.

Alcón, E. and Guzman, J. (1995) The relationship between content knowledge and practise opportunities in non-native learners' interaction. *Australian Review of Applied Linguistics*, 18(2), 19–32.

Beach, W. A. (1993) Transitional regularities for 'casual' "okay" usages. *Journal of Pragmatics*, 19(4), 325–352.

Beaman, K. (1984) Coordination and subordination revisited: syntactic complexity in spoken and written narrative discourse. In: D. Tannen (ed.). *Coherence in spoken and written discourse*. Norwood, NJ: Ablex, 45–80.

Bennett, M. and Jarvis, J. (1991) The communicative function of minimal responses in everyday conversation. *The Journal of Social Psychology*, 131(4), 519–523.

Benus, S., Gravano, A. and Hirschberg, J. (2007) The prosody of backchannels in American English. [Online]. In: *Proceedings of ICPhS*, Saarbrücken, Germany, 6–10 August 2007. Available from: http://www1.cs.columbia.edu/nlp/papers/2007/benus_al_07b.pdf [Accessed 29 June 2012]

Bevacqua, E., de Sevin, E., Hyniewska, S. J. and Pelachaud, C. (2012) A listener model: introducing personality traits. *Journal on Multimodal User Interfaces*, 6(1–2), 27–38.

Bjørge, A. K. (2010) Conflict or cooperation: the use of backchannelling in ELF negotiations. *English for Specific Purposes*, 29(3), 191–203.

Boyle, E. A., Anderson, A. H. and Newlands, A. (1994) The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech*, 37(1), 1–20.

Bray, J. H. and Maxwell, S. E. (1985) *Multivariate analysis of variance*. London: Sage.

British Council (n.d.) *What is IELTS?* [Online]. Available from: http://takeielts.britishcouncil.org/choose-ielts/what-ielts [Accessed 8 December 2015]

Brown, A. and Hill, K. (2007) Interviewer style and candidate performance in the IELTS oral interview. In: L. Taylor and P. Falvey (eds.). *IELTS collected papers: research in speaking and writing assessment*. Cambridge: Cambridge University Press, 37–61.

Brown, J. D. (2012) Classical test theory. In: G. Fulcher and F. Davidson (eds.). *The Routledge handbook of language testing*. London: Routledge, 323–335.

Brunner, L. J. (1979) Smiles can be back channels. *Journal of Personality and Social Psychology*, 37(5), 728–734.

Buchweitz, A., Keller, T. A., Meyler, A. and Just, M. A. (2012) Brain activation for language dual-tasking: listening to two people speak at the same time and a change in network timing. *Human Brain Mapping*, 33(8), 1868–1882.

Bulté, B. and Housen, A. (2012) Defining and operationalising L2 complexity. In: A. Housen, F. Kuiken and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 21–46.

Cardiff University (n.d.) *English language requirements*. [Online]. Available from: http://www.cardiff.ac.uk/study/international/english-language-requirements [Accessed 1 September 2016]

Casakin, H., Ball, L. J., Christensen, B. T. and Badke-Schaub, P. (2015) How do analogizing and mental simulation influence team dynamics in innovative product design? *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, 29(2), 173–183.

Caspers, J. (2000) Melodic characteristics of backchannels in Dutch map task dialogues. [Online]. In: *Sixth international conference on spoken language processing*, Beijing, 16–20 October 2000. Available from: https://openaccess.leidenuniv.nl/bitstream/handle/1887/14339/ICSLP2000b.pdf [Accessed 28 June 2012]

Cathcart, N., Carletta, J. and Klein, E. (2003) A shallow model of backchannel continuers in spoken dialogue. [Online]. In: *Proceedings of the 10th conference of the European chapter of the Association for Computational Linguistics*, Budapest, 12–17 April 2003. Available from: http://acl.ldc.upenn.edu/eacl2003/papers/main/p6.pdf [Accessed 30 June 2012]

Clancy, P. M., Thompson, S. A., Suzuki, R. and Tao, H. (1996) The conversational use of reactive tokens in English, Japanese, and Mandarin. *Journal of Pragmatics*, 26(3), 355–387.

Clark, H. H. and Fox Tree, J. E. (2002) Using uh and um in spontaneous speaking. *Cognition*, 84(1), 73–111.

Coates, J. (1989) Women's speech, women's strength? *York Papers in Linguistics*, 13, 65–76.

Coates, J. (1986) *Women, men, and language: a sociolinguistic account of sex differences in language*. London: Longman.

Covner, B. J. (1944) Studies in phonographic recordings of verbal material: III. The completeness and accuracy of counseling interview reports. *The Journal of General Psychology*, 30(2), 181–203.

Creswell, J. W. (2014) *Research design: qualitative, quantitative, and mixed methods approaches*. 4th ed., Thousand Oaks, CA: Sage.

Crookes, G. (1990) The utterance, and other basic units for second language discourse analysis. *Applied Linguistics*, 11(2), 183–199.

Cumming, G. (2012) *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. London: Routledge.

Cutrone, P. (2014) A cross-cultural examination of the backchannel behavior of Japanese and Americans: considerations for Japanese EFL learners. *Intercultural Pragmatics*, 11(1), 83–120.

Cutrone, P. (2013) *Assessing pragmatic competence in the Japanese EFL context: towards the learning of listener responses*. Newcastle upon Tyne: Cambridge Scholars.

Cutrone, P. (2005) A case study examining backchannels in conversations between Japanese-British dyads. *Multilingua*, 24(3), 237–274.

Dattalo, P. (2013) *Analysis of multiple dependent variables*. New York: Oxford University Press.

de Jong, N. H. and Bosker, H. R. (2013) Choosing a threshold for silent pauses to measure second language fluency. In: *The 6th workshop of disfluency in spontaneous speech*, Stockholm 2013. Stockholm: Royal Institute of Technology (KTH), 17–20.

de Jong, N. H., Groenhout, R., Schoonen, R. and Hulstijn, J. H. (2015) Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 36(2), 223–243.

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R. and Hulstijn, J. H. (2012) The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performances of native and non-native speakers. In: A. Housen, F. Kuiken

and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 121–142.

de Jong, N. H. and Wempe, T. (2009) Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385–390.

de Kok, I. and Heylen, D. (2011) The MultiLis Corpus – dealing with individual differences in nonverbal listening behavior. In: A. Esposito, A. M. Esposito, R. Martone, V. C. Müller and G. Scarpetta (eds.). *Toward autonomous, adaptive, and context-aware multimodal interfaces: theoretical and practical issues*. London: Springer, 362–375.

Demo, D. A. (2006) *Interactional competence in gatekeeping encounters: a discourse analysis of cross-cultural employment interviews*. PhD, Georgetown University.

Deng, X. (2009) Listener response. In: S. D'hondt, J. Östman and J. Verschueren (eds.). *The pragmatics of interaction*. Amsterdam: John Benjamins, 104–124.

Deng, X. (2008) The use of listener responses in Mandarin Chinese and Australian English conversations. *Pragmatics*, 18(2), 303–328.

Dethlefs, N., Hastie, H., Cuayáhuitl, H., Yu, Y., Rieser, V. and Lemon, O. (2016) Information density and overlap in spoken dialogue. *Computer Speech and Language*, 37, 82–97.

Deutschmann, M. and Panichi, L. (2009) Talking into empty space? Signalling involvement in a virtual language classroom in second life. *Language Awareness*, 18(3–4), 310–328.

DiNardo, A. C., Schober, M. F. and Stuart, J. (2005) Chair and couch discourse: a study of visual copresence in psychoanalysis. *Discourse Processes*, 40(3), 209–238.

Dittmann, A. T. and Llewellyn, L. G. (1968) Relationship between vocalizations and head nods as listener responses. *Journal of Personality and Social Psychology*, 9(1), 79–84.

Dittmann, A. T. and Llewellyn, L. G. (1967) The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology*, 6(3), 341–349.

Dixon, J. A. and Foster, D. H. (1998) Gender, social context, and backchannel responses. *The Journal of Social Psychology*, 138(1), 134–136.

Dörnyei, Z. (2005) *The psychology of the language learner: individual differences in second language acquisition*. Mahwah: Lawrence Erlbaum Associates.

Dörnyei, Z. and Taguchi, T. (2010) *Questionnaires in second language research: construction, administration, and processing*. 2nd ed., London: Routledge.

Drummond, K. and Hopper, R. (1993) Some uses of yeah. *Research on Language and Social Interaction*, 26(2), 203–212.

Ducasse, A. M. and Brown, A. (2009) Assessing paired orals: raters' orientation to interaction. *Language Testing*, 26(3), 423–443.

Duncan, S. (1974) On the structure of speaker-auditor interaction during speaking turns. *Language in Society*, 3(2), 161–180.

Duncan, S. and Fiske, D. (1977) *Face-to-face interaction: research, methods, and theory*. Hillsdale, NJ: Lawrence Erlbaum.

Dungan, S. (1993) Control strategies in conferencing: a sociolinguistic analysis of micropolitical strategies in supervision. [Online]. In: *Annual meeting of the American Educational Research Association*, Atlanta, GA, 12–16 April 1993. Available from: http://www.eric.ed.gov/PDFS/ED360286.pdf [Accessed 27 June 2012]

Eckert, P. (2013) Ethics in linguistic research. In: R. J. Podesva and D. Sharma (eds.). *Research methods in linguistics*. Cambridge: Cambridge University Press, 11–26.

Edlund, J., Gustafson, J., Heldner, M. and Hjalmarrson, A. (2008) Towards human-like spoken dialogue systems. *Speech Communication*, 50(8–9), 630–645.

Ellis, R. and Barkhuizen, G. (2005) *Analysing learner language*. Oxford: Oxford University Press.

ETS (2015) *Examinee Handbook: Listening & Reading*. [Online]. Available from: https://www.ets.org/Media/Tests/TOEIC/pdf/TOEIC_LR_examinee_handbook.pdf [Accessed 1 September 2016]

ETS (2010) *Linking TOEFL iBT scores to IELTS scores: a research report*. [Online]. Available from: https://www.ets.org/s/toefl/pdf/linking_toefl_ibt_scores_to_ielts_scores.pdf [Accessed 1 September 2016]

Farr, F. (2003) Engaged listenership in spoken academic discourse: the case of student-tutor meetings. *Journal of English for Academic Purposes*, 2(1), 67–85.

Feke, M. S. (2003) Effects of native language and sex on back channel behavior. [Online]. In: *Selected proceedings of the first workshop on Spanish sociolinguistics*, New York, 14–15 March 2002. Available from: http://www.lingref.com/cpp/wss/1/paper1012.pdf [Accessed 28 June 2012]

Fellegy, A. M. (1995) Patterns and functions of minimal response. *American Speech*, 70(2), 186–199.

Feng, G. C. (2014) Intercoder reliability indices: disuse, misuse, and abuse. *Quality and Quantity*, 48(3), 1803–1815.

Ferrari, S. (2012) A longitudinal study of complexity, accuracy and fluency variation in second language development. In: A. Housen, F. Kuiken and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 277–298.

Field, A. P. (2013) *Discovering statistics using IBM SPSS statistics*. 4th ed., London: Sage.

Flint, A. (2013) Alex Housen, Folkert Kuiken and Ineke Vedder (eds.), 2012. Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA.

Amsterdam/Philadelphia: John Benjamins, xii + 305 pages, ISBN 978-90-272-1306-8 (paperback). *International Journal of Applied Linguistics*, 23(2), 269–274.

Flint, A. (2012) *A systematic review of verbal backchannel rates*. MSc, University of Oxford.

Flint, A. (2010) *Interlocutor backchannels and L2 oral fluency*. MA, University of Leeds.

Foster, P. and Skehan, P. (1996) The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18(3), 299–323.

Foster, P., Tonkyn, A. and Wigglesworth, G. (2000) Measuring spoken language: a unit for all reasons. *Applied Linguistics*, 21(3), 354–375.

Foster, P. and Wigglesworth, G. (2016) Capturing accuracy in second language performance: the case for a weighted clause ratio. *Annual Review of Applied Linguistics*, 36, 98–116.

Fox, B. A., Maschler, Y. and Uhmann, S. (2010) A cross-linguistic study of self-repair: evidence from English, German, and Hebrew. *Journal of Pragmatics*, 42(9), 2487–2505.

Fox, J. (2004) Biasing for the best in language testing and learning: an interview with Merrill Swain. *Language Assessment Quarterly*, 1(4), 235–251.

Freedman, S. W. and Sperling, M. (1983) Teacher student interaction in the writing conference: response and teaching. [Online]. In: *Annual meeting of the American Educational Research Association*, Montreal, 11–15 April 1983. Available from: http://www.eric.ed.gov/PDFS/ED229754.pdf [Accessed 5 July 2012]

Fries, C. C. (1952) *The structure of English: an introduction to the construction of English sentences*. London: Longmans.

Fujii, Y. (2008) 'You must have a wealth of stories': cross-linguistic differences between addressee support behaviour in Australian and Japanese. *Multilingua*, 27(4), 325–370.

Fujimoto, D. T. (2007) Listener responses in interaction: a case for abandoning the term, backchannel. *Journal of Osaka Jogakuin College*, 37, 35–54.

Fujimura-Wilson, K. (2005) *A sociolinguistic study of Japanese conversation: analysis of three conversational features*. PhD, Birkbeck College, University of London.

Furo, H. (2001) *Turn-taking in English and Japanese: projectability in grammar, intonation, and semantics*. New York: Routledge.

Furo, H. (2000) Listening responses in Japanese and US English: gender and social interaction. In: S. Swierzbin, F. Morris, M. E. Anderson, C. A. Klee and E. Tarone (eds.). *Social and cognitive factors in second language acquisition*. Somerville, MA: Cascadilla Press, 445–457.

Fulcher, G. and Davidson, F. (2007) *Language testing and assessment: an advanced resource book*. London: Routledge.

Furo, H. (2002) Listener responses in telephone and face-to-face conversations: how do non-verbal behaviors affect Japanese and English interactions? In: *Japanese/Korean Linguistics, Volume 10*. Los Angeles: CSLI, 164–177.

Gardner, R. (2001) *When listeners talk: response tokens and listener stance*. Philadelphia: John Benjamins.

Gardner, R. (1998) Between speaking and listening: the vocalisation of understandings. *Applied Linguistics*, 19(2), 204–224.

Gardner, R. C. (1985) *Social psychology and second language learning: the role of attitudes and motivation*. London: Arnold.

Gilabert, R. (2007) The simultaneous manipulation of task complexity along planning time and (+/- here-and-now): effects on L2 oral production. In: M. P. García Mayo (ed.). *Investigating tasks in formal language learning*. Clevedon: Multilingual Matters, 44–68.

Gilabert, R., Barón, J. and Levkina, M. (2011) Manipulating task complexity across task types and modes. In: P. Robinson (ed.). *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, 105–138.

Ginther, A., Dimova, S. and Yang, R. (2010) Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, 27(3), 379–399.

Goldman-Eisler, F. (1968) *Psycholinguistics: experiments in spontaneous speech*. London: Academic Press.

Goldman-Eisler, F. (1961) A comparative study of two hesitation phenomena. *Language and Speech*, 4(1), 18–26.

Goodwin, C. (1986) Between and within: alternative sequential treatments of continuers and assessments. *Human Studies*, 9(2/3), 205–217.

Goodwin, C. (1981) *Conversational organization: interaction between speakers and hearers*. London: Academic Press.

Gravano, A. (2009) *Turn-taking and affirmative cue words in task-oriented dialogue*. PhD, Columbia University.

Gravano, A. and Hirschberg, J. (2011) Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25, 601–634.

Gustafson, J., Heldner, M. and Edlund, J. (2008) Potential benefits of human-like dialogue behaviour in the call routing domain. In: E. André, L. Dybkjær, W. Minker, H. Neumann, R. Pieraccini and M. Weber (eds.). *Perception in multimodal dialogue systems*. Berlin: Springer, 240–251.

Hall, J. A., Irish, J. T., Roter, D. L., Ehrlich, C. M. and Miller, L. H. (1994) Gender in medical encounters: an analysis of physician and patient communication in a primary care setting. *Health Psychology*, 13(5), 384–392.

Hannah, A. and Murachver, T. (1999) Gender and conversational style as predictors of conversational behavior. *Journal of Language and Social Psychology*, 18(2), 153–174.

Hari, R., Henriksson, L., Malinen, S. and Parkkonen, L. (2015) Centrality of social interaction in human brain function. *Neuron*, 88(1), 181–193.

Harrigan, J. A. (1980) Methods of turn-taking in group interaction. In: J. Kreiman and A. E. Ojeda (eds.). *Papers from the sixteenth regional meeting Chicago Linguistic Society*. Chicago, IL: Chicago Linguistic Society, 102–111.

Heinz, B. (2003) Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics*, 35(7), 1113–1142.

Heitman, C. (1999) A comparison of the use of back channeling gambits in intermediate ESL students before and after instruction: a preliminary report. *The ORTESOL Journal*, 20, 75–83.

Hess, L. J. and Johnston, J. R. (1988) Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, 11(3), 319–335.

Hieke, A. E., Kowal, S. and O'Connell, D. C. (1983) The trouble with "articulatory" pauses. *Language and Speech*, 26(3), 203–214.

Hirokawa, K. (1995) *The expressions of culture in the conversational styles of Japanese and Americans*. PhD, University of Michigan.

Housen, A. and Kuiken, F. (2009) Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461–473.

Housen, A., Kuiken, F. and Vedder, I. (eds.) (2012a) *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins.

Housen, A., Kuiken, F. and Vedder, I. (2012b) Complexity, accuracy and fluency: definitions, measurement and research. In: A. Housen, F. Kuiken and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 1–20.

Housen, A., Kuiken, F. and Vedder, I. (2012c) Epilogue. In: A. Housen, F. Kuiken and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 299–302.

Husband, E. M. (2015) Self-repairs as right node raising constructions. *Lingua*, 160, 20–37.

IELTS (n.d. a) *Test taker performance 2015*. [Online]. Available from: https://www.ielts.org/teaching-and-research/test-taker-performance [Accessed 1 September 2016]

IELTS (n.d. b) *Information for candidates*. [Online]. Available from: http://www.ielts.org/pdf/Information%20for%20Candidates_2013.pdf [Accessed 10 June 2013]

IELTS (n.d. c) *Test Performance 2015*. [Online]. Available from:https://www.ielts.org/teaching-and-research/test-performance [Accessed 1 September 2016]

IIBC (2009a) *TOEIC Test Data & Analysis 2008*. [Online]. Available from: https://web.archive.org/web/20100202182517/http://www.toeic.or.jp/toeic_en/pdf/data/TOEIC_DAA2008.pdf [Accessed 1 September 2016]

IIBC (2009b) *TOEIC Newsletter, No. 105*. [Online]. Available from: http://www.iibc-global.org/library/redirect_only/library/toeic_data/toeic_en/pdf/newsletter/newsletterdigest105.pdf [Accessed 1 September 2016]

Ishikawa, T. (2011) Examining the influence of intentional reasoning demands on learner perceptions of task difficulty and L2 monologic speech. In: P. Robinson (ed.). *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, 307–330.

Iwashita, N., McNamara, T. and Elder, C. (2001) Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning*, 51(3), 401–436.

Jackson, D. O. and Suethanapornkul, S. (2013) The cognition hypothesis: a synthesis and meta-analysis of research on second language task complexity. *Language Learning*, 63(2), 330–367.

Japan Times (1983a) Japanese mannerism is key point in IBM case, 29 January, 2.

Japan Times (1983b) Mitsubishi avoids embarrassing trial with plea of no contest. 22 October, 1.

Jenkins, J. (2006) Points of view and blind spots: ELF and SLA. *International Journal of Applied Linguistics*, 16(2), 137–162.

Jenks, C. J. (2011) *Transcribing talk and interaction: issues in the representation of communication data*. Amsterdam: John Benjamins.

Jurafsky, D., Shriberg, E., Fox, B. and Curl, T. (1998) Lexical, prosodic, and syntactic cues for dialog acts. [Online]. In: *Proceedings of ACL/COLING-98 workshop on discourse relations and discourse markers*, Montreal, 15 August 1998. Available from: http://acl.ldc.upenn.edu/W/W98/W98-0319.pdf [Accessed 5 July 2012]

Kahng, J. (2014) Exploring utterance and cognitive fluency of L1 and L2 English speakers: temporal measures and stimulated recall. *Language Learning*, 64(4), 809–854.

Kanfer, F. H. and McBrearty, J. F. (1962) Minimal social reinforcement and interview content. *Journal of Clinical Psychology*, 18(2), 210–215.

Kim, J. S. (2012) Within-subjects design. In: N. J. Salkind (ed.). *Encyclopedia of research design*. Thousand Oaks, CA: Sage, 1639–1644.

Kirsner, K., Dunn, J. and Hird, K. (2003) Fluency: time for a paradigm shift. In: R. Eklund (ed.). *Gothenburg Papers in Theoretical Linguistics 90*. Gothenburg: University of Gothenburg, 13–16.

Kjellmer, G. (2009) Where do we backchannel? On the use of mm, mhm, uh huh and such like. *International Journal of Corpus Linguistics*, 14(1), 81–112.

Knight, D. (2009) *A multi-modal corpus approach to the analysis of backchanneling behaviour*. PhD, University of Nottingham.

Knight, D. and Adolphs, S. (2008) Multi-modal corpus pragmatics: the case of active listenership. In: J. Romero-Trillo (ed.). *Pragmatics and corpus linguistics: a mutualistic entente*. Berlin: Mouton de Gruyter, 175–190.

Kogure, M. (2003) *Gender differences in the use of backchannels: do Japanese men and women accommodate to each other?* PhD, University of Arizona.

Koike, T., Tanabe, H. C. and Sadato, N. (2015) Hyperscanning neuroimaging technique to reveal the "two-in-one" system in social interactions. *Neuroscience Research*, 90, 25–32.

Kormos, J. (2006) *Speech production and second language acquisition*. Mahwah: Lawrence Erlbaum Associates.

Kormos, J. and Dénes, M. (2004) Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145–164.

Kormos, J. and Trebits, A. (2011) Working memory capacity and narrative task performance. In: P. Robinson (ed.). *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, 267–286.

Koudenburg, N., Postmes, T. and Gordijn, E. H. (2016) Beyond content of conversation: the role of conversational form in the emergence and regulation of social structure.

[Online]. *Personality and Social Psychology Review*, online first, 12 February 2016. Available from: http://psr.sagepub.com/ [Accessed 22 August 2016]

Kraut, R. E., Lewis, S. H. and Swezey, L. W. (1982) Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology*, 43(4), 718–731.

Kubota, M. (1991) *The use of back channel behaviors by Japanese and American bilingual persons*. PhD, Indiana University.

Kuiken, F. and Vedder, I. (2011) Task complexity and linguistic performance in L2 writing and speaking: the effect of mode. In: P. Robinson (ed.). *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, 91–104.

Labov, W. (1972) *Language in the inner city: studies in the black vernacular*. Philadelphia: University of Pennsylvania Press.

Lambert, C., Kormos, J. and Minn, D. (2016) Task repetition and second language speech processing. [Online]. *Studies in Second Language Acquisition*, first view articles, 18 March 2016. Available from: http://journals.cambridge.org/action/displayJournal?jid=SLA [Accessed 24 March 2016]

Larson-Hall, J. (2015) *A guide to doing statistics in second language research using SPSS and R*. 2nd ed., London: Routledge.

Larson-Hall, J. and Herrington, R. (2009) Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics*, 31(3), 368–390.

Larson-Hall, J. and Plonsky, L. (2015) Reporting and interpreting quantitative research findings: what gets reported and recommendations for the field. *Language Learning*, 65(Supp. 1), 127–159.

Leacock, S. (1912) *Nonsense novels*. London: John Lane.

Lebra, T. S. (1976) *Japanese patterns of behavior*. Honolulu: University of Hawaii Press.

Lee, D. and Mukai, C. (1998) A study of Japanese back channels. *Australian Review of Applied Linguistics*, 21(Supplement 15), 77–92.

Leech, G. (2000) Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning*, 50(4), 675–724.

Lehtonen, J. and Sajavaara, K. (1985) The silent Finn. In: D. Tannen and M. Saville-Troike (eds.). *Perspectives on silence*. Norwood, NJ: Ablex, 193–201.

Lennon, P. (2000) The lexical element in spoken second language fluency. In: H. Riggenbach (ed.). *Perspectives on fluency*. Ann Arbor: University of Michigan Press, 25–42.

Lennon, P. (1990) Investigating fluency in EFL: a quantitative approach. *Language Learning*, 40(3), 387–417.

Levelt, W. J. M. (2001) Relations between speech production and speech perception: some behavioral and neurological observations. In: E. Dupoux (ed.). *Language, brain, and cognitive development: essays in honor of Jacques Mehler*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1999) Producing spoken language: a blueprint of the speaker. In: C. M. Brown and P. Hagoort (eds.). *The neurocognition of language*. Oxford: Oxford University Press, 83–122.

Levelt, W. J. M. (1989) *Speaking: from intention to articulation*. London: MIT Press.

Levkina, M. and Gilabert, R. (2012) The effects of cognitive task complexity on L2 oral production. In: A. Housen, F. Kuiken and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 171–198.

Levow, G. A., Duncan, S. and King, E. T. (2010) Cross-cultural investigation of prosody in verbal feedback in interactional rapport. [Online]. In: *Interspeech 2010*, Makuhari, Japan, 26–30 September 2010. Available from: http://mcneilllab.uchicago.edu/pdfs/Levow_etal_InterSpeech_2010.pdf [Accessed 5 July 2012]

Li, H. Z. (2006) Backchannel responses as misleading feedback in intercultural discourse. *Journal of Intercultural Communication Research*, 35(2), 99–116.

LoCastro, V. (1987) Aizuchi: a Japanese conversational routine. In: L. E. Smith (ed.). *Discourse across cultures: strategies in world Englishes*. Hemel Hempstead: Prentice Hall, 101–113.

Macaulay, R. K. S. (2005) *Talk that counts: age, gender, and social class differences in discourse*. Oxford: Oxford University Press.

MacIntyre, P. D. and Gardner, R. C. (1991) Language anxiety: its relation to other anxieties and to processing in native and second languages. *Language Learning*, 41(4), 513–534.

Mackey, A. and Gass, S. M. (2016) *Second language research: methodology and design*. 2nd ed., London: Routledge.

Margetts, A. and Margetts, A. (2012) Audio and video recording techniques for linguistic research. In: N. Thieberger (ed.). *The Oxford handbook of linguistic fieldwork*. Oxford: Oxford University Press, 13–53.

Maynard, S. K. (1997) Analyzing interactional management in native/non-native English conversation: a case of listener response. *International Review of Applied Linguistics in Language Teaching*, 35(1), 37–60.

Maynard, S. K. (1989) *Japanese conversation: self-contextualization through structure and interactional management*. Norwood, NJ: Ablex.

McKelvie, D. (1998) The syntax of disfluency in spontaneous spoken language. *HCRC Research Papers*, 95.

McLachlan, A. (1991) The effects of agreement, disagreement, gender and familiarity on patterns of dyadic interaction. *Journal of Language and Social Psychology*, 10(3), 205–212.

McNamara, T. and Roever, C. (2006) *Language testing: the social dimension*. Malden: Blackwell.

Mehnert, U. (1998) The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20(1), 83–108.

Mesch, J. (2013) Tactile signing with one-handed perception. *Sign Language Studies*, 13(2), 238–263.

Meyer, A. S., Huettig, F. and Levelt, W. J. M. (2016) Same, different, or closely related: what is the relationship between language production and comprehension? *Journal of Memory and Language*, 89, 1–7.

Michel, M. C. (2011) Effects of task complexity and interaction on L2 performance. In: P. Robinson (ed.). *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins, 141–174.

Michel, M. C., Kuiken, F. and Vedder, I. (2007) The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics in Language Teaching*, 45(3), 241–259.

Miyazaki, S. (2005) *Japanese women's listening behavior in face-to-face conversation: the use of reactive tokens and nods*. PhD, Michigan State University.

Morton, J., Wigglesworth, G. and Williams, D. (1997) Approaches to the evaluation of interviewer behaviour in oral tests. In: G. Brindley and G. Wigglesworth (eds.).

Access: issues in language test design and delivery. Sydney: National Centre for English Language Teaching and Research, 175–196.

Mott, H. and Petrie, H. (1995) Workplace interactions: women's linguistic behavior. *Journal of Language and Social Psychology*, 14(3), 324–336.

Mulder, K. and Hulstijn, J. H. (2011) Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics*, 32(5), 475–494.

Müller, F. E. (1996) Affiliating and disaffiliating with continuers: prosodic aspects of recipiency. In: E. Couper-Kuhlen and M. Selting (eds.). *Prosody in conversation: interactional studies*. Cambridge: Cambridge University Press, 131–176.

Nordenstam, K. (1992) Male and female conversation style. *International Journal of the Sociology of Language*, 94(1), 75–98.

Norris J. M. (2015) Statistical significance testing in second language research: basic problems and suggestions for reform. *Language Learning*, 65(Supp. 1), 97–126.

Norris, J. M. and Ortega, L. (2009) Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics*, 30(4), 555–578.

O'Brien, R. G. and Kaiser, M. K. (1985) MANOVA method for analyzing repeated measures designs: an extensive primer. *Psychological Bulletin*, 97(2), 316–333.

O'Conaill, B., Whittaker, S. and Wilbur, S. (1993) Conversations over video conferences: an evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, 8(4), 389–428.

Ohira, K. (1998) *Have you changed? Pragmatic transfer of back-channel behavior by Japanese bilingual speakers*. PhD, University of Illinois at Urbana-Champaign.

Ohtaki, S., Ohtaki, T. and Fetters, M. D. (2003) Doctor-patient communication: a comparison of the USA and Japan. *Family Practice*, 20(3), 276–282.

Okamoto, S. and Sato, S. (2008) Culture and interactional styles: the interpretation of reactive tokens in Japanese conversations. In: *Japanese/Korean linguistics, volume 13*. Michigan, 1–3 August 2003. Stanford, CA: CSLI, 445–456.

O'Keeffe, A. and Adolphs, S. (2008) Response tokens in British and Irish discourse: corpus, context and variational pragmatics. In: K. P. Schneider and A. Barron (eds.). *Variational pragmatics: a focus on regional varieties in pluricentric languages*. Amsterdam: John Benjamins, 69–98.

Oreström, B. (1983) *Turn-taking in English conversation*. Lund: CWK Gleerup.

Ortega, L. and Byrnes, H. (2008) Theorizing advancedness, setting up the longitudinal research agenda. In: L. Ortega and H. Byrnes (eds.). *The longitudinal study of advanced L2 capacities*. Abingdon: Routledge, 281–300.

O'Sullivan, B. (2002) Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277–295.

O'Sullivan, B. and Lu, Y. (2006) The impact on candidate language of examiner deviation from a set interlocutor frame in the IELTS Speaking Test. *IELTS research reports*, 6.

O'Sullivan, B. and Porter, D. (1996) Speech style, gender, and oral proficiency interview performance. [Online]. In: *Southeast Asian Ministers of Education Organization regional language center seminar*, Singapore, 1996. Available from: http://www.eric.ed.gov/PDFS/ED403744.pdf [Accessed 26 June 2012]

Pallotti, G. (2009) CAF: defining, refining and differentiating constructs. *Applied Linguistics*, 30(4), 590–601.

Pillon, A., Degauquier, C. and Duquesne, F. (1992) Males' and females' conversational behavior in cross-sex dyads: from gender differences to gender similarities. *Journal of Psycholinguistic Research*, 21(3), 147–172.

Plonsky, L. and Derrick, D. J. (2016) A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, 100(2), 538–553.

Plonsky, L., Egbert, J. and LaFlair, G. T. (2015) Bootstrapping in applied linguistics: assessing its potential using shared data. *Applied Linguistics*, 36(5), 591–610.

Plonsky, L. and Oswald, F. L. (2014) How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912.

Plough, I. and Gass, S. M. (1993) Interlocutor and task familiarity: effects on interactional structure. In: G. Crookes and S. M. Gass (eds.). *Tasks and language learning: Integrating theory and practice*. Clevedon: Multilingual Matters, 35–56.

Podesva, R. J. and Zsiga, E. (2013) Sound recordings: acoustic and articulatory data. In: R. J. Podesva and D. Sharma (eds.). *Research methods in linguistics*. Cambridge: Cambridge University Press, 169–194.

Poppe, R., Truong, K. P. and Heylen, D. (2011) Backchannel: quantity, type and timing matters. In: H. H. Vilhjálmsson, S. Kopp, S. Marsella and K. R. Thórisson (eds.). *Intelligent virtual agents*. London: Springer, 228–239.

Powers, W. R. (2005) *Transcription techniques for the spoken word*. Oxford: AltaMira.

Quan, L. and Weisser, M. (2015) A study of 'self-repair' operations in conversation by Chinese English learners. *System*, 49, 39–49.

Rahimpour, M. (1999) Task complexity and variation in interlanguage. In: N. O. Jungheim and P. Robinson (eds.). *Pragmatics and pedagogy: proceedings of the 3rd Pacific Second Language Research Forum*. Tokyo: The Pacific Second Language Research Forum, 115–134.

Reid, A. D. (1980) *The effect of gender on counseling interview skills*. PhD, Arizona State University.

Révész, A., Ekiert, M. and Torgersen, E. N. (2016) The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, 37(6), 828–848.

Richards, K. (2003) *Qualitative inquiry in TESOL*. Basingstoke: Palgrave MacMillan.

Rivera, A. G. and Ward, N. G. (2006) Prosodic features that lead to back-channel feedback in Northern Mexican Spanish. In: *Proceedings of the seventh annual High Desert Linguistics Society conference*, University of New Mexico, 2006. Albuquerque, NM: High Desert Linguistics Society, 19–26.

Robey, E. B., Canary, D. J. and Burggraf, C. S. (1998) Conversational maintenance behaviors of husbands and wives: an observational analysis. In: D. J. Canary and K. Dindia (eds.). *Sex differences and similarities in communication: critical essays and empirical investigations of sex and gender in interaction*. Mahwah, NJ: Lawrence Erlbaum, 373–392.

Robinson, P. (ed.) (2011) *Second language task complexity: researching the cognition hypothesis of language learning and performance*. Amsterdam: John Benjamins.

Rosenthal, R. (1966) *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts.

Ross, S. J. (2007) A comparative task-in-interaction analysis of OPI backsliding. *Journal of Pragmatics*, 39(11), 2017–2044.

Rost, M. (2011) *Teaching and researching listening*. 2nd ed., Harlow: Longman.

Sakuragi, T. (2011) The construct validity of the measures of complexity, accuracy, and fluency: analyzing the speaking performance of learners of Japanese. *JALT Journal*, 33(2), 157–173.

Sannomiya, M., Kawaguchi, A., Yamakawa, I. and Morita, Y. (2003) Effect of backchannel utterances on facilitating idea-generation in Japanese think-aloud tasks. *Psychological Reports*, 93(1), 41–46.

Sardegna, V. G. and Molle, D. (2010) Videoconferencing with strangers: teaching Japanese EFL students verbal backchannel signals and reactive expressions. *Intercultural Pragmatics*, **7**(2), 279–310.

Sawaki, Y. (2012) Technology in language testing. In: G. Fulcher and F. Davidson (eds.). *The Routledge handbook of language testing*. London: Routledge, 426–437.

Schegloff, E. A. (1982) Discourse as an interactional achievement: some uses of 'uh huh' and other things that come between sentences. In: D. Tannen (ed.). *Analyzing discourse: text and talk*. Washington, D.C.: Georgetown University Press, 71–93.

Schiffrin, D. (1981) Tense variation in narrative. *Language*, 57(1), 45–62.

Schleef, E. (2009) A cross-cultural investigation of German and American academic style. *Journal of Pragmatics*, 41(6), 1104–1124.

Schoot, L., Hagoort, P. and Segaert, K. (2016) What can we learn from a two-brain approach to verbal interaction? *Neuroscience and Biobehavioral Reviews*, 68, 454–459.

Seedhouse, P. and Egbert, M. (2006) The interactional organisation of the IELTS Speaking Test. *IELTS research reports*, 6.

Seedhouse, P. and Harris, A. (2011) Topic development in the IELTS Speaking Test. *IELTS research reports*, 12.

Segalowitz, N. (2016) Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, 54(2), 79–95.

Sharpley, C. F., Fairnie, E., Tabary-Collins, E., Bates, R. and Lee, P. (2000) The use of counsellor verbal response modes and client-perceived rapport. *Counselling Psychology Quarterly*, 13(1), 99–116.

Siegman, A. W. (1976) Do noncontingent interviewer mm-hmms facilitate interviewee productivity? *Journal of Consulting and Clinical Psychology*, 44(2), 171–182.

Skehan, P. (2014) The context for researching a processing perspective on task performance. In: P. Skehan (ed.). *Processing perspectives on task performance*. Amsterdam: John Benjamins, 1–26.

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

Skehan, P. (2003) Task-based instruction. *Language Teaching*, 36(1), 1–14.

Spicer, K. S. C. (2005) *Identifying power differentials in nonviolent heterosexual couples in counseling through discourse analysis*. PhD, University of Florida.

Stenström, A.-B. (1994) *An introduction to spoken interaction*. London: Longman.

Stubbe, M. (1998) Are you listening? Cultural influences on the use of supportive verbal feedback in conversation. *Journal of Pragmatics*, 29(3), 257–289.

Tabachnick, B. G. and Fidell, L. S. (2012) *Using multivariate statistics*. 6th ed., London: Pearson.

Tachakra, S. and Rajani, R. (2002) Social presence in telemedicine. *Journal of Telemedicine and Telecare*, 8(4), 226–230.

Tanaka, K. (2007) *Differential use of reactive tokens in Japanese in turn management and by gender*. EdD, Temple University.

Tao, H. and Thompson, S. A. (1991) English backchannels in Mandarin conversations: a case study of superstratum pragmatic "interference". *Journal of Pragmatics*, 16(3), 209–223.

Tavakoli, P. (2016) Fluency in monologic and dialogic task performance: challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, 54(2), 133–150.

Tavakoli, P. and Foster, P. (2008) Task design and second language performance: the effect of narrative type on learner output. *Language Learning*, 58(2), 439–473.

Thonus, T. (2007) Listener responses as a pragmatic resource for learners of English. *The CATESOL Journal*, 19(1), 132–145.

Thonus, T. (1998) *What makes a writing tutorial successful: an analysis of linguistic variables and social context*. PhD, Indiana University.

Tolins, J. and Fox Tree, J. E. (2014) Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70, 152–164.

Tonkyn, A. (2012) Measuring and perceiving changes in oral complexity, accuracy and fluency: examining instructed learners' short-term gains. In: A. Housen, F. Kuiken and I. Vedder (eds.). *Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 221–244.

Tottie, G. (1991) Conversational style in British and American English: the case of backchannels. In: K. Aijmer and B. Altenberg (eds.). *English corpus linguistics: studies in honour of Jan Svartvik*. Harlow: Longman, 254–271.

Towell, R., Hawkins, R. and Bazergui, N. (1996) The development of fluency in advanced learners of French. *Applied Linguistics*, 17(1), 84–119.

Truong, K. P., Poppe, R. and Heylen, D. (2010) A rule-based backchannel prediction model using pitch and pause information. [Online]. In: *Interspeech 2010*, Makuhari, Japan, 26–30 September 2010. Available from: http://doc.utwente.nl/74048/1/IS100566.PDF [Accessed 27 June 2012]

University of Pennsylvania (2002) *Switchboard: a user's manual*. [Online]. Available from: http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html [Accessed 30 June 2012]

University of Sheffield (n.d.) *Comparison of English language test scores*. [Online]. Available from: https://www.sheffield.ac.uk/eltc/englishtests/test-comparisons [Accessed 1 September 2016]

Van Daele, S., Housen, A., Kuiken, F., Pierrard, M. and Vedder, I. (2007) Preface. In: S. Van Daele, A. Housen, F. Kuiken, M. Pierrard and I. Vedder (eds.). *Complexity, accuracy and fluency in second language use, learning and teaching*. Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, 11–14.

van Son, R. J. J. H., Wesseling, W., Sanders, E. and van den Heuvel, H. (2008) The IFADV corpus: a free dialog video corpus. [Online]. In: *Language resources and evaluation conference*, Marrakech, Morocco, 28–30 May 2008. Available from: http://www.lrec-conf.org/proceedings/lrec2008/pdf/132_paper.pdf [Accessed 10 June 2012]

Vercellotti, M. L. (2015) The development of complexity, accuracy, and fluency in second language performance: a longitudinal study. [Online]. *Applied Linguistics*, advance access, 13 March 2015. Available from: http://applij.oxfordjournals.org [Accessed 9 July 2016]

Verma, J. P. (2015) *Repeated Measures Design for Empirical Researchers*. Hoboken, NJ: Wiley.

Wannaruk, A. (1997) *Back-channel behavior in Thai and American casual telephone conversations*. PhD, University of Illinois at Urbana-Champaign.

Ward, N. (1996) Using prosodic clues to decide when to produce back-channel utterances. [Online]. In: *International conference on spoken language processing*, Philadelphia,

PA, October 1996. Available from: http://www.asel.udel.edu/icslp/cdrom/vol3/062/a062.pdf [Accessed 25 June 2012]

Ward, N. G. and Al Bayyari, Y. (2010) American and Arab perceptions of an Arabic turn-taking cue. *Journal of Cross-Cultural Psychology*, 41(2), 270–275.

Ward, N. and Al Bayyari, Y. (2007) A prosodic feature that invites back-channels in Egyptian Arabic. In: M. A. Mughazy (ed.). *Perspectives on Arabic linguistics xx: papers from the twentieth annual symposium on Arabic linguistics*. Kalamazoo, MI, March 2006. Amsterdam: John Benjamins, 187–206.

Ward, N. G., Escalante, R., Al Bayyari, Y. and Solorio, T. (2007) Learning to show you're listening. *Computer Assisted Language Learning*, 20(4), 385–407.

Ward, N. and Tsukahara, W. (2000) Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8), 1177–1207.

Weir, C., O'Sullivan, B. and Horai, T. (2006) Exploring difficulty in speaking tasks: an intra-task perspective. *IELTS research reports*, 6.

White, R. (1997) Back channelling, repair, pausing, and private speech. *Applied Linguistics*, 18(3), 314–344.

White, S. (1989) Backchannels across cultures: a study of Americans and Japanese. *Language in Society*, 18(1), 59–76.

White, S. (1986) *Functions of backchannels in English: a cross-cultural analysis of Americans and Japanese*. PhD, Georgetown University.

Wolf, J. P. (2008) The effects of backchannels on fluency in L2 oral task production. *System*, 36(2), 279–294.

Wolfson, N. (1979) The conversational historical present alternation. *Language*, 55(1), 168–182.

Wood, D. (2007) Mastering the English formula: fluency development of Japanese learners in a study abroad context. *JALT Journal*, 29(2), 209–230.

Xi, X. (2010) How do we go about investigating test fairness? *Language Testing*, 27(2), 147–170.

Yamada, H. (1992) *American and Japanese business discourse: a comparison of interactional styles*. Norwood, NJ: Ablex.

Yngve, V. H. (1970) On getting a word in edgewise. In: *Papers from the Sixth Regional Meeting, Chicago Linguistic Society*. Chicago: Chicago Linguistic Society, 567–578.

Young, R. F. and Lee, J. (2004) Identifying units in interaction: reactive tokens in Korean and English conversations. *Journal of Sociolinguistics*, 8(3), 380–407.

Yule, G. (2010) *The study of language*. 4th ed., Cambridge: Cambridge University Press.

Zuengler, J. (1993) Encouraging learners' conversational participation: the effect of content knowledge. *Language Learning*, 43(3), 403–432.

Zuengler, J. and Bent, B. (1991) Relative knowledge of content domain: an influence on native-non-native conversations. *Applied Linguistics*, 12(4), 397–415.